

Luka Hribar¹

APPLICABILITY OF GENERAL-PURPOSE LANGUAGE MODELS FOR ARCHIVAL DESCRIPTION: PRACTICAL INSIGHTS AND LESSONS LEARNED

Abstract

Purpose: *To test the usability and examine the limitations of general-purpose large language models (LLMs) in archival description. The study was designed as a quantitative/qualitative assessment to monitor trends in this rapidly evolving field.*

Methodology: *The experiment involved testing five AI services on a set of archival records. The set of questions and tasks was divided into two categories: technical tasks (page counting, structure recognition, optical character recognition – OCR) and content-related tasks, such as language detection, content summarization, and title suggestions. Performance was evaluated using quantitative and qualitative methods, along with archivists' assessments.*

Results: *A significant discrepancy was found between the models' performance across different types of tasks. The tested models proved unreliable in seemingly simple technical tasks, such as determining the number of pages or detecting graphical elements, while showing greater utility in complex content-related tasks.*

Discussion: *The analysis highlights that the tested LLMs are currently unsuitable for automating precise technical description processes but represent a useful analytical and generative tool for producing content summaries and descriptions. By observing how AI systems perform, archivists also gain better insight into potential difficulties faced by users.*

Keywords: *Archival records, artificial intelligence (AI), large language models (LLM), digital humanities, archival cataloguing and description, OCR.*

¹ Luka Hribar, študent doktorskega študija Arhivske znanosti na Univerzi Alma Mater Europaea, mail: luka.hribar@almamater.si.

APPLICABILITÀ DEI MODELLI LINGUISTICI GENERICI ALLA DESCRIZIONE ARCHIVISTICA: APPROFONDIMENTI PRATICI E LEZIONI APPRESE

Abstract

Scopo: *testare l'usabilità ed esaminare i limiti dei modelli linguistici generici di grandi dimensioni (LLM) nella descrizione archivistica. Lo studio è stato concepito come una valutazione quantitativa/qualitativa per monitorare le tendenze in questo campo in rapida evoluzione.*

Metodologia: *l'esperimento ha previsto la sperimentazione di cinque servizi di intelligenza artificiale su una serie di documenti d'archivio. La serie di domande e compiti è stata suddivisa in due categorie: compiti tecnici (conteggio delle pagine, riconoscimento della struttura, riconoscimento ottico dei caratteri - OCR) e compiti relativi al contenuto, come il rilevamento della lingua, la sintesi del contenuto e i suggerimenti per i titoli. Le prestazioni sono state valutate utilizzando metodi quantitativi e qualitativi, insieme alle valutazioni degli archivisti.*

Risultati: *È stata riscontrata una discrepanza significativa tra le prestazioni dei modelli nei diversi tipi di attività. I modelli testati si sono dimostrati inaffidabili in attività tecniche apparentemente semplici, come determinare il numero di pagine o rilevare elementi grafici, mentre hanno mostrato una maggiore utilità in attività complesse legate ai contenuti.*

Discussione: *L'analisi evidenzia che gli LLM testati sono attualmente inadatti per automatizzare processi di descrizione tecnica precisi, ma rappresentano uno strumento analitico e generativo utile per produrre sintesi e descrizioni dei contenuti. Osservando le prestazioni dei sistemi di IA, gli archivisti acquisiscono anche una migliore comprensione delle potenziali difficoltà incontrate dagli utenti.*

Parole chiave: *Documenti d'archivio, intelligenza artificiale (IA), modelli linguistici di grandi dimensioni (LLM), discipline umanistiche digitali, catalogazione e descrizione archivistica, OCR.*

UPORABNOST VELIKIH JEZIKOVNIH MODELOV PRI POPIŠOVANJU ARHIVSKEGA GRADIVA: PRAKTIČNI VPOGLEDI IN PRIDOBLENE IZKUŠNJE

Izvleček

Namen: *Preizkus uporabnosti in preverjanje omejitev splošnih velikih jezikovnih modelov (angl. Large Language Models, LLM) pri popisovanju arhivskega gradiva. Raziskava je zasnovana kot kvantitativna/kvalitativna ocena za spremljanje trenda hitro spreminjajočega se področja.*

Metodologija: *Preizkus je zajemal testiranje petih storitev na naboru arhivskega gradiva. Sklop vprašanj oz. nalog je bil razdeljen v dve kategoriji: tehnične naloge, kot so štetje strani, prepoznavanje strukture in optično prepoznavanje znakov (OCR), ter vsebinske naloge, kot so prepoznavanje jezika, povzemanje vsebine in predlaganje naslovov. Uspešnost je bila ovrednotena s kvantitativnimi in kvalitativnimi metodami in oceno arhivistov.*

Rezultati: *Ugotovljeno je bilo pomembno neskladje med uspešnostjo modelov pri različnih tipih nalog. Modeli so se izkazali za nezanesljive pri navidezno preprostih tehničnih nalogah, kot sta določanje števila strani ali zaznavanje grafičnih elementov, medtem ko so dosegli večjo stopnjo uporabnosti pri kompleksnih vsebinskih nalogah.*

Razprava: *Analiza poudarja, da preizkušani modeli LLM trenutno niso primerni za avtomatizacijo natančnih tehničnih popisnih postopkov, predstavljajo pa uporaben analitični in generativni pripomoček za pripravo vsebinskih povzetkov in opisov. Arhivisti z opazovanjem delovanja UI bolje razumemo morebitne težave naših uporabnikov.*

Ključne besede: *Arhivsko gradivo, umetna inteligenca (UI), veliki jezikovni modeli (LLM), digitalna humanistika, popisovanje in opisovanje, OCR.*

1 UVOD

Digitalna transformacija je v zadnjem desetletju korenito preoblikovala delovanje arhivov, knjižnic in muzejev (GLAM sektor). S pospešeno digitalizacijo in nastajanjem obsežnega izvorno digitalnega gradiva (angl. *born-digital*) se ustanove soočajo z eksponentno rastjo podatkov, ki presega zmožnosti tradicionalnih, ročnih postopkov popisovanja in urejanja. V tem kontekstu umetna inteligenca (UI), zlasti veliki jezikovni modeli (ang. Large Language Models, LLM), ponuja potencialne rešitve. Integracija UI v arhivsko prakso, kot je transkripcija, popisovanje in analiza vsebine, je eden izmed osrednjih izzivov in priložnosti, s katerimi se sooča stroka. Razvoj UI obeta tudi odpiranje temnih arhivov (Decker et al., 2022), ki so zaradi pomanjkljivih metapodatkov, občutljivosti vsebine ali neurejenosti, nedostopni javnosti.

Namen te raziskave je izvesti sistematičen preizkus zmožnosti nekaterih splošno dostopnih komercialnih LLM pri izvajanju specifičnih nalog, značilnih za popisovanje arhivskega gradiva. Glavni cilj je bil raziskati v kolikšni meri so te storitve že uporabne za arhiviste in zunanje uporabnike ter identificirati njihove prednosti in omejitve. Poseben poudarek je bil na razumevanju, kje modeli delujejo učinkovito in kje se pojavljajo sistemske pomanjkljivosti, ki bi lahko vplivale na zaupanje in zanesljivost. Končni cilj ni bil natančno identificirati vseh pravilnih in napačnih navedb in določanje zmagovalca, ampak postaviti temelje za spremljanje prihodnjih trendov in ponuditi smernice za vključitev UI v arhivsko prakso.

2 POPISOVANJE ARHIVSKEGA GRADIVA

Namen popisovanja arhivskega gradiva presega zgolj ustvarjanje nekakšnega seznama oz. inventarja. Njegova tradicionalna osrednja naloga je vzpostaviti intelektualna orodja za dostop, ki lahko delujejo tudi kot deskriptivni nadomestki za fizično gradivo (Pezzica 2023). Ta orodja omogočajo uporabnikom odkrivanje in razumevanje arhivskega gradiva s pomočjo natančnega opisa njegove vsebine in konteksta. Arhivsko popisovanje predstavlja ključno nalogo arhivista, saj preoblikuje oz. dopolni gradivo v dostopen vir znanja. S prihodom spleta in ponujanjem digitaliziranega gradiva, se je sicer ta vloga začela tudi spreminjati. Mlajše generacije zunanjih, laičnih uporabnikov, se pogosto slabo zavedajo obstoja nekaterih arhivskih pripomočkov (Hankins, 2019), npr. t. i. arhivskih popisov, ki so včasih

pripeti na višje nivoje (fonde, zbirke in serije), ne pa na nivo same popisne enote. Zaradi jasnosti, se moramo dotakniti tudi terminov popisovanje in opisovanje. Medtem ko popisovanje (v angl. *cataloguing*) nakazuje na strukturiran, predmetni popis, ki se osredotoča na osnovne podatke (npr. obseg, oznake, osnovni podatki, ...), opisovanje (v angl. *description*) zajema bogatejše, narativne opise, ki zagotavljajo nujno potrebne povzetke in kontekst. V okviru arhivske znanosti oz. stroke v slovenščini za ti združeni nalogi največkrat uporabljamo kar termin popisovanje, v angleščini pa včasih izraz *cataloguing and description* in včasih samo *description*.

Da bi zagotovili doslednost in interoperabilnost, se arhivisti po svetu zanašajo na mednarodne standarde, kot so ISAD(G), RIC, ipd. Ti standardi določajo skupen okvir in strukturo za popisovanje arhivskega gradiva ne glede na njegovo obliko (pisna, tiskana, elektronska). Standardi sledijo osnovnim načelom, ki so več-nivojsko popisovanje, upoštevanje relevantnosti informacij na vsaki ravni, povezovanje opisov in izogibanje ponavljanju informacij.

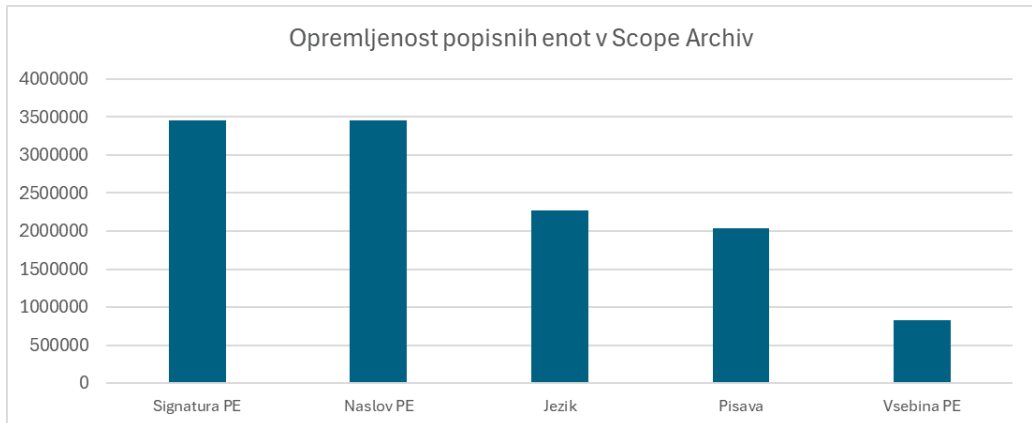
V našem preizkusu od storitev UI nismo zahtevali podajanja informacij v obliki katerega od že omenjenih standardov. Pri postavljanju vprašanj UI pa smo se potrudili, da bi dosegli vsaj razmeroma enostavne preslikave rezultatov v ustrezne kategorije. To nam je tudi omogočilo boljše in enostavnejše ocenjevanje zmožnosti UI pri ustvarjanju popisov, ki bi bili v skladu s strokovnimi zahtevami.

Popisovanje arhivskega gradiva se sicer sooča s številnimi izzivi. Osnovno težavo, ki se jo skuša nasloviti z uporabo UI, je četverček virov: pomanjkanje ustrezno usposobljenega osebja, specialističnega znanja, sredstev in časa (Bingham in Byrne, 2021). Ti dejavniki vodijo v obsežene zaostanke pri obdelavi gradiva v arhivih po vsem svetu. Posledično veliko dragocenega in zanimivega gradiva ostaja nepopisanega ali pa je popisano le na osnovni ravni, kar otežuje odkrivanje in uporabo.

2.1 STANJE POPISOVANJA V SLOVENSКИH ARHIVIH

V tem poglavju so na kratko predstavljeni podatki o stanju popisovanja v slovenskih arhivih. Ti podatki, ki smo jih črpali iz programske opreme Scope Archiv, ki jo uporablja Arhiv Republike Slovenije in regionalni arhivi, prikazujejo obseg zaostankov in pomanjkanje virov, kar še dodatno utemeljuje nujnost raziskav možnosti rabe novih metod in orodij, kot je UI.

V programski opremi Scope Archiv je trenutno (avgust 2025) v vseh slovenskih arhivih skupaj vnesenih 3.455.690 zapisov s signaturo, torej popisnih enot (PE). Naslov PE je vnesen pri 3.449.376 zapisih (99,82 %), jezik(i) so vneseni pri 2.266.651 zapisih (65,59 %), pisava(e) so vnesene pri 2.032.542 zapisih (58,82 %), vsebina PE (opis vsebine) pa je vnesena pri 825.316 zapisih (23,88 %).



Slika 1: Opremljenost popisnih enot v Scope Archiv

Kot lahko vidimo, je polje Vsebina PE izpolnjeno pri manj kot četrtini vseh PE. To ne pomeni, da kjer to polje ni izpolnjeno, ni podatkov o vsebini. Mnogokrat se npr. ti podatki, če gre za dokument ali združene dokumente, nahajajo na višjem nivoju tektonike arhiva, torej pri opisu vsebine serije, zbirke ali celo fonda ali pa so ti podatki v pridruženih iskalnih pripomočkih. A kljub naštetemu, je videti, da je pri opisovanju vsebine še prostor za izboljšave. Opis vsebine je tudi dragocen podatek za vse uporabnike virtualne čitalnice (VAČ), ki se včasih slabo znajdejo pri poizvedbah in slabo poznajo zakonitosti arhivskega popisovanja in dostopanja do podatkov.

3 UMETNA INTELIGENCA V ARHIVISTIKI

3.1 KRATEK PREGLED RAZVOJA UMETNE INTELIGENCE

Širši razvoj UI se je začel s poskusi, ki so temeljili na simboličnih in na pravilih temelječih sistemih, znanih kot ekspertni sistemi. Ti sistemi so bili zasnovani za posnemanje človeškega sklepanja z uporabo vnaprej določenih logičnih pravil in baz znanja. Razvili so se predvsem za reševanje specifičnih problemov, npr. v medicini ali inženirstvu (Brock, 2018).

Zgodnji sistemi so bili primarno zasnovani za obdelavo strukturiranih² podatkov. Z razvojem strojnega učenja se je fokus UI premaknil z vnaprej določenih pravil na statistične metode, ki so omogočale računalnikom, da se učijo iz velikih količin podatkov. V tem okviru se je razvila tudi obdelava naravnega jezika (angl. *Natural Language Processing, NLP*), področje, ki računalnikom omogoča razumevanje, interpretacijo in generiranje človeškega jezika. NLP se je izkazal kot ključna tehnologija za obdelavo nestrukturiranega besedilnega gradiva, kar je v veliki meri odpravilo omejitve, ki so jih imeli zgodnji sistemi (Feng et al., 2024). Revolucionaren preboj v obdelavi naravnega jezika pa se je zgodil z arhitekturo „transformerjev“, ki je omogočila razvoj LLM-jev z več sto milijardami parametrov. Ta arhitektura je šele omogočila modelom učinkovito obdelavo dolgih zaporedij podatkov, kar je prineslo preboje v konverzacijski UI, strojnem prevajanju, analizi in generiranju besedil (Feng et al., 2024). To tudi odpira vrata avtomatizaciji arhivskih procesov kot so transkripcija, popisovanje in analiza vsebine.

3.2 UMETNA INTELIGENCA IN POPISOVANJE

Poleg samega skeniranja, optično prepoznavanje znakov (angl. *Optical Character Recognition, OCR*) predstavlja ključno osnovo za večino arhivskih digitalizacijskih projektov, saj pretvarja slikovno gradivo v strojno berljivo besedilo. Kakovost OCR je neposredno povezana z učinkovitostjo kasnejših postopkov, vključno z luščenjem informacij in iskanjem. LLM-ji nadgrajujejo to osnovo, saj ne zgolj prepoznavajo besedilo, temveč ga tudi na nek način razumejo in analizirajo v kontekstu.

Raziskave o uporabi LLM-jev za transkripcijo zgodovinskih rokopisov navajajo, da le-za dosega bistveno večjo natančnost kot tradicionalna specializirana programska oprema. Vendar pa se pri uporabi LLM-jev pojavljajo tudi težave, zlasti pri nizko zastopanih (angl. *low-resource*) jezikih. Čeprav so nekateri modeli, dosegli izjemne rezultate tudi v nizko zastopanih jezikih, obstajajo težave pri starejšem arhivskem gradivu, ki je napisano v arhaičnem jeziku, ki ni ustrezno zastopan v učnih podatkovnih zbirkah. To lahko vodi v netočnosti pri prepoznavi arhaičnega besedišča in slovničnih struktur, kar v nadaljevanju zahteva v skrbno človeško preverjanje rezultatov (Kaluvilla et al., 2025; Khan et al., 2024; Humphries et al., 2024).

2 Večina arhivskega gradiva, kot so pisma, poročila in drugi dokumenti je nestrukturirana.

Kljub temu sposobnost velikih LLM-jev, da združijo različne koncepte in izraze iz več jezikov, nudi veliko obetov za delo z večjezičnimi arhivskimi fondi. Poleg tega se LLM-ji zelo dobro obnesejo pri povzemanju in prepoznavanju imenovanih entitet, s katerim identificirajo ključne informacije (teme, osebe, kraje, organizacije) (Zhang in Colavizza, 2025). Te lastnosti oz. zmožnosti so temelj za avtomatizirano ustvarjanje arhivskih opisov. Strojno učenje in obdelavo naravnega jezika je mogoče izkoristiti za lažje odkrivanje povezav znotraj obsežnih korpusov, s čimer se izboljša organizacija in kontekstualizacija arhivskega gradiva (Cushing in Osti, 2023). Študije primerov, kjer je bila UI uspešno uvedena v različnih institucijah, nakazujejo na njeno primernost tudi pri avtomatizaciji arhivskih popisov (Arias Hernandez et al., 2024).

Poleg tega integracija tehnologij UI obljublja napredek pri ustvarjanju in izboljšavanju metapodatkov, ki so ključni pri zagotavljanju dostopnosti arhivskega gradiva. UI lahko ponovno analizirajo zapise, ko se npr. dodaja novo gradivo, kar omogoča ponavljajoče se posodobitve in s tem ustreznost (Fan et al., 2020). To je še posebej dragoceno v hitro razvijajočih se situacijah, npr. kriznih razmerah, kjer je potrebno hitro ukrepanje.

Kljub jasnim prednostim UI pri arhivskem opisu, obstajajo tudi pomisleki glede njene uporabe. Učinkovitost UI je v veliki meri odvisna od specifičnega usposabljanja modelov, ki ustrezajo arhivskim kontekstom. Zaradi nepregledne narave številnih algoritmov strojnega učenja je ključna tudi potreba po razložljivosti, zlasti pri določanju konteksta in pomena zapisov (Hou et al., 2022). Ta skrb odražajo raziskave etičnih posledic in morebitnih pristranskosti, vgrajenih v okvire obdelave UI, zlasti glede na zapleteno naravo arhivskega gradiva (Tenzer et al., 2024).

3.3 MEDNARODNI PROJEKTI IN ZDRUŽENJA

V zadnjih letih je nastalo več pomembnih mednarodnih projektov, ki se osredotočajo na raziskovanje, etične vidike in praktično uporabo UI pri upravljanju in dostopnosti digitalne kulturne dediščine. Ti projekti poudarjajo ključno vlogo meddisciplinarnega sodelovanja med računalniškimi strokovnjaki, arhivisti in humanisti. Rešitve se za arhivistiko pogosto razvijajo znotraj specializiranih skupnosti in ne zgolj v okviru velikih pridobitno naravnanih podjetij. V nadaljevanju so na kratko navedene nekatere, ki predstavljajo obsežna izhodišča in vire znanja.

InterPARES Trust AI je eden izmed temeljnih projektov, ki se ukvarja z vprašanji uporabnosti in zanesljivosti UI v arhivistiki. V okviru projekta potekajo študije, kot je npr. *Teachable AI for the Archival Professions*. AEOLIAN (2025) in sorodna mreža AURA (2025) se osredotočata na izziv „temnih“ digitalnih arhivov, ki so zaradi občutljivosti vsebine, avtorskih pravic ali drugih razlogov zaprti za javnost. Z UI podprto metodo poskušajo omogočiti selektivno odpiranje zbirk in izboljšanje dostopnosti, ne da bi bilo potrebno ročno prebirati ogromne količine podatkov. LUSTRE je projekt, ki se osredotoča na izvorno digitalno gradivo. Projekt raziskuje potencialen vpliv UI na te zapise in na delo arhivistov. AI4LAM je odprta skupnostna iniciativa, ki služi kot stičišče za strokovnjake in navdušence.

4. PREIZKUS IN REZULTATI

4.1 NAMEN

Namen tega preizkusa je bil sistematično preveriti zmožnosti nekaterih splošno dostopnih LLM pri popisovanju gradiva Arhiva Republike Slovenije, kar je morda celo prva tovrstna raziskava v slovenskem prostoru. Utemeljitev za to raziskavo izhaja iz naslednjih dejavnikov:

- Gradivo je prepleteno z besedili iz slovanskih, germanskih in romanskih jezikov.
- Slovenščina je jezik, ki ni močno zastopan v velikih jezikovnih modelih.
- Poseben poudarek je na starejšem gradivu, ki ni močno zastopano v učnih zbirkah LLM in je pogosto tudi slabo razumljivo zunanjim uporabnikom.
- Seznanitev arhivistov kako trenutno najbolj priljubljene UI storitve odgovarjajo uporabnikom, da bi bolje razumeli problematiko ter morda prilagodili načine in strategije popisovanja.

4.2 IZBIRA GRADIVA

Izbira gradiva je bila izvedena na podlagi kriterijev, ki omogočajo neodvisne ponovitve raziskave in pravno ter etično nesporno obravnavo. Za preizkus izbrano gradivo je javno dostopno preko virtualne arhivske čitalnice (VAČ) in ne vsebuje zaščitene občutljivih podatkov. S tem se zagotavlja, da lahko drugi raziskovalci ponovijo analizo in preizkusijo druge storitve.

4.2.1 Gradivo preizkusa

Za gradivo preizkusa smo izbrali pet popisnih enot od katerih je imela vsaka kakšno posebnost, ki jo srečamo v realnem svetu (nepopoln ali neobstoječi OCR, mešanje rokopisnega in tiskanega dela, presevanje vsebine z hrbtne strani, različne pisave in jeziki zastarelost jezika oz. terminologije, združevanje različnih PE v eno datoteko, ...).

1. SI AS 730/2/1/1 Odloki, patenti, razglasi, okrožnice od 1725 do 1792 (datoteka: SI_AS_730_2_1_1.pdf).



Slika 2: SI AS 730/2/1/1; prvih 8 strani 34 strani dolge PDF datoteke.

2. SI AS 1073/II-37r Vinogorski red iz leta 1543 (datoteka: SI_AS_1073_495_II-37r_Vinogorski_redi_iz_leta_1543_(Gorske_Regelze_al_Artikelni ...)-18.stol-1.pdf).



Slika 3: SI AS 1073/II-37r; prvih 8 strani 63 strani dolge PDF datoteke.

3. SI AS 1080/I/2,3,7 (3 datoteke so bile združene v eno): SI AS 1080/I/2 Kapucinski provincial na Štajerskem, Koroškem in Kranjskem Silvester de Polcenico sprejme Burkharda Hitzinga [Hitzingkh] in njegovo ženo Sidonijo med duhovne otroke kapucinov, 1625 (datoteka: SI_AS_1080_I_2.pdf); SI AS 1080/I/3 Papež Inocenc X. podeli odpustke bratovščini sv. Rešnjega telesa v župnijski cerkvi sv. Egidija v Višnji Gori, 1647 (datoteka: SI_AS_1080_I_3.pdf); SI AS 1080/I/7 Kranjski deželni glavar Janez Gašper grof Cobenzl razsodi v zadevi zapuščine po pokojnem Francu Engelbrehtu pl. Zetschkerju, 1718 (datoteka: SI_AS_1080_I_7.pdf).



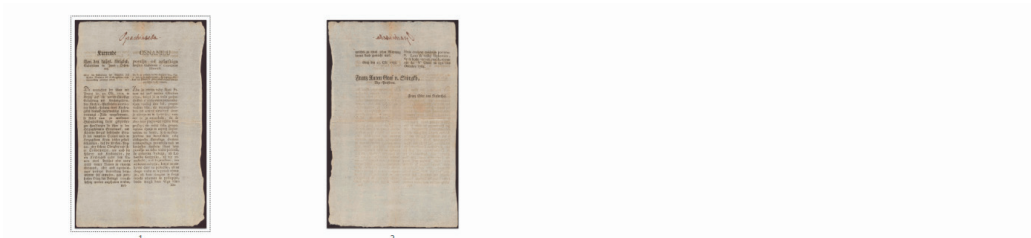
Slika 4: SI AS 1080/I/2,3,7; vse strani PDF datoteke.

4. SI AS 2048/HR DARI/2 Zadeve mesta Reka (Rijeka), 1560-1714 (datoteka: SI_AS_2048_HR_DARI_2_273_08.pdf).



Slika 5: SI AS 2048/HR DARI/2; vse strani datoteke PDF datoteke.

5. SI AS 2048/IV/1 Okrožnica Notranjeavstrijskega gubernija v Gradcu, s katero se zapoveduje, da cerkveni upravitelji ne smejo samovoljno ravnati s cerkvenim denarjem, 1788 (datoteka: SI_AS_2048_IV_1_in_Okroznica_in_1788.pdf).



Slika 6: SI AS 2048/IV/1, vse strani PDF datoteke

6. SI AS 2058/6 Stenografske beležke Senata Kraljevine Jugoslavije, redni sklic, 16. rednega sestanka do 27. rednega sestanka, od 14. februarja 1933 do 10. marca 1933 (datoteka: II-241_1933_knjiga_II.pdf).



Slika 7: SI AS 2058/6; prvih 8 strani 148 strani dolge PDF datoteke.

4.2.2 Pred obdelava gradiva

Pred obdelava gradiva je vključevala naslednje korake:

- Odstranjevanje nekaterih strani npr. skenov platnic z nalepkami in drugimi pojasnili, ki bi lahko zmotili postopek prepoznavanja in analize.
- Ločljivost nekaterih večjih datotek je bila zmanjšana na raven, ki še vedno omogoča odlično optično prepoznavo besedila, a je omogočala hitrejšo obdelavo.
- Iz imen datotek so bili odstranjeni vsebinski podatki, ohranjena je bila le oznaka signature. S tem se smo dosegli preizkus sposobnosti LLM, da vsebino in kontekst gradiva določijo izključno na podlagi samega gradiva, šele nato lahko za podatki poberskajo po spletu³.

³ Spletišče VAC – Virtualna arhivska čitalnica je opremljeno s posebnimi navodili, ki spletnim pajkom, ki so sestavina velikih iskalnikov, prepovedujejo brskanje. Če se teh navodil držijo tudi pajki, ki pripravljajo vsebine za učenje LLM-jev, ni znano.

4.3 IZBIRA IN PRIMERJAVA STORITEV

Za testiranje so bile izbrane storitve, ki so trenutno priljubljene in pogosto že uporabljane s strani splošne javnosti:

- **OpenAI ChatGPT 5** (v nadaljevanju CGPT5): Ta storitev temelji na modelu GPT 5, predstavljenem sredi leta 2025. Ponuja napredno obdelavo, ki naj bi vključevala tudi izboljšano „razumsko“ sposobnost za analizo kompleksnih dokumentov in manj halucinacij pri interpretaciji vsebin. Podpira obdelavo besedila, slik in drugih vnosov za naloge, kot so povzemanje in kontekstualna analiza. Dostopno na: <https://chatgpt.com/>
- **Microsoft Copilot** (v nadaljevanju COPLT): UI asistent omogoča kontekstualno razumevanje in avtomatizacijo obdelave dokumentov, vključno z analizo podatkov in povzemanjem vsebin v orodjih, kot so npr. Word in Excel. Temelji na tehnologiji podjetja OpenAI, ki jo podjetje Microsoft dopolnjuje z lastnimi modeli in rešitvami. Dostopno na: <https://copilot.microsoft.com/>
- **Google Notebook LM** (v nadaljevanju GNLM): UI orodje za raziskave, ki analizira vire (PDF-ji, spletne strani, ...) ter generira povzetke, vprašanja z navedbami, generira avdio pogovore in miselne zemljevide za boljše razumevanje gradiva. Osredotoča se predvsem na uporabniško naložene vire. Dostopno na: <https://notebooklm.google.com/>
- **Google AI Studio** (v nadaljevanju GAIST): Platforma za gradnjo multimodalnih UI aplikacij z modeli, kot sta Gemini 2.5 Pro in Gemma, ki omogoča generiranje in obdelavo vsebin iz besedila, slik in drugih vnosov. Storitve poudarja varnostne funkcije za odgovorno uporabo pri analizi in podajanju rezultatov. Dostopno na: <https://aistudio.google.com/>
- **Google Gemini** (v nadaljevanju GGEM): Multimodalni model z različicami, kot sta 2.5 Flash in 2.5 Pro. Vključuje funkcije za globoko raziskovanje in integracijo z orodji za spletno brskanje in drugimi storitvami podjetja Google. Večkrat se ga omenja kot na Googlov odgovor na storitve podjetja Open AI. Dostopno na: <https://gemini.google.com/>

Te storitve so bile izbrane tudi zaradi predhodnih izkušenj, ki so pokazale, da so zmožne samostojno izvesti OCR, če ga datoteka še nima⁴. Vsaka od teh storitev ima različne tehnične značilnosti, vsem pa je skupno da so multimodalne (spreje-

4 Zaradi tega razloga so bile izpuščene nekatere ostale popularne storitve.

majo besedilne, slikovne in druge vnose), podpirajo množico jezikov in pri optični prepoznavi besedila podpirajo tudi rokopis, kar je posebej koristno za obdelavo arhivskega gradiva. Omenjene tri Googlove storitve (GNLM, GAIST in GGEM) temeljijo na podobni osnovni tehnologiji, kot je model Gemini, vendar so zasnovane za različne potrebe: GNLM za raziskovalno analizo, GAIST za razvoj aplikacij in GGEM za splošno interakcijo in ustvarjanje, zato smo preizkusili vse tri.

4.4 POTEK PREIZKUSA

V tem preizkusu bistvo ni bilo v natančni merljivosti pravih in napačnih odgovorov, saj se modeli razvijajo izjemno hitro in bi bila takšna dolgotrajna, natančna meritev, že ob objavi zastarela. Namesto tega je bil poudarek na kvantitativno/kvalitativni oceni in vzpostaviti referenčnega izhodišča, ki bo omogočal spremljane trendov primernosti in uspešnosti v prihodnosti. Morebitne ponovitve preizkusov s prihodnjimi generacijami modelov bodo torej pokazale ali se storitve izboljšujejo pri reševanju specifičnih arhivskih izzivov.

Da bi zagotovili konsistentnost in ponovljivost eksperimenta, so bili za vse storitve uporabljeni enaki protokoli in vprašanja. Proces se je začel z skrbno premišljenim pozivom⁵. Vsakemu LLM je bilo najprej posredovano splošno navodilo, ki je nakazovalo pričakovanja in določalo, da naj model ustvari kratke, jasne in informativne odgovore. Sledila so specifična vprašanja, ki so usmerjala LLM k pridobivanju točno določenih podatkov, ki so pomembni za arhivski popis. S tem so bili izhodi usmerjeni v ustvarjanje popisnih elementov, ki jih je mogoče brez večjih zadreg preslikati v podatkovno shemo kot je npr. ISAD(G).

Naloge so bile razdeljene v dva sklopa:

- **Tehnične naloge (1-6):** Določanje imena datoteke, števila strani, prisotnosti besedila/slik, ocena in izvedba OCR.
- **Vsebinske naloge (7-16):** Prepoznavanje jezika in pisav, povzetek vsebine, predlaganje naslova, ekstrakcija ključnih entitet (osebe, kraji, časovna obdobja). Uspešnost pri vsebinskih nalogah je bila ocenjena s strani arhivistov. Ocene so bile razdeljene v dve kategoriji:

⁵ Inženiring pozivov (angl. *prompt engineering*) je nova večšina pri uporabi UI. Določa, kako se modeli odzivajo na navodila.

A. Pomoč arhivistu pri popisovanju gradiva.

B. Pomoč zunanjemu, neveščemu uporabniku.

Pri vsebinskih nalogah smo uporabili lestvico: (1) – nikakor, (2) – malo, (3) – dokaj in (4) – zelo.

Pri vsaki izmed storitev smo zahtevali uporabo najbolj sodobnega modela, ki je bil na voljo. Če je bilo mogoče smo vklopili načina za poglobljeno razmišljanje in storitvi dovolili brskanje po spletu. Nekatere storitve dovolijo te izbire, nekatere ne.

Vsaka izmed storitev je v povezavi s posamezno datoteko dobila naslednji poziv. V ležeči pisavi so navedeni tudi naši razmisleki zakaj.

----- ZAČETEK POZIVA -----

Splošno navodilo:

- Podrobno razišči, če lahko.

To navodilo spodbuja storitev k uporabi t. i. poglobljenega razmišljanja.

- Pomagaj si z iskanjem preko spleta, če lahko in če je potrebno.

To navodilo spodbuja storitev, k oblikovanju odgovora tudi s pomočjo informacij, ki jih pridobi na spletu.

- Uporabi agente, če lahko in je potrebno.

To navodilo spodbuja storitev, da za izvedbo neke naloge pokliče t. i. agenta, storitev npr. za prevod, prepoznavo, ...

- Če o svojih navedbah nisi prepričana, napiši opozorilo. Želimo informacije, a želimo tudi vedeti ali gre pri odgovorih za ugibanje ali gre za jasno razpoznane in zanesljive navedbe.

Storitve smo poskušali pozvati k nekakšni samoocentivi oz. kritičnosti do podanih odgovorov.

- Pri iskanju po spletu se NE omejuj na slovenske spletne vire, pomagaj si s prevodi ključnih besed in izrazov, če je potrebno.

Iz preteklih preizkušanj smo ugotovili, da storitve, če je poziv zapisan v slovenščini, za dodatnimi informacijami potem tudi iščejo zgolj po spletnih vsebinah v slovenskem jeziku. Besedo „ne“ smo namenoma poudarili.

- Na koncu obvezno navedi spletne vire iz katerih si črpala informacije. Vse spletne vire navedi skupaj s povezavami (linki).

Storitve UI, ki temeljijo na LLM, so glede virov informacij pogosto zelo netransparentne. S to zahtevo smo jih skušali prepričati k ustreznemu delovanju pri

izdelavi načrta akcij (da naj bo vanj vključeno tudi navajanje virov) v fazi poglobljenega razmišljanja.

- Na koncu ne ponujaj nadaljnjih predlogov.

S to zahtevo smo se želeli izogniti nadaljnjim vprašanjem po prvem odzivu storitve. V preizkusih se nismo spuščali v interakcije s storitvami.

Naloga:

V povezavi s pripeto datoteko navedi naslednje:

1. Navedi ime pripete datoteke.
2. Navedi število strani pripete datoteke.
3. Navedi na katerih straneh je besedilo, če je v datoteki besedilo. Navedi številke strani, na katerih ni besedilnih elementov.
4. Navedi na katerih straneh so slike ali fotografije, če so v datoteki slike ali fotografije.
5. Ali je v datoteki besedilo, ki je bilo razpoznano z OCR? Oceni kvaliteto OCR v datoteki.
6. Napravi OCR, če je potrebno.
7. Navedi jezike v katerih je napisano besedilo, če jih razpoznaš.
8. Navedi pisave, v katerih je napisano besedilo, če jih razpoznaš.
9. Ali datoteka predstavlja en dokument ali je gre za več dokumentov? Koliko, če jih je več?
10. V nekaj odstavkih predstavi vsebino datoteke.
11. Predlagaj naslov oz. ime, ki bi bilo primerno za datoteko, če bi bila namenjena zgodovinarju.
12. Navedi časovna obdobja ali letnice, ki se omenjajo v datoteki (za vsak dokument posebej, če domnevaš, da je v datoteki več dokumentov).
13. Navedi osebe, ki se omenjajo v datoteki (za vsak dokument posebej, če domnevaš, da je v datoteki več dokumentov).
14. Navedi ustanove, podjetja in organizacije, ki se omenjajo v datoteki (za vsak dokument posebej, če domnevaš, da je v datoteki več dokumentov).
15. Navedi kraje ali lokacije, ki se omenjajo v datoteki (za vsak dokument posebej, če domnevaš, da je v datoteki več dokumentov).
16. Navedi morebitne nenavadnosti ali zanimivosti, ki si jih opazila v datoteki.

Za vsako od točk (1-16) navedi ali si jo naredila (DA/NE), in kako močno si prepričana v pravilnost svojih trditev (v odstotkih %).

Na koncu ne ponujaj nadaljnjih predlogov.

----- KONEC POZIVA -----

4.5 REZULTATI PREIZKUSA

Odgovore vseh storitev smo za vsako od preizkušenih datotek kopirali v en dokument, katerega obseg je narastel na 168 strani oz. skoraj pol milijona znakov s presledki. Ker smo povsod, kjer je bilo pri storitvah mogoče vklopili načina razmišljanja, so nekatere storitve potrebovale tudi do 15 minut, da so pripravile odgovor. Pri storitvi GNLM smo zastavili samo vprašanja od 1–16 saj celotnega poziva storitev ni želela sprejeti.

Dokument smo delno uredili in rezultate kar najbolj smiselno poskušali povzeti v tabelarično obliko. Pri povzemanju, ki se je izkazalo za izjemno časovno zahtevno opravilo, ni šlo brez zadreg, saj včasih storitve ne vrnejo zgolj odgovora, ki smo ga zahtevali, ampak zraven dopišejo še kakšno zanimivost, ki razširja odgovor, a ga včasih naredi tudi dvoumnega ali pa so dodatne navedbe napačne. Rezultate navajamo v nadaljevanju skupaj z manjšimi sprotnimi komentarji.

4.5.1 Navedi ime pripete datoteke

To vprašanje nas je zanimalo, ker smo želeli preveriti ali določevanje tega parametra, ki je pomemben predvsem pri sklicevanju, storitvam povzroča probleme. Če UI obdeluje zgolj eno datoteko naenkrat, problemov pri sklicevanju ni, če pa obdeluje sveženj datotek, je zelo pomembno, da pri sklicevanju in pojasnjevanju vedno navede pravilno ime datoteke.

Tabela 1: Odgovori na vprašanje o imenu datoteke.

Datoteka	CGPT5	COPLT	GNLM	GAIST	GGEM
SIAS730/2/1/1	Pravilno	Pravilno	Pravilno	NP	Pravilno
SIAS1073/II-37r	–	Pravilno	Pravilno	NP	Pravilno
SIAS1080/1/2,3,7	Pravilno	Pravilno	Pravilno	NP	Pravilno
SIAS2048/HRDARI/2	Pravilno	Pravilno	Pravilno	NP	Pravilno
SI AS 2048/IV/1	Pravilno	Pravilno	Pravilno	NP	Pravilno
SI AS2058/6	Pravilno	Pravilno	Pravilno	NP	Pravilno

Opazanja: Storitve CGPT5 datoteke povezane s SIAS1073/II/-37r ni uspela sprejeti v obdelavo. Vsakič, ko smo jo poskušali naložiti je storitev sporočila „Iz te datoteke ni bilo mogoče izvleči nobenega besedila“. Zato je v rezultatih te kom-

binacije v vseh naslednjih tabelah navedeno „–“. Storitvev GAIST pa je pri vseh datotekah vedno odgovorila: „Ime datoteke ni bilo neposredno navedeno v zahtevku, vendar se nanaša na digitalno datoteko, ki vsebuje skenirane zgodovinske dokumente“, v tabeli smo to označili z „NP“.

Splošne skupne ugotovitve glede zastavljenega vprašanja: Na videz enostavna naloga je pokazala da včasih LLM ne dobi podatka o imenu datoteke. To dejstvo je pomembno upoštevati predvsem v primerih, ko je naslov datoteke eden ključnih podatkov gradiva. Takšen primer so npr. izpiski raznih meritev senzorjev, kjer datoteke niso nujno opremljene z ustrezno glavo.

4.5.2 Navedi število strani pripete datoteke.

To vprašanje nas je zanimalo, ker smo želeli preveriti ali določevanje tega parametra, ki je pomemben predvsem pri ocenjevanju oz. določanju obsežnosti gradiva, storitvam povzroča težave.

Tabela 2: Odgovori na vprašanje o številu strani datoteke.

Datoteka	CGPT5	COPLT	GNLM	GAIST	GGEM
SIAS730/2/1/1	34	45	68	34	34
SIAS1073/II-37r	–	61	63	63	18
SIAS1080/1/2,3,7	6	5	6	6	6
SIAS2048/HRDARI/2	4	NP	4	4	4
SI AS 2048/IV/1	2	1	2	2	2
SI AS2058/6	148	NP	„strani od 7-134“	55	134

SIAS730/2/1/1: Število strani datoteke je 34. Nekateri skenogrami vsebujejo tako levo, kot desno stran. V preteklosti je nekdo s svinčnikom dopisal paginacijo. Številke si sledijo od 1-45. Stran 7 je enaka strani 6, le da je nalepljen pečatni del razprt in nima ročno dopisane paginacije. **Opažanja:** Naštete značilnosti datoteke so večkrat močno zmotile UI pri določanju rezultata. Vprašanje je bilo sicer jasno zastavljeno, zanimalo nas je število strani datoteke in ne oštevilčenje strani. COPLT je navedla najvišjo označbo strani (45), kar lahko le pogojno vzamemo kot pravilno. GNLM je pojasnila da je število strani 68 določila glede na „preštevanje izvlečkov vseh virov“ in za to trditvev ne najdemo nekega razumnega pojasnila.

SIAS1073/II-37r: Število strani datoteke je 63. Pri tej datoteki smo sicer še pred preizkusom odrezali prvo in zadnje tri strani, ki so vsebovale neželene elemente

(platnice, nalepke), ki jih storitvam nismo želeli pokazati. **Opazanja:** Tudi pri tem dokumentu je nekdo v preteklosti s svinčnikom dopisal številke strani (od 1 do 61, naša odrezana datoteka se začne s paginacijo 1). Posebnost tega dokumenta so vmesne strani, ki niso ročno oštevilčene in na eni izmed njih se nahaja star notni zapis. Tudi tu so našteje značilnosti datoteke zmotile UI pri določanju rezultata. Nekatero storitve so zaupale paginaciji, nekatere dejanskim PDF stranim. GGEM je navedla zelo napačen podatek in ne poznamo razloga zakaj.

SIAS1080/1/2,3,7: Število strani datoteke je 6. Posebnost te datoteke je, da gre za tri, v eno datoteko združene popisne enote. **Opazanja:** Razen v enem primeru, so storitve vrnilo pravilno vrednost. Zakaj je COPLT navedla eno stran manj, ni znano oz. pojasnjeno.

SIAS2048/HRDARI/2: Število strani datoteke je 4. Posebnost te datoteke je, da sta strani 2 in 3 prazni, a na njih preseva besedilo s hrbtni strani. **Opazanja:** Storitve so se pri tej štiri stranski datoteki dobro odrezale. Samo COPLT ni znala navesti tega podatka.

SI AS 2048/IV/1: Število strani datoteke je 2. Posebnih drugih značilnosti, ki bi utegnili vplivati na določanje števila strani nismo opazili. **Opazanja:** Storitev COPLT, ki je edina vrnila napačen rezultat, je sicer navedla, da je v to prepričana zgolj 70 %.

SI AS2058/6: Število strani datoteke je 148. Gre za tiskano in vezano gradivo. Prvih 10 strani predstavljajo naslovne strani in kolofoni. Sledi kazalo, ki je oštevilčeno z rimskimi številkami. Med temi uvodnimi stranmi je tudi nekaj praznih. Nato sledi vsebina (zapisniki), pri katerih so strani oštevilčene levo oz. desno zgoraj od 1 do 133. Na koncu datoteke so 4 popolnoma prazne strani. **Opazanja:** Storitev COPLT zaradi velikosti datoteke, ki je ni bilo možno še bolj zmanjšati, ni uspela obdelati datoteke. Storitev GNML je navedla da „priloženi izseki zajemajo strani 7 do 134 izvirnega dokumenta, kar ni v skladu z jasno zastavljenim vprašanjem in tudi napačno. Storitev GAIST je napisala popolnoma napačen podatek, storitev GGEM pa je zaupala napisani paginaciji.

Splošne skupne ugotovitve glede zastavljenega vprašanja: Na videz enostavna naloga je pokazala, da štetje strani, četudi zapovemo štetje strani PDF datoteke, testirani LLM-ji ne razumejo enoznačno. Opazamo zadrege, kadar so strani oštevilčene. Nekaterih rezultatov ni mogoče pojasniti, morda so napake pri največji

datoteki povezane z obsežnostjo (tako velikosti v bajtih, kot število strani). Zanimivo je, da so nekatere storitve število strani navedle z besedo. Če bi torej želeli, da storitve vedno vračajo nek podatek v številski obliki (npr. pri avtomatiziranih postopkih), bi jim očitno morali to posebej naročiti.

4.5.3 Navedi na katerih straneh je besedilo, če je v datoteki besedilo. Navedi številke strani, na katerih ni besedilnih elementov

To vprašanje nas je zanimalo, ker smo želeli preveriti ali storitve lahko izvedejo razpoznavo zgradbe dokumenta. To arhivistu pomaga pri vprašanjih ki se tičejo obsega besedila in razmerja med besedilom in slikovnim gradivom in je pereče pri gradivu, kjer OCR ni bil napravljen in je težko ugotoviti količino besedila v znakih in posledično določiti potrebne vire za nadaljnje obdelave. Tretje v seriji vprašanj je pravzaprav razširitev prvih dveh, s katerimi smo želeli preveriti uporabnost teh storitev pri določanju strukture in oblik skeniranega gradiva.

Tabela 3: Odgovori na vprašanje glede prisotnosti besedila.

Datoteka	CGPT5	COPLT	GNLM	GAIST	GGEM
SIAS730/2/1/1	„Vse“	„Vse“	„Vse od 1–68“	„Vse“	„Na večini“
SIAS1073/II-37r	–	„Vse“	„Večina“	1–42, 45–63	1–10
SIAS1080/1/2,3,7	NP; „Ni strojno berljivega besedila“	„Vse“	1, 3, 5 obsežnejše; 2,4,6 koščki	1, 3, 5 obsežnejše; 2,4,6 koščki	„Vse“ in kratka pojasnila vsebine
SIAS2048/HRDARI/2	NP, „Ni digitalno prepoznanega besedila“	NP; „Ni plasti z OCR“	1, 2, in 4	Besedilo na 1, 4; 2 in 3 sta prazni	Besedilo na 1, 4; 2 in 3 sta prazni
SIAS2048/IV/1	„Obe strani“	1	„Obe strani“	„Obe strani“	„Obe strani“
SIAS2058/6	„Na večini“; nato so vse prazne strani pravilno navedene	NP	„Ni praznih strani“	„Prazne ali zgolj z žigi so strani 2, 4, 6, 10“	„Prazne so strani 2, 4, 6, 8“

SIAS730/2/1/1: Gre za skenirano gradivo. Besedilo se pojavlja na vseh straneh datoteke. Strani 1–4, 6–7, 22, 27–28 in 34 se glede pisave pomembno razlikujejo od preostanka dokumenta. **Opažanja:** COPLT navaja, da je na vseh straneh besedilo in v kratkem pojasnilu tudi da „ne more navesti strani brez besedila, ker manjka paginacija“. GNLM narobe navaja, da je besedilo na vseh straneh od 1–68 (skladno sicer z napačnim odgovorom na vprašanje 1), pri pojasnjevanju pa pravilno ugotavlja, da obstajajo strani s tabelami (čeprav tega nismo vprašali). GGEM navaja, da je besedilo na večini strani in navede, da besedila ni na straneh 17, 22, 25, 27 in 28, kar je sicer napačno, a na teh straneh je res zelo malo besedila.

Datoteka, ki se začne z rokopisnim delom (brez OCR) in nadaljuje s tiskanim (ki ima OCR) je očitno trd oreh glede razpoznavanja strukture.

SIAS1073/II-37r: Vse strani te PDF datoteke vsebujejo besedilne elemente. Stran 43, ki prikazuje star notni zapis vsebuje besedilo samo na nalepki v levem kotu zgoraj. Stran 44 datoteke je glede razpoznave prisotnosti besedila zelo zahtevna saj je besedilo obrnjeno za 90 stopinj in na robovih. **Opazanja:** Storitvev CGPT5 ni podala podatka, ker ni uspela sprejeti datoteke. Storitvev GNML je identificirala stran 43 kot nenavadno stran z notnim zapisom in stran 44 kot stran s sledovi vezave (oboje pravilno). Storitvev GAIST je navedla stran 43 z notnim zapisom in opozorila na večinoma prazno stran 44. Navedla je tudi da strani 1, 14, 23 in 45 predstavljajo začetke posameznih dokumentov (pravilno). Storitvev GGEM se je edina res močno motila, navedla je, da je besedilo na straneh 1-10, strani od 11–18 pa naj bi bile prazne in to naj bi tudi pomenilo, „da je bil rokopis del večjega zvezka, kjer so strani ostale nepopisane“ (napačno).

SIAS1080/1/2,3,7: Datoteka vsebuje 6 strani. Na vseh je besedilo. Ker gre za združitev treh popisnih enot je glavna vsebina na straneh 1, 3 in 5, strani 2, 4 in 6 vsebujejo samo nekaj koščkov besedila. **Opazanja:** Storitvev CGPT5 je navedla, da v datoteki ni strojno berljivega besedila (kar je pravilno), a tudi ni navedla, da besedilo na teh straneh obstaja. Ostale storitve so pravilno odgovorile in tudi pojasnile zgradbo.

SIAS2048/HRDARI/2: Besedilo te štiri stranske datoteke je na prvi in zadnji strani (1 in 4). Na zadnji strani je rotirano za 90 stopinj. Prazni strani 2 in 3 močno presevata besedilo s hrbtne strani. **Opazanja:** Storitvi CGPT in COPLT sta ugotovili, da OCR ni izveden in v zvezi s tem vprašanjem nista dali jasnega odgovora. Storitvev GNLM je našla besedilo na 2 strani (napačno), morda jo je zmotilo močno presevanje s hrbtne stani. Samo storitvi GAIS in GGEM sta pravilno ugotovili na katerih straneh je besedilo.

SIAS 2048/IV/1: Datoteko sestavljata dve strani z besedilom v dveh stolpcih. Levi stolpec je v nemščini, desni stolpec v slovenščini in predstavlja prevod levega stolpca. Gradivo je tiskano, jezik in različni pisavi sta pogojeni času in običajem. **Opazanja:** Storitvev COPLT je edina navedla napačen podatek, da je besedilo le na eni strani. Ugotovitev je v skladu z odgovorom o številu strani, kjer je tudi tu navedla zgolj eno stran.

SI AS2058/6: Datoteka vsebuje 148 strani. Na strani 2 je naknadno dodan žig in signirne oznake, popolnoma prazne pa so strani 4, 6, 10, 18, 28, 56 66, 96, 116 ter 145–148. **Opazanja:** Samo storitev CGPT5 je z odliko opravila nalogo identifikacije praznih strani. Za COPLT je bila datoteka prevelika. Vse ostale storitve so dale napačne ali pomanjkljive rezultate. Izid je presenetljiv, saj je to edina datoteka na preizkusu, ki je imela v popolnosti izveden OCR in tako slab rezultat ni bil pričakovan.

Splošne skupne ugotovitve glede zastavljenega vprašanja: Tudi ta na videz enostavna naloga je mnogim storitvam predstavljala trd oreh. Morda je storitvam pri zadnji datoteki ponagajala obsežnost gradiva, ki je morda preseгла vhodno zmogljivost LLM, ali pa storitve niso uspele poklicati ustreznih t. i. agentov za opravilo naloge. A vsaj ena izmed storitev jo je opravila z odliko. Upajmo, da je zgolj vprašanje časa, kdaj ji bodo sledile druge.

4.5.4 Navedi na katerih straneh so slike ali fotografije, če so v datoteki slike ali fotografije

To vprašanje nas je zanimalo iz podobnih razlogov, kot predhodno. Obstajajo namreč situacije, ko je še posebej pri dolgih dokumentih, koristno če vemo ali vsebujejo tudi slikovno gradivo in koliko. Tako lahko ocenimo potrebne postopke pri nadaljnji obdelavi gradiva, recimo njegovi pripravi za ljudi z okvarami vida, ko je potrebno gradivo opremiti z metapodatki za bralne pripomočke ali kako drugače interpretirati.

Tabela 4: Odgovori na vprašanja glede slikovnega gradiva.

	CGPT5	COPLT	GNLM	GAIST	GGEM
SIAS730/2/1/1	„Ni fotografij ali ilustracij“	„Ni fotografij, ilustracij ali figur“	„Ni slik ali fotografij“; „okrasne črte, simboli in pečati“	„Na nekaterih straneh so grafični elementi“	„Obstajajo grafični elementi“
SIAS1073/II-37r	–	„Možni žigi in okrasni elementi“	„Ni slik ali fotografij“	„Ni slik ali fotografij“; „vidne so kaligrafske linije in okraski“; „stran 43 vsebuje notni zapis“	„Ni slik ali fotografij“
SIAS1080/1/2,3,7	„Vse strani so digitalizirani posnetki (slike)“; „dodane so kalibracijske kartice“	„Vse strani so skenirane barvne slike“; „dodana so barvne karte“	„Vsaka stran je fotografija dokumenta z barvno karto“	„Datoteka je sestavljena iz fotografij dokumentov“; „vidne so okrasne začetnice in žigi“	„Prisotne so tehnične oznake in žig na več straneh“

SIAS2048/ HRDARI/2	„Vse strani so skeni dokumentov; „verjetno slika na str. 4“	„Ni razvidno ali vsebuje slike ali ne“	„Slike skeniranih dokumentov“; „ni ločenih slik“	„Skeni dokumenta“; „ni dodatnih slikovnih elementov“	„Ni slik ali fotografij v sodobnem pomenu“
SIAS2048/IV/1	„Ni ločenih slik ali fotografij“	„Fotografij ni“; „morda okrasni znaki“	„Datoteka je sestavljena iz slik dokumenta“; „ni ločenih fotografij“	„Celotna datoteka je skenogram“; „viden grafični okras na prvi strani“	„Dokument je izključno besedilne narave“
SIAS2058/6	„Skenirane strani besedila“	NP	„Besedilo ne vsebuje sklicev na slike ali fotografije“	„V datoteki ni fotografij, prisoten je grb KJ in uradni žig“	„Datoteka ne vsebuje slik ali fotografij“

SIAS730/2/1/1: Datoteka ne vsebuje fotografij, med slikovno gradivo bi morda lahko šteli žiga na 7 in 22 strani (v slepem tisku) in podobo ptice na 13 strani. Okrasne iniciiale so sicer prave male umetnine. **Opazanja:** Trditve vseh storitev so glede na nerodno zastavljeno vprašanje (za pojasnila glej konec razdelka) pogojno pravilne. Storitve GAIS je našla grafične elemente, a jih je narobe identificirala (okrasne iniciiale je navedla kot grbe, oznake LS v krogu (mesto predvidenega pečata) pa kot pečate. Storitve GGEM je tudi zaznala grafične elemente in glede na sicer nepopolne navedene številke strani gre verjetno za okrasne iniciiale, oznake LS ali dele besedila. Nobena izmed storitev ni opozorila na podobe ptice na 13. strani.

SIAS1073/II-37r: Datoteka vsebuje žiga na strani 14, okrasno črto na strani 22, žiga na strani 23, notni zapis in znamko z rokopisom na strani 43, okrasno črto in žiga na strani 45 ter okrasno črto na strani 54. Razen že omenjenih okrasnih črt so grafični elementi dodani naknadno in niso del originalnega gradiva. **Opazanja:** Trditve storitev so glede na nerodno zastavljeno vprašanje (za pojasnila glej konec razdelka) pogojno pravilne. Pri tej popisni enoti se je najbolje odrezala storitev GAIST, ki je v povezavi s tem vprašanjem edina opozorila na notni zapis na strani 43, ki ga je možno v nekaterih pogledih razumeti kot grafični element.

SIAS1080/1/2,3,7: Datoteka razen okrasnih elementov besedila in naknadno dodanih žigov, ne vsebuje grafičnih elementov. Na vseh straneh je na vrhu dodana kalibracijska barvna kartica. **Opazanja:** Trditve storitev so glede na nerodno zastavljeno vprašanje (za pojasnila glej konec razdelka) pogojno pravilne.

SIAS2048/HRDARI/2: Datoteka ne vsebuje nobenih grafičnih elementov. **Opazanja:** Vse storitve razen COPLT, ki je navedla „Ni razvidno ali vsebuje slike ali ne“, so pravilno navedle podatek, da slikovno gradivo ni prisotno.

SIAS2048/IV/1: Datoteka ne vsebuje nobenih grafičnih elementov. **Opazanja:** Ročen pripis na prvi strani sta COPLT in GAIS razpoznal kot okrasni element (napačno).

SIAS2058/6: Datoteka, ki jo sestavlja 148 strani, razen žigov in grbov na prvih treh straneh ne vsebuje drugega slikovnega gradiva. **Opazanja:** Storitvev GNML je navedla, da „besedilo ne vsebuje informacij ali sklicevanja na slike na fotografije“, podatek o slikah ali fotografijah je očitno skušala pridobiti preko iskanja sklicev. Domnevamo, da je to nekakšna bližnjica do rezultata, ko ima obdelovano gradivo narejen razmeroma kvaliteten OCR. Storitvev GAIST je kot edina opozorila na prisotnost grbov in žigov.

Splošne skupne ugotovitve glede zastavljenega vprašanja: Pri obdelavi rezultatov se je pokazalo, da smo to vprašanje zastavili nekoliko nerodno. Storitve bi morali morda vprašati po prisotnosti grafičnih elementov (fotografije, slike, ilustracije, simboli, ...) Storitve so sicer nalogo opravile razmeroma dobro. Enkrat smo dobili odgovor, da ni razvidno ali gradivo vsebuje slike, enkrat si je storitev skušala pomagati z nekakšno bližnjico. Storitvev GAIST je najbolj konsistentno in pravilno odgovarjala in opozarjala na stvari, ki jih druge storitve niso.

4.5.5 Ali je v datoteki besedilo, ki je bilo razpoznano z OCR? Oceni kvaliteto OCR v datoteki. Napravi OCR, če je potrebno

To vprašanje nas je zanimalo, ker je iskanje po arhivskem gradivu veliko bolj učinkovito, če si lahko pomagamo tudi z iskanjem po vsebini popisnih enot in ne zgolj po podatkih popisa, ki jih pripravil arhivist. Gradivo na fizičnih nosilcih se pospešeno skenira v mnogih arhivskih ustanovah, kvaliteta optične prepoznavne znakov (če je izvedena) pa zelo niha. Najtrši oreh so rokopisi, starejše pisave in jeziki. Le redki arhivi najdejo ustrezne vire (osebje, denar in čas), da izvajajo ročno korekcijo prepoznavne. Približno oceno kakovosti OCR lahko ugotovimo že s preprostim pregledom izluščenega besedila. Če se v njem pojavlja veliko „zatipkanih“ besed, je OCR verjetno slab. Za bolj kvalitetno oceno kakovosti pa je OCR potrebno ponoviti z boljšim orodjem in rezultata medsebojno primerjati. Zanimalo nas je, kako se bodo storitve lotile te naloge. Od vseh šestih dokumentov sta imela OCR narejen le prvi (in to le delno) ter zadnji dokument. V tem razdelku smo združili rezultate na 5. vprašanje in 6 nalogo.

Tabela 5: Odgovori na vprašanja o obstoju OCR. Napačne navedbe so označene krepko.

	CGPT5	COPLT	GNLM	GAIST	GGEM
SIAS730/2/1/1	Da	Da	Da	Da	Da
SIAS1073/II-37r	–	Da	Ne	Da	?
SIAS1080/1/2,3,7	Ne	Da	Ne	Da	Da
SIAS2048/HRDARI/2	Ne	Da	Ne	Da	Ne
SIAS2048/IV/1	Ne	Da	Ne	Da	Da
SIAS2058/6	Da	NP (velikost datoteke)	Da	Da	Da

SIAS730/2/1/1: Optična razpoznavna je bila na datoteki že predhodno narejena na tiskanih delih dokumenta na straneh 5, 10-21, 23–24, 26, 29–33. Rokopisni deli na začetku niso bili optično prepoznani.

SIAS1073/II-37r: Noben del te 63 stranske datoteke ni imel narejenega OCR.

SIAS1080/1/2,3,7: Noben del te 6 stranske datoteke ni imel narejenega OCR.

SIAS2048/HRDARI/2: Noben del te 4 stranske datoteke ni imel predhodno narejenega OCR.

SIAS2048/IV/1: Noben del te 2 stranske datoteke ni imel narejenega OCR. Besedilo je tiskano.

SIAS2058/6: OCR v tej 148 strani dolgi PDF datoteki je bil OCR narejen na vseh straneh z besedilom.

Splošne skupne ugotovitve glede zastavljenega vprašanja in naloge

Skupni obseg odgovorov storitev glede obstoja OCR, njegove kakovosti in navodila naj OCR izvedejo, če ga datoteka še nima, je nanesele dobrih 25.000 znakov, veliko več, kot pa je bilo pričakovano. Tukaj navajamo le nekaj zanimivejših ugotovitev.

Storitve so pri datoteki, ki je imela delno narejen OCR večinoma izpustile analizo na straneh z rokopisom, kjer OCR ni bil narejen. Osredotočile so se samo na obstoječi OCR, kjer so ugotovile, da je pomanjkljiv in ga poskušale popraviti. Storitve CGPT5 nismo uspeli prepričati, naj analizira datoteko SIAS1073/II-37r. Ta storitev se občasno „odloči“, da ne bo izvajala analize PDF dokumentov v katerih ni niti enega znaka strojno berljivega besedila. Storitve so bili pri datotekah, ki niso vsebovale OCR plasti večinoma zmedene. Trdile so, da sicer pomanjkljiv OCR že obstaja. Nekatere storitve so navedle da so poskušale izvesti še enkrat ali

popraviti že obstoječega. Nekatere storitve so zmedeno navajale kvaliteto OCR-a, ki v datotekah ni bil prisoten. Nekatere storitve niti niso prepričane ali so izvedle OCR ali ne. Nekatere storitve so navajale izvedbo ročnega prepisa, a ni jasno kaj naj bi to bilo.

Navajanje rezultatov v prepričanost odgovora, tako glede obstoja, kot kakovosti OCR je bilo dvoumno. Trditve UI modelov, da OCR obstaja, čeprav ga datoteka zanesljivo nima, so nenavadne. Uganko je pomagalo razrešiti pregledovanje poteka razmišljanja, ki ga lahko opazujemo pri nekaterih storitvah. Zdi se, da mnoge storitve zaženejo nekakšno pred obdelavo naloženih datotek še preden uporabnik napiše svoje želje oz. poda nalogo. Zgodí se torej, da LLM niti ne ve (ta podatek mu ni posredovan), da je bil OCR narejen samodejno, v sklopu storitve, ob nalaganju gradiva. To pojasnjuje tudi zavračanje kakršne koli obdelave, če ta prvi korak ni uspešen (npr. pri kombinaciji SIAS1073/II-37r in CGPT5, kjer storitev „ve“, da ne bo mogla opraviti naloge, ker je brez podatkov). Na tak način morda lahko razložimo tudi uganko zakaj včasih kakšna storitev navede, da ni 100 % prepričana ali je OCR izvedla ali ne.

Zanimivo je opazovati, kako se storitve lotevajo razreševanja slabe kakovosti OCR. Nekatere ga poskušajo popraviti z ugibanjem iz konteksta in nekaj predznaja o tem kakšne so pogoste napake (npr. zamenjave znakov) in kar nekajkrat so bile pri tem celo zelo uspešne. Nekatere storitve trdijo, da so dokument „ročno“ pregledale oz. napravile novo, kakovostnejšo izvedbo OCR. Ali to pomeni, da so npr. na problematičnih delih gradiva izvedle še eno analizo, žal ne moremo z gotovostjo potrditi.

Tretji del tega opažanj pa se navezuje na navajanje prepričanosti. V splošnih navodilih smo želeli s stavkom: „Če o svojih navedbah nisi prepričana, napiši opozorilo. Želimo informacije, a želimo vedeti ali gre pri tvojih odgovorih za ugibanje ali gre za jasno razpoznanje in zanesljive navedbe.“ To navodilo bi morali očitno ubesediti drugače, ker včasih pri navajanju rezultatov ni jasno ali to pomeni, da je bilo vprašanje zgolj naslovljeno ali pa tudi opravljeno⁶. Pri nalogi izdelave OCR smo dobili velike vrednosti v prepričanost izvedbe naloge (pravilno), čeprav nas je bolj zanimala samoocena kakovosti OCR, ki ga je storitev poskušala izvesti.

⁶ Med preizkušanjem smo ugotovili, da bi bilo morda bolje, če bi kakšno izmed vprašanj in nalog ubesedili rahlo drugače. Mikalo nas je, da bi preizkuse ponovili. Na koncu smo se odločili, da dragocen prispevek raziskave lahko tudi navajanje in pojasnjevanje tistega, kar se ni izkazalo in bo morda koristno drugim raziskovalcem.

4.5.6 Vsebinske naloge (vprašanja 7 do 11)

V tem razdelku smo poskušali oceniti ustreznost odgovorov na vprašanja o jeziku, sestavi, vsebini in predlaganem naslovu za gradivo v datotekah⁷. Vseh pet storitev je za vsako od šestih datotek odgovorilo na naslednja vprašanja:

- Navedi jezike v katerih je napisano besedilo, če jih razpoznaš.
- Ali datoteka predstavlja en dokument ali je gre za več dokumentov? Koliko, če jih je več?
- V nekaj odstavkih predstavi vsebino datoteke.
- Predlagaj naslov oz. ime, ki bi bilo primerno za datoteko, če bi bila namenjena zgodovinarju.

Kakovost odgovorov smo dali preveriti arhivistom, ki poznajo gradivo⁸. Njihova naloga ni bila natančno iskanje pravilnih in napačnih navedb, ampak so morali podati dve oceni za vsa štiri vprašanja skupaj za gradivo, ki jim je bilo posredovano v oceno.

Storitve so na ta vprašanja sicer odgovarjale različno obsežno. Vse odgovore smo zbrali v eno datoteko, ki je obsegala 40 strani (cc 85.000 znakov) in jo tehnično uredili ter poenotili, da so arhivisti lažje izpeljali postopke ocenjevanja⁹.

Vprašnji arhivistom

A: Ali menite, da bi bil opis, ki ga je pripravila UI kakorkoli v pomoč arhivistu pri popisovanju gradiva?

Ocenite z ocenami od 1 do 4 pri čemer ocene pomenijo: (1) Nikakor: Opis ni v pomoč, saj je vsebinsko napačen, nepovezan ali popolnoma ne relevanten za popisovanje gradiva; (2) Malo: Opis vsebuje nekaj pravilnih elementov, ki bi arhivistu lahko dali izhodišče, a zaradi napak ali površnosti ni posebej uporaben; (3) Dokaj: Opis je večinoma smiseln in uporaben, a vsebuje nekaj napak; arhivist ga lahko uporabi kot osnovo, a ga mora preveriti in dopolniti; (4) Zelo: Opis je jasen, pravilen in dovolj izčrpen, da arhivistu bistveno olajša popisovanje gradiva.

B: Ali menite, da bi bil opis, ki ga je pripravila UI kakorkoli v pomoč zunanjemu uporabniku arhivskega gradiva, če ne more ali ne zna drugje pridobiti informacij?

⁷ Obravnavo odgovorov na vprašanja od 12 do 16 smo zaradi obsežnosti prihranili za naslednje prispevke.

⁸ Za neprecenljivo pomoč in vložen trud pri ocenjevanju odgovorov UI se zahvaljujem sodelavkam in sodelavcem: Danijeli Juričić Čargo, Žigi Koncilja, Branku Radulović, Juretu Volčjak in Lilijani Žnidaršič Golec.

⁹ Ugotovili smo, da so odgovori na vprašanje v zvezi s pisavami ali zelo generični (starinska, mešana) ali pa zelo specifični in bi njihovo preverjanje vzelo zelo veliko časa, zato smo dolgotrajno analizo teh rezultatov prihranili za naslednje prispevke.

Ocenite z ocenami od 1 do 4 pri čemer ocene pomenijo: (1) Nikakor: Opis vsebuje veliko napačnih trditev, pravilnih je malo ali nič. Napačne informacije so zavajajoče; (2) Malo: Opis ima nekaj pravilnih elementov, a vsebuje veliko napak. Napačne informacije so moteče in zmanjšajo uporabnost; (3) Dokaj: Opis je večinoma pravilen, vsebuje nekaj netočnosti. Napačne informacije ne zakrijejo pravilnosti celotne slike; (4) Zelo: Opis je vsebinsko pravilen, brez pomembnih napačnih informacij.

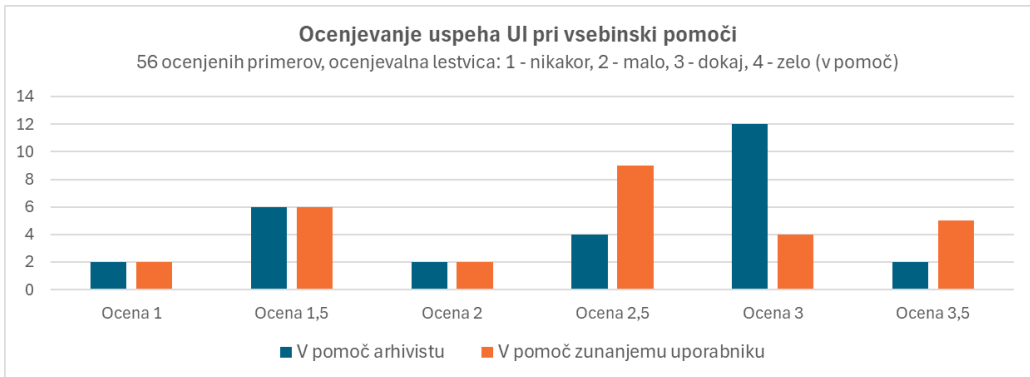
REZULTATI

Povprečne vrednosti so zbrane v spodnji tabeli. Prvi podatek je povprečna ocena odgovorov na vprašanje A, drugi je povprečna ocena odgovorov na vprašanje B. Najvišje ocenjeni odgovori za posamezno datoteko so označeni z zeleno. Vsako vprašanje ima oceni dveh arhivistov.

Tabela 6: *povprečne ocene za vsebinska vprašanja.*

Datoteka	CGPT5 A	CGPT5 B	COPLT A	COPLT B	GNLM A	GNLM B	GAIST A	GAIST B	GGEM A	GGEM B
SIAS730/2/1/1	2,5	2,5	2,5	2,5	3	2,5	3	2,5	3	2,5
SIAS1073/II-37r	–	–	3	2,5	2,5	2,5	3	2,5	3	3
SIAS1080/1/2,3,7	2	2	1,5	1,5	1,5	1,5	2	2	1,5	1,5
SIAS2048/ HRDARI/2	1	1	1	1	1,5	1,5	1,5	1,5	1,5	1,5
SIAS2048/IV/1	3	3	3	3	2,5	2,5	3,5	3,5	3	3
SIAS2058/6	3	3	–	–	3	3,5	3	3,5	3,5	3,5

V štirih primerih UI ni bila prav v *nikakršno pomoč* (ocena 1). V dvanajstih primerih se je začela približevati *pragu, da je malo v pomoč* (ocena 1,5). Dvakrat je bila dosežena vrednost *malo v pomoč* (2), štirikrat je bila *dokaj v pomoč*, trinajstkrat je bila *med malo in do dokaj v pomoč*, šestnajstkrat *dokaj v pomoč* in sedemkrat *med dokaj in zelo v pomoč*. Razmerje odgovorov, kjer ni bila v pomoč proti odgovorom kjer je bila v pomoč je 1:2,5. Storitev GAIST je pobrala največ najboljših ocen.



Slika 8: Porazdelitev povprečnih ocen.

Povprečna ocena glede na ocenjevano datoteko so: SIAS730/2/1/1 (2,65); SIAS1073/II-37r (2,75), SIAS1080/1/2,3,7 (1,7); SIAS2048/HRDARI/2 (1,3); SIAS2048/IV/1 (3); SIAS2058/6 (3,25). Takoj lahko ugotovimo, da so najboljše ocene dobile datoteke, ki so imele ali že izveden OCR ali pa ga je storitev uspela dobro napraviti sama.

5. RAZPRAVA

5.1 ŠIRŠI POGLED NA REZULTATE PREIZKUSA

Analiza rezultatov razkriva temeljno protislovje v delovanju modelov, ki smo jih preizkusili. Medtem ko so se izkazali za kompetentne pri semantično bogatih nalogah (npr. povzemanje vsebine, predlaganje naslova), so imeli presenetljive težave z navidezno preprostimi tehničnimi nalogami, kot so štetje strani ali prepoznavanje praznih listov. Predvidevamo, da je ta paradoks posledica dejstva, da so splošni LLM-ji optimizirani za obdelavo in generiranje jezika, ne pa za natančno vizualno in strukturalno analizo dokumentov, ki je značilna za arhivsko popisovanje. Modeli delujejo najboljše, ko lahko uporabijo svoje osnovne zmožnosti semantičnega sklepanja, medtem ko se borijo z nizko-nivojskimi vizualnimi detajli, ki so za človeka intuitivni. Povedano drugače, modeli ne „vidijo“ dokumenta na enak način kot človek, ampak se zanašajo na plast besedila in poskušajo iz nje izluščiti metapodatke, kar pogosto vodi do netočnosti ali sistemskih napak, kot so npr. neupravičene trditve o obstoju OCR.

Testiranje LLM-jev ne služi le kot evalvacija tehnologije, ampak tudi kot izkušnja za arhiviste. Z opazovanjem, kako splošni modeli obravnavajo arhivsko gradivo, arhivisti pridobivamo vpogled v informacijske potrebe in logiko UI. Kateri pozivi

in podatki so modelom najpomembnejši za generiranje odgovorov? Kateri jih zmedejo? Neuspešni poskusi prepoznavanja specifičnih elementov ali paleografskih posebnosti, opozarjajo arhiviste na področja, kjer je potrebna človeška strokovnost. Rezultati jasno kažejo, da obravnavane, *ne specializirane storitve*, ne morejo nadomestiti vloge arhivista. Lahko pa služijo kot podporno orodje avtomatizacije generativnih nalog. Uporaba LLM-jev za pripravo začetnih vsebinskih povzetkov ali predlogov za naslove lahko arhivistu prihrani čas, ki ga lahko preusmeri v bolj poglobljene analize in v urejanje gradiva, kjer je potrebna visoka natančnost (natančna transkripcija rokopisov, urejanje strokovnih metapodatkov, ...). Ključni izziv za prihodnost je razvoj specializiranih, domensko specifičnih modelov, ki bodo trenirani na specifičnem arhivskem gradivu (npr. tip, obdobje, ustvarjalec, ...) in bodo s tem premagovali omejitve splošnih modelov, ki so bile opazne v tej raziskavi.

Ugotovitve raziskave so posnetek stanja v določenem trenutku. Modeli se razvijajo izjemno hitro in že v kratkem času morda lahko pričakujemo izboljšave, ki bodo odpravile nekatere sedanje pomanjkljivosti. Na podlagi opažanj je mogoče napovedati, da se bo v prihodnosti potegnila jasna ločnica med splošnimi LLM-ji, namenjenimi laičnim uporabnikom in specializiranimi modeli, razvitimi v okviru akademskih in skupnostnih projektov. Splošni modeli bodo služili za začetno ali splošno razumevanje vsebine, medtem ko bodo domensko specifični modeli ključni za poglobljeno raziskovalno delo in profesionalno popisovanje. To bo vodilo k večji natančnosti, zanesljivosti in transparentnosti, kar je nujno za ohranjanje zaupanja v avtomatizirane arhivske procese.

Preizkus je razkril, da so veliki jezikovni modeli v svoji trenutni obliki že razmeroma uporabni za generiranje vsebinskih povzetkov in opisov starejšega, raznolikega rokopisnega arhivskega gradiva, kar potrjujejo ocene s strani arhivistov. Hkrati so izjemno nezanesljivi pri tehničnih in natančnih nalogah, kot sta prepoznavanje števila strani in pravilno izvajanje OCR. Ključna pomanjkljivost se kaže v ne transparentnosti delovanja modelov, kjer zgolj ugibamo, kako je bila obdelava izvedena in kako bi lahko dosegli boljše.

5.2 PRIPOROČILA, OMEJITVE RAZISKAVE IN PREDLOGI ZA NADALJNJE RAZISKAV

Na podlagi rezultatov in pridobljenih izkušenj bi lahko priporočila arhivistom in zunanjim uporabnikom strnili v nekaj naslednjih misli:

- Uporabite splošne LLM-je le kot podporno orodje za avtomatizacijo rutinskih, generativnih nalog (npr. priprava osnutkov vsebinskih opisov).
- Na splošne LLM-je se ne zanašajte pri tehničnih nalogah, ki zahtevajo visoko natančnost (npr. štetje strani, verifikacija metapodatkov). Te naloge je priporočljivo opraviti z bolj specializiranimi (klasičnimi) orodji ali jih preveriti in popraviti ročno.
- Zavedajte se, da so odgovori vseh modelov lahko nepopolni ali netočni in zahtevajo preverjanje s strani strokovnjaka.
- Če imate namen uporabiti LLM modele v verigi avtomatiziranih postopkov, skrbno oblikujte in temeljito preverite pozive, da bodo le-ti vračali zelene podatke v pričakovanih formatih in ne nekih drugih, dodatnih, motečih ali odvečnih podatkov.

Priporočamo ponovitev podobnega preizkusa v prihodnosti, da se ugotovi, kako hiter razvoj UI vpliva na uspešnost pri arhivskih nalogah. V prihodnjih raziskavah je smiselno vključiti neposredno evalvacijo končnih uporabnikov, da se natančneje ugotovi vpliv UI na njihovo izkušnjo iskanja in dostopa do arhivske dediščine.

Raziskava ima določene omejitve. Osredotočila se je na splošne, javno dostopne LLM-je in ne na modele, ki bi bili specializirani za arhivsko gradivo. Obravnavani korpus je specifičen in razširitev rezultatov na celotno slovensko arhivsko gradivo zahteva nadaljnje raziskave. Prihodnje raziskave bi se lahko osredotočile na naslednja področja:

- Usposabljanje ali urjenje specializiranih LLM-jev na slovenskem arhivskem gradivu bi lahko bistveno izboljšalo natančnost pri obravnavi.
- Razvoj in morebitna poenotenje nabora pozivov (ukazov storitvam) bi lahko arhivistom poenostavila delo in zagotovila boljše rezultate.
- Ključno je raziskati, kako zagotoviti dolgoročno ohranjanje metapodatkov, ustvarjenih z UI v skladu s standardi in arhivskimi principi, da se zagotovi njihova trajnostna uporabnost v prihodnosti.
- Vse večji poudarek na zasebnosti in varnosti podatkov zahteva raziskovanje odprtokodnih LLM-jev, ki jih je mogoče namestiti in upravljati v internem omrežju arhivov.

5.3 ETIČNE IN PRAVNE IMPLIKACIJE

Uporaba splošno dostopnih LLM-jev prinaša tudi etične in pravne izzive, ki jih arhivska skupnost ne sme prezreti. Prvi med njimi je vprašanje varnosti in za-

sebnosti podatkov. Uporaba javnih storitev predstavlja tveganje, da bi lahko občutljivi ali osebni podatki postali del učne zbirke ponudnika in bi bili nenamerno izpostavljeni. Zato je za arhive ključnega pomena, da uporabljajo storitve, ki zagotavljajo ničelno zadrževanje podatkov (angl. *zero data retention policy*) in imajo robustne varnostne protokole. Alternativa so zasebne rešitve (angl. *on-premise*), ki zagotavljajo popoln nadzor nad podatki.

Drugi pomemben izziv je problem črne škatle (angl. *black box problem*). Javni komercialni modeli so ne transparentni in njihovi mehanizmi odločanja so razmeroma neznani. Varnostni mehanizmi LLM-jev lahko celo ne transparentno filtrirajo oz. blokirajo dele zgodovinskih besedil (npr. ker so se ustvarjalcu modela zdeli škodljivi). To lahko privede do nenamernih pomanjkljivosti ali cenzure v prevodih in povzetkih in je v nasprotju z arhivsko etiko, ki zahteva popolno preglednost in odgovornost.

Nazadnje je treba obravnavati vprašanje avtorstva in avtorskih pravic. Marsikje zakonodaja o avtorskih pravicah zahteva človeško avtorstvo. Vsebine, ki jih ustvari UI brez zadostnega človeškega nadzora, ne morejo biti zaščitene z avtorskimi pravicami. V kontekstu arhivskega popisovanja (če ga želimo šteti kot za avtorsko delo) to pomeni, da je vloga arhivista, ki oblikuje pozive, kurira izhode in ureja končne opise, ključna. Ta človeška intervencija določa ustvarjalni nadzor (angl. *creative control*), ki je potreben za priznanje avtorstva in intelektualne lastnine nad končnim popisnim produktom.

VIRI IN LITERATURA

AEOLIAN Network – Artificial Intelligence for Cultural Organisations. (2025). Pridobljeno na <https://www.aeolian-network.net/> (dostop 11.9.2025).

AI4LAM. (s. d.), Pridobljeno na <https://sites.google.com/view/ai4lam> (dostop 11.9.2025).

Hernandez, A. R., Fewster, K., in Penniman, S. (2024). Artificial Intelligence and Machine Learning Competencies for the Archival Professions. *Proceedings of the Association for Information Science and Technology*, 61(1), 36–43. Pridobljeno na <https://doi.org/10.1002/pr2.1006> (dostop 11.9.2025).

AURA Network – AURA Network: Archives and AI in the UK and Ireland. (2025). Pridobljeno na <https://www.aura-network.net/> (dostop 11.9.2025).

- Bingham, N. J. in Byrne, H. (2021). Archival Strategies for Contemporary Collecting in a World of Big Data: Challenges and Opportunities with Curating the UK Web Archive. *Big Data & Society*, 8(1). Pridobljeno na <https://doi.org/10.1177/2053951721990409> (dostop 11.9.2025).
- Brock, D. C. (2018). Learning from Artificial Intelligence's Previous Awakenings: The History of Expert Systems. *AI Magazine*, 39(3), 3–15. Pridobljeno na <https://doi.org/10.1609/aimag.v39i3.2809> (dostop 11.9.2025).
- Cushing, A. L. in Osti, G. (2023). "So How Do We Balance All of These Needs?": How the Concept of AI Technology Impacts Digital Archival Expertise". *Journal of Documentation*, 79(7), 12–29. Pridobljeno na <https://doi.org/10.1108/JD-08-2022-0170> (dostop 11.9.2025).
- Decker, S., Kirsch, D. A., Kuppili Venkata, S. in Nix, A. (2022). Finding Light in Dark Archives: Using AI to Connect Context and Content in Email. *AI & SOCIETY*, 37(3), 859–72. Pridobljeno na <https://doi.org/10.1007/s00146-021-01369-9> (dostop 11.9.2025).
- Fan, L., Yin, Z., Yu, H., & Gilliland, A. (10. 11. 2020). Using Machine Learning to Enhance Archival Processing of Social Media Archives". Preprint, LIS Scholarship Archive. Pridobljeno na <https://doi.org/10.31229/osf.io/gkydm> (dostop 11.9.2025).
- Feng, C., Yifan, L., Chen, Z., & Guo, L. (2024, October 18). The evolution and breakthrough of natural language processing: The revolution from rules to deep learning. In *Proceedings of the 2024 5th International Conference on Computer Science and Management Technology* (pgs. 307–311). New York, USA: Association for Computing Machinery. Pridobljeno na <https://doi.org/10.1145/3708036.3708089> (dostop 11.9.2025).
- Hankins, R. (2019). Information Literacy and Instruction: Embracing Informational and Archival Literacies: Challenges and Successes. *RUSQ: A Journal of Reference and User Experience*, 58(3), 153–57. Pridobljeno na <https://doi.org/10.5860/rusq.58.3.7042> (dostop 11.9.2025).
- Yumeng, H., Kenderdine, S., Picca, D., Egloff, M. in Adamou, A. (2022). Digitizing Intangible Cultural Heritage Embodied: State of the Art. *Journal on Computing and Cultural Heritage*, 15(3), 1–20. Pridobljeno na <https://doi.org/10.1145/3494837> (dostop 11.9.2025).

- Humphries, M., Leddy, L. C., Downton, Q., Legace, M., McConnell, J., Murray, I., in Spence, E. (2024). Unlocking the archives: Using large language models to transcribe handwritten historical documents (Version 1) [Preprint]. arXiv. Pridobljeno na <https://doi.org/10.48550/ARXIV.2411.03340> (dostop 11.9.2025).
- InterPARES Trust AI - Artificial Intelligence. (s. d.). Pridobljeno na https://interparestrustai.org/trust/about_research/workinggroups (dostop 11.9.2025).
- International Council on Archives (ICA). (2000). *ISAD(G): General international standard archival description (2nd ed.)*. International Council on Archives. Pridobljeno na <https://www.ica.org/resource/isadg-general-international-standard-archival-description-second-edition/> (dostop 11.9.2025).
- Kaluvilla, B. B., Kalarikkal, S. A., & Thamilvanan, G. (8.9.2025). AI-driven extraction and intelligent retrieval of missionary archives in Malabar: Advancing preservation and accessibility with machine learning. *Performance Measurement and Metrics*, 1–15. Pridobljeno na <https://doi.org/10.1108/PMM-02-2025-0008> (dostop 11.9.2025).
- Khan, A., Rai, U., Singh, S. S., & et al. (2024). OCR approaches for humanities: Applications of artificial intelligence/machine learning on transcription and transliteration of historical documents. *Digital Studies in Language and Literature*, 1(1–2), 85–112. Pridobljeno na <https://doi.org/10.1515/dsll-2024-0013> (dostop 11.9.2025).
- LUSTRE – Unlocking our Digital Past with Artificial Intelligence. (s. d.). Pridobljeno na <https://lustre-network.net/> (dostop 11.9.2025).
- Pezzica, L. (2023). Archival Inventories as a Profession. *JLIS.It*, 14(3), 64–71. Pridobljeno na <https://doi.org/10.36253/jlis.it-563> (dostop 11.9.2025).
- International Council on Archives. (s. d.). *Records in Contexts (RiC)*. International Council on Archives. Pridobljeno na <https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-ric/> (dostop 11.9.2025).
- Tenzer, M., Pistilli, G., Brandsen, A., & Shenfield, A. (2024). Debating AI in Archaeology: applications, implications, and ethical considerations. *Internet Archaeology*, 67(marec). Pridobljeno na <https://doi.org/10.11141/ia.67.8> (dostop 11.9.2025).
- Zhang, S., & Colavizza, G. (2025). Named entity recognition of historical text via large language model (Version 1) [Preprint]. arXiv. Pridobljeno na <https://doi.org/10.48550/ARXIV.2508.18090> (dostop 11.9.2025).

Summary

This study examines the applicability of general-purpose large language models (LLMs) for archival description, emphasizing both their potential and their current limitations in professional practice. Five widely used LLM-based services (OpenAI ChatGPT-5, Microsoft Copilot, Google NotebookLM, Google AI Studio, and Google Gemini) were tested on a selection of archival documents from the Archives of the Republic of Slovenia. The chosen materials reflected real-world challenges, including incomplete or absent OCR, mixtures of handwritten and printed text, bleed-through, archaic orthographies, and multilingual content spanning several centuries.

The evaluation focused on two groups of tasks. The first involved technical operations such as page counting, distinguishing between textual and graphic elements, and assessing OCR quality. The second addressed content-oriented tasks, including language and script recognition, summarization, title generation, and the extraction of named entities. Outputs were assessed through both quantitative comparison and qualitative review by professional archivists.

The findings revealed a disparity between the models' technical and content-related performance. LLMs frequently struggled with seemingly simple technical tasks, such as reporting the correct number of pages or identifying blank versus text-bearing pages. OCR detection was particularly inconsistent, with some models incorrectly asserting the existence of OCR layers. By contrast, results improved markedly for content-related tasks. The models showed considerable potential in language identification, title suggestion, and summarization, producing coherent descriptions that could serve as initial drafts for archivists. Although inaccuracies and hallucinations remained present, several outputs were judged useful, particularly for external, non-specialist users navigating complex archival sources.

The study underscores that current tested LLMs cannot yet be trusted for the automation of precise technical cataloguing, where reliability and consistency are essential. Nevertheless, they hold promise as analytical and generative aids for creating descriptive summaries, contextual overviews, and draft metadata suitable for professional refinement. In this capacity, LLMs could help mitigate persistent backlogs in archival description, a challenge exacerbated by shortages of staff, expertise, and resources in Slovenian archives.

Beyond its immediate findings, the case study contributes a replicable framework for testing LLM performance on archival material. Establishing such a baseline is valuable for monitoring technological progress over time and for guiding future integration of AI tools into archival workflows. The research also provides practical insights and strategic considerations for institutions contemplating AI adoption. Ultimately, while LLMs cannot replace the expertise of professional archivists, they can function as supportive tools that enhance accessibility, reduce processing delays, and open new opportunities for engaging diverse audiences with archival heritage.

Typology: 1.01. Original scientific research

VARIE /MISCELLANEOUS/ RAZNO

