ESZTERHÁZY KÁROLY KATOLIKUS EGYETEM
INFORMATIKA KAR • DIGITÁLIS TECHNOLÓGIA INTÉZET
AGRIA MÉDIA KONFERENCIA 2023

https://doi.org/10.17048/AM.2023.49

https://videotorium.hu/hu/recordings/51353

Ujhelyi Gábor: Interaktív hangoskönyvek az oktatásban

Ujhelyi Gábor
Eszterházy Károly Katolikus Egyetem, Neveléstudományi Doktori Iskola
ug@mensa.hu

Absztrakt: A hangoskönyvek oktatási alkalmazására már sok esetben tettek próbálkozásokat, azonban egyelőre nem épült be szervesen a módszerekbe. Annak ellenére, hogy létrejöttük nem újkeletű, kevés fejlődésen mentek keresztül, azonban az utóbbi évek technológiai fejlesztései számos új alkalmazási, bővítési lehetőséget nyitnak meg. Az írott ismeretanyagok hangon keresztül befogadhatóságához több út vezet. Ebből a megközelítésből a figyelem középpontjába elsősorban a tudományos, ismeretterjesztő, szak- illetve tankönyvek kerülnek. Ebben az esetben arányaiban kisebb jelentősége van az audio előadás módjának, stílusának, hangszínének, dominánsabb az információ jól érthető átadásának szándéka, aminek köszönhetően a maga hiányosságaival együtt is van létjogosultsága a gépi felolvasásnak, beszédszintézisnek.

Kutatásom célja annak igazolása, hogy a jelenleg széles körben elérhető és egyre fejlettebb nagy nyelvi modellek, valamint felhő szolgáltatásokon keresztül elérhető text-to-speech (gépi felolvasó) és speech-to-text (leiratozó) megoldások segítségével megvalósítható, hogy egy digitálisan rendelkezésre álló írott könyv automatikusan előállhat hangoskönyvként úgy, hogy ne csak felolvastatható és meghallgatható, de gépileg interaktívvá tehető, összefoglaltatható, magyaráztatható és kérdezhető legyen, ami nagy mértékben hozzájárulhat ahhoz, hogy az oktatási anyagok az auditív tanulótípusokhoz is közelebb kerüljenek és segítsék az anyag megértését és elmélyülését, valamint segítséget nyújthat az olvasási nehézségekkel küzdő tanulók számára.

A kutatás megvalósítása során a jelenleg elérhető fejlett felhő alapú és helyben futtatható nagy nyelvi modellek tulajdonságait, finomhangolhatóságát, különböző méretű dokumentumok feldolgozhatóságát és azokról egy minta alkalmazásban írott vagy szóbeli formában természetes nyelven feltett kérdésekre általuk adott válaszok tartalmi validitását vizsgálom.

Kulcsszavak: hangoskönyv, interaktivitás, mesterséges intelligencia, nagy nyelvi modellek, auditív tanulótípus, olvasási nehézségek

INTERACTIVE AUDIOBOOKS IN EDUCATION

Abstract: Attempts have already been made to use audio books in education in many cases, but so far it has not been integrated organically into the methods. Despite the fact that their creation is not new, they have undergone little development, however, the technological developments of recent years open up many new application and expansion possibilities. There are several ways to absorb written knowledge materials through sound. From this approach, the focus of attention is primarily on scientific, informative, specialist and textbooks. In this case, the method, style, and timbre of the audio presentation are relatively less important, the intent of the easy-to-understand transfer of information is more dominant, thanks to which even with its shortcomings, machine reading and speech synthesis have a right to exist.

The aim of my research is to prove that with the help of currently widely available and increasingly advanced large language models, as well as text-to-speech (machine reading) and speech-to-text (transcribing) solutions available through cloud services, it is possible to implement a digitally available written book to automatically appear as an audio book in such a way that it can not only be read aloud and listened to, but also automatically made interactive, summarized, explained and asked questions, which can greatly contribute to bringing educational materials closer to auditory types of students and helping them to understand and deepen the material, and can help students with reading difficulties.

During the implementation of the research, I am examining the properties, fine-tuning, document processing capabilites of currently available large lannguage models of different sizes and the content validity of the answers they give to questions asked in natural language in a sample application.

Keywords: audiobook, interactivity, artificial intelligence, large language models, auditory learner type, reading difficulties

1. Bevezetés

A hangoskönyvek szerepét és felhasználási lehetőségét méltatlanul kevés figyelem övezi. Létrejöttük óta az azóta eltelt időhöz képest kevés fejlődésen mentek keresztül, azonban az utóbbi időszak technológiai fejlődése számos új lehetőséget nyit meg az előállításuk és felhasználhatóságuk terén. A kutatás motivációja, hogy kiterjeszthető legyen a hangoskönyvek eddigi használati gyakorlata, szinergiákat teremtve feltárjam az adaptációs lehetőségeket a mesterséges intelligencia, a nagy nyelvi modellek és a szoftver infrastruktúra fejlődéséhez, alkalmazkodva a megváltozott szokásokhoz és eszköz használathoz, valamint olyan lehetőséget keressek, ami a kihívásokkal küzdő oktatás területén is segítséget nyújthat.

Az első teljes hosszúságú hangoskönyvet 1930-ban vették fel, majd a második világháborút követően kezdték támogatással terjeszteni, elsősorban a látássérült háborús veteránok megsegítése érdekében, de segítséget nyújtva a más okból kifolyólag olvasási nehézségekkel küzdők számára is (Smith, 2022). A teljes terjedelmében felolvasott könyveket követve hamarosan megjelentek az átdolgozott, rövidített változatú kiadványok is. A zeneipart is megelőző terjedési ütemet értek el lemezen, majd az eleinte erős helyhez kötöttséggel szakítva az adathordozók és a lejátszó berendezések fejlődésével a kompaktabb szalag, kazetta, cd lemez, illetve a hordozható magnók, walkmanek, discmanek széleskörű térhódításával jutottak el egyre szélesebb közönséghez. A terjedés ütemének szignifikáns növekedését a digitális audio adatformátumok, adathordozók, lejátszó eszközök, valamint az internet megjelenése és gyors elterjedése tette lehetővé. A papír alapú könyvek mellett a könyvtárak is bővíteni kezdték gyűjteményüket analóg, majd digitális hangoskönyvekkel. A "beszélő könyvek"-nek is hívott alkotásokat időnként éri olyan kritika, hogy az a lusta emberek olvasási módja, azonban könnyen belátható, hogy az irodalom fogyasztás ezen módját az olvasásban tartós fogyatékosság vagy ideiglenes egészségügyi ok miatt fogyatékkal élők mellett az egészséges emberek is előnyben részesíthetik olyan élethelyzetekben, amikor átmenetileg korlátozott az olvasási lehetőségük valamilyen szituációból eredően, ami lehet akár csak egy figyelmet enyhén megosztó, de az érzékszerveket lefoglaló monoton cselekvés, mint a főzés, takarítás, vagy autóvezetés. Utóbbi esetben a figyelem fenntartása a balesetveszély elkerülésében is segítség lehet a hosszú, ingerszegény utakon nem ritka elalvás megelőzésével.

A szórakozás, időtöltés mellett nagy jelentősége van a hangoskönyveknek az információszerzés és tanulás területén, nem csak az autodidakta ismeretelsajátítás, hanem a szervezett keretek közötti oktatás során is jól használhatóak. Az élménynyújtás helyébe lépő információközvetítési cél sokkal tágabb teret enged az audio könyvek létrehozása terén, lényegesen kevésbé domináns a felolvasással szembeni elvárás a szórakoztatásban szerepet játszó hanglejtés, hangszín, hangsúlyozás jellemzői terén. Ez nyitja meg alapvetően annak a lehetőségét, hogy a hangoskönyvet kiterjeszthessük a klasszikus statikusan előre felolvasás által kialakított korlátok közül. Az önálló alkalmazása része lehet az otthoni felkészülésnek, de komplexebb e-learning kurzusokba is beágyazhatók írott és video alapú tartalmak helyett vagy azok kiegészítéseként, különös tekintettel az azonos témát feldolgozó alternatív tananyag változatokra, aminek segítségével a fejlesztett tananyag jobban adaptálódhat a különböző tanulótípusok preferenciáihoz, segítve a befogadást, a megértést az auditív típusú tanulóknak, a figyelem fenntartását, flow élmény megteremtési lehetőségét (Csíkszentmihályi, 1991). További felhasználási lehetőségei a frontális oktatás során való alkalmazás akár önállóan, akár írott anyag kiegészítéseként (Serrano, 2023). Az írott anyagok és audio változatuk egyidejű használata (Singh & Alexander, 2022) nagy könnyebbséget jelent az olvasási nehézséggel küzdők számára, de kutatások azt is igazolták, hogy a csoportok további tagjai körében is segített az értő olvasás könnyebb elsajátításában (Nash, 2023), az olvasás fejlesztésében (Chen, 2004). Külön kiemelendő felhasználási lehetőséget képvisel az idegen nyelv tanulás és oktatás, ahol a tanulással elsajátított nyelv továbbadásához képest az anyanyelvi vagy azt megközelítő felolvasás nagyban elősegíti a helyes kiejtés elsajátítását (Kartal & Simsek, 2017). Mindezek megvalósíthatók a hangoskönyv hagyományosan értelmezett keretein belül, előre elkészített hanganyagokkal. A teljes átjárhatóság korlátai között meg kell említeni, hogy az előállítási módtól függetlenül problémát okoz komplett alkotások esetében a nem felolvasható tartalom reprezentálása, mint képek, ábrák, táblázatok, azonban vannak olyan esetek, amikor a hangoskönyv kifejezőképessége magasabb az írott változatnál, ilyenek a dialektusok, vagy a hangeffektusok közvetítési lehetősége.

2. *-learning

Az elektronikus tanulási környezeteknek több olyan elnevezése van, amelynek egymáshoz való kapcsolatáról megoszlik a szakirodalom. Már az elnevezések is megosztók, míg a *-learning kifejezések alapvetően tanulást jelentenek, a használatuk sokkal inkább tanulási környezetre, rendszerre, illetve platformra irányul. Az egyes fogalmakat szokták egymásba ágyazott halmazokként és egymás mellett jelenlévő, egymással metszeteket képező kategóriákként is ábrázolni. E-learning alatt tágan értelmezve olyan tanulási környezeteket értünk, amelyek valamilyen elektronikai eszközt alkalmaznak (elektronikus eszközzel támogatott tanulás). A napi szóhasználatban azonban általában olyan online rendszereket értünk alatta, amely valamilyen internetre kötött számítástechnikai berendezés (számítógép, tablet, okostelefon) használatával igénybe vehetően biztosít digitális tananyagot, illetve komplexebb esetben komplett kurzus hiearchiákat számonkérési lehetőségekkel együtt. A d-learning a digitális tanulási környezetre utal, mai viszonylatban, amikor szinte minden eszközünk digitális, az e-learning alatt is d-learninget értünk, azonban a szó szoros értelmében tanulást támogató elektronikus eszközök lehetnének analóg eszközök, mint pl. a hagyományos lemezjátszók (amik még nem tartalmaztak digitális elektronikát). M-learning (mobile-learning) alatt a mobil eszközzel támogatott (egyes értelmezések szerint kizárólag mobil eszközön végzett) tanulást értjük. Ez alapvetően az e-learning és a d-learning részhalmazaként értendő, mind tágabb értelemben, a digitális (elektronikus) eszközök közül a hordozható készülékeket értve alatta, mind pedig az e-learning hétköznapibb jelentését tekintve, amikor az online e-learning rendszerek által nyújtott szolgáltatásokat mobil eszközről (jellemzően okostelefonról) veszik igénybe (Basak, Wotto, & Bélanger, 2018). Az u-learning (ubiquitous-learning) egy absztraktabb fogalom, az embert körülvevő világ kiterjesztése komplex tanulási környezetté, ami a szó konkrét jelentése alapján nem kellene, hogy feltétlenül digitális tanulásra vonatkozzon, beleérthető az "unplugged" világ minden eleme, azonban a kifejezést mégis inkább az embereket minden területen körülvevő digitális elemek tanulásba bevonására használják (Zhang, 2008). A szintén gyakran használt blended learning pedig ismét egy más megközelítésű csoportosítás, alapvetően nem a használt eszköz határozza meg technológiája alapján, hanem (itt is szemben a learning szó szerinti tanulás jelentésével) olyan hibrid oktatási módszert jelöl, amelyben a hagyományos tantermi oktatást kombinálják a digitális eszközök és online elérhető szolgáltatások alkalmazásával (Sharma, 2010). Ezen módszerek és eszközök mindegyikében közös, hogy a hangoskönyvek valamilyen formájának szerepeltetése kiválóan illeszkedik az eszköztárukba függően az adatforrástól, médiától, lejátszó eszköztől és a beépítés módjától.

3. Hangoskönyvek és felolvasások előállítása

A hangoskönyveket megjelenésükkor kizárólag emberi felolvasás rögzítésével állították elő, amely egyrészről a mai napig a legjobb minőséget, az élvezhetőség, az élmény legmagasabb szintjét biztosítja, ugyanakkor a leginkább költség- és időigényes. A felvételek minősége és tárolási módjai a technikai fejlődés előrehaladtával sokat változtak, a digitális technológiai megoldásokkal számos paraméter (pl. sebesség, hangmagasság) utólag is változtathatóvá vált. A paraméterek megfelelő megválasztása nem csak minőség, hanem felhasználási cél és lejátszó eszköz függvényében is eltérő lehet, ilyenek a mintavételi frekvencia, bitmélység, veszteséges és veszteségmentes tömörítési módok, bitráta. Ezen paraméterek nagy mértékben befolyásolják az audio anyagok tárolásához és átviteléhez szükséges erőforrásokat, így mindig fontosak a kompromisszumok. Példaként említve a hagyományos telefonhálózatokon az emberi fül által hallható frekvenciatartomány csak egy részének átjuttatását biztosítják, mivel az is elegendő az emberi beszéd megértéséhez.

Annak ellenére, hogy az emberi beszéd utánzására voltak korai próbálkozások mechanikus-akusztikus módszerekkel, amelyek az emberi hangképző szervek működését próbálták másolni, vagy a kiadott hangokhoz hasonló effektusokat más forrásokból előállítani, ezek a módszerek nem voltak alkalmasak arra, hogy gyakorlati használathoz elegendő színvonalú szintetizált beszédet állítsanak elő vele. A hangfelvételi és reprodukciós eljárások fejlődésével nyílt meg a lehetőség a beszéd felvett hangokból való összeillesztésére, azonban automatizálása a nyelvtani és hangtani szabályok bonyolult implementációját igényelte, és a fonémák szekvenciáján túl az emberi beszédre jellemző további paraméterek (hangsúly, hanglejtés) megvalósítása további nehézségeket okozott. Igazi rohamos fejlődést a tanuló algoritmusok, illetve kifejezetten a neurális hálózatok megjelenése és alkalmazásának szélesebb körű elterjedése hozott. A neurális hálózatok, vagy mesterséges neuronhálók alapvetően az emberi agy működését hivatottak utánozni, ahol az agyban lévő neuronokat kis funkcionális programegységek reprezentálják, amik között a szinapszisokat súlyozott összeköttetéseknek feleltetik meg. Ezek a neuronok rétegekbe vannak csoportosítva, és a gyakorlati megvalósításban a bemeneti és kimeneti rétegek között nagy számosságú köztes rejtett réteg van, a szomszédos rétegekben elhelyezkedő neuronok közötti összeköttetések súlyai pedig a modell paraméterei. A hálózatokat tanító nagy mennyiségű, tanító és ellenőrző adatcsoportra osztott adathalmazokkal tanítják,

ami leegyszerűsítve annyit jelent, hogy iterálva újra és újra kiértékelik a modellt az aktuális paraméterekkel a kiértékelések között módosítva a paramétereket, amíg a modell által adott eredmény pontossága el nem ér egy előre definiált mértéket. Ez a technológia generikusan implementált megoldásokkal sok komplex célalgoritmus leprogramozását tudja kiváltani úgy, hogy nagy mennyiségű adattal tanítva széleskörűen alkalmazható nagy pontosságú eredményt tud biztosítani. A mai korszerű beszédszintetizátorok ilyen megvalósítást alkalmaznak, melyek bizonyos korlátok között a beszéd egészen sok emberi aspektusát képesek reprodukálni és sok szabadsági fokot biztosítanak az előállított beszéd tulajdonságainak meghatározására, úgy mint hangszín, sebesség, nyelv, hanglejtés, mely az érthetőségen túl közvetve az információ befogadására is hatással lehet, ugyanis az ember nem képes objektíven elválasztani az információt a közlőtől (Cialdini, 2009). Segítségükkel az emberi felolvasáshoz képest nagyságrenddel alacsonyabb idő alatt és költség mellett állítható elő írott anyagok audio reprezentációja. Ennek köszönhetően elérhetővé váltak az igény szerinti, akár valós idejű felolvasások, melyre széles körben találunk megoldásokat az asztali és mobil operációs rendszerek szolgáltatásaitól kezdve vastagklienses célalkalmazásokon keresztül felhő szolgáltatásként elérhető webalkalamazásokig és programozható interfészekig. Ezek minősége és paraméterezhetősége nagy szórást mutat, azonban rendkívül gyors fejlődést tanúsítanak, különösen a felhő alapú megoldások terén. Megemlítendő a hangszín és stílus befolyásolásánál a hangminta alapú tanítás, amely lehetővé teszi tetszőleges személy hangjának, beszédmódjának utánzását tetszőleges szöveg beszéddé konvertálásával, annak előnyeivel és legfőképp etikai aggályaival együtt. Egyes szolgáltatások az élményt olyan emberszerűséget növelő hangeffektusok beépítésével is növelik, mint a nevetés, éneklés, vagy az "ö-zés", a mondaton belüli változó beszédsebesség és a pillanatnyi szünetek beiktatása.

4. Gépi felolvasás jelenlegi problémái és megoldási irányok

Bár a korábbi beszédszintetizátorok problémái sokkal komolyabbak és zavaróbbak voltak, a mai rohamosan fejlődő TTS (text-to-speech) megoldások meglepően jó teljesítményük ellenére is küzdenek néhány visszatérő kihívással. Elsőként megemlítendőek a nyelvi korlátok. Ugyan az igazán nagy mintán tanított modellek nagyon tág határok között multilinguálisak, a ritkább nyelveken elérhető jó minőségű tanító minták alacsonyabb számossága miatt jelentős különbség van a világnyelvek és a kis nyelvek közötti beszédszintézis minőségében. Míg angol nyelven léteznek a valós emberi beszédet esetenként megtévesztésig megközelítő megoldások, addig például a magyar nyelvvel a legjobb modellek is inkább csak botladoznak, ami leginkább a hangsúlyozási hibákban tetten érhető. Vannak azonban nyelvfüggetlen problémák is, amelyek kiejtése a legtöbb modellnek a mai napig problémát okoz, ilyenek a különlegesebb tulajdonnevek, a rövidítések, mozaikszavak, számok, római számok, dátumok, időpontok, képletek. Ezen akadályok elhárításán nagy erőkkel dolgoznak a fejlesztők, és nagy valószínűséggel túlnyomó többségükre a közeljövőben jó megoldások fognak születni. A megoldási irányok közé tartozik az egyre nagyobb és jobb minőségű tanító halmaz gyűjtése, az egyre nagyobb paraméterszámú modellek létrehozása, az emberi visszajelzés általi megerősített tanulás, azaz RLHF – reinforcement learning from human feedback (Casper, és mtsai., 2023). Egyes hiányosságok kompenzálására alkalmazhatók kerülő megoldások, mint a szinonimák használata, a fonetikus átirat készítése, vagy a kiejtési mód annotálása (pl. SSML – Speech Synthesis Markup Laguage), amely esetekben a beszédszintézisre átadott adattartalom nagyobb (szavak cseréje) vagy kisebb (kiejtési hibás karaktertöbbesek cseréje, formátum jelölés) mértékben eltér az eredeti tartalomtól a helyes felolvasás érdekében (Jin, Lee, & Park, 2004).

5. Interaktivitás

A beszédszintetizátorok fejlődése ugyan nagy előrelépés volt az ember-gép közötti kommunikációban, de talán még nagyobb lépést jelentett a természetesnyelv feldolgozás és értés előretörése. A hangoskönyvek lejátszása során a kezdeti meghallgatási lehetőségekhez képest a szó szoros értelmében interakcióba léphetünk a közvetítő eszközök fejlődésének köszönhetően a lejátszási paraméterek (lejátszási pozíció, sebesség, hangmagasság) szabad és dinamikus változtatásával matematikai módszerek segítségével (Veldhuis & He, 1998), amire az olvasással összevetve jogosan is merül fel az igény, tekintve, hogy elveszítjük a gyors áttekintés, keresés, "szkennelve" olvasás, F-minta szerinti feldolgozás lehetőségét (Shrestha, Lenz, & Owens, 2007). Az emberi beszéd írottá konvertálásával (STT - speech-to-text) és természetes szöveg értelmezését és válasz generálását lehetővé tévő generatív nagy nyelvi modellek térhódításával azonban olyan interakció vált lehetővé a tartalomra vonatkozóan, aminek köszönhetően élő szóban is kapcsolatba léphetünk egy hangoskönyvvel, ami akár annak illúzióját is megközelítheti, mintha egyúttal annak szerzőjével teremtenénk kontaktust. Lehetővé vált párbeszédes kommunikáció során a hangoskönyv tartalmára vonatkozó kérdések feltevése és automatikus megválaszoltatása. Ez kimerülhet egyes

részek megkeresésében és lejátszásában, de lehetőség nyílik komplex megfogalmazások átfogalmaztatására, magyaráztatására, hosszabb tartalmi részek vagy teljes művek összefoglaltatására is. A legújabb nagy nyelvi modellek multimodalitása átjárást biztosít az információk különböző reprezentációi között, így a hangoskönyvben átadott tartalomból képek állíthatók elő, vagy fordított irányban létrehozható olyan hangoskönyv vagy hangoskönyv részlet, mely egy annak forrásául szolgáló írott könyvben szereplő képeket, ábrákat szóban írja le, ezáltal jelentősen kitágítva az ez esetben találóbb kifejezéssel élve beszélő könyvek határait. Ezek a képességek egészen új felhasználási területeket hoznak létre a hangoskönyvek számára az oktatás területén belül is. Segíthetik az otthoni felkészülést ellenőrző kérdések automatikus feltevésével vagy válaszok kiértékelésével, helyesség ellenőrzésével, de idővel megvalósíthatóvá válik akár szóbeli vizsgáztatás is. A generációról generációra, illetve generáción belül is változó felhasználói szokásokhoz és technológiai környezethez alkalmazkodva a pedagógia mind a diákokkal való újabb kapcsolódási pontok megtalálásában, mind az oktatás hatékonyságának növelésében profitálni tudni a technológia fejlődésének köszönhetően a hangoskönyvek kiterjesztett és interaktív alkalmazásával mind a szervezett tanítás, mind az önálló tanulás keretein belül.

6. Nagy nyelvi modellek (LLM)

A jelenleg szöveg generálásra és természetes nyelvi válaszadásra használt, az elmúlt egy évben robbanásszerűen elterjedt nagy nyelvi modellek a generatív előtanított transzformer (GPT – generative pretrained transformer) modellek közé tartoznak. Az ezek mögött működő neurális hálózatok százmilliárdos nagyságrendű paraméterszámmal dolgoznak és hasonló nagyságrendű szóból álló szövegkorpuszon tanították be azokat. Napi-heti szinten hozzák ki a nagy gyártók és a kisebb kutatócsoportok az újabb és újabb megoldásaikat, modelljeiket. Számos elérhető modell közül vannak felhő szolgáltatásban elérhetőek és letölthető modellekkel és súlyokkal on-premise telepíthető és üzemeltethető (Sun, Zhang, Chen, Zhang, & Liang, 2007), nyílt forráskódú változatok, amelyek paraméterszáma és tanítókorpuszuk mérete 2-3 nagyságrenddel is eltérhet egymástól. Jelenleg a legnagyobb figyelmet az OpenAI ChatGPT szolgáltatása kapja, amely megjelenése után már két hónappal minden rekordot megdöntve 100 millió felhasználóval rendelkezett (Hu, 2023). A modellek a szöveget token egységekben kezelik, amelyek megfeleltetése nyelvenként eltérő, angol esetén közelít a szavakkal való megfeleltetéshez, magyar nyelv esetén egy szót általában több token reprezentál. Alapvető működésük szerint úgy állítanak elő komplex szöveges tartalmat, hogy az előre betanított neurális hálózat kiértékelésével a mindenkor soron következő legnagyobb valószínűségű tokent helyezik el folytatólagosan. A promptban megadott szövegrészletet egészítik ki, így formálnak választ kérdésre, vagy adnak megoldást egy feladatra. A ChatGPT változatok nem csak az aktuálisan megadott promptot veszik figyelembe, hanem rendelkeznek egy kontextusablakkal, a megelőző kérdés-válasz párokat is felhasználják a tartalom generálásakor. Az adekvát tartalom minőségében nagy jelentősége van a figyelem (attention) funkcionalitásnak, amely eltérő súlyokkal veszi számításba a bemeneti tokeneket (Xu, Liang, Huang, & Xiang, 2021). Ezek tudatában meglepőnek tűnhet, hogy milyen emberszerűen megfogalmazott, értelmes, tartalmas és összetett szövegeket tudnak előállítani, ugyanakkor ez ad magyarázatot a tipikus hibáira, felhasználhatósági korlátaira. A modellek ömagukban az előtanítás miatt hatalmas, de véges adatforrásból dolgoznak. Ebből kifolyólag csak olyan információk alapján tudnak tartalmat előállítani, ami a betanító korpuszuk részét képezte, nem rendelkeznek naprakész tudással, és a kontextus ablakukon kívül a felhasználói aktivitás nem hat automatikusan vissza a tudásbázisukra kellemetlen hatások begyűrűzésének veszélye miatt (Davis, 2016). A nyelvi modell így magától nem képes egy megadott specifikus információ forrás alapján válaszolni kérdésekre. Nem rendelkeznek értelemmel, a szöveg mögött nincs absztrakt fogalmi reprezentáció, nincsenek érzései és normái, emiatt külső korlátozásokat kell beépíteni az általa generálható tartalmak összetételére. A tanító halmaz részét képező, eredetileg emberi forrásból származó szövegminták alapján minden esetben adódik egy mindenkori legvalószínűbb soron következő szó, ezért abban az esetben is előállít látszólag értelmes és koherens szöveget, amikor a kérdezett témáról nincs információja, ez sok esetben valótlan lexikális adatok előállítását okozza. Ezt a jelenséget szokták hallucinációnak hívni és ez volt a modellek széleskörű nyilvános használatba kerülését követő legélesebb és leggyakoribb kritika. Kontextus ablakuk limitált, ami erős határt szab az egy lépésben feldolgozható információ mennyiségének a párbeszédes alkalmazás során. Ezen hiányosságok pótlása jelenleg is nagy erőforrások ráfordításával van folyamatban, és komoly eredmények is születtek. Legnagyobb részük olyan komplex architektúrák építésével orvosolható, ami interfész hívásokkal egészíti ki a modellt és többlépéses feldolgozást biztosít. Így ma már lehetőség van a modellek élő online tartalommal való összekapcsolására, dokumentum elemzésre, egyedi komplex rendszerek építésére. A kész szolgáltatások által még nem megoldott problémák áthidalására több lehetőség kínálkozik. A finomhangolás során prompt-completion párokat “tanítanak rá” a modellre, amelyek hatására a mély neurális

hálózat felső rétegei között szereplő paraméterek módosulnak, ezáltal hatást gyakorolva a generált válaszok adattartalmára és jellegére. Hibrid megoldásokkal kombinálhatók felhő szolgáltatásban elérhető nagyobb modellek kisebb specifikus on-premise modellekkel, valamint több lépésben elő- és utófeldolgozás, ezek segítségével prompt újrafogalmazás, többszörösen előállított és abból kiválasztott válaszképzés, válasz validálás valósítható meg, nagyobb adatmennyiségek darabolást követő iterációkkal vagy hierarchiába szervezett vektorizálással, kontextus tömörítéssel, részenkénti hasonlóságkereséssel kezelhetők. Transzkripció és beszédszintézis hozzáillesztésével olyan élőszavas párbeszéd alakítható ki, amely az ember és gép közötti kapcsolat eddigi legfejlettebb szintjét hozza el.

7. Kutatási eredmény

A kutatásom során megismert információk felhasználásával meg terveztem mutatni, hogy a tágabban értelmezett hangoskönyveknek a mesterséges intelligencia felhasználásával olyan új lehetőséget nyitnak meg az oktatásban, amelyek egyaránt segítséget nyújtanak a tanároknak és a diákoknak a könyvekben szereplő információ befogadása és értelmezése során. A kutatás során elkészült egy minta alkalmazás, melyben két jelentősen eltérő tematikájú és stílusú könyv, Csepeli György: Ember 2.0 és Simon Harris, James Ross: Kezdőkönyv az algoritmusokról című művének tartalma került feldolgozásra. A vékonyklienses megoldás felhasználói felületén a számítógép vagy mobiltelefon mikrofonján keresztül szóbeli kérdéseket tehetünk fel, melyre a könyvek tartalma alapján szóbeli választ kapunk. Az alkalmazás angolul és magyarul automatikusan felismeri a kérdező nyelvét és azon válaszol, miközben felületén az információ forrását is feltünteti. Az volt az elvárás, hogy a program megtalálja a kérdezett információt, helyes választ adjon, és ismerje fel, ha olyan információt kérdezünk tőle, ami nem szerepel a könyvekben. A működést biztosító hibrid szoftver architektúra gradio prototipizáló keretrendszerben készült python nyelvben, Azure felhő szolgáltatáson keresztül TTS és STT szolgáltatásokat, valamint az OpenAI chatGPT interfészét veszi igénybe kiegészítve linux környezetben futtatott LangChain, Weavieate komponensekkel és chromadb adatbázissal. A működéséről egy vágatlan videofelvétel megtekinthető itt: https://youtu.be/Y72JsRD4cQA.

A törekvés sikeres volt, a százas nagyságrendben elvégzett angol és magyar nyelvű tesztek során néhány kivételtől eltekintve elvárt helyes választ adott, helyesen találta meg a rendelkezésre álló információkat, közölte a hiányzó adatok meg nem találhatóságát, és ellenőrző kérdéseket adott kívánt releváns témában. A hanglejtés magyar nyelv esetén kissé természetellenes volt, valamint a válaszidők hosszabbak voltak, mint egy élő szereplős párbeszéd esetén megszokott. A néhány sikertelen tesztet minden esetben az okozta, hogy az artikuláció hiányossága vagy a beszéd-szünet váltások pontatlan érzékelése miatt a transzkripció során elhagyott egy szót vagy szórészt, vagy tévesen ismerte fel a kérdés nyelvét. Következtetéseim szerint a nyelvfelismerés nagyon érzékeny a kiejtésre és a háttérzajra (ez kompenzálható manuális nyelv kiválasztási lehetőséggel), a szünet érzékelési és transzkripciós hibák jelentősen megváltoztathatják a kérdés eredeti szándékát (ez elkerülhető írásos kérdés bevitellel). Továbbfejlesztési lehetőségekként azonosítottam a fejlett zajszűrés alkalmazását, diarizáció beépítését háttérzajból származó másodlagos beszélő leválasztására, szájról olvasás lehetőségét a zaj kompenzálására (Fernandez-Lopez & Sukno, 2018), a válaszidő csökkentésre folyamatos streaming beépítését a folyamatba (TTS megkezdése és hanglejátszás elindítása a teljes válasz elkészülése előtt) és tokenszám optimalizálását, multimodalitás alkalmazását, valamint komplex verbális e-learning keretrendszer köréépítését.

8. Összegzés

A kutatás során nagy nyelvi modell alkalmazásával sikerült a hangoskönyvek alkalmazási területének egy olyan kiterjesztési lehetőségét azonosítanom és megvalósíthatóságát validálnom, ami képes az oktatásban alkalmazható szakmai tartalmak élőszóban interkatívvá tételére és ezáltal segítséget nyújthat a tanároknak a tananyag közvetítésére és a diákoknak a tudásanyag jobb megértésére és könnyebb befogadására. A mesterséges intelligenciára épülő technológiai megoldások jelenlegi fejlődési üteme alapján várhatóan hónapokon belül sokkal pontosabbá és élményszerűbbé tehetők az ehhez hasonló alkalmazások, de a tanárok szerepe továbbra is nélkülözhetetlen (Csepeli, 2020).

Ezúton szeretném megköszönni Gyöngyössy Natabarának a szoftver implementációban és promptképzésben nyújtott segítségét, valamint a Mynds.ai Kft.-nek az infrastruktúra biztosítását.

Irodalomjegyzék

Basak, S., Wotto, M., Bélanger, P. (2018). E-learning, M-learning and D-learning: Conceptual definition and comparative analysis. *Sage Journals*.

Casper, S., Davies, X., Shi, C., Krendl Gilbert, T., Scheurer, J., Rando, J. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv*.

Chen, S.H. (2004). Improving Reading Skills through Audiobooks. *School Library Media Activities Monthly* (21), 22-25.

Cialdini, R. (2009). Hatás. Budapest, HVG Könyvek

Csepeli, G. (2020). Ember 2.0: A mesterséges intelligencia gazdasági és társadalmi hatásai. Budapest, Kossuth

Csíkszentmihályi, M. (1991). Flow: The Psychology of Optimal Experience. New York, Harper & Row

Davis, E. (2016). AI amusements: the tragic tale of Tay the chatbot. *AI Matters*.

Fernandez-Lopez, A., Sukno, F. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*.

Hu, K. (2023. 02 02). ChatGPT sets record for fastest-growing user base - analyst note. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Jin, L., Lee, H.-J., Park, J. (2004). Constructing SSML Documents with Automatically Generated Intonation Information in a Combinatory Categorial Grammar Framework. *International Journal of Computer Processing of Languages*.

Kartal, G., Simsek, H. (2017). The Effects of Audiobooks on EFL Students' Listening Comprehension. *The Reading Matrix: An International Online Journal* (17).

Nash, B. (2023). Attending to the Sounds of Stories: The Affordances of Audiobooks in the English Classroom. *Changing English: Studies in Culture & Education* (30), 99-106.

Lengyelné, M. T. (2020). Future of Libraries in the Cyber-Physical Society. *US-China Foreign Language* (18) 9, 283-290.

Lengyelné, M. T. (2011). A pedagógiai mérés és értékelés feladataira való felkészítés az árnyalt tanulói értékelés módszertanának tükrében. Estefánné, Varga Magdolna (szerk.) Megújuló tananyagtartalmak, módszerek a kompetencialapú tanárképzésben. Eger, Eszterházy Károly Főiskola 83-105.

Serrano, R. (2023). Extensive Reading and Science Vocabulary Learning in L2: Comparing Reading-Only and Reading-While-Listening. *Education Sciences*(13).

Sharma, P. (2010). Blended learning. *ELT Journal*, 456–458.

Shrestha, S., Lenz, K., Owens, J. (2007). "F" Pattern Scanning of Text and Images in Web Pages. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* .

Singh, A., Alexander, P. (2022). Audiobooks, Print, and Comprehension: What We Know and What We Need to Know. *Educational Psychology Review*

Smith, K. (2022). All Ears: An Examination of the Documentality of Audiobooks, Podcasts, and Oral Histories with Extended Research into the London History Workshop Centre Oral History Collection in Collaboration with the Museum of London. London

Sun, W., Zhang, K., Chen, S.-K., Zhang, X., Liang, H. (2007). Software as a Service: An Integration Perspective. *Service-Oriented Computing*

Tóthné, P. L., Lengyelné, M. T., Kis-Tóth, Lajos (2014). Statisztikai programrendszerek, Eger, EKF Líceum Kiadó

Veldhuis, R., He, H. (1998). Time-scale and pitch modi cations of speech signals and resynthesis from the discrete short-time Fourier transform. *Speech Communication*

Xu, P., Liang, D., Huang, Z., & Xiang, B. (2021). Attention-guided Generative Models for Extractive Question Answering. *arXiv*.

Zhang, J.-P. (2008). Hybrid Learning and Ubiquitous Learning. *Lecture Notes in Computer Science*