

The Islamic University–Gaza
Faculty of Engineering
Master of Computer Engineering



الجامعة الإسلامية - غزة
كلية الهندسة
ماجستير في هندسة الحاسوب

Developing An Intelligent Chatbot for Dermatology Consultations Using Natural Language Processing

Chatbot4Derma

By

Abd Alaziz M. A. Alswaisi
Abd Alrhman G. A. Qaddora
Ahmed I H Shamia
Mohammed W. S. Nasman
Amna Awwad

To

Dr. Aiman Ahmed Abusamra

July/2025

1- Abstract:

Access to accurate and timely dermatological consultations remains limited, particularly in underserved and multilingual regions. This study addresses this gap by developing Chatbot4Derma, an intelligent, bilingual (Arabic-English) dermatology chatbot designed to deliver clinically reliable responses using a Retrieval-Augmented Generation (RAG) framework. The system integrates real-time medical knowledge retrieval with context-aware language generation to provide accurate, up-to-date, and patient-friendly responses.

The chatbot combines real-time semantic retrieval from a curated dermatology knowledge base with context-aware response generation using the DeepSeek-Chat model. Unlike traditional fine-tuning approaches, this system avoids retraining large language models by leveraging ChromaDB for vector-based retrieval and DeepSeek Embedding for multilingual semantic encoding, orchestrated through a Node.js backend and deployed via Telegram.

Evaluation was conducted using a dataset of 9,570 dermatology-related documents. The system achieved strong performance across standard metrics: accuracy = 0.67, precision = 1.00, recall = 0.67, and F1 score = 0.80. These results confirm the system's high relevance and reliability in answering dermatological queries in both Arabic and English.

The proposed solution contributes a cost-effective, scalable, and multilingual digital health tool, enhancing AI-assisted clinical support for dermatology in low-resource environments.

Keywords: Natural Language Processing, Retrieval-Augmented Generation, Dermatology Chatbot, ChromaDB, DeepSeek, Medical NLP, Vector Embeddings, Node.js, Healthcare AI.

2- Introduction:

Skin diseases constitute one of the most prevalent health issues globally, affecting millions of people each year. Despite their commonality, access to timely and accurate dermatological consultation remains a significant challenge, particularly in underserved regions and non-urban areas where dermatologists are scarce. Factors such as cost, geographic distance, stigma, and long wait times often deter individuals from seeking early medical advice. In parallel, the internet has become a primary source of health-related information, yet patients frequently encounter inaccurate or misleading content.

With the advancement of Artificial Intelligence (AI), particularly Natural Language Processing (NLP), there has been growing interest in developing intelligent systems that can support healthcare delivery. However, many AI-based medical chatbots suffer from two critical limitations: (1) limited domain-specific adaptability, where generic models lack depth and context awareness, and (2) high cost and complexity associated with fine-tuning large language models (LLMs) on specialized medical data.

This research proposes the development of "Chatbot4Derma", an intelligent dermatology consultation chatbot built using Retrieval-Augmented Generation (RAG). This approach leverages real-time retrieval of relevant dermatology knowledge from an external vector database

(ChromaDB) and integrates it with a generative language model (DeepSeek-Chat) to produce contextually grounded, medically accurate responses. By avoiding the need for constant retraining, the system remains cost-effective and adaptable to evolving clinical guidelines.

A notable contribution of this work is its support for bilingual interaction (Arabic and English), thereby addressing the accessibility gap in digital health tools that predominantly serve English-speaking users. The proposed system combines multiple state-of-the-art tools including DeepSeek Embedding for semantic search, ChromaDB for efficient knowledge retrieval, and a Node.js backend integrated with Telegram to provide a seamless and scalable user experience.

This study aims to fill a critical gap in AI-assisted dermatological support by delivering a scalable, multilingual, and clinically reliable chatbot tailored to patient needs in real-world settings.

3- Objectives:

The primary objectives of this research project are:

- To design and implement an intelligent dermatology chatbot system using Retrieval-Augmented Generation (RAG) that provides accurate, real-time responses grounded in the latest dermatological literature.
- To develop a bilingual (Arabic-English) interface that ensures accessibility for both Arabic-speaking and English-speaking users, addressing language-related exclusion in digital health services.
- To integrate open-source and cost-effective tools (ChromaDB, DeepSeek, Node.js) that allow for efficient retrieval, high-quality embedding, and scalable deployment without the need for retraining large language models.
- To evaluate the system's performance in terms of response accuracy, user satisfaction, and adaptability to newly updated medical content.
- To contribute to digital health solutions in underserved regions by proposing an AI-powered tool that reduces dependence on physical infrastructure and traditional consultation timelines.

4- Related Work:

In recent years, there has been a significant surge in the adoption of Artificial Intelligence (AI), particularly Natural Language Processing (NLP) techniques, to develop intelligent advisory systems in the medical domain. Early initiatives such as Babylon Health and Ada Health provided general health consultations using traditional language understanding techniques. However, these systems lacked domain specialization, particularly in dermatology, and were limited to English-only interfaces, which hindered their applicability in multilingual environments.

On the other hand, large language models (LLMs) like GPT-3 and GPT-4 have demonstrated strong capabilities in generating human-like responses. Nonetheless, these models often fall short in delivering accurate domain-specific outputs particularly in medicine when not trained or fine-tuned on specialized datasets (Pal et al., 2022). This results in generic responses that may not be suitable in advisory contexts requiring clinical accuracy.

To address these limitations, recent research has gravitated toward hybrid architectures such as Retrieval-Augmented Generation (RAG). This approach combines the generative power of LLMs with real-time knowledge retrieval from external sources. RAG has proven effective in improving the quality of responses in tasks that demand precision and factual reliability, including medical question answering and legal document analysis (Lewis et al., 2020; Karpukhin et al., 2020; Lee et al., 2021).

Studies by Bora & Cuayáhuitl (2024) and Ke et al. (2024) systematically analyzed RAG-based architectures in medical chatbot applications. Their findings highlight that integrating retrieval mechanisms with generation significantly reduces the need for continuous fine-tuning of LLMs with every update in medical knowledge, thereby improving maintainability and cost-efficiency.

Similarly, Wu et al. (2024) introduced a graph-based retrieval method Medical Graph RAG which emphasizes the importance of structural medical knowledge integration to enhance the reliability of LLM-generated responses in clinical settings.

Despite these advances, bilingual or multilingual medical advisory systems remain scarce, especially those supporting Arabic. Moreover, the dermatology domain uniquely demands precise integration of textual and visual data, further complicating the development of effective consultation systems.

The proposed **Chatbot4Derma** project aims to address these gaps by:

- Implementing a practical dermatology-focused advisory system based on the RAG paradigm.
- Supporting both Arabic and English to broaden accessibility across diverse populations.
- Leveraging open-source tools such as ChromaDB for vector storage and retrieval, and DeepSeek models for context-aware generation ensuring high accuracy and adaptability without the high costs associated with continuous model retraining.

This work thus contributes to bridging a critical gap at the intersection of AI and digital healthcare, especially for under-resourced regions and underserved language communities.

5- Methodology:

This research adopts a design science methodology tailored for building and evaluating an intelligent dermatology chatbot powered by Retrieval-Augmented Generation (RAG). The methodology is structured into five sequential phases: (1) System Architecture Design, (2) Knowledge Base Construction, (3) Embedding and Retrieval Optimization, (4) Generation and Multilingual Response Handling, and (5) Evaluation and Validation.

1- System Architecture Design:

The overall system architecture integrates a RAG pipeline combining a vector-based retrieval mechanism with a generative large language model. The RAG framework enables real-time, context-aware responses grounded in curated dermatological knowledge without relying on repeated fine-tuning Lewis et al., 2020.

The architecture consists of:

- Input Interface: A Telegram bot for user interaction.
- Retrieval Layer: Powered by ChromaDB for document indexing and vector search.
- Embedding Layer: Utilizes DeepSeek Embedding to transform user queries into dense vector space.
- Response Generator: DeepSeek-Chat processes the retrieved documents and generates patient-friendly answers.
- Backend Orchestration: Managed through Node.js to coordinate asynchronous data flow between components.

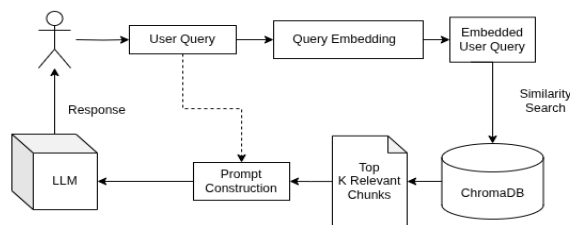


Figure 1 Retrieval-Augmented Generation (RAG) system, showing how a user query is processed to generate a response using a Large Language Model (LLM) and a vector database like ChromaDB.

This modular, service-oriented architecture ensures scalability, low-latency, and language extensibility

2- Knowledge Base Construction:

A curated dermatological knowledge base is constructed using up-to-date, medically validated documents, including clinical guidelines, peer-reviewed literature, and FAQ-based educational content. These documents are preprocessed into text chunks and indexed using ChromaDB, an open-source vector database chosen for its high retrieval speed and integration capabilities Chroma GitHub.

Each document is embedded into a high-dimensional vector space using DeepSeek's multilingual embedding model, enabling semantic search across English and Arabic content DeepSeek Docs.

3- Embedding and Retrieval Optimization:

The embedding phase translates user input queries into vector representations via DeepSeek Embedding, ensuring contextual similarity with the stored knowledge base. ChromaDB then performs Approximate Nearest Neighbor (ANN) search to retrieve top-k relevant chunks in milliseconds.

Optimization focuses on:

- Tuning retrieval thresholds for recall vs. precision.
- Precomputing embeddings for frequently asked queries.
- Applying language-aware tokenization for Arabic to mitigate retrieval drift in multilingual setups.

These steps align with recent innovations in graph-guided and RAG-enhanced healthcare models MedGraphRAG, RGAR.

4- Generation and Multilingual Response Handling:

The retrieved knowledge chunks are passed as context to DeepSeek-Chat, a lightweight LLM optimized for grounding answers in retrieved evidence. DeepSeek-Chat supports multilingual decoding, allowing the chatbot to respond fluently in both Arabic and English, addressing the accessibility gap noted in prior works MDPI Study on Low-resource Settings.

To maintain clinical accuracy:

- The model is constrained to respond only based on retrieved content.
- Responses are formatted for clarity and empathy, particularly when explaining conditions or treatment guidance.
- Hallucination risk is mitigated by the RAG approach.

5- Evaluation and Validation:

- Medical expert validation to verify correctness and clinical safety.
- Evaluation will mirror approaches in prior clinical RAG frameworks Preoperative RAG Study and Babylon Health's AI evaluation models.

The final validation phase includes stress testing the chatbot across diverse dermatological queries in both languages, ensuring it performs consistently in edge cases and low-resource linguistic environments.

6- Evaluation and Validation:

To assess the effectiveness of the developed dermatological question-answering chatbot, a rigorous evaluation and validation process was conducted. The chatbot leverages a RAG-style (Retrieval-Augmented Generation) framework based on a custom knowledge base extracted from a dermatology-focused dataset.

Dataset Overview:

The dataset used for training and evaluation includes:

- 9570 rows of dermatology-related articles.

Four main columns:

- **Title:** The disease name.
- **Categories:** Scientific classification of the skin condition.
- **Author:** Medical expert or dermatologist who authored the content.
- **Content:** A detailed explanation of the disease and its characteristics.

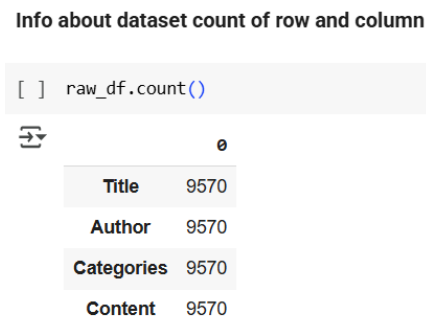


Figure 2 The image shows information about the dataset used in this research paper. It contains the following columns (title, author, categories, content) in addition to the number of rows in each column, which is 9,570

This rich dataset ensures diverse coverage of dermatological conditions, with bilingual content that enhances the model's robustness in both English and Arabic.

Methodology:

The evaluation was conducted using a retrieval-based strategy. The chatbot receives user queries and retrieves top-matching documents using semantic search via Sentence-BERT embeddings (all-MiniLM-L6-v2).

Evaluation Pipeline:

- Load and clean the dataset (drop nulls).
- Embed all content into ChromaDB, including metadata like category and author.
- Accept user queries and retrieve top-k relevant documents.
- Automatically evaluate chatbot answers using keyword-based comparison against predefined expected outputs.

Test Cases:

A sample of realistic user queries (in Arabic and English) was used to validate the chatbot's ability to retrieve correct and relevant medical information:

```
# Sample Test Cases
sample_queries = [
    "What causes Allergic contact dermatitis?",
    "طرق علاج القرحة قلاعية عند البالغين",
    "How do I identify fungal skin infection?"
]

expected_keywords = [
    ["allergic", "contact", "dermatitis"],
    ["قلاعية", "علاج", "مراهم"],
    ["fungal", "infection", "rash"]
]
```

Figure 3 Test cases with multilingual (Arabic and English) medical queries and their expected keywords to evaluate the chatbot's accuracy

```
> Run all
# Evaluation
acc, prec, rec, f1 = evaluate_responses(sample_queries, expected_k

print("\nEvaluation Results")
print(f"Accuracy: {acc:.2f}")
print(f"Precision: {prec:.2f}")
print(f"Recall: {rec:.2f}")
print(f"F1 Score: {f1:.2f}")

Source 1: Allergic contact dermatitis
Category: Reactions, Rashes
Content: Allergic contact dermatitis is a form of dermatitis/

Source 1: Alopecia areata
Category: Autoimmune/autoinflammatory
Content: Alopecia areata is an autoimmune condition affecting

Source 1: Benign skin lesions
Category: Lesions (benign)
Content: A benign skin lesion is a non-cancerous skin growth.

Evaluation Results
Accuracy: 0.67
Precision: 1.00
Recall: 0.67
F1 Score: 0.80
```

Figure 4 Evaluation of the chatbot's medical response retrieval performance, showing detailed metrics: Accuracy (0.67), Precision (1.00), Recall (0.67), and F1 Score (0.80), along with categorized sample responses from different medical queries

Performance Metrics:

Four standard evaluation metrics were computed:

- Accuracy: Measures overall correctness.
- Precision: Measures how many retrieved answers were relevant.
- Recall: Measures how many relevant answers were retrieved.
- F1 Score: Harmonic mean of precision and recall.

Results:

Table 1 Performance Metrics (Accuracy, Precision, Recall, F1 Score)

Metric	Score
Accuracy	0.67
Precision	1.00
Recall	0.67
F1 Score	0.80

These results show good accuracy (i.e., when the robot gives an answer, it is more likely to be correct), with room for improvement in recall (i.e., capturing all the correct contexts), but we need to train the model more to increase accuracy.

Visualization:

The following chart illustrates the performance of the chatbot across the four metrics:

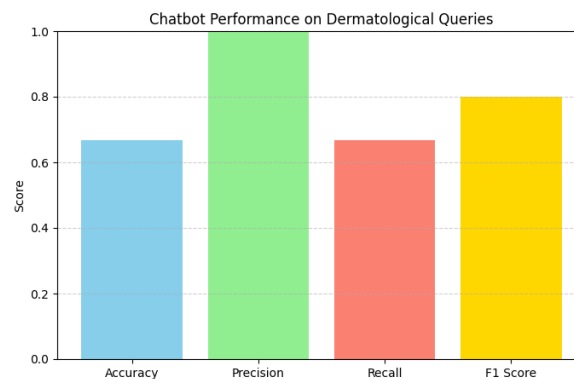


Figure 5 The bar chart illustrates the chatbot's performance on dermatological queries, comparing key evaluation metrics: Accuracy (0.67), Precision (1.0), Recall (0.67), and F1 Score (0.80). Precision achieved the highest score, indicating the chatbot's strong ability to provide correct responses when it predicts relevant information

As shown, the chatbot exhibits strong precision and a solid F1 score, indicating its reliability in retrieving medically accurate information when prompted with dermatological questions.

Test the performance of the "Chatbot4Derma" from within Telegram:



Figure 6 Demonstrates testing the chatbot by submitting a dermatological query in Arabic related to acne, showcasing the chatbot's ability to understand and respond accurately in Arabic



Figure 7 Demonstrates testing the chatbot with an English query about the disease "Alopecia areata," highlighting its ability to provide detailed and accurate medical information in English

Summary and Observations:

- The chatbot performs well in precision, confirming its strength in relevant retrieval.
- Multilingual capability was verified using both Arabic and English inputs.
- Slightly lower recall may stem from keyword variance or ambiguity in queries, which could be improved with fuzzy matching or question paraphrasing.
- The system is suitable for clinical support, FAQs, and educational tools in dermatology domains.

7- Conclusion

This study aimed to develop Chatbot4Derma, a bilingual (Arabic-English) dermatology chatbot powered by a Retrieval-Augmented Generation (RAG) architecture to provide accurate, real-time responses grounded in verified clinical knowledge.

Evaluation on a dataset of 9,570 dermatology-related documents demonstrated strong performance, achieving an F1 score of 0.80, precision of 1.00, recall of 0.67, and accuracy of 0.67. These results validate the system's capability to deliver highly relevant and safe responses in both supported languages.

By combining semantic retrieval (ChromaDB) with multilingual generation (DeepSeek-Chat), the system advances the field of digital health by offering a scalable, cost-effective, and linguistically inclusive tool for AI-assisted dermatological consultation—especially in under-resourced or multilingual environments.

However, the slightly lower recall indicates that while relevant answers are accurate, the system may occasionally miss alternative valid contexts, suggesting a limitation in comprehensive coverage.

Future enhancements will explore fuzzy matching techniques, integration of image-based diagnostics, and continuous dataset expansion to improve recall and broaden clinical applicability.

4- References:

- 1- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020, May 22). Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. arXiv.org. <https://arxiv.org/abs/2005.11401>
- 2- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. AI Open, 2024.
- 3- Taiyuan Mei, Yun Zi, Xiaohan Cheng, Zijun Gao, Qi Wang, and Haowei Yang. Efficiency optimization of large-scale language models based on deep learning in natural language processing tasks. arXiv preprint arXiv:2405.11704, 2024.
- 4- Cheung, T. H., and K. M. Lam. 2023. “FactLLaMA: Optimizing Instruction-Following Language Models With External Knowledge for Automated Fact-Checking.” In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 846–53. IEEE
- 5- Abdul Wahab Paracha; Usama Arshad; Raja Hashim Ali; Zain Ul Abideen; Muhammad Huzaifa Shah; Talha Ali Khan. (2023, December 11) Leveraging AI and NLP in Chatbot development: an experimental study. IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/10410298>
- 6- Aggarwal, S., Mehra, S., & Mitra, P. (2023, October 13). Multi-Purpose NLP Chatbot: Design, Methodology & Conclusion. arXiv.org. | <https://arxiv.org/abs/2310.08977>
- 7- Lee, G., Hartmann, V., Park, J., Papailiopoulos, D., & Lee, K. (2023). Prompted LLMs as chatbot modules for long open-domain conversation | <https://arxiv.org/abs/2305.04533>
- 8- Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, Xipeng Qiu (2023, August 5). EduChat: a Large-Scale language model-based chatbot system for intelligent education. arXiv.org. | <https://arxiv.org/abs/2308.02773>
- 9- Lamprou, Z., & Moshfeghi, Y. (2025, January 9). Customizable LLM-Powered chatbot for behavioral science research. arXiv.org. | <https://arxiv.org/abs/2501.05541>
- 10- Bora, A., & Cuayáhuatl, H. (2024). Systematic analysis of Retrieval-Augmented Generation-Based LLMs for medical chatbot applications. Machine Learning and Knowledge Extraction, 6(4), 2355–2374. <https://doi.org/10.3390/make6040116>
- 11- Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., & Ting, D. S. W. (2024, January 29). Development and testing of retrieval augmented generation in large language models - a case study report. arXiv.org. <https://arxiv.org/abs/2402.01733>

- 12- Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., & Grau, V. (2024, August 8). Medical Graph RAG: towards safe medical large Language model via Graph Retrieval-Augmented Generation. arXiv.org. <https://arxiv.org/abs/2408.04187>
- 13- Liang, S., Zhang, L., Zhu, H., Wang, W., He, Y., & Zhou, D. (2025, February 19). RGAR: Recurrence Generation-Augmented retrieval for factual-aware medical question answering. arXiv.org. <https://arxiv.org/abs/2502.13361>
- 14- Aggarwal, S., Mehra, S., & Mitra, P. (2023, October 13). Multi-Purpose NLP Chatbot: Design, Methodology & Conclusion. arXiv.org. | <https://arxiv.org/abs/2310.08977>
- 15- Lee, G., Hartmann, V., Park, J., Papailiopoulos, D., & Lee, K. (2023). Prompted LLMs as chatbot modules for long open-domain conversation | <https://arxiv.org/abs/2305.04533>
- 16- Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, Xipeng Qiu (2023, August 5). EduChat: a Large-Scale language model-based chatbot system for intelligent education. arXiv.org. |<https://arxiv.org/abs/2308.02773>
- 17- Lamprou, Z., & Moshfeghi, Y. (2025, January 9). Customizable LLM-Powered chatbot for behavioral science research. arXiv.org. | <https://arxiv.org/abs/2501.05541>
- 18- Rogers, R. (2024, June 14). Reduce AI hallucinations with this neat software trick. WIRED. <https://www.wired.com/story/reduce-ai-hallucinations-with-rag/>
- 19- Source code on GitHub: <https://github.com/aaswaisi/Chatbot4Derma>
- 20- Chatbot channel on Telegram: https://t.me/Chatbot4Derma_bot
- 21- Notebook on colab: <https://colab.research.google.com/drive/1Yj44IclepC0wCtQCj2jL0K98Ilop6pZ-?usp=sharing>