UDC: 616-002.5:004.89:004.75
DOI: 10.32902/2663-0338-9-2026-3-5-9

A multi-agent artificial intelligence system to support decision-making in the routing and treatment of patients with tuberculosis

I.O. Semianiv[1], Ya.I. Vyklyuk[2], B.M. Luginin[2]
1. Bukovinian State Medical University, Chernivtsi, Ukraine
2. Lviv Polytechnic National University, Lviv, Ukraine

Conflict of interest: none

BACKGROUND. Tuberculosis (TB) remains one of the greatest global threats to public health. The referral of a patient with suspected TB, particularly in the presence of HIV co-infection, is a multi-stage process requiring coordinated collaboration between the general practitioner, the TB specialist and the infectious diseases specialist; the fragmented exchange of clinical information between them increases the risk of diagnostic errors and delays the start of specific treatment.
OBJECTIVE. To develop and clinically validate a multi-agent artificial intelligence (AI) information system to support patient referral and decision-making for TB patients, based on approved medical protocols.
MATERIALS AND METHODS. The 'MedCouncil AI' system was developed based on five specialised AI agents (general practitioner, TB specialist, infectious diseases specialist, medical coordinator, medical analyst), integrated with a retrieval-augmented generation (RAG) subsystem, which is based on protocols from the Ministry of Health of Ukraine and World Health Organization recommendations. Validation was carried out on 50 synthetic clinical cases (including TB/HIV co-infection and multidrug-resistant TB) using eight large language models, evaluated using the ROUGE-L and BERTScore metrics in both RAG-enabled and RAG-disabled modes.
RESULTS. The integration of RAG improved the BERTScore for all tested models; the most pronounced effect was observed for compact local models (Qwen3:4b: from 0.619 to 0.818; Josiefied 8b: from 0.500 to 0.812), accompanied by a reduction in the risk of clinically dangerous 'hallucinations' from a critical to a minimal level. The Gemini 3 Flash model with RAG (BERTScore = 0.805) demonstrated the optimal balance between accuracy, response generation speed (3-4 s) and cost.
CONCLUSIONS. A multi-agent system based on RAG increases the reliability of automatically generated medical conclusions in the management of TB patients and, provided a data de-identification module is used, can serve as an auxiliary tool ('second opinion') to support patient routing in the context of staff shortages within the TB service.

KEY WORDS: tuberculosis, artificial intelligence, large language models, clinical decision support system, multi-agent systems, patient routing.

Багатоагентна система штучного інтелекту для підтримки прийняття рішень при маршрутизації та лікуванні хворих на туберкульоз

І.О. Сем'янів[1], Я.І. Виклюк[2], Б.М. Лугінін[2]
1. Буковинський державний медичний університет, м. Чернівці, Україна
2. Національний університет «Львівська політехніка», м. Львів, Україна

Конфлікт інтересів: відсутній

ОБҐРУНТУВАННЯ. Туберкульоз (ТБ) залишається однією з найбільших глобальних загроз громадському здоров'ю. Маршрутизація хворого з підозрою на ТБ, особливо за наявності коінфекції з ВІЛ, є багатоетапним процесом, що потребує узгодженої взаємодії сімейного лікаря, фтизіатра й інфекціоніста; фрагментованість обміну клінічною інформацією між ними підвищує ризик діагностичних помилок і затримує старт специфічної терапії.
МЕТА. Розроблення та клінічна апробація багатоагентної інформаційної системи штучного інтелекту (ШІ) для підтримки маршрутизації та прийняття рішень у хворих на ТБ з опорою на затверджені медичні протоколи.

МАТЕРІАЛИ ТА МЕТОДИ. Розроблено систему MedCouncil AI на основі п'яти спеціалізованих ШІ-агентів (сімейний лікар, фтизіатр, інфекціоніст, медичний координатор, медичний аналітик), інтегровану з підсистемою retrieval-augmented generation (RAG), що спирається на протоколи Міністерства охорони здоров'я України та рекомендації Всесвітньої організації охорони здоров'я. Апробацію проведено на 50 синтетичних клінічних випадках (включно з коінфекцією ТБ/ВІЛ і мультирезистентним ТБ) із застосуванням восьми великих мовних моделей, оцінених за метриками ROUGE-L і BERTScore у режимах з RAG і без нього.
РЕЗУЛЬТАТИ. Інтеграція RAG підвищувала BERTScore для всіх протестованих моделей; найвираженіший ефект зафіксовано для компактних локальних моделей (Qwen3:4b: з 0,619 до 0,818; Josiefied 8b: з 0,500 до 0,812), що супроводжувалося зниженням ризику клінічно небезпечних «галюцинацій» із критичного до мінімального рівня. Оптимальне співвідношення точності, швидкості генерації відповіді (3-4 с) та вартості продемонструвала модель Gemini 3 Flash з RAG (BERTScore = 0,805).
ВИСНОВКИ. Багатоагентна система з опорою на RAG підвищує достовірність автоматично згенерованих медичних висновків при веденні хворих на ТБ і, за умови застосування модуля деперсоніфікації даних, може використовуватися як допоміжний інструмент («друга думка») для підтримки маршрутизації пацієнтів в умовах кадрового дефіциту фтизіатричної служби.

КЛЮЧОВІ СЛОВА: туберкульоз, штучний інтелект, великі мовні моделі, система підтримки прийняття клінічних рішень, багатоагентні системи, маршрутизація пацієнтів.

Tuberculosis (TB) remains one of the greatest global threats to public health: approximately 10.8 million new cases are recorded worldwide each year [1]. Ukraine consistently ranks among the countries with a high burden of multidrug-resistant TB (MDR-TB), the treatment of which requires an individualised selection of second-line drugs and long-term monitoring of adverse reactions. The full-scale war has significantly exacerbated the epidemiological situation: around 2,000 healthcare facilities, including TB clinics, have been damaged or destroyed, whilst the shortage of TB specialists in certain regions has reached 16-40 per cent [2, 3].

The referral pathway for a patient with suspected TB – particularly where there is co-infection with HIV – is a multi-stage process requiring coordinated collaboration between the general practitioner, the TB specialist and the infectious diseases specialist. In practice, the exchange of clinical information between these specialists is asynchronous and fragmented (primarily via formal discharge summaries and electronic referrals lacking full clinical context), which increases the risk of diagnostic errors and delays the start of specific treatment; every day of such a delay for a patient who is shedding bacteria poses an additional risk of infecting those around them.

Existing Clinical Decision Support Systems based on rigid rules or classical machine learning are unable to adequately process unstructured clinical text (complaints, medical history) and cannot justify their own recommendations (the 'black box' problem), which undermines doctors' trust [4, 5]. The emergence of Large Language Models (LLMs) has opened up new possibilities for analysing free-form medical text and generating clinically sound conclusions [6, 7]. At the same time, the use of a single (monolithic) language model for complex clinical scenarios (for example, the co-infection of TB and HIV) carries the risk of generating factually inaccurate recommendations ('hallucinations') and requires a solution to the problem of medical data confidentiality when using cloud services [8].

Objective. To develop and clinically validate a multi-agent artificial intelligence (AI) information system to support patient pathway management and decision-making for TB patients, capable of processing unstructured clinical data based on approved medical protocols and ensuring the protection of medical confidentiality.

Materials and methods

The 'MedCouncil AI' information system has been developed, which simulates the work of a medical consultation by means of sequential interaction between five specialised, autonomous AI agents: 'General Practitioner' (initial assessment of symptoms and referral), 'TB Specialist' (assessment of the likelihood of TB and MDR/XDR resistance profiles), 'Infectious Diseases Specialist' (differential diagnosis with other infections, monitoring of HIV status and antiretroviral therapy), 'Chief Medical Coordinator' (consolidation of specialists' conclusions in 'Consultation' or 'Diagnosis' mode) and the background agent 'Medical Analyst' (automatic generation of a dynamic clinical profile for the patient). A general diagram of how the agents interact is shown in fig. 1.

To minimise the risk of generating factually unsubstantiated recommendations, a retrieval-augmented generation (RAG) subsystem based on a vector database has been integrated into the architecture; this database contains the official protocols of the Ministry of Health of Ukraine and World Health Organization (WHO) recommendations regarding the diagnosis and treatment of TB and co-infections. Before generating a response, specialist agents perform a semantic search for relevant fragments of the protocols, which are added to the generation context, thereby limiting the response to the available evidence base.

The system was tested on a dataset of 50 synthetic clinical cases (including TB/HIV co-infection and MDR-TB), generated on the basis of current clinical guidelines, as the use of real, de-identified medical records is restricted by data protection legislation. Eight LLMs across three architectural

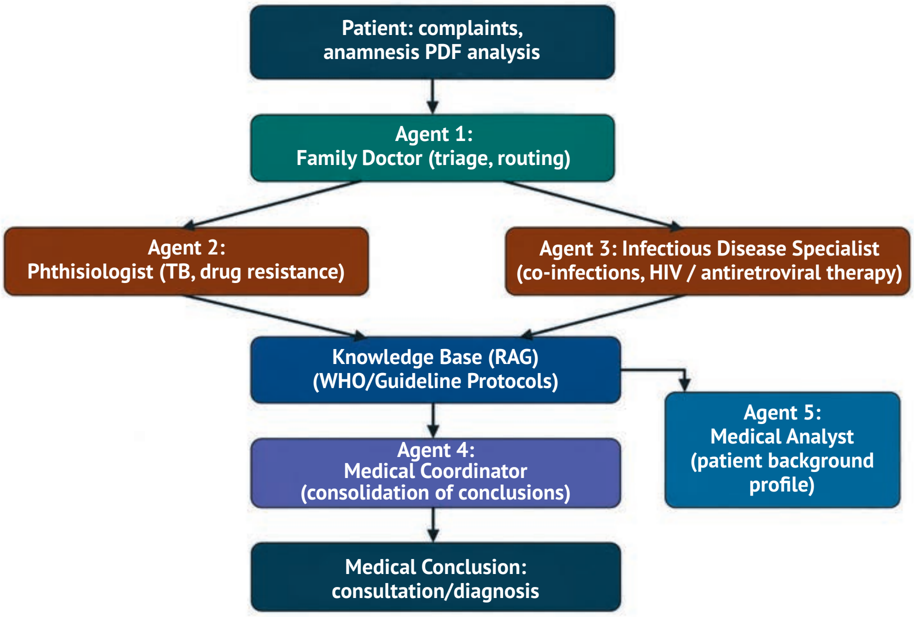

Fig. 1. Diagram of the interaction between AI agents in the 'MedCouncil AI' decision support system

categories were tested: flagship cloud-based models (Gemini 3.1 Pro, GPT-4o, Claude 3.5 Sonnet), cloud-based open models (Llama 3.1 70B, Mixtral 8x7B), an optimised cloud model (Gemini 3 Flash) and on-premises models for deployment on healthcare facility servers (Qwen3:4b, Josiefied 8b). Each model was evaluated in two modes – with and without the RAG subsystem.

The quality of the generated medical conclusions was compared with the responses of an expert doctor (reference text) using two complementary natural language processing metrics: ROUGE-L, which assesses the match of key lexical units (in particular, drug names and dosages) whilst taking word order into account, and BERTScore, which assesses the deep semantic correspondence of the text regardless of its literal form (resilient to synonymy, for example 'fever' and 'elevated temperature'). In addition, we qualitatively assessed the risk of clinically dangerous 'hallucinations', the response generation speed (latency) and the approximate cost of using the model.

To comply with medical data confidentiality requirements (in line with GDPR/HIPAA), a local de-identification module has been developed: the patient's personal identifiers (surname, first name, patronymic, date of birth, address) are processed and stored exclusively within the facility's secure network, whilst only de-identified clinical descriptions (age, gender, symptoms, test results) are transmitted to the cloud-based model.

Results and discussion

The Gemini 3.1 Pro model, combined with RAG, demonstrated the highest clinical accuracy of the generated conclusions (BERTScore F1 = 0.835), which corresponds to a level acceptable for considering the system as a 'second opinion' tool. However, the most telling result was that the integration of RAG had the most significant impact on compact local models, which can be deployed directly on a healthcare facility's servers without transferring data to external networks. For example, for the Qwen3:4b model, the BERTScore increased from 0.619 (without RAG, which corresponded to a critical risk of 'hallucinations', in particular the generation of non-existent treatment regimens) to 0.818 (with RAG, minimal risk); for the Josiefied 8b model – from 0.500 to 0.812 respectively (table, fig. 2).

Table. Accuracy of medical conclusions from the AI models studied (N=50)

Model	ROUGE-L (F1) without/with RAG	BERTScore (F1) without/with RAG	Risk of critical hallucinations
Gemini 3.1 Pro	0.645/0.775	0.765/0.835	Moderate → None
GPT-4o	0.630/0.770	0.755/0.825	Moderate → None
Claude 3.5 Sonnet	0.625/0.765	0.745/0.815	Moderate → None
Gemini 3 Flash	0.610/0.760	0.730/0.805	Moderate → Minimal
Llama 3.1 70B	0.500/0.690	0.680/0.760	High → Low
Mixtral 8x7B	0.490/0.680	0.650/0.740	High → Low
Qwen3:4b	0.450/0.700	0.619/0.818	Critical → Minimal
Josiefied 8b	0.400/0.650	0.500/0.812	Critical → Minimal

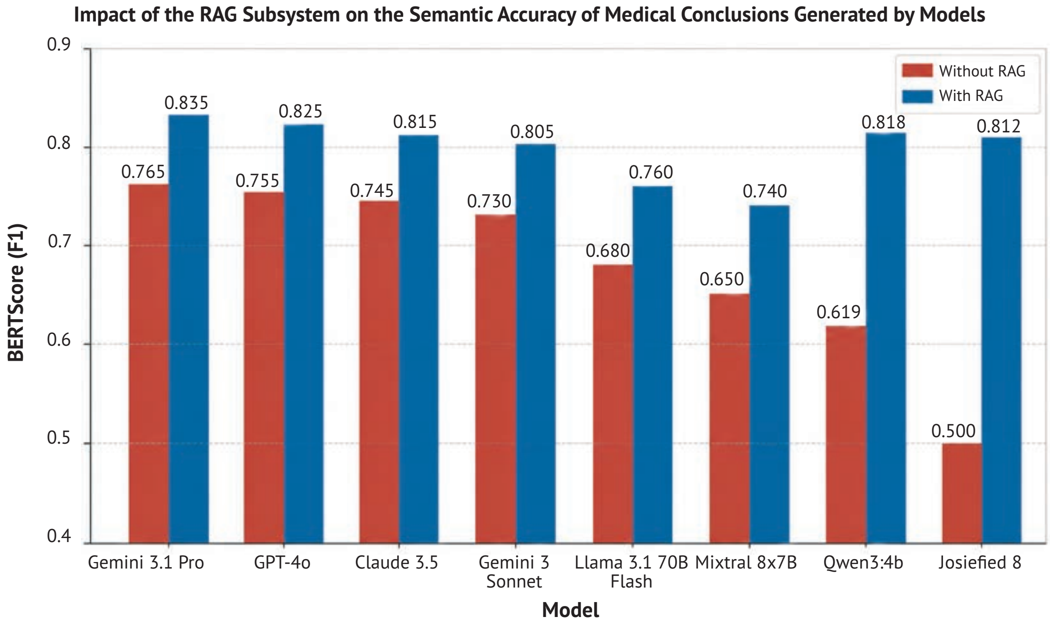

Fig. 2. The impact of the RAG subsystem on the semantic accuracy (BERTScore) of medical conclusions generated by the models under investigation

A similar, albeit less pronounced, trend was observed for cloud-based models: adding RAG increased the BERTScore by between 0.07 and 0.08 points for all tested LLMs, whilst the risk of clinically dangerous hallucinations decreased from 'medium/high' to 'none/minimal' for virtually all models.

Based on a combination of accuracy metrics, response generation speed (3-4 s versus 8-10 s for heavy cloud-based models) and cost of use, the optimised cloud-based Gemini 3 Flash model combined with RAG (BERTScore = 0.805, ROUGE-L = 0.760), which, in terms of clinical accuracy, comes close to flagship models, lagging behind by only a few percentage points.

The results confirm that aligning medical text generation with the approved evidence base (protocols of the Ministry of Health of Ukraine and WHO recommendations) is a key factor in the safe use of language models in clinical practice, particularly for compact models with a limited number of parameters. This has practical significance for healthcare facilities with limited technical resources: deploying a local model with a RAG module allows an acceptable level of clinical accuracy to be achieved without transferring confidential patient data outside the facility's secure perimeter.

The de-identification module developed – which replaces the patient's identifying data with an anonymised clinical vector before the data is sent to the cloud-based model – allows the benefits of optimised cloud solutions (speed, accuracy) to be combined with compliance with medical confidentiality requirements.

Practical testing of the system in a simulated consultation setting has shown that the automatic background extraction of clinically significant data (complaints, medical history, risk factors, allergies) into a dynamic patient profile has the potential to reduce the time a doctor spends re-summarising information when referring a patient between stages (general practitioner → TB specialist → infectious diseases specialist), which directly addresses the issue of the 'documentation burden', which is estimated to account for up to 45 per cent of a doctor's working time [9].

Conclusions

The developed multi-agent AI system, which simulates the work of a medical consultation and relies on retrieval-augmented generation technology to align with approved clinical protocols, is capable of significantly improving the reliability of automatically generated medical conclusions in the management of TB patients, particularly in cases of co-infection with HIV and MDR-TB. The integration of the RAG subsystem reduces the risk of clinically dangerous 'hallucinations' by language models, especially compact ones suitable for on-premises deployment without transferring personal data outside the healthcare facility. The combination of an optimised language model with a local de-identification module ensures an acceptable balance between clinical accuracy, response speed and compliance with confidentiality requirements, which provides grounds for considering the developed approach as a supportive tool ('second opinion') to support referral and decision-making in TB cases, particularly in the context of staff shortages within the TB service.

INFUSION & CHEMOTHERAPY • ISSN 2663-0338

Література/References

1. World Health Organization. Global tuberculosis report 2024. Geneva: WHO; 2024.
2. Ministry of Health of Ukraine. Over two years of full-scale war: Russia destroyed 206 and damaged 1,546 healthcare facilities [Internet]. 2024.
3. Alliance for Public Health (APH) Ukraine. Fighting TB under fire: Ukraine's resilient and innovative response. 2025.
4. Singhal K., Azizi S., Tu T., et al. Large language models encode clinical knowledge. *Nature*. 2023; 620: 172-180.
5. Thirunavukarasu A.J., Ting D.S.J., Elangovan K., et al. Large language models in medicine. *Nat. Med*. 2023; 29 (8): 1930-1940.
6. Nori H., King N., McKinney S.M., et al. Can generalist foundation models outperform special-purpose tuning? A case study in medicine. arXiv: 2311.16452, 2023.
7. Wu Q., Bansal G., Zhang J., et al. AutoGen: enabling next-generation LLM applications via multi-agent conversation. arXiv: 2308.08155, 2023.
8. Lewis P., Perez E., Piktus A., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural. Inf. Process. Syst.* 2020; 33: 9459-9474.
9. Kuziemsky C., Maeder A.J., John O., et al. The role of artificial intelligence in the telehealth sector. *Yearb. Med. Inform*. 2023; 32 (01): 42-50.
10. Zhang T., Kishore V., Wu F., Weinberger K.Q., Artzi Y. BERTScore: evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR); 2020.
11. Lin C.Y. ROUGE: a package for the automatic evaluation of summaries. In: Text Summarisation Branches Out. Barcelona: Association for Computational Linguistics; 2004. Pp. 74-81.

ВІДОМОСТІ ПРО АВТОРІВ / INFORMATION ABOUT AUTHORS

Сем'янів Ігор Олександрович
Доцент закладу вищої освіти, кафедра фтизіатрії та пульмонології Буковинського державного медичного університету.
Канд. мед. наук, доцент.
2, пл. Театральна, м. Чернівці, 58002, Україна.
ORCID iD: orcid.org/0000-0003-0340-0766

Виклюк Ярослав Ігорович
Професор кафедри систем штучного інтелекту Національного університету «Львівська політехніка».
Д-р мед. наук, професор.
12, вул. С. Бандери, м. Львів, 79000, Україна.
ORCID iD: orcid.org/0000-0003-4766-4659

Лугінін Богдан Миколайович
Студент інституту комп'ютерних наук та інформаційних технологій, Національний університет «Львівська політехніка».
12, вул. С. Бандери, м. Львів, 79000, Україна.

Semianiv Ihor Oleksandrovych
Associate Professor, Department of Phthisiology and Pulmonology, Bukovinian State Medical University.
PhD in Medicine, Associate Professor.
2, Teatralna sq., Chernivtsi, 58002, Ukraine.
ORCID iD: orcid.org/0000-0003-0340-0766

Vyklyuk Yaroslav Ihorovych
Professor of the Department of Artificial Intelligence Systems, Lviv Polytechnic National University.
MD, Professor.
12, S. Bandery st., Lviv, 79000, Ukraine.
ORCID iD: orcid.org/0000-0003-4766-4659

Luginin Bohdan Mykolaiovych
Student of the Institute of Computer Science and Information Technologies, Lviv Polytechnic National University.
12, S. Bandery st., Lviv, 79000, Ukraine.

КОНТАКТНА ІНФОРМАЦІЯ / CORRESPONDENCE TO
Сем'янів Ігор Олександрович
2, пл. Театральна, м. Чернівці, 58002, Україна.
E-mail: igor_semianiv@bsmu.edu.ua

Authors' contributions

Data collection, data processing – I.O. Semianiv, Ya.I. Vyklyuk; research concept and design – Ya.I. Vyklyuk, I.O. Semianiv, B.M. Luginin; text drafting – I.O. Semianiv, B.M. Luginin; software implementation and experimental studies – Ya.I. Vyklyuk, B.M. Luginin; text editing – I.O. Semianiv, Ya.I. Vyklyuk.

Надійшла до редакції / Received: 21.07.2026 ▪ Прийнято до друку / Accepted: 04.08.2026 ▪ Опубліковано / Published: 24.09.2026