LuminaURO: A comprehensive Artificial Intelligence Driven Assistant for enhancing urological diagnostics and patient care

LuminaURO: Ürolojik tanı ve hasta bakımını geliştirmek için kapsamlı bir Yapay Zeka Destekli Asistan

Abstract

Aim: This study aims to develop and validate LuminaURO, a Retrieval-Augmented Generation (RAG)-based AI Assistant specifically designed for urological healthcare, addressing the limitations of conventional Large Language Models (LLMs) in healthcare applications.

Methods: We developed LuminaURO using a specialized repository of urological documents and implemented a novel pooling methodology to search multilingual documents and aggregate information for response generation. The system was evaluated using multiple similarity algorithms (OESM, Spacy, T5, and BERTScore) and expert assessment by urologists (n=3).

Results: LuminaURO generates responses within 8-15 seconds from multilingual documents and enhances user interaction by providing two contextually relevant follow-up questions per query. The architecture demonstrates significant improvements in search latency, memory requirements, and similarity metrics compared to state-of-the-art approaches. Validation shows similarity scores of 0.6756, 0.7206, 0.9296, 0.9223, and 0.9183 for English responses, and 0.6686, 0.7166, 0.8119, 0.9220, 0.9315, and 0.9086 for Turkish responses. Expert evaluation by urologists revealed similarity scores of 0.9444 and 0.9408 for English and Turkish responses, respectively.

Conclusion: LuminaURO successfully addresses the limitations of conventional LLM implementations in healthcare by utilizing specialized urological documents and our innovative pooling methodology for multilanguage document processing. The high similarity scores across multiple evaluation metrics and strong expert validation confirm the system's effectiveness in providing accurate and relevant urological information. Future research will focus on expanding this approach to other medical specialties, with the ultimate goal of developing LuminaHealth, a comprehensive healthcare assistant covering all medical domains.

Keywords: Artificial intelligence; decision support systems; natural language processing; medical informatics; urology

Öz

Amaç: Bu çalışma, ürolojik sağlık hizmetleri için özel olarak tasarlanmış, Erişim-Güçlendirilmiş Üretim (RAG) tabanlı bir yapay zeka asistanı olan LuminaURO'yu geliştirmeyi ve doğrulamayı amaçlamaktadır. Bu sistem, sağlık uygulamalarında geleneksel Büyük Dil Modellerinin (LLM) sınırlamalarını ele almaktadır.

Yöntemler: LuminaURO'yu ürolojik dokümanların özel bir deposunu kullanarak geliştirdik ve çok dilli dokümanları aramak ve yanıt üretimi için bilgileri toplamak amacıyla yenilikçi bir havuzlama metodolojisi uyguladık. Sistem, çoklu benzerlik algoritmaları (OESM, Spacy, T5 ve BERTScore) ve ürologlar tarafından uzman değerlendirmesi (n=3) kullanılarak değerlendirildi.

Bulgular: LuminaURO, çok dilli dokümanlardan 8-15 saniye içinde yanıtlar üretmekte ve her sorgu için bağlamsal olarak ilgili iki takip sorusu sunarak kullanıcı etkileşimini geliştirmektedir. Mimari, son teknoloji yaklaşımlara kıyasla arama gecikmesi, bellek gereksinimleri ve benzerlik metrikleri açısından önemli iyileştirmeler göstermektedir. Doğrulama, İngilizce yanıtlar için 0,6756, 0,7206, 0,9296, 0,9223 ve 0,9183, Türkçe yanıtlar için ise 0,6686, 0,7166, 0,8119, 0,9220, 0,9315 ve 0,9086 benzerlik puanları göstermektedir. Ürologlar tarafından yapılan uzman değerlendirmesi, sırasıyla İngilizce ve Türkçe yanıtlar için 0,9444 ve 0,9408 benzerlik puanları ortaya koymuştur.

Sonuç: LuminaURO, özel ürolojik dokümanları ve çok dilli doküman işleme için yenilikçi havuzlama metodolojimizi kullanarak sağlık hizmetlerinde geleneksel LLM uygulamalarının sınırlamalarını başarıyla ele almaktadır. Çoklu değerlendirme metriklerinde elde edilen yüksek benzerlik puanları ve güçlü uzman doğrulaması, sistemin doğru ve ilgili ürolojik bilgileri sağlama konusundaki etkinliğini teyit etmektedir. Gelecekteki araştırmalar, bu yaklaşımı diğer tıbbi uzmanlık alanlarına genişletmeye odaklanacak ve nihai hedef olarak tüm tıbbi alanları kapsayan kapsamlı bir sağlık asistanı olan LuminaHealth'i geliştirmek olacaktır.

Anahtar Sözcükler: Doğal lisan işleme; karar destek sistemleri; tıbbi bilişim; üroloji; yapay zeka

Tuncay Soylu[1], Ibrahim Topcu[2], Muhammet Ihsan Karaman[3], Esra Melis Tuzcu[4], Abdullah Harun Kinik[5], Mustafa Sacit Guneren[6], Zeynep Salman[7], Perihan Demir[7], Beyzanur Kac[7]

[1] Department of Occupational Health and Safety Program, University of Health Sciences

[2] Department of History of Medicine and Ethics, University of Health Sciences

[3] Department of Medical Ethics and History of Medicine, İstanbul Health and Technology University

[4] Faculty of Engineering and Natural Sciences, Division of Biomedical Engineering, Işık University

[5] Department of Urology, Gaziosmanpaşa Training and Research Hospital

[6] Department of Urology, Bakırköy Dr. Sadi Konuk Training and Research Hospital

[7] Department of Medical History and Ethics, Hamidiye Health Sciences Institute, Health Sciences University

Received/*Geliş* : 07.03.2025
Accepted/*Kabul*: 28.04.2025

DOI: 10.21673/anadoluklin.1653335

Corresponding author/*Yazışma yazarı*

Tuncay Soylu

Sağlık Bilimleri Üniversitesi, Hamidiye Sağlık Hizmetleri Meslek Yüksekokulu, İş Sağlığı ve Güvenliği Programı, İstanbul, Türkiye
E-mail: tuncay.soylu@sbu.edu.tr

ORCID
Tuncay Soylu: 0000-0001-7595-8879
İbrahim Topçu: 0000-0001-7685-8597
M. İhsan Karaman: 0000-0001-5700-0835
Esra Melis Tuzcu: 0009-0003-2194-1563
Abdullah Harun Kınık: 0000-0001-8121-1480
Mustafa Sacit Güneren: 0009-0000-2411-1096
Zeynep Salman: 0000-0002-9864-3156
Perihan Demir: 0000-0002-0490-8353
Beyzanur Kaç: 0000-0002-7461-1942

INTRODUCTION

Artificial Intelligence (AI) comprises computational systems that simulate cognitive functions and execute tasks. These systems replicate human intelligence through learning, reasoning, perception, and language processing. Machine Learning (ML) and Deep Learning (DL) are sub-disciplines of AI, representing their most effective forms (1). ML enables systems to learn from data with statistical techniques and algorithms to develop models (1). DL, a subset of ML, uses multilayered artificial neural networks (ANN) inspired by biological neural architecture (1). This approach is successful in pattern recognition and feature extraction with large datasets (2). DL algorithms advance fields including image recognition speech processing and natural language processing (3-5). Natural language processing (NLP), a subfield of DL, focuses on understanding and generating human language (5,6). This discipline exists at the intersection of linguistics, computer science, and cognitive psychology. NLP applications include text classification translation and question-answering systems (7-10). The Transformer architecture marks a milestone in NLP evolution (6). This advancement enabled large language models (LLMs) improving language processing (5).

LLMs are complex ANN models trained on extensive text corpora with billions of parameters (11,12). The Generative Pretrained Transformer (GPT) series demonstrates capabilities in contextual understanding language generation and language tasks with human-like performance. These models perform high-level generalization and have applications across diverse fields (13-16). LLMs present limitations and ethical concerns (17–21). LLM projects require substantial computational resources can show biases and can generate incorrect information or “hallucinations” (22-24). Researchers must remain vigilant about these issues (25). LLMs are increasingly used in AI assistants chatbots ,translation and content generation (26-29). Examples include ChatGPT by OpenAI Google’s Gemini (31), Meta’s LLama (32), and Claude by Anthropic (30-33). While these are general-purpose LLMs, domain-specific models are developed from these frameworks.

AI is increasingly prevalent in healthcare (34). AI models efficiently execute repetitive tasks, minimize human errors, and operate without fatigue-related limitations (35,36). Growing patient populations and healthcare professional shortages accelerate AI adoption (34). AI applications are increasingly used in research for diseases like cancer where treatment protocols are still developing (35). AI applications in healthcare address unresolved problems while saving time and costs (34). The rise in digital healthcare data has motivated further research in this area (36).

Patients and families lack access to reliable medical resources for their healthcare questions (37). This need is critical after hospital discharge or when doctor communication is limited. There is a growing demand for systems that can address ongoing healthcare questions (38). These systems help patients and families make informed health decisions and manage treatments effectively. Patients and families often seek health solutions online (37). GPT-based AI models like ChatGPT, Claude, and Gemini are increasingly used for health questions (38). However, these platforms often provide medically unreliable responses without proper oversight (38), potentially confusing patients. Overcoming these challenges necessitates the development of domain-specific solutions (39). Retrieval-Augmented Generation (RAG) applications with LLMs have emerged as a solution for specialized healthcare responses (40).

This paper introduces LuminaURO, an AI-powered urology health assistant developed to provide specialized medical responses. LuminaURO generates personalized and medically validated responses by integrating LLM and RAG technologies.

This paper reveals the following major contributions:

- LuminaURO: RAG-based Urology Assistant: This study presents LuminaURO, an AI system for urology using the RAG framework with curated medical datasets. The system features a document pooling methodology that processes multilingual medical documents, enhancing response accuracy while maintaining medical context (Section 3, 3.4).
- Multilingual Retrieval Architecture: The implementation uses FAISS technology for rapid similarity searches across multilingual documents, generating responses within 8-15 seconds. The sys-

tem supports English and Turkish interfaces, with capacity for language expansion. It maintains high accuracy through integrated results from original and translated queries (Section 3.4).

- Comprehensive Validation Framework: The assistant's performance is validated through computational metrics (OESM, SpaCy, T5, BERTScore) and expert evaluation by urologists, achieving over 90% similarity scores across languages for clinical reliability (Section 4).
- Interactive Query Enhancement: Through LangChain PromptTemplate integration, the assistant generates two domain-specific follow-up questions per user query for user engagement in urological discussions (Section 3.5).

LuminaURO introduces an innovative AI solution that enhances patient care, clinical decision-making, and medical education in urology (18). The web-based interface provides broader public access and improved clinical testability. This study provides insights into AI applications in healthcare, offering a foundation for future research. The system improves patient outcomes, healthcare efficiency, and access to medical knowledge in urology. The system educates patients and their relatives about urological conditions and treatment options. This approach leads to informed decision-making and increased confidence in treatment. This assistant advances AI applications in medicine, enabling more personalized patient care.

BACKGROUND

AI and LLM systems have fostered significant advancements across a broad spectrum of the healthcare sector, including clinical decision support systems, patient education, medical research, diagnosis, and patient management. In this section, we will comprehensively examine these technological developments in healthcare by systematically categorizing studies from the literature, emphasizing specific metrics and key features.

Development of healthcare-specific LLMs

The complexity and specialized language of medical texts emphasize the critical need for healthcare-specific LLMs. Peng et al. developed GatorTronGPT, architected on GPT-3 and trained with 277 billion clinical and general English corpus (41). Physician-conducted Turing evaluations demonstrated clinical parity between GatorTronGPT generated and authentic medical documentation. Lee et al. developed BioBERT pretraining the BERT model on biomedical texts, which led to improvements of 0.62%, 2.80%, and 12.24% in tasks such as Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA) tasks, respectively (39).

In another study, an Otolaryngology-specific LLM, ChatENT, was developed by Long et al. which outperformed ChatGPT-4.0 on Canadian and U.S. ENT board examination questions by improving answer consistency and reducing hallucinations (42). Therefore, greater accuracy and proficiency were achieved. Luo et al. adopted a similar approach with ChatZOC, a Chinese ophthalmology-specific model, and compared it with 10 different LLMs, including GPT-4 (43). ChatZOC demonstrated performance comparable to GPT-4 regarding human ranking scores, alignment with scientific consensus, and reduced error rates. Zheng et al. Introduced MOPH, another Chinese ophthalmology specific LLM, which matched the clinical proficiency of ophthalmology residents, achieving an average exam score of 64.7% and 81.1% accuracy in clinical case diagnoses (44). Li et al. Introduced ChatDoctor, a LLaMA-based medical chatbot fine-tuned on patient-doctor dialogues, demonstrating superior BERTScore metrics compared to ChatGPT in delivering clinical information (45).

LLMs in clinical diagnosis and decision support

The effectiveness of LLMs in clinical diagnosis and decision support has been rigorously explored in numerous studies. Huang et al. Demonstrated GPT-4's superior performance in ophthalmological diagnostics, surpassing experts in glaucoma identification (506.2 vs 403.4, p<0.001) and achieving parity or superiority in retinal pathology assessment (235.3 vs 216.1, p=0.17) (46). This rigorous evaluation, employing a decadic Likert scale, elucidated GPT-4's significant potential in specialized medical diagnostics. Haider et al. Demonstrated that the Gemini model significantly outperformed ChatGPT-4 in breast disease classification, achieving an overall accuracy of 98%, compared to ChatGPT-4's 71% (47). Notably, the Gemini model excelled in the Fischer Grade, Kajava, and Regnault classifications. Similarly, Kim et al. Reported that LLMs reached a diagnostic accuracy of 96.1% in identifying obsessive-compulsive disorder (OCD), with ChatGPT-4 accurately diagnosing all cases (48).

Upadhyaya et al. Implemented a Gemini-based LLM for diagnosing moderate/severe amblyopia, mild cases, and nystagmus identification, achieving accuracy of 83%, 81%, and 85%, respectively (49). Dou et al. Evaluated ShennongGPT against GPT-4, BentsaoGPT, HuatuoGPT, ChatGLM, and NewBing in drug guidance, achieving superior efficacy with 97 points in expert evaluations (50). Furthermore, Chang and Chang Introduced SocraHealth, which integrates models such as GPT-4 and Bard to reduce diagnostic errors, leading to more precise clinical assessments (51).

Ge et al. Engineered LiVersa, a RAG-based liver specific LLM integrating AASLD guidelines, demonstrating superior performance in clinical decision support for hepatitis B management and hepatocellular carcinoma screening (52). Mukherjee et al. Demonstrated that the privacy-preserving Vicuna-13B model achieved moderate agreement with annotators in labeling radiology reports (median k = 0.52-0.64), matching human performance in 9 out of 11 findings (53). Kresevic et al. Mplemented a GPT-4 Turbo framework, achieving 99% accuracy in clinical guideline processing, optimizing decision-making protocols for chronic Hepatitis C management (54).

LLMs for patient education and communication

The effectiveness of LLMs in patient education and communication has also been investigated.

Kozaily et al. Observed that ChatGPT achieved 90% appropriateness and 93% consistency when addressing heart failure-related questions, surpassing Bard, which reached 77% appropriateness (55). Evaluations by two heart failure specialists revealed concordance rates of 83% for ChatGPT and 67% for Bard. Bernstein et al. Reported that ChatGPT's responses in ophthalmology showed no statistically significant difference from human responses concerning incorrect content, potential harm, and alignment with medical consensus (56). Yalamanchili et al. Demonstrated LLMs surpassed expert responses in radiation oncology care, achieving accuracy of 94%, 77%, and 91% for accuracy, completeness, and conciseness, respectively, though noting elevated readability levels exceeding middle school comprehension (57). Warren et al. Demonstrated that use of prompts with LLMs significantly enhanced the quality of responses for patient education on Peyronie's disease, raising them from medium to high quality, although only 42.5% of the cited sources were accurate (58).

Challenges and ethical considerations in healthcare LLMs

Healthcare LLMs face substantial constraints that impede their widespread clinical adoption. These limitations encompass contextual understanding deficiencies, interpretability challenges, and multifaceted ethical considerations, necessitating rigorous multidisciplinary collaboration. Benary et al.'s empirical analysis demonstrates their current insufficiency for routine clinical implementation, particularly in personalized oncology, where performance metrics remain suboptimal (17). Eckrich et al.'s investigation in urology applications revealed significantly inferior medical appropriateness compared to expert assessments, with concerning misinformation rates ranging from 2.8% to 18.9% (18). Lu et al.'s research highlighted these systems' propensity to generate unverified information termed "hallucinations" and produce biased outcomes due to inherent training data limitations (19). Nerella et al.'s work identified further critical challenges, including substantial computational overhead and complex ethical implications while Alonso et al.'s findings revealed notable performance degradation in non-English medical contexts (20, 21).

Future directions and conclusion

Research demonstrates significant potential in the development and application of LLMs in healthcare. Substantial progress has been made in creating domain-specific models, achieving high performance in clinical tasks, and advancing patient education. Nevertheless, ethical concerns, misinformation risks, and interpretability issues must be addressed, necessitating further research. Future studies should focus on enhancing the reliability of these models and facilitating their integration into clinical practice. Our study, introducing LuminaURO as a specialized AI-Powered Urology Assistant, represents a significant step toward addressing these challenges while providing dependable, domain-specific healthcare information.

LUMINAURO: AI-POWERED UROLOGY HEALTH ASSISTANT

In this paper, present a succinct overview of LuminaURO, an AI-Powered Urology Assistant engineered to address urological concerns of patients and their families. The called "LuminaURO" derives from the Latin "Lumina", denoting "illumination" or "enlightenment", aptly encapsulating its function within urology.

LuminaURO embodies a sophisticated synthesis of advanced NLP and ML algorithms, meticulously calibrated for urological applications. The assistant's knowledge repository is predicated upon a comprehensive urology corpus, assiduously curated by eminent specialists in the field. It employs a state-of-the-art vectorization methodology, leveraging cutting-edge embedding techniques to process and operationalize this specialized knowledge with unprecedented efficacy.

The assistant's architectural framework is constructed upon an OpenAI-developed LLM, augmented through the implementation of Retrieval-Augmented Generation (RAG) methodology. This sophisticated approach facilitates efficient information retrieval and the generation of contextually pertinent responses.

A publicly accessible web-based LuminaURO Application has been developed, featuring a user-centric interface that enables straightforward submission of patient queries. The application's sophisticated processing mechanism evaluates user inquiries comprehensively, generating personalized and medically accurate responses. Through this integrated approach, an avant-garde position in urology is established. Conventional question-answering boundaries are surpassed, whereby significant improvements in patient care quality and accessibility to specialized urological information are enabled. A substantial advancement in patient-centered healthcare technology is represented through the delivery of tailored, evidence-based responses, effectively bridging the gap between complex medical knowledge and patient comprehension in urology. The developmental process encompasses several critical phases, each contributing to the system's robust functionality and clinical relevance, ensuring efficacy and reliability in providing valuable urological insights to users.

Dataset data preprocessing and cleansing

This section details the methodology encompassed within step 1 in Figure 1. In the development of LuminaURO, a dataset comprising publicly accessible documents, predominantly books published in urology, was utilized. The dataset compilation process involved a comprehensive literature review conducted by three urology specialists, who identified seminal source documents fundamental to the dataset. These documents underwent a rigorous evaluation process to assess their congruence with the study's objectives. During this assessment, meticulous attention was paid to ensuring that the documents were mutually supportive and collectively encompassed the entire urology. Furthermore, redundant documents and topic headings were systematically eliminated from the corpus to enhance its coherence and efficiency.

The dataset architecture was methodically designed to encompass the complete spectrum of urological practice, from core principles to advanced subspecialties. This comprehensive framework incorporates critical domains including pediatric urology, female urology, and neuro-urology, alongside prevalent urological conditions such as urinary incontinence and urolithiasis. The inclusion of male and female sexual health, coupled with contemporary advances in uro-oncology, enhances the dataset's clinical relevance and modern applicability. This extensive corpus addresses the urological pathologies across diverse demographic cohorts, from pediatric to geriatric populations. The dataset has been rigorously curated to include emergent urological conditions and state-of-the-art clinical practices, ensuring coverage of the discipline. This systematic approach synthesizes cutting-edge therapeutic modalities and current clinical knowledge, while simultaneously illuminating future research trajectories. The dataset comprises twelve meticulously selected source documents, constituting eight in Turkish and

Table 1. The semantic similarity values

		Hypothesis sentences					Max similarity
		H_1	H_2	H_3	H_4	H_5	
Reference sentences	R_1	59.80	59.99	84.28	74.69	65.82	84.28
	R_2	86.51	92.21	74.11	80.89	91.03	92.21
	R_3	64.98	66.94	59.98	61.86	73.98	73.98
	R_4	87.41	85.14	79.90	84.76	91.24	91.24
				Overall similarity			***85.43***

R: Reference, **H:** Hypothesis

Table 2. Evaluation and comparison of similarity scores between algorithms and expert assessments

Qs	OESM-Sent.		OESM-Par.		Spacy		T5		BERT-Sent.		BERT-Par.		Urologist	
	EN	TR	EN	TR	EN	TR	EN	TR	EN	TR	EN	TR	EN	TR
Q1	0.6770	0.6679	0.7154	0.7241	0.9458	0.8355	0.9696	0.9477	0.9249	0.9236	0.9262	0.8826	0.9767	0.9500
Q2	0.7880	0.7317	0.7528	0.7087	0.9590	0.8073	0.9086	0.9466	0.9457	0.9217	0.9210	0.9310	0.9833	0.9933
Q3	0.6600	0.6915	0.7251	0.6560	0.9312	0.8264	0.9363	0.9556	0.9327	0.9452	0.9447	0.8810	0.9300	0.9033
Q4	0.6380	0.6128	0.6779	0.7000	0.9499	0.8295	0.9653	0.9413	0.9571	0.9235	0.9237	0.8895	0.9300	0.9333
Q5	0.6480	0.6695	0.6933	0.7421	0.9488	0.7709	0.9242	0.9078	0.8937	0.9604	0.9532	0.8904	0.9333	0.9667
Q6	0.6730	0.6787	0.6886	0.7262	0.9331	0.8353	0.9787	0.9444	0.9211	0.9346	0.9526	0.8545	0.9633	0.9000
Q7	0.7410	0.6874	0.7497	0.6747	0.9371	0.7628	0.9364	0.8868	0.9169	0.8897	0.8760	0.9076	0.9267	0.9733
Q8	0.7020	0.6267	0.6502	0.7207	0.9129	0.8151	0.9463	0.9023	0.9038	0.9106	0.9007	0.8774	0.9467	0.9633
Q9	0.6510	0.6763	0.5382	0.5577	0.9110	0.7738	0.9843	0.9284	0.9273	0.9579	0.9775	0.8898	0.9467	0.9633
Q10	0.6830	0.6625	0.7375	0.6188	0.9466	0.7870	0.9699	0.9562	0.8973	0.9576	0.9295	0.9159	1.0000	1.0000
Q11	0.7060	0.6573	0.7543	0.7708	0.9316	0.8028	0.9015	0.9014	0.8994	0.9628	0.9095	0.9192	0.9800	0.9700
Q12	0.6650	0.7173	0.8186	0.8110	0.9305	0.8168	0.8968	0.9169	0.9730	0.9586	0.8764	0.9168	0.9667	0.9000
Q13	0.7010	0.6588	0.7244	0.7054	0.8780	0.8204	0.9740	0.9318	0.9066	0.8780	0.9723	0.8952	0.9833	0.9833
Q14	0.6520	0.6579	0.7320	0.7148	0.9600	0.7858	0.8615	0.9375	0.9290	0.9402	0.9253	0.8721	0.9667	0.9333
Q15	0.6840	0.6704	0.6416	0.6976	0.9165	0.8527	0.9131	0.8850	0.9648	0.9251	0.9358	0.9183	0.9833	0.9867
Q16	0.6350	0.6160	0.7295	0.6856	0.9344	0.8022	0.8535	0.9172	0.9690	0.9125	0.8587	0.8668	0.9300	0.7667
Q17	0.6470	0.6562	0.7956	0.7770	0.9429	0.8076	0.8628	0.9186	0.9694	0.9107	0.9263	0.9294	0.9767	0.9633
Q18	0.6820	0.6523	0.8103	0.8528	0.9115	0.8024	0.8974	0.8813	0.9252	0.9185	0.9269	0.9121	0.9833	0.9833
Q19	0.6760	0.6804	0.7381	0.7332	0.9098	0.8163	0.9094	0.9727	0.9661	0.9604	0.8924	0.9034	0.9367	0.8700
Q20	0.6730	0.6825	0.6822	0.7063	0.9240	0.7959	0.9230	0.9809	0.8869	0.9679	0.8810	0.9153	0.8933	0.8667
Q21	0.6940	0.6961	0.8042	0.7319	0.9251	0.8395	0.9589	0.9121	0.9559	0.9197	0.9286	0.9127	0.9667	0.9567
Q22	0.6730	0.6893	0.8598	0.8073	0.9234	0.8063	0.9057	0.9665	0.9213	0.8868	0.9216	0.9794	0.9667	0.9933
Q23	0.6620	0.6301	0.6027	0.5760	0.9393	0.8390	0.9107	0.8860	0.9686	0.9696	0.9324	0.9537	0.9267	0.8667
Q24	0.6640	0.6451	0.7474	0.7870	0.9126	0.8223	0.8866	0.8856	0.9501	0.9620	0.9138	0.9116	0.9567	0.9033
Q25	0.6110	0.6439	0.6578	0.7237	0.9248	0.8243	0.9581	0.9189	0.9444	0.9134	0.9095	0.9115	0.9267	0.9133
Q26	0.6300	0.6578	0.7475	0.6834	0.9299	0.8347	0.8807	0.8883	0.9257	0.9624	0.9249	0.9669	0.9167	0.9333
Q27	0.6510	0.6516	0.6183	0.6754	0.9373	0.8143	0.8706	0.8835	0.9163	0.8790	0.8886	0.8574	0.7700	0.9500
Q28	0.7110	0.7091	0.7154	0.7872	0.9257	0.8211	0.9234	0.8872	0.9645	0.9082	0.9667	0.9164	0.9167	0.9700
Q29	0.7120	0.7033	0.6835	0.6813	0.9376	0.7788	0.9388	0.9682	0.9357	0.9165	0.8831	0.9338	0.8833	0.9833
Q30	0.6790	0.6790	0.8252	0.7616	0.9173	0.8291	0.9233	0.9025	0.9675	0.9686	0.8700	0.9470	0.9667	0.9833
Avg.	**67.56%**	**66.86%**	**72.06%**	**71.66%**	**92.96%**	**81.19%**	**92.23%**	**92.20%**	**93.53%**	**93.15%**	**91.83%**	**90.86%**	**94.44%**	**94.08%**

***Q:** Question, **Avg:** Average, **EN:** English, **TR:** Turkish, **Sent:** Sentences, **Par:** Paragraph

OESM (OpenAI Embedding Similarity, BERT (Bidirectional Encoder Representations from Transformers) Methodology)

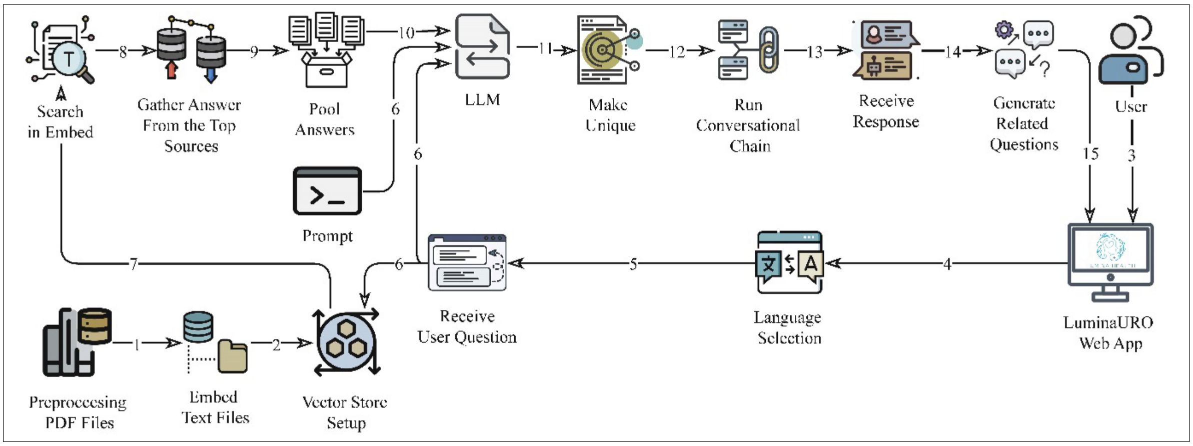


Figure 1. LLM: Large language model

four in English, thereby establishing a robust bilingual foundation.

Subsequent to the initial selection process, all documents in the dataset underwent a comprehensive preprocessing procedure. This phase involved the careful removal of extraneous elements, images, pagination, redundant sections, and header and footer information from the source materials. The refined documents were then converted to plain text format, with each document preserved under its respective nomenclature. The entire preprocessing workflow was executed through a custom-written Python script. Furthermore, to ensure the highest standards of data fidelity, each generated document was subjected to individual scrutiny and verification.

Embedding

The processes described herein are schematically represented between steps 2 and 3 in Figure 1. Following the preprocessing phase, the resultant dataset was transformed into vectorial representations utilizing OpenAI's Embeddings API (59). The 'text–embedding–3–small' model was selected for this process, predicated on a rigorous evaluation of performance metrics, efficiency parameters, and cost-effectiveness criteria. The vector generation process was executed via the LangChain library ensuring consistency and efficacy throughout the procedure (60).

The embedding process generated high-dimensional vectorial representations, subsequently preserved in a pickle-formatted database architecture to optimize accessibility and indexing efficiency. The database infrastructure was further enhanced through the implementation of FAISS (Facebook AI Similarity Search) library representing a cutting-edge solution for similarity search operations and dense vector clustering (61). This sophisticated indexing framework establishes an optimized architecture for executing similarity searches and vector-based queries, substantially augmenting the assistant's capacity for rapid and precise information retrieval. Through this methodological approach, both computational efficiency and retrieval accuracy are significantly enhanced.

The implementation of advanced embedding techniques and sophisticated indexing methodologies represents a crucial step in preparing the dataset for integration into the LuminaURO system. Complex urological query processing is enhanced through this approach, whereby improvements in response time and accuracy are attained. Through the combination of OpenAI's cutting-edge embedding model and FAISS's efficient indexing capabilities, significant advancements in AI-powered medical assistance tools are achieved, particularly in specialized urology. This technical framework establishes the foundation for future enhancements and system scalability. As urology evolves and novel information emerges, the infrastructure facilitates seamless integration of additional data, ensuring the assistant remains a current and comprehensive resource for urological information.

In the document segmentation process, the optimal chunk_size for the text_splitter was ascertained to

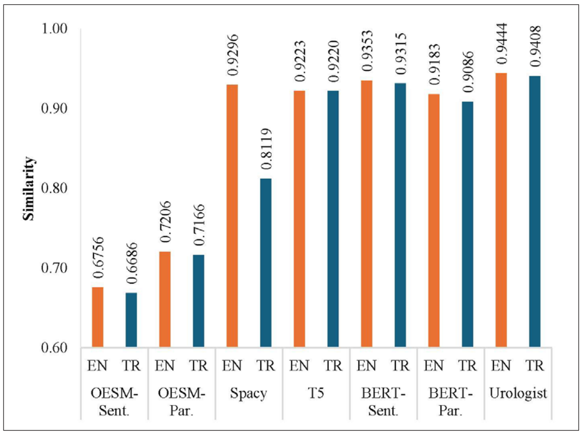


Figure 2. OESM (OpenAI embedding similarity methodology)
T5 (Text-to-text transfer transformer),
BERT (Bidirectional encoder representations from transformers)
EN: English, **TR**: Turkish, **Sent**: Sentence, **Par**: Paragraph

be 4000 characters through extensive experimentation with values ranging from 1000 to 8000 characters. Maintain semantic integrity while enabling effective vector representation. For the embedding process, batch_size = 10 was selected. These parameters were meticulously calibrated to optimize the equilibrium between preserving complex urological concepts and ensuring efficient processing, thereby enabling the provision of accurate and contextually relevant responses.

Selection and optimization of the LLM

The processes detailed in this section are represented in steps 10-13 in Figure 1. In the development of LuminaURO, OpenAI's gpt– 4o– mini model was selected as the LLM, predicated on considerations of cost-effectiveness, operational compatibility with the online web application, and performance balance. This model offers a cost-efficient solution compared to alternatives while providing performance levels that meet the project requirements. It exhibits robust performance capabilities both locally and on web platforms.

To enhance the model's efficacy and tailor it to the project's specific exigencies, comprehensive Prompt Engineering, tailored to both English and Turkish linguistic structures, has been integrated into the LLM model. The prompt engineering process is particularly focused on a target audience consisting of patients. Consequently, prompts were designed to reduce the complexity of medical terminology and ensure that generated responses maintain clarity and simplicity comprehensible to patients.

User experience optimization and dialogue continuity are ensured through the implementation of a memory mechanism. For this purpose, the ConversationBufferMemory model was selected. This model effectively manages the interaction history between the user and the assistant, thereby maintaining contextual consistency and facilitating more natural and fluid communication.

This configuration aspires to engender an effective AI Powered Urology Assistant by optimizing critical factors such as cost-effectiveness, language compatibility, patient focused communication, and consistent dialogue management. The synergistic integration of these elements is designed to enhance the assistant's ability to provide relevant, accessible, and contextually appropriate responses in urological consultations.

Retrieval-augmented generation (RAG)

This process is schematically represented in steps 8-10 of Figure 1. In the development of LuminaURO, a novel retrieval methodology has been devised and implemented, diverging from conventional approaches prevalent in extant literature. This approach adopts a multilingual strategy, aiming to optimize the information retrieval process.

The methodology's operation begins with a user query input to the assistant, proceeding with automatic language detection and translation. Queries entered in English are translated to Turkish, while those in Turkish are translated to English. (ChatOpenAI is employed for language translation, with parameters set to temperature = 0 and max_tokens = 1024). This bidirectional translation process expands the search scope, enabling the utilization of resources in both languages. Post-translation, the derived queries are simultaneously searched within previously embedded sources in the respective languages. This parallel search strategy is applied across a total of 12 sources (4 English and 8 Turkish). Search results are evaluated using a similarity algorithm, with the 4 highest-similarity sources from each language (8 sources in total) selected. Text fragments obtained from these selected sources are then aggregated in a temporary data pool.

In the subsequent phase, based on the user's query and predefined semantic compatibility criteria, the most appropriate text fragments are selected and integrated from the generated data pool. This process aims to engender a coherent and comprehensive response chain. The texts included in the data pool may contain similar content from different sources. To address this, a de-duplication process is applied to the selected text fragments. As a result of this operation, each text fragment is transformed into a unique value, thereby preventing redundancy in the response chain and optimizing information diversity. Finally. the constructed answer chain is formatted in accordance with the context and language of the user's original query, prepared for display on the user interface.

LuminaURO additionally generates a source list indicating which references were used to formulate the answer to the user's query, as well as a response_time metric quantifying the duration required to produce the result. Both of these are maintained as logs. The purpose of keeping a source list is to enable traceability of the generated answer. Responses produced by the assistant undergo rigorous scrutiny, potential errors are checked, and evaluations can be conducted by referring back to the source documents. Through this process, accuracy and reliability of the information provided are ensured.

This innovative retriever method aims to provide users with more comprehensive and accurate answers by effectively extracting information from multilingual sources. Furthermore, through transcending language barriers and facilitating information resource integration in different languages, significant knowledge base expansion is accomplished. This approach not only enhances the breadth and depth of information retrieval but also ensures a more inclusive and diverse range of responses, catering to a multilingual user base in urology.

Generation of related questions

The developed protocol is schematically represented between steps 14 and 15 of Figure 1. During the process of answering a user's query, LuminaURO generates two additional related questions associated with the primary inquiry. These related questions are produced using the PromptTemplate feature of the LangChain library. The process is executed through a specialized template meticulously crafted from the perspective of a urology expert, based on the original question and answer. The generated questions are formulated as concise and unambiguous interrogative sentences, strictly confined to urology. These related questions are prominently displayed at the bottom of the user interface immediately after the answer to the user's original question is presented on the screen. This feature is designed to stimulate users to acquire more comprehensive knowledge about the subject matter, thereby transforming the querying process into an interactive and exhaustive experience (Figure 3.e).

Web application and user interaction

The implemented framework is delineated in steps 3-6 of Figure 1. A salient feature distinguishing LuminaURO from analogous studies in the literature is its accessibility through a web-based application open to a broad user base. This approach represents an important step in transforming academic research into practical applications and providing societal benefits.

The assistant web application was architected utilizing the Streamlit platform and deployed on GitHub Cloud infrastructure, Streamlit, an open-source Python framework, efficiently transforming data science and machine learning implementations into web-based applications, executing Python processes in real-time (62). This platform was chosen for its rapid development capabilities and user experience optimization features. The application, hosted on GitHub Cloud, enables seamless access via effectively transitioning the assistant from concept to practical implementation (63). Access is provided through standard web browsers via cloud-based deployment, whereby local installation and specialized hardware requirements are eliminated. The integration of Streamlit and GitHub Cloud infrastructure enhances the project's scientific merit while enabling real world testing and continuous improvement based on user feedback. This architectural approach also aligns with and promotes open science principles and healthcare technology innovation.

User query processing and application response mechanism

The LuminaURO AI-Powered Urology Assistant employs a sophisticated and efficacious methodology in processing user queries and generating responses. The application is primarily optimized to address urological health-related queries and concerns of end-users in patient status. A user-centric approach to medical information dissemination is achieved through specialized tailoring to meet the specific exigencies of individuals seeking information about urological health issues.

The process commences with users accessing the application via any web browser, followed by the selection of their preferred language (currently English or Turkish) (Figure 3.a) (Figure 1: step 3-5). The health query, inputted in text format by the user, is processed through the Streamlit-based LuminaURO Web Application interface (Figure 3.b). The query, initially received by the application interface, is transmitted to the vector repository, where it undergoes vectorial representation (Figure 1: steps 5-7). Subsequently, a processing chain is established using the LangChain library.

The vectorized query is then searched within the existing database, following the methodology detailed in section 3.4 of the article (Figure 1: steps 7-8). The highest similarity scoring matching patterns from each document (k=8, comprising k=4E (English) + k=4T (Turkish)) are selected and transferred to a temporary data pool (Figure 1: steps 8-10). These pooled matching patterns are evaluated in accordance with the user query and predefined prompt content. The most appropriate information fragments are selected to form a coherent and comprehensive response chain (Figure 1: steps 10-13).

The final response, having undergone optimization, is transmitted to the user interface, allowing users to view detailed and personalized answers via the assistant Web Application (Figure 3.d) (Figure 1: steps 13-14). In alignment with the method elaborated in Section 3.5. two additional questions related to the user's original query are generated (Figure 1: steps 14-15). These related questions are presented to the user in two discrete boxes immediately subsequent to the display of the answer to the original question. Users can initiate a new search by engaging with the suggested questions, thereby perpetuating the interactive information acquisition process. This feature enriches the user's experience and provides an opportunity for more comprehensive research on the topic.

The assistant aims to provide expeditious, accurate, and comprehensible answers to patients' health questions by symbiotically integrating advanced NLP and ML techniques with a user-centric interface design. Enhanced methods in health information access and optimized user experience are established through this integrated approach. Through the seamless combination of sophisticated technological capabilities and an intuitive user interface, a significant advancement in AI-assisted healthcare information systems is represented, potentially transforming how urological health information is interacted with and understood by patients.

PERFORMANCE EVALUATION

This section delineates the methodologies employed to evaluate the model's performance, with an emphasis on the relevant evaluation metrics and comparative approaches.

Experimental setup

A comprehensive corpus comprising 30 urological queries and their corresponding responses were systematically developed by a panel of urological specialists, as presented in Table 3. Each subspecialty domain is represented by three questions that either typically emerge in clinical settings or are commonly posed by patients and their relatives to AI models. This methodologically rigorous selection process ensured comprehensive coverage across the spectrum of urological subspecialties. The query set was strategically structured to encompass three representative questions from each of ten pivotal domains: Bladder Cancer, Renal Tumors, Urinary Stone Disease, Nocturnal Enuresis (Bedwetting), Undescended Testis (Cryptorchidism), Erectile Dysfunction, Male Infertility, Urinary Incontinence in Women, Benign Prostatic Hyperplasia (BPH), and Prostate Cancer. This systematic stratification yielded a total of 30 clinically relevant inquiries, ensuring balanced representation across the entire field of urological practice.

Evaluation metrics

The evaluation of validation metrics presents one of the most intricate challenges in LLM implementations.

While conventional ML and DL applications, such as image or text classification, facilitate classical metric calculations, our study's complexity, extending beyond simple one-to-one translation tasks, necessitates sophisticated, semantic-based evaluation methodologies. Traditional LLM accuracy and similarity metrics, including BLEU (Bilingual Evaluation Understudy) (64), METEOR (Metric for Evaluation of Translation with Explicit Ordering) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) prove inadequate in capturing the nuanced semantic dimensions critical to our research objectives (65,66). In tasks where semantic preservation is paramount, paragraph level semantic evaluation demonstrates superior efficacy compared to word or sentence-level assessments. Consequently, the evaluation framework transcends surface-level metrics by implementing sophisticated semantic measures: OESM (OpenAI Embedding Similarity Methodology) (67), T5 Semantic (Text-To-Text Transfer Transformer) (15), BERTScore Semantic (Bidirectional Encoder Representations from Transformers) (5) and SpaCy Semantic (68). This methodological approach facilitates more comprehensive and nuanced evaluations of semantic coherence and accuracy.

These semantic evaluation methodologies represent distinct approaches to measuring textual similarity through advanced NLP techniques. OESM leverages OpenAI's embedding models to transform text into high-dimensional vector representations, enabling semantic comparison through cosine similarity calculations. SpaCy Semantic utilizes industrial-strength NLP capabilities with pre-trained word vectors, incorporating linguistic features and statistical models for semantic similarity assessment. Each methodology systematically advances semantic comprehension through the implementation of sophisticated embedding techniques, neural architectures, and contextual evaluation frameworks, exemplifying the progressive evolution of NLP approaches in capturing subtle linguistic nuances. T5 Semantic employs a unified transformer architecture that converts all NLP tasks into text-to-text formats, utilizing pre-trained models for comprehensive semantic assessment. BERTScore Semantic implements BERT's contextual embeddings and bidirectional transformer architecture, computing token-level matching scores through cross-attention mechanisms and multiple transformer layers. Furthermore, a panel of three independent urological specialists, distinct from those who formulated the initial corpus of questions and responses, conducted comprehensive evaluations of LuminaURO's outputs, with their aggregated assessment metrics documented in Table 2 under the "Urologist" designation. Urological specialist validations serve as a paramount benchmark, providing an authoritative reference standard against which the semantic evaluation methodologies can be assessed for precision and clinical relevance.

Quantitative assessment methodology

The evaluation methodology employs advanced semantic similarity algorithms (OESM, T5, SpaCy, and BERT) that utilize the Cosine Similarity model as their core computational framework. The evaluation methodology was implemented utilizing the following systematic framework:

- ***Step 1: Cosine similarity calculation***

For each pair of sentences between the reference text and the hypothesis text , compute the cosine similarity score. Cosine similarity between a reference sentence and a hypothesis sentence is given by:

$$CosineSimilarity(r_i . h_j) = \frac{\vec{r_i} \cdot \vec{h_j}}{\left\| \vec{r_i} \right\| \cdot \left\| \vec{h_j} \right\|}$$

where:

- r_i and h_j are the vector representations of the reference and hypothesis sentences, respectively.
- denotes the dot product.
- $\left\| \vec{r_i} \right\|$ and $\left\| \vec{h_j} \right\|$ are the magnitudes (or norms) of the sentence vectors.
- ***Step 2: Maximum similarity score selection***

For each reference sentence , calculate the cosine similarity with all hypothesis sentences . Identify the maximum cosine similarity score for each reference sentence r_i :

$$MaxSimilarity(r_i) = \max_{j \in \{1.2.....m\}} \left(CosineSimilarity(r_i . h_j) \right)$$

- ***Step 3: Overall Similarity***

The Overall Similarity (OS) was calculated using the following formula:

$$OS = \frac{\sum_{i=1}^{n} MaxSimilarity(}{n}$$

- ***Step 4: Example***

Table 1 presents the semantic similarity values. that quantify the alignment between each reference and hypothesis sentence. The quantitative similarity metrics were systematically derived through pairwise comparative analysis between reference sentences and their corresponding hypothesis sentences, with the resultant evaluation scores methodically documented in sequential progression within Table 1.

The maximum similarity score for each reference sentence is recorded documented in the Max Similarity column, representing the optimal semantic alignment. The Overall Similarity metric is then derived by computing the arithmetic mean of these maximum similarity values, providing a comprehensive measure of semantic concordance. In this example, the maximum similarity values for R_1. R_2. R_3. and R_4 are 84.28, 92.21, 73.98, and 91.24, respectively. Consequently, the computed Overall Similarity value of 85.43 demonstrates the degree of semantic concordance between the reference and hypothesis texts.

LuminaURO response similarity metrics and evaluation

Table 2 presents evaluation scores for responses to the 30 questions listed in Table 3, incorporating assessments through OESM (Sentence and Paragraph), SpaCy, T5, BERTScore (Sentence and Paragraph) algorithms, and expert ratings provided by urological specialists. These comprehensive evaluations were conducted independently for both English and Turkish responses. The evaluation methodology was executed utilizing the following systematic algorithmic framework: Questions and corresponding answers formulated by urological specialists (n=3) served as reference responses. These questions were presented to LuminaURO, and their generated responses were recorded as hypothesis answers. Each hypothesis response against its corresponding reference response was evaluated using the specified algorithms to generate similarity scores. This systematic assessment enabled quantitative measurement of semantic alignment between expert-formulated and AI-generated responses. The resultant similarity scores were documented in Table 2. The evaluation framework was implemented independently for responses in both English and Turkish. This maintains methodological consistency while accommodating linguistic specificity.

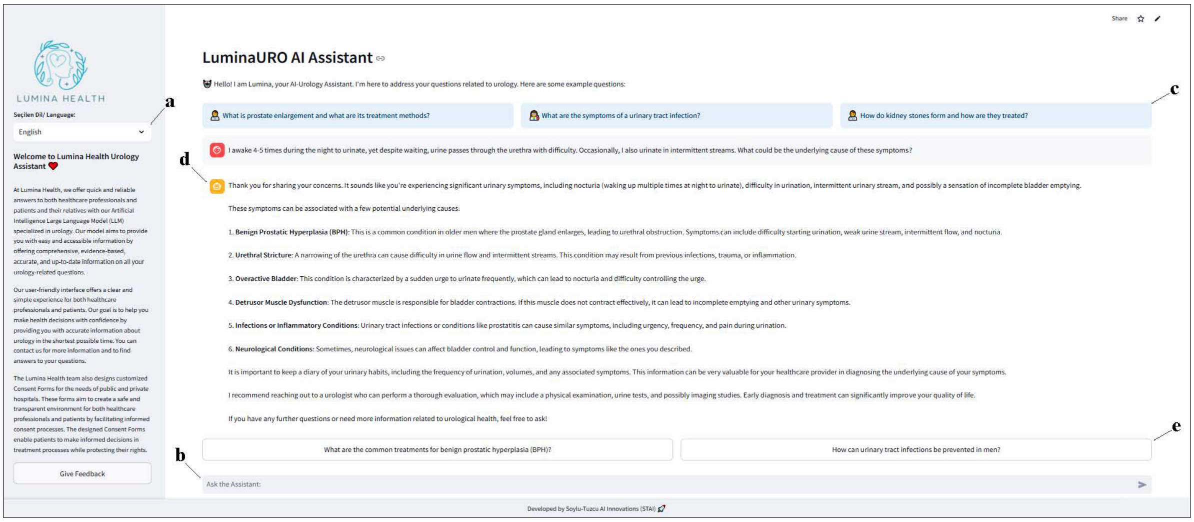
Three independent urological specialists evaluated the assistant's responses (the hypothesis) based on predetermined criteria: Scientific Accuracy, Patient Comprehension Appropriateness, Comprehensiveness and Clinical Relevance, Language and Communication, Ethical Considerations, Safety, and Patient Guidance. The evaluating specialists (n = 3), distinct from those who prepared the reference answers, conducted their assessments independently, with final scores calculated by averaging individual ratings. Furthermore, the evaluating experts were blinded to the algorithm-generated responses to ensure an unbiased assessment process. This methodological approach facilitated an unbiased evaluation of the assistant's performance. This systematic evaluation protocol was implemented identically for both English and Turkish responses, maintaining consistency across languages.

Word-based methods (BLEU, METEOR, and ROUGE) yield markedly low similarity scores (0.15 to 0.40), demonstrating their inherent limitations in capturing semantic nuances beyond lexical matching. Operating exclusively at the lexical level without leveraging embeddings, these methods showed limited capability in comprehending contextual and semantic relationships. Consequently. these results were excluded from the evaluation Table 2, Figure 2a comprehensive evaluation of the similarity metrics, revealing a hierarchical progression in assessment effectiveness across different methodological approaches. The baseline OESM metrics, both at sentence and paragraph levels, demonstrate relatively modest performance (0.66-0.72), while more sophisticated algorithms exhibit markedly superior results. Notably, advanced semantic measures, including Spacy, T5, and BERT demonstrate robust performance, with similarity scores consistently exceeding 0.90 in both English and Turkish implementations. The BERT-based metrics, in particular, achieve an exceptional consistency between sentence-level (0.93) and paragraph-level (0.90-0.91) evaluations. Most significantly, these computational evaluations demonstrate remarkable alignment with expert urologist assessments (0.94), validating the efficacy of advanced semantic metrics in addressing medical communication complexities and substantiating their reliability in medical NLP applications.

A.1. Systematic review of comprehensive urological questions

Table 3. Representative urological queries across major subspecialties

Bladder cancer	
Q1	How can I ascertain whether I have bladder cancer?
Q2	Does my smoking habit elevate the risk of developing bladder cancer?
Q3	What therapeutic approaches are employed in the treatment of bladder cancer?
Renal tumors	
Q4	How can renal tumors be detected?
Q5	How can benign and malignant renal tumors be distinguished?
Q6	Is surgical intervention always necessary for the treatment of renal tumors?
Urinary stone disease	
Q7	What are the etiological factors contributing to the formation of kidney stones?
Q8	What interventions can I undertake to facilitate the passage of a kidney stone?
Q9	How can I prevent the recurrence of nephrolithiasis?
Nocturnal enuresis (bedwetting)	
Q10	Our daughter is 4 years old and wets the bed at night; until what age is nocturnal enuresis considered normal in children?
Q11	What are the underlying causes of bedwetting?
Q12	How is nocturnal enuresis treated, and do the medications used in its treatment lead to infertility?
Undescended testis (cryptorchidism)	
Q13	At what age should an undescended testis be medically addressed?
Q14	What potential future health complications can arise if cryptorchidism remains untreated?
Q15	Is infertility an inevitable outcome if my child does not undergo surgical intervention for cryptorchidism?
Erectile dysfunction	
Q16	I initiated sexual intercourse, but my erection diminishes before penetration; this did not happen previously. What could be the underlying cause of this?
Q17	We used to engage in sexual intercourse three times per week, but now it occurs less than once a month. What interventions can address my erectile dysfunction?
Q18	As soon as penetration occurs, I experience involuntary ejaculation; I am unable to comprehend this. What could be the reason for such premature ejaculation?
Male infertility	
Q19	I have been married for five years and have not been able to conceive a child; if the problem lies with me, what could be the potential causes?
Q20	I have two children aged eight and fifteen; we wish to have another child, but it has not happened for the past two years. What might be the reason?
Q21	My examining physician informed me that there is a dilation in my spermatic veins. What does this mean, and how is it treated?
Urinary incontinence in women	
Q22	I experience urine leakage when I laugh, cough, or sneeze; what could be the underlying causes and potential treatment methods?
Q23	During personal hygiene, I detect a mass with my hand; sometimes I have difficulty urinating, and other times it leaks involuntarily. What might be my condition?
Q24	I go to the restroom 20 times a day and practically live there; even if I drink half a glass of water, I immediately need to rush to the toilet, making it impossible for me to visit friends. What are the possible causes of this?
Benign prostatic hyperplasia (BPH)	
Q25	I awake 4-5 times during the night to urinate, yet despite waiting, urine passes through the urethra with difficulty. Occasionally, I also urinate in intermittent streams. What could be the underlying cause of these symptoms?
Q26	I experience difficulty during urination; my urine dribbles and intermittently stops, and I feel as though I cannot fully empty my bladder. The doctor mentioned that I have prostate enlargement. What are the treatment options for this condition?
Q27	I have been diagnosed with prostate enlargement and a minimally invasive prostate surgery was recommended. What are the potential risks associated with this surgical procedure?
Prostate cancer	
Q28	Although I have no urinary symptoms, my blood test revealed elevated PSA levels. What could be the potential reasons for this?
Q29	An MRI of my prostate was conducted due to the high PSA levels, and biopsies from suspicious areas have been recommended. What are the risks associated with this procedure?
Q30	A prostate biopsy was performed, and the results indicated cancer. What are the available treatment options?



Seçilen Dil/ Language:

English

Welcome to Lumina Health Urology Assistant ❤

At Lumina Health, we offer quick and reliable answers to both healthcare professionals and patients and their relatives with our Artificial Intelligence Large Language Model (LLM) specialized in urology. Our model aims to provide you with easy and accessible information by offering comprehensive, evidence-based, accurate, and up-to-date information on all your urology-related questions.

Our user-friendly interface offers a clear and simple experience for both healthcare professionals and patients. Our goal is to help you make health decisions with confidence by providing you with accurate information about urology in the shortest possible time. You can contact us for more information and to find answers to your questions.

The Lumina Health team also designs customized Consent Forms for the needs of public and private hospitals. These forms aim to create a safe and transparent environment for both healthcare professionals and patients by facilitating informed consent processes. The designed Consent Forms enable patients to make informed decisions in treatment processes while protecting their rights.

LuminaURO AI Assistant

Hello! I am Lumina, your AI-Urology Assistant. I'm here to address your questions related to urology. Here are some example questions:

What is prostate enlargement and what are its treatment methods?

What are the symptoms of a urinary tract infection?

How do kidney stones form and how are they treated?

I awake 4-5 times during the night to urinate, yet despite waiting, urine passes through the urethra with difficulty. Occasionally, I also urinate in intermittent streams. What could be the underlying cause of these symptoms?

Thank you for sharing your concerns. It sounds like you're experiencing significant urinary symptoms, including nocturia (waking up multiple times at night to urinate), difficulty in urination, intermittent urinary stream, and possibly a sensation of incomplete bladder emptying.

These symptoms can be associated with a few potential underlying causes:

1. **Benign Prostatic Hyperplasia (BPH)**: This is a common condition in older men where the prostate gland enlarges, leading to urethral obstruction. Symptoms can include difficulty starting urination, weak urine stream, intermittent flow, and nocturia.

2. **Urethral Stricture**: A narrowing of the urethra can cause difficulty in urine flow and intermittent streams. This condition may result from previous infections, trauma, or inflammation.

3. **Overactive Bladder**: This condition is characterized by a sudden urge to urinate frequently, which can lead to nocturia and difficulty controlling the urge.

4. **Detrusor Muscle Dysfunction**: The detrusor muscle is responsible for bladder contractions. If this muscle does not contract effectively, it can lead to incomplete emptying and other urinary symptoms.

5. **Infections or Inflammatory Conditions**: Urinary tract infections or conditions like prostatitis can cause similar symptoms, including urgency, frequency, and pain during urination.

6. **Neurological Conditions**: Sometimes, neurological issues can affect bladder control and function, leading to symptoms like the ones you described.

It is important to keep a diary of your urinary habits, including the frequency of urination, volumes, and any associated symptoms. This information can be very valuable for your healthcare provider in diagnosing the underlying cause of your symptoms.

I recommend reaching out to a urologist who can perform a thorough evaluation, which may include a physical examination, urine tests, and possibly imaging studies. Early diagnosis and treatment can significantly improve your quality of life.

If you have any further questions or need more information related to urological health, feel free to ask!

What are the common treatments for benign prostatic hyperplasia (BPH)?

How can urinary tract infections be prevented in men?

Developed by Soylu-Tuzcu AI Innovations (STAI)

A.2. Overview of LuminaURO Interface

Figure 3. Overview of LuminaURO interface: a. Language selection and introductory information about LuminaURO, b. Input area where users can ask specific urology-related questions, c. Suggested questions for quick access to common urological topics, d. Detailed response area providing explanations and potential causes of symptoms, e. Related questions, offering contextually relevant follow-up queries.

Evaluation of quantitative metrics (Table 2 and Figure 2) demonstrates LuminaURO's superior performance in English compared to Turkish response generation, consistently evident across algorithmic and expert evaluations. This linguistic performance disparity is empirically validated through both computational methodologies and specialist assessments. Evaluation of Table 2 reveals that urologist evaluations demonstrate relatively lower assessment scores for question 16 in Turkish responses and question 27 in English responses compared to other answers. This variation is attributed to query comprehension challenges and minor interpretative discrepancies by the assistant. Given that our assistant is designed for use by patients and their relatives across diverse educational backgrounds, such nuanced variations are anticipated. To address this limitation and enhance response accuracy, we have implemented a Related Questions feature, detailed in Section 3.5, which aims to facilitate more precise user interactions and improved response generation.

CONCLUSION

In this study, we have developed a novel RAG-based AI model that addresses the limitations of conventional LLM implementations by utilizing a specialized dataset of urological documents. The introduction of our innovative pooling methodology enables simultaneous multi-language document processing, significantly enhancing the system's capability to provide accurate and contextually relevant responses. LuminaURO validation demonstrates the effectiveness of our approach, with high similarity scores across multiple evaluation metrics. Through expert evaluations, particularly strong performance is validated, whereby the accuracy and relevance of responses in both urologists confirm English and Turkish languages. These results indicate that the assistant appears to bridges the gap between advanced language models and reliable medical information delivery in urology.

Future research could aim to expand the assistant's applications across other medical specialties, building upon our successful implementation in urology. The development strategy may involve systematic validation and implementation in individual medical fields. Following this approach, we envision LuminaHealth, a comprehensive healthcare assistant encompassing all medical specialties.

Conflict-of-interest and financial disclosure

The authors declare that they have no conflict of interest to disclose. The authors also declare that they did not receive any financial support for the study.

Acknowledgements

Not applicable. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

REFERENCES

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
2. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85-117.
3. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–778.
4. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2013. p. 6645–6649.
5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT); 2019. p. 4171–4186.
6. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS); 2017. p. 5998–6008.
7. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems (NeurIPS); 2015. p. 649–657.
8. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT); 2016. p. 1480–1489.
9. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NeurIPS); 2014. p. 3104–3112.
10. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2016. p. 2383–2392.
11. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS); 2020. p. 1877–1901.
12. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res. 2023;24(240):1–113.
13. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
14. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. 2021.
15. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
16. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. 2022.
17. Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Netw Open. 2023;6(11):e2343689.
18. Eckrich J, Ellinger J, Cox A, et al. Urology consultants versus large language models: potentials and hazards for medical advice in urology. BJUI Compass. 2024;5(5):438–44.
19. Lu Z, Peng Y, Cohen T, Ghassemi M, Weng C, Tian S. Large language models in biomedicine and health: current research landscape and future directions. J Am Med Inform Assoc. 2024;31(9):1801-11.
20. Nerella S, Bandyopadhyay S, Zhang J, et al. Transformers and large language models in healthcare: a review. Artif Intell Med. 2024;154:102900.
21. Alonso I, Oronoz M, Agerri R. MedExpQA: multilingual benchmarking of large language models for medical question answering. Artif Intell Med. 2024;155:102938.
22. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361. 2020.
23. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT); 2021. p. 610–623.
24. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38.

25. Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359. 2021.
26. Thoppilan R, De Freitas D, Hall J, et al. LaMDA: language models for dialog applications. arXiv preprint arXiv:2201.08239. 2022.
27. Zhou H, Liu F, Gu B, et al. A survey of large language models in medicine: progress, application, and challenge. arXiv preprint arXiv:2311.05112. 2023.
28. Fan A, Bhosale S, Schwenk H, et al. Beyond English-centric multilingual machine translation. J Mach Learn Res. 2021;22:1–48.
29. Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. In: Proceedings of the 11th International Conference on Learning Representations (ICLR); 2023.
30. ChatGPT [Internet]. Available from: https://chatgpt.com
31. Gemini [Internet]. Available from: https://gemini.google.com/app
32. Llama 3.2 [Internet]. Available from: https://www.llama.com
33. Claude [Internet]. Available from: https://claude.ai/login?returnTo=%2F%3F
34. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719-31.
35. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94.
36. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
37. Fox S, Duggan M. Health online 2013. Pew Research Center [Internet]. 2013 [cited 2024 Nov 5]. Available from: https://www.ordinedeimedici.com/documenti/Docs7-cybercondria-PIP-HealthOnline.pdf
38. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198.
39. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
40. RAG [Internet]. Available from: https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview
41. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023;6(1):210.
42. Long C, Subburam D, Lowe K, et al. ChatENT: augmented large language model for expert knowledge retrieval in otolaryngology–head and neck surgery. Otolaryngol Head Neck Surg. 2024;171(4):1042–51.
43. Luo MJ, Pang J, Bi S, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. 2024;142(9):798–805.
44. Zheng C, Ye H, Guo J, et al. Development and evaluation of a large language model of ophthalmology in Chinese. Br J Ophthalmol. 2024;108(10):1390–7.
45. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023;15(6):e40895.
46. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol. 2024;142(4):371–5.
47. Haider SA, Pressman SM, Borna S, et al. Evaluating large language model (LLM) performance on established breast classification systems. Diagnostics (Basel). 2024;14(14):1491.
48. Kim J, Leonte KG, Chen ML, et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. NPJ Digit Med. 2024;7(1):193.
49. Upadhyaya D, Shaikh A, Cakir G, et al. A 360 degree view for large language models: early detection of amblyopia in children using multi-view eye movement recordings. medRxiv [Preprint]. 2024.
50. Dou Y, Huang Y, Zhao X, et al. ShennongMGS: an LLM-based Chinese medication guidance system. ACM Trans Manag Inf Syst. 2024;16(2):1-14.
51. Chang JJ, Chang EY. SocraHealth: enhancing medical diagnosis and correcting historical records. In: Proceedings of the 10th International Conference on Computational Science and Computational Intelligence (CSCI); 2023.
52. Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology. 2024;80(5):1158-68.
53. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. Radiology. 2023;309:e231147.
54. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit

Med. 2024;7(1):102.

55. Kozaily E, Geagea M, Akdogan ER, et al. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. Int J Cardiol. 2024;408:132115.
56. Bernstein IA, Zhang Y, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open. 2023;6(8):e2330320.
57. Yalamanchili A, Sengupta B, Song J, et al. Quality of large language model responses to radiation oncology patient care questions. JAMA Netw Open. 2024;7(4):e244630.
58. Warren CJ, Edmonds VS, Payne NG, et al. Prompt matters: evaluation of large language model chatbot responses related to Peyronie's disease. Sex Med. 2024;12(4):qfae055.
59. OpenAI [Internet]. 2024 [cited 2024 Dec 16]. Available from: https://platform.openai.com/docs/overview
60. LangChain [Internet]. 2024 [cited 2025 Feb 6]. Available from: https://www.langchain.com/
61. FAISS [Internet]. 2024. Available from: https://ai.meta.com/tools/faiss
62. Streamlit [Internet]. 2024. Available from: https://streamlit.io
63. Lumina [Internet]. Available from: https://lumina-healthstai.streamlit.app
64. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL); 2002. p. 311–318.
65. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; 2005. p. 65–72.
66. Lin CY, Och FJ. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04); 2004. p. 605–612.
67. Neelakantan A, Xu T, Puri R, et al. Text and code embeddings by contrastive pre-training. arXiv preprint arXiv:2201. 2022.
68. spaCy [Internet]. [cited 2025 Feb 22]. Available from: https://spacy.io/