Sveučilište u Rijeci
Fakultet informatike
i digitalnih tehnologija

Reliable Biomedical NLP via Ontology integration: Leveraging UMLS to Align Language Models with Clinical Terminologies and Use Cases

Doktorand: Luka Blašković

Mentor: prof. dr. sc. Ivo Ipšić

Komentor: doc. dr. sc. Nikola Tanković

Znanstveno područje istraživanja

- **Area:** Artificial intelligence
  - **Problem area:** Natural Language Processing (NLP) in Biomedical domain
    - **Topic:** Leveraging UMLS to allign language models with clinical (biomedical) terminologies

Motivacija

Reliable Biomedical NLP via Ontology integration: Leveraging UMLS to Align Language Models with Clinical Terminologies and Use Cases

- **Trenutno stanje zdravstvenog sustava u RH**
  - Sporo provođenje reformi i digitalizacije zdravstvenog IS
  - Nedostatak stručnog zdravstvenog kadra te njegova neravnomjerna raspodjela – **preopterećenje kapaciteta zdravstvenih ustanova** [1] [3]
  - **Duge liste čekanja** za specijalističko-konzilijarnu i bolničku zdravstvenu zaštitu među najvećim čimbenikom nezadovoljstva zdravstvenim sustavom RH [1] [2] [3]
  - **Neslavne statistike** OECD, WHO, Eurostata i drugih nezavisnih organizacija

[1] Buljan, A. i Šimović, H. (2022). Učinkovitost hrvatskog zdravstvenog sustava - usporedba sa zemljama Europske unije. *Revija za socijalnu politiku, 29* (3), 321-354. https://doi.org/10.3935/rsp.v29i3.1933

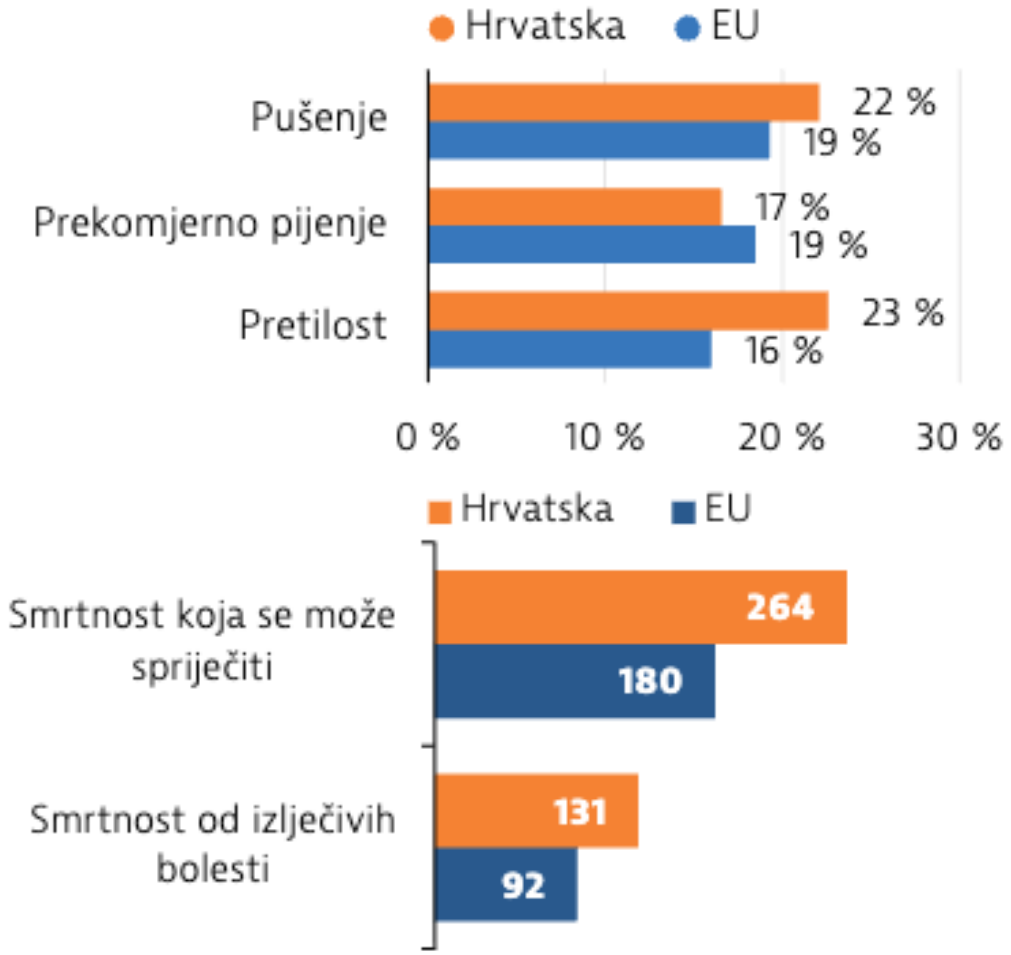
[2] Bobinac A. Access to Healthcare and Health Literacy in Croatia: Empirical Investigation. Healthcare (Basel). 2023

[3] Popović, Stjepka. "Odrednice stavova i zadovoljstva građana hrvatskim zdravstvenim sustavom." *Medicina Fluminensis*, vol. 53, br. 1, 2017, str. 85-100. https://doi.org/10.21860/medflum2017_173385.

Motivacija

Efektivnost zdravstvenog sustava [4]

- Konzumacija alkohola, duhanskih proizvoda, loše prehrambene navike te poremećaji mentalnog zdravlja iznad su prosjeka EU
- Očekivani životni vijek je 77.7 godina što je 3 godine kraće od prosjeka EU-a
- Povećane stope smrtnosti koja se može spriječiti – primjer: stopa smrtnosti od raka pluća u RH druga po veličini u EU



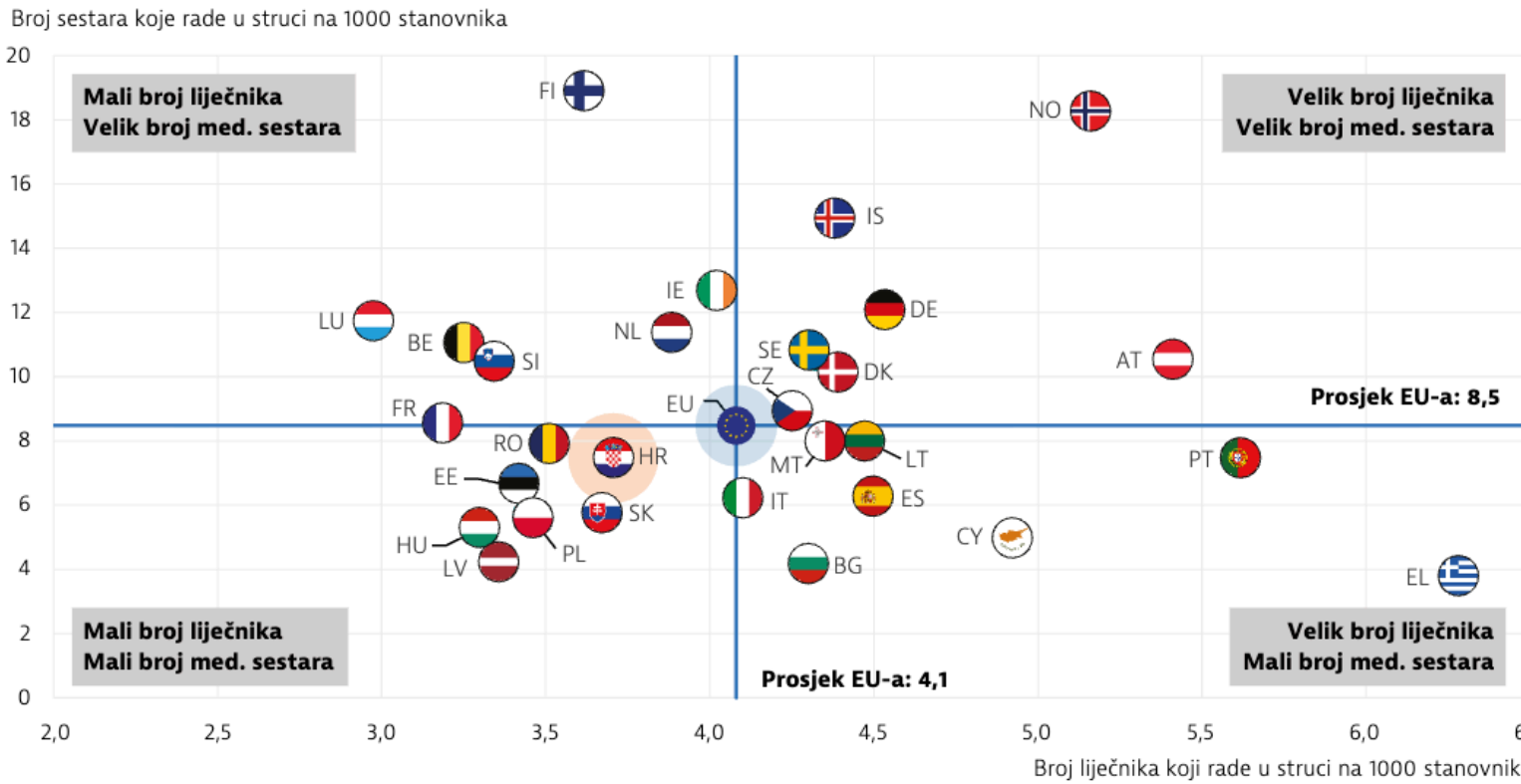
Ilustracije iz Eurostata [4]

[4] Eurostat – State of Health in the EU: „Hrvatska: Pregled stanja zdravlja i zdravstvene zaštite 2023”

Motivacija

Dostupnost zdravstvenog sustava [4]

- Broj liječnika i medicinskih sestara u Hrvatskoj (3,7 liječnika i 7,5 medicinskih sestara na 1000 stanovnika) **manji je od prosjeka Europske unije**, koji iznosi 8,5 medicinskih sestara i 4,1 doktora na 1000 stanovnika.



Ilustracija iz Eurostata [4]

[4] Eurostat – State of Health in the EU: „Hrvatska: Pregled stanja zdravlja i zdravstvene zaštite 2023”

Pilot study [5]

Reliable Biomedical NLP via Ontology integration: Leveraging UMLS to Align Language Models with Clinical Terminologies and Use Cases

Ideja: Provesti nekoliko eksperimenata primjene velikih jezičnih modela na konkretnim zadacima obrade prirodnog jezika u simuliranom kliničkom okruženju.

Motivacija: pokazati kako se primjenom AI-a mogu ubrzati/pojednostaviti određeni administrativni zadaci s krajnjim ciljem rasterećenja bolničkog sustava

1. **Natural language to SQL generation** (MIMIC-III [6] / MIMIC-SQL [7])
   - ➡ Transformacija nestrukturiranog teksta (NL upit) → strukturirani zapis (SQL kod)
2. **Question answering over synthetic electronic health records** (EHR) using Retrieval augmented generation (RAG [8])
   - ➡ Transformacija nestrukturiranog teksta (pitanje) → nestrukturirani tekst (odgovor)

[5] Blašković, L.; Tanković, N.; Lorencin, I.; Baressi Šegota, S. Robust Clinical Querying with Local LLMs: Lexical Challenges in NL2SQL and Retrieval-Augmented QA on EHRs. Big Data Cogn. Comput. 2025, 9, 256.

[6] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3.1 (2016): 1-9.

[7] Wang, Ping, Tian Shi, and Chandan K. Reddy. "Text-to-SQL generation for question answering on electronic medical records." *Proceedings of The Web Conference 2020*. 2020.

[8] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.

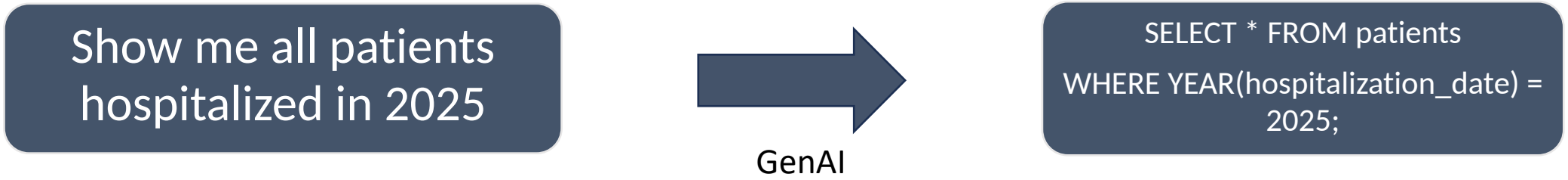
Pilot study

1. Natural language to SQL generation (Text-to-SQL)

Podskup problema: Natural language understanding (NLU) [9]

- cilj: preslikavanje nestrukturiranih podataka u formalnu semantički-bogatu i reprezentativnu strukturu pogodnu za daljnju računalnu obradu.

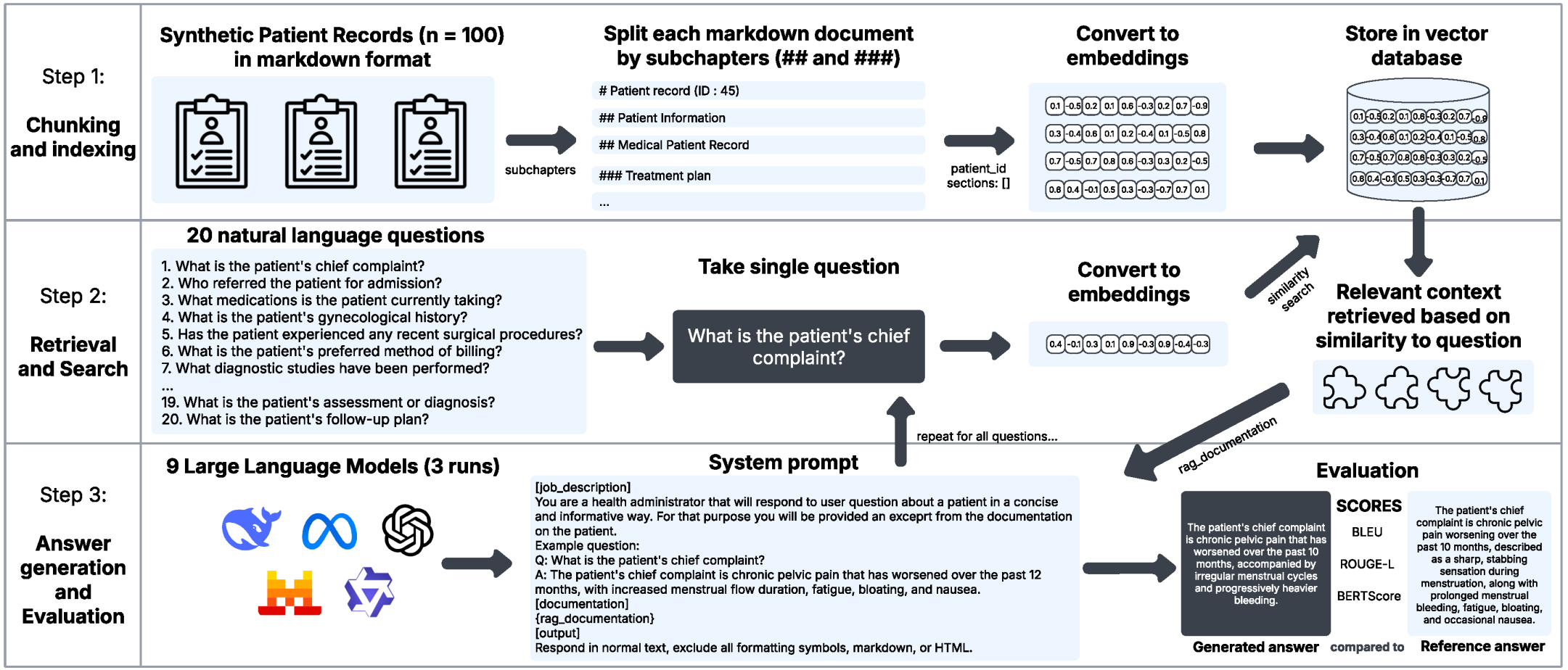
Konkretno: *sequence-to-sequence* transformacija prirodnog jezika u izvršivi SQL kod koji dohvaća ispravne podatke iz relacijske baze podataka.



[9] Natural language understanding. *Wikipedia, pristupljeno 15. 10. 2025., dostupno na: https://en.wikipedia.org/wiki/Natural_language_understanding*

Pilot study

2. Question answering over synthetic electronic health records



[5] Blašković, L.; Tanković, N.; Lorencin, I.; Baressi Šegota, S. Robust Clinical Querying with Local LLMs: Lexical Challenges in NL2SQL and Retrieval-Augmented QA on EHRs. Big Data Cogn. Comput. 2025, 9, 256.

Pilot study

Što je **elektronički zapis pacijenta (EHR)**?

- **Elektronički zapis pacijenata** (elektronički zdravstveni karton - **eKarton**) predstavlja dio medicinske dokumentacije pacijenta koji objedinjava zdravstvene podatke o pacijentu (U HR kroz CEZIH).
- U RH djeluje 49 javnih zdravstvenih ustanova, među kojima su opće bolnice, klinički bolnički centri, specijalne bolnice i lječilišta. [10]
  - 42 ustanove imaju vlastiti bolnički informacijski sustav, a 36 posjeduje interne IT odjele.
  - Okvirni omjer strukturiranih/nestrukturiranih podataka u zdravstvenim IS iznosi ~20/80 [11] [12]
  - Primjenjivost provedenih metoda ne može se generalizirati na sve IS.
  - Regulativa *European Health Data Space 2025/327* – na snazi od 26. 3. 2025. [13]

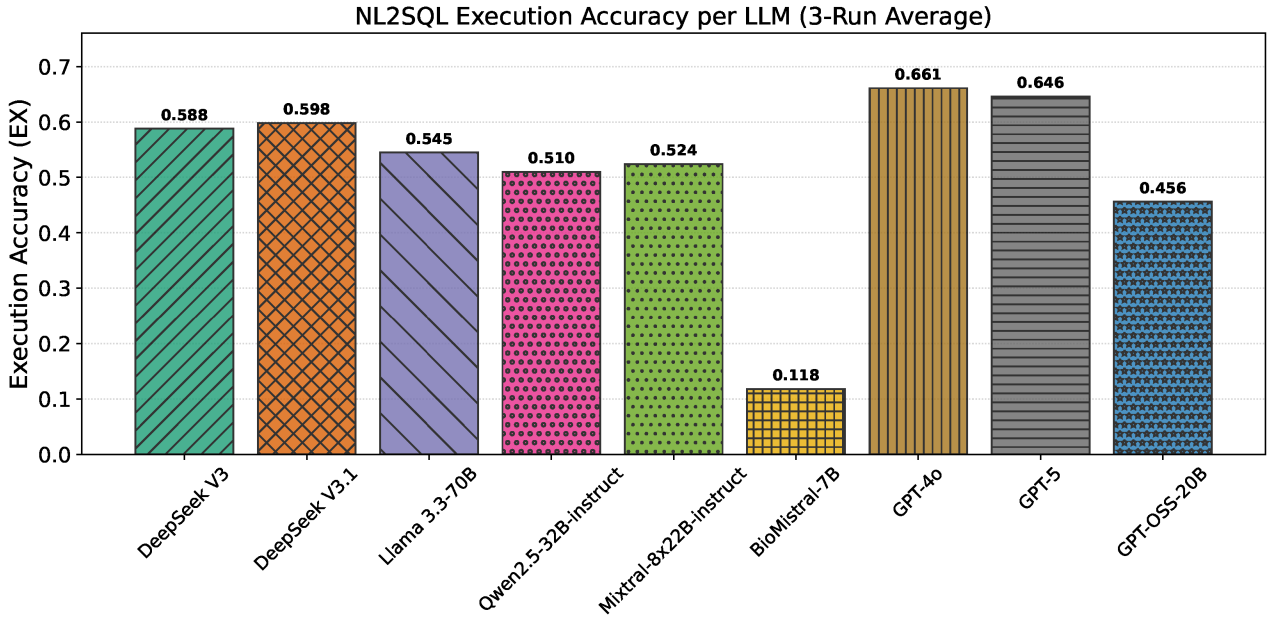
[10] Ministarstvo zdravstva Republike Hrvatske – Bolničke zdravstvene ustanove. Dostupno na: https://zdravlje.gov.hr/kontakti/kontakti-zdravstvenih-ustanova/bolnicke-zdravstvene-ustanove/2722
[11] Osvaldić, Josipa. "Information system implementation in healthcare: case study of Croatia." *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy* 12.2 (2021): 114-124.
[12] Kong H. J. (2019). Managing Unstructured Big Data in Healthcare System. Healthcare informatics research, 25(1), 1–2.
Capurro, D., Yetisgen, M., van Eaton, E., Black, R., & Tarczy-Hornoch, P. (2014). Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement:
[13] Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847

Pilot study

Rezultati 1/2

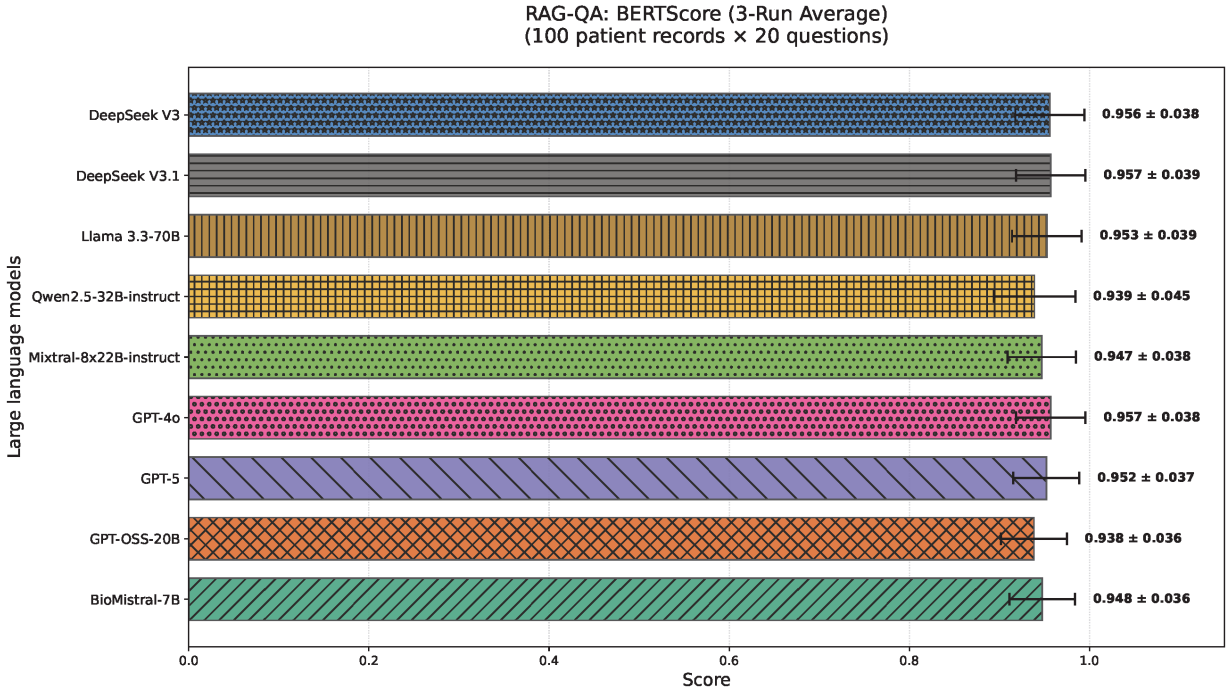
1. Natural language to SQL generation

Metrika: Execution accuracy



2. Question answering over synthetic EHRs

Metrike: BLEU, ROUGE-L, BERTScore

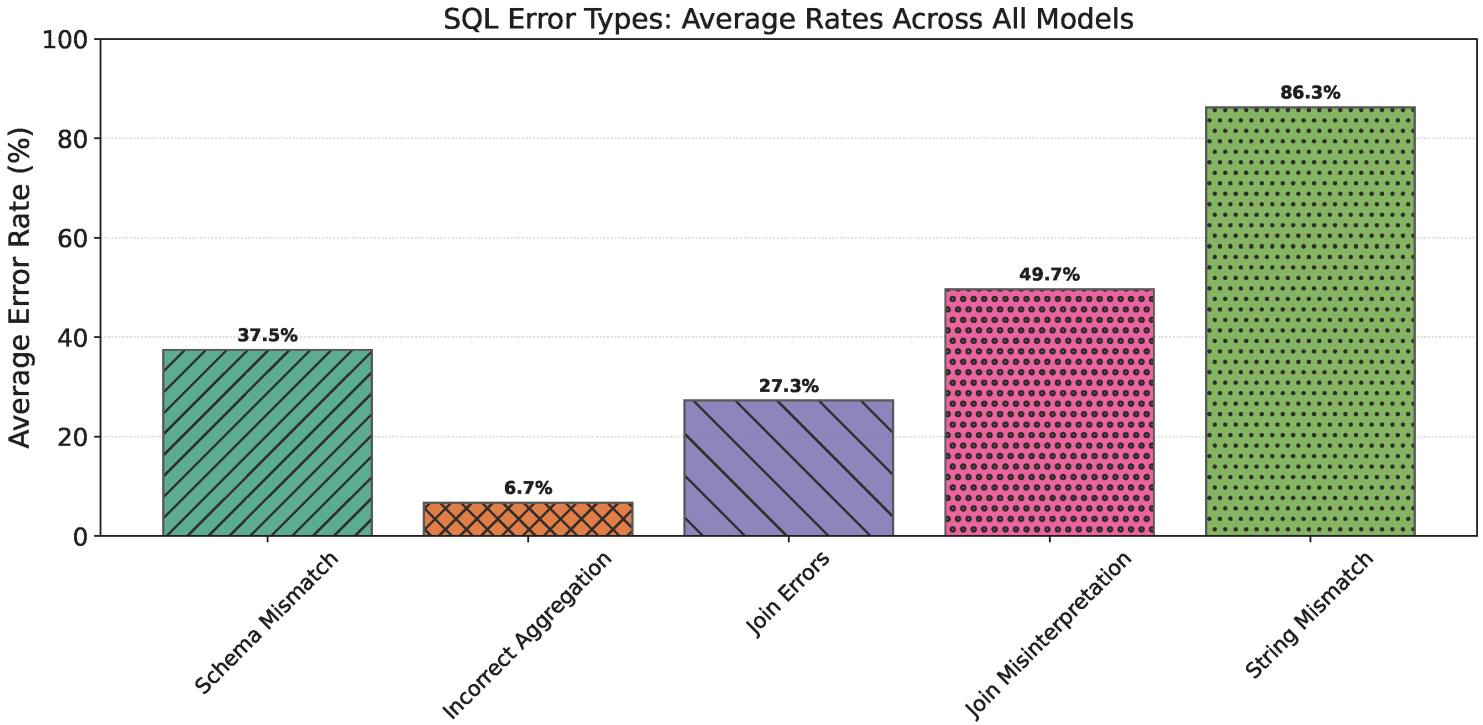


[5] Blašković, L.; Tanković, N.; Lorencin, I.; Baressi Šegota, S. Robust Clinical Querying with Local LLMs: Lexical Challenges in NL2SQL and Retrieval-Augmented QA on EHRs. Big Data Cogn. Comput. 2025, 9, 256.

Pilot study

Rezultati 2/2

- Za NL2SQL provedena dodatna analiza pogrešno generiranih SQL upita
- *Multi-label classification task*
- **Najčešća greška**: *String mismatch prilikom pretraživanja podataka* (*WHERE*)
  - neusklađenost između znakovnih nizova tj. usporedba dvaju normaliziranih nizova ne daje rezultat koji se podudara



[5] Blašković, L.; Tanković, Nq.; Lorencin, I.; Baressi Šegota, S. Robust Clinical Querying with Local LLMs: Lexical Challenges in NL2SQL and Retrieval-Augmented QA on EHRs. Big Data Cogn. Comput. 2025, 9, 256.

Pilot study

Reliable Biomedical NLP via Ontology integration: Leveraging UMLS to Align Language Models with Clinical Terminologies and Use Cases

- Većina uočenih nepodudaranja proizlazi iz **leksičkih varijacija**, **sinonimije**, **polisemije** te **uporabe kratica** među kliničkim medicinskim konceptima (dijagnoze, simptomi, procedure, lijekovi, uređaji/alati, patološki, laboratorijski i drugi rezultati, naplatni podaci).

➡ Daljnje istraživanje usmjerava se na **standardizaciju** i **normalizaciju** medicinske terminologije.

Primjer normalizacije entiteta:

Upit: "Show all patients who had a heart attack"

Pohranjena dijagnoza: *Acute myocardial infarction*

Program treba:

- prepoznati da izrazi „heart attack" i "Acute myocardial infarction" predstavljaju isti medicinski koncept
- preslikati koncept na određeni jedinstveni identifikator koji baza podataka koristi (npr. ICD-9/11 kod)
- generirati izvršiv SQL upit koji prikazuje točne podatke

Primjer: Diabetes mellitus (bolesti)

Sinonimi

- Type 2 diabetes
- Insulin-independent diabetes
- Adult-onset diabetes
- Sugar disease
- Diabetes mellitus
- non-insulin-dependent diabetes mellitus

Kratice

- DM
- T2DM
- NIDDM
- IDDM

Primjer: Infarction (patološki process)

Sinonimi

- Infarction
- Cardiac infarction
- Heart attack
- Acute myocardial infarction (AMI)
- Myocardial infarction (MI)
- Myocardial necrosis

Polisemija

- Myocardial infarction
- Cerebral infarction
- Pulmonary infarction

Primjer: ASA (lijek)

Sinonimi

- Acetylsalicylic acid
- Aspirin
- Acetyl salicylate
- 2-acetoxybenzoic acid
- Salicylic acid acetate
- monoacetic acid ester of salicylic acid

Slobodni nazivi u prodaji

- Bufferin
- Ecotrin
- Anacin
- Aspro
- Disprin
- Empirin
- St. Joseph Aspirin

Unified Medical Language System (UMLS)

Reliable Biomedical NLP via Ontology integration: Leveraging UMLS to Align Language Models with Clinical Terminologies and Use Cases

- Opsežan skup biomedicinskih vokabulara i alata koji omogućava integraciju, razmjenu i standardizaciju medicinskih informacija između različitih IS (U.S. National Library of Medicine) [14]
- U aktivnom razvoju od 1986. godine, posljednja verzija **2025AA**
  - Sadrži ~**3.5 milijuna** jedinstvenih medicinskih koncepta
  - Sadrži ~**17 milijuna** različitih naziva pojedinog koncepta
  - Sadrži ~**75 milijuna** odnosa između koncepata
  - Objedinjuje 190 različitih medicinskih vokabulara/terminologija kroz 29 jezika
- Sastoji se od:
  - *Metathesaurus* (leksikon koji objedinjuje medicinske koncepte)
  - *Semantic Network* (semantičke kategorije i odnosi (veze) između koncepata) – ~120 različitih semantičkih tipova i 54 veza
  - Dodatni alati i resursi za lakši rad s podacima
- Ukupno veličina iznosi oko ~35 GB strukturiranih podataka
- Razvijen prvenstveno za istraživačke svrhe

[14] National Library of Medicine. (2024). *Unified Medical Language System (UMLS)*. U.S. Department of Health and Human Services

Unified Medical Language System (UMLS)

Što je **medicinska terminologija (vokabular)?** [14]

- **Skup stručnih izraza (entiteta) i odnosa** koji se u medicini koristi za precizno označavanje i opisivanje bolesti, stanja, dijagnoza, postupaka, anatomskih struktura i drugih biomedicinskih pojmova, s ciljem smanjenja ili potpunog uklanjanja dvosmislenosti u komunikaciji.
- Sastoji se od:
  - **Jedinstvenih identifikatora** (često u obliku kodova, npr. ICD-9,10,11, LOINC Code, SNOMED CT ID)
  - **Naziva koncepata** (preferirani nazivi biomedicinskih koncepata)
  - **Velikog broja sinonima**
- **Primjeri poznatih terminologija**:
  - *MeSH* – koristi se u PubMedu i drugim bazama podataka za indeksiranje i pretraživanje biomedicinske literature.
  - *SNOMED CT* – klinička terminologija koja obuhvaća dijagnoze, simptome, postupke; često se koristi u EHR
  - *ICD 10* – koristi se za klasifikaciju bolesti i srodnih zdravstvenih problema
  - *LOINC* – standard za kodiranje laboratorijskih i kliničkih mjerenja
  - *RxNorm* – standard za lijekove i njihove doze

[14] National Library of Medicine. (2024). *Unified Medical Language System (UMLS)*. U.S. Department of Health and Human Services

Unified Medical Language System (UMLS)

Problemi vezani uz biomedicinske terminologije [14] [15]

- **Ne postoji jedinstvena, sveobuhvatna terminologija za biomedicinsko područje**
  - Postoji velik broj biomedicinskih terminologija koje se koriste u različite svrhe, u različitim državama i organizacijama, te se razlikuju prema podatkovnim formatima, identifikatorima, vezama između entiteta i dr.
- **Terminologije se često ne nadovezuju jedna na drugu**, a pristup mnogima od njih nije javno dostupan.

[14] National Library of Medicine. (2024). *Unified Medical Language System (UMLS)*. U.S. Department of Health and Human Services

[15] Awaysheh, Abdullah, et al. "A review of medical terminology standards and structured reporting." *Journal of veterinary diagnostic investigation* 30.1 (2018): 17-25.

Unified Medical Language System (UMLS)

Primjer: Diabetes mellitus

- **CUI** (*Concept Unique Identifier*): jedinstveni UMLS identifikator općeg biomedicinskog koncepta
  - *C0011849 – Diabetes Mellitus* (preferirani naziv)
- *Primjer upita nad e-kartonima* [16] : “Return all patients ona GLP-1 medication with an HBAic > 10.”
  - Cilj: **dohvat svih „pacijenata s dijabetesom” prema navedenom lijeku koji uzimaju i laboratorijskom rezultatu**
  - Problem: Postoji veliki broj **specifičnih dijagnoza** vezanih uz medicinski koncept „Diabetes mellitus”

[14] National Library of Medicine. (2024). *Unified Medical Language System (UMLS)*. U.S. Department of Health and Human Services

[16] National Library of Medicine. (2016, July 29). *UMLS Quick Start Guide, pristupljeno: 22. 10. 2025., dostupno na: https://www.nlm.nih.gov/research/umls/quickstart.html*

Unified Medical Language System (UMLS)

Primjer: Diabetes mellitus

CUI: *C0011849 – Diabetes Mellitus* (preferirani naziv medicinskog koncepta)

- *Primjer upita nad e-kartonima* [16] "Return all patients on a GLP-1 medication with an HBAic > 10."
  - Cilj: **dohvat svih „pacijenata s dijabetesom" prema navedenom lijeku koji uzimaju i laboratorijskom rezultatu**
  - Problem: Postoji veliki broj **specifičnih dijagnoza** vezanih uz medicinski koncept *Diabetes mellitus*

Iz skupa je dohvaćeno je preko **2666 različitih dijagnoza** koje opisuju isti medicinski koncept *Diabetes mellitus*, kroz višestruki broj terminologija.

- Acidosis due to type 1 diabetes mellitus
- Acidosis due to type 2 diabetes mellitus
- Acute complication with diabetes mellitus
- Diabetic dyslipidemia associated with type 2 diabetes mellitus
- Diabetic hyperosmolar non-ketotic state
- Diabetic lumbosacral radiculoplexus neuropathy
- Diabetic mastopathy
- Diabetic severe hyperglycemia
- Disorder of nerve co-occurrent and due to type 1 diabetes mellitus
- Dyslipidemia due to type 1 diabetes mellitus
- Hyperglycemia due to type 1 diabetes mellitus
- Hyperglycemia due to type 2 diabetes mellitus
- Hyperglycemic crisis in diabetes mellitus
- Hyperlipidemia due to type 1 diabetes mellitus
- Hyperlipidemia due to type 2 diabetes mellitus
- Hyperosmolality due to uncontrolled type 1 diabetes mellitus
- Hyperosmolar coma associated with diabetes mellitus
- Hyperosmolarity co-occurrent and due to drug induced diabetes mellitus
- Hypoglycemia due to type 1 diabetes mellitus
- Hypoglycemia due to type 2 diabetes mellitus
- Hypoglycemic coma in diabetes mellitus
- Hypoglycemic state in diabetes
- Ketoacidosis due to secondary diabetes mellitus
- Lactic acidosis with diabetes mellitus
- Malnutrition-related diabetes mellitus with multiple complications
- Metabolic acidosis with diabetes mellitus
- Mixed hyperlipidemia due to type 1 diabetes mellitus
- Mixed hyperlipidemia due to type 2 diabetes mellitus
- Multiple complications due to diabetes mellitus
- Peripheral neuropathy due to type 1 diabetes mellitus
- Radiculoplexoneuropathy due to diabetes mellitus
- Type 1 diabetes mellitus with hyperosmolar coma
- Type 2 diabetes mellitus with hyperosmolar coma
- Diabetic acute painful polyneuropathy
- Coma associated with malnutrition-related diabetes mellitus
- Diabetic coma with ketoacidosis
- Hyperosmolar coma associated with diabetes mellitus
- Hypoglycemic coma co-occurrent and due to diabetes mellitus type II
- ETC

[14] National Library of Medicine. (2024). *Unified Medical Language System (UMLS)*. U.S. Department of Health and Human Services

[16] National Library of Medicine. (2016, July 29). *UMLS Quick Start Guide, pristupljeno: 22. 10. 2025., dostupno na: https://www.nlm.nih.gov/research/umls/quickstart.html*

Named entity recognition (NER)

- **Named-entity recognition** (NER) predstavlja podzadatak ekstrakcije informacija (*engl. information extraction*) iz teksta, čiji je cilj **prepoznavanje** i **klasifikacija** imenovanih entiteta u nestrukturiranom tekstu u unaprijed definirane kategorije. [17]

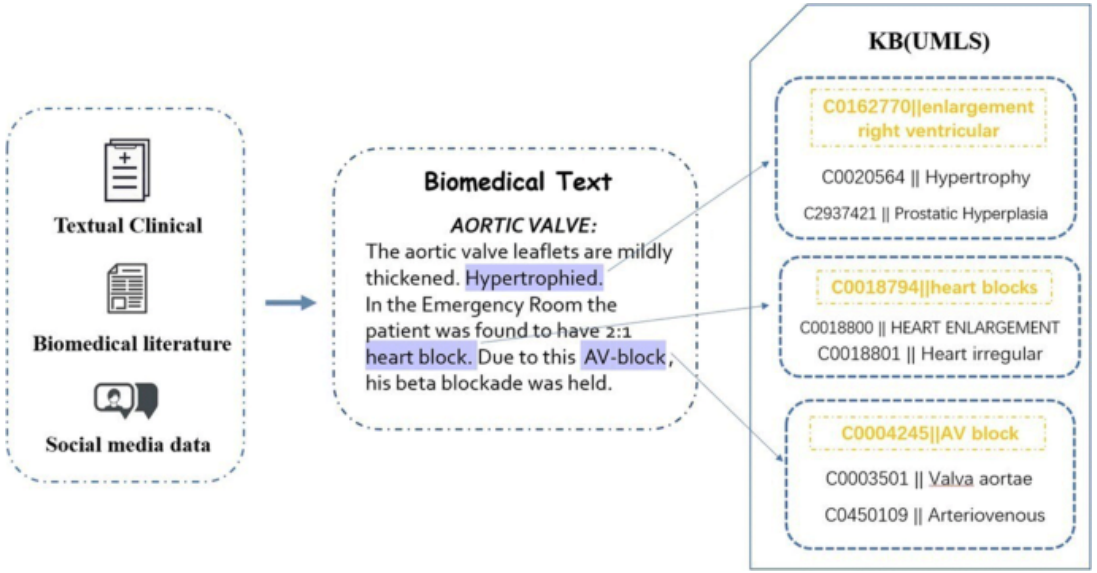
Dr. Smith prescribed 500mg of Amoxicilin to treat John's bacterial infection.

Dr. Smith PERSON	500mg DOSAGE	Amoxicilin DRUG	John PATIENT	Bacterial infection DISEASE

[17] Named-entity recognition (NER), Wikipedia, pristupljeno: 19. 10. 2025. dostupno na: https://en.wikipedia.org/wiki/Named-entity_recognition

Entity linking (EL)

- **Entity linking** (EL) predstavlja složeniji oblik NER-a koji ne samo da identificira i klasificira pojmove unutar teksta, već ih i **povezuje s jedinstvenim identifikatorima** odnosno konceptima unutar određene baze znanja (primjer: Wikipedia) [18]
- **Biomedical entity linking (BEL)** odnosi se na proces povezivanja medicinskih koncepata, prepoznatih u tekstu, s odgovarajućim jedinstvenim identifikatorima iz relevantnih baza znanja, poput **UMLS CUI** sustava [19]



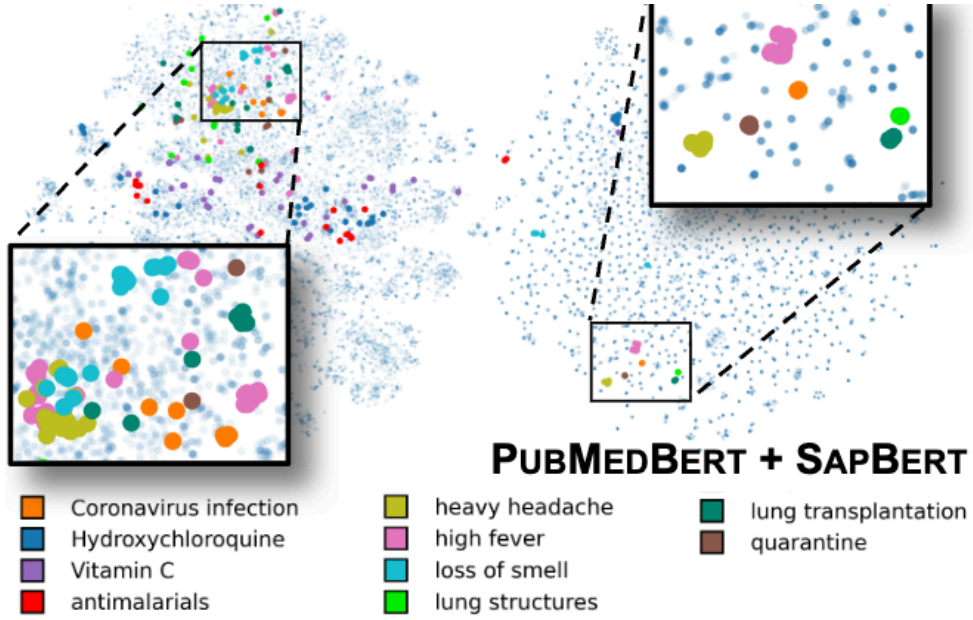
[18] Entity linking (EL), Wikipedia, pristupljeno: 21. 10. 2025. dostupno na: https://en.wikipedia.org/wiki/Entity_linking

[19] French, E. (2023). *An overview of biomedical entity linking throughout the years*. *Journal of Biomedical Informatics*, 132

Biomedical entity linking (BEL)

Rana faza istraživanja

- Rani sustavi za povezivanje biomedicinskih entiteta, poput *MetaMapa*[20] i *cTAKES*-a[21], oslanjali su se na **ručno pravila, heuristike i leksičko podudaranje** za prepoznavanje i preslikavanje medicinskih entiteta na UMLS koncepte.
- Alati poput *QuickUMLS*-a[22] i *ScispaCy*-a[23] unaprijedili su taj pristup uvođenjem **aproksimativnog podudaranja nizova** i **TF–IDF reprezentacija** riječi, čime su postigli veću brzinu i usporedivu preciznost.
- Razvojem **dubokog učenja**, *fine-tuned* BERT modeli poput *BioBERT*-a[24], SapBERT-a[25] i BioSyn-a[26] ostvaruju značajan napredak u **preciznosti i semantičkom razumijevanju biomedicinskih pojmova.**
- Rješavanje dvosmislenosti, usklađivanje teksta s strukturiranom bazom znanja i standardizacija *benchmarkova* i dalje su uglavnom nedovoljno istraženi u ovoj domeni.



[20] Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American medical informatics association*, *17*(3), 229-236.
[21] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507-513.
[22] Soldaini, L., & Goharian, N. (2016, July). Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir* (pp. 1-4).
[23] Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.
[24] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.
[25] Liu, Fangyu, et al. "Self-alignment pretraining for biomedical entity representations." arXiv preprint arXiv:2010.11784 (2020).
[26] Sung, M., Jeon, H., Lee, J., & Kang, J. (2020). Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*.

Zaključak i daljnja istraživanja

- Povezivanje biomedicinskih pojmova iz nestrukturiranih tekstova s jedinstvenim identifikatorima koncepata predstavlja složen izazov zbog velike **granularnosti podataka**, **raznolikosti medicinskih terminologija**, prisutnosti **sinonimije** i **polisemije** te različitih oblika šuma u podacima – primjerice, nestandardiziranih kratica ili izraza preuzetih s društvenih mreža.
- **Ne postoji jedinstvena ontologija koja obuhvaća sve biomedicinske koncepte**, no postoje baze znanja, poput **UMLS**-a, koje integriraju i povezuju medicinske terminologije.
- **Biomedical entity linking - svrha:** *Information Extraction*, *Knowledge graph contruction*, C*linical decision support*, indeksiranje i pretraživanje biomedicinskih publikacija.
- **Daljnje istraživanje** temeljit će se na analizi suvremenih jezičnih modela, skupova podataka i drugih pristupa za povezivanje biomedicinskih termina u tekstu s medicinskim konceptima
- Posebna će se pažnja usmjeriti na istraživanje **primjene velikih jezičnih modela** i **agentskih sustava** koji mogu pristupati bazama znanja, provoditi inferenciju klasičnih BIO-NLP modela (npr. za sažimanje ili proširivanje skupa kandidata) te integrirati različite alate u svrhu postizanja preciznije identifikacije medicinskih termina.

Literatura

[1] Buljan, A. i Šimović, H. (2022). Učinkovitost hrvatskog zdravstvenog sustava - usporedba sa zemljama Europske unije. *Revija za socijalnu politiku, 29* (3), 321-354. https://doi.org/10.3935/rsp.v29i3.1933

[2] Bobinac A. Access to Healthcare and Health Literacy in Croatia: Empirical Investigation. Healthcare (Basel). 2023

[3] Popović, Stjepka. "Odrednice stavova i zadovoljstva građana hrvatskim zdravstvenim sustavom." *Medicina Fluminensis*, vol. 53, br. 1, 2017, str. 85-100. https://doi.org/10.21860/medflum2017_173385.

[4] Eurostat – State of Health in the EU: „Hrvatska: Pregled stanja zdravlja i zdravstvene zaštite 2023”

[5] Blašković, L.; Tanković, N.; Lorencin, I.; Baressi Šegota, S. Robust Clinical Querying with Local LLMs: Lexical Challenges in NL2SQL and Retrieval-Augmented QA on EHRs. Big Data Cogn. Comput. 2025, 9, 256.

[6] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3.1 (2016): 1-9.

[7] Wang, Ping, Tian Shi, and Chandan K. Reddy. "Text-to-SQL generation for question answering on electronic medical records." *Proceedings of The Web Conference 2020*. 2020.

[8] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.

[9]Natural language understanding. *Wikipedia, pristupljeno 15. 10. 2025., dostupno na: https://en.wikipedia.org/wiki/Natural_language_understanding*

[10] Ministarstvo zdravstva Republike Hrvatske – Bolničke zdravstvene ustanove. Dostupno na: https://zdravlje.gov.hr/kontakti/kontakti-zdravstvenih-ustanova/bolnicke-zdravstvene-ustanove/2722

[11] Osvaldić, Josipa. "Information system implementation in healthcare: case study of Croatia." *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy* 12.2 (2021): 114-124.

[12] Kong H. J. (2019). Managing Unstructured Big Data in Healthcare System. Healthcare informatics research, 25(1), 1–2.

Capurro, D., Yetisgen, M., van Eaton, E., Black, R., & Tarczy-Hornoch, P. (2014). Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement:

[13] Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847

[14] National Library of Medicine. (2024). *Unified Medical Language System (UMLS)*. U.S. Department of Health and Human Services

[15] Awaysheh, Abdullah, et al. "A review of medical terminology standards and structured reporting." *Journal of veterinary diagnostic investigation* 30.1 (2018): 17-25.

[16] National Library of Medicine. (2016, July 29). *UMLS Quick Start Guide, pristupljeno: 22. 10. 2025., dostupno na:* https://www.nlm.nih.gov/research/umls/quickstart.html

[17] Named-entity recognition (NER), Wikipedia, pristupljeno: 19. 10. 2025. dostupno na: https://en.wikipedia.org/wiki/Named-entity_recognition

[18] Entity linking (EL), Wikipedia, pristupljeno: 21. 10. 2025. dostupno na: https://en.wikipedia.org/wiki/Entity_linking

[19] French, E. (2023). *An overview of biomedical entity linking throughout the years*. *Journal of Biomedical Informatics*, 132

[20] Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American medical informatics association*, *17*(3), 229-236.

[21] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507-513.

[22] Soldaini, L., & Goharian, N. (2016, July). Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir* (pp. 1-4).

[23] Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

[24] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

[25] Liu, Fangyu, et al. "Self-alignment pretraining for biomedical entity representations." arXiv preprint arXiv:2010.11784 (2020).

[26] Sung, M., Jeon, H., Lee, J., & Kang, J. (2020). Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*.

Sveučilište u Rijeci
Fakultet informatike i digitalnih tehnologija

uniri

Reliable Biomedical NLP via Ontology integration: Leveraging UMLS to Align Language Models with Clinical Terminologies and Use Cases

Doktorand: Luka Blašković

Hvala na pažnji!

INFCON25 – 4. Znanstveni skup doktorskih studenata informatike

Rijeka, 24. 10. 2025.