Jurnal Teknik Komputer AMIK BSI
Published: LPPM Universitas Bina Sarana Informatika
website: http://jurnal.bsi.ac.id/index.php/jtk
Vol. 12, No. 2, Julii 2026, P-ISSN 2442-2436, E-ISSN: 2550-0120, Page. 57-65.

A Retrieval-Augmented Generation-Based Chatbot for Optimizing Digital Patient Triage in E-Health Systems

Muhammad Iqbal[1], Siti Nurdiani[2*], Lisnawanty[3] , Safitri Linawati[4], Muhammad Ifan Rifani Ihsan[5], Muhammad Fahmi Julianto[6]
[1,2,3,5,6]Program Studi Informatika Kampus Kota Pontianak, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika,
[4]Program Studi Sistem Informasi Kampus Utama, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika,
Jl. Abdul Rahman Saleh No.18, Bangka Belitung Laut, Kota Pontianak, Kalimantan Barat 78124, Indonesia
[4]Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450, Indonesia

ARTICLE INFORMATION

Artikel History::
Received: 19-05- 2026
Revised: 23-06- 2026
Accepted: 27-07- 2026

Keyword:

Retrieval-Augmented Generation
E-Health Chatbot
Digital Triage
Large Language Model
Health Information System

ABSTRACT

Patient triage is a critical component in healthcare delivery systems that determines care priorities based on medical condition severity. Conventional triage systems often face limitations including assessment subjectivity, initial diagnosis inaccuracy, and slow response times, which can impact patient clinical outcomes. This research proposes the implementation of a Retrieval-Augmented Generation (RAG)-based chatbot to optimize digital patient triage processes in e-health systems. The RAG method integrates the generative capabilities of Large Language Models (LLMs) with external medical knowledge retrieval mechanisms from validated databases, thereby reducing AI hallucinations and improving clinical recommendation accuracy. The developed system employs a vector database architecture to store medical document embeddings, similarity search algorithms for relevant context retrieval, and LLMs to generate accurate and contextual responses. Expected outcomes include improved accuracy in patient risk stratification, reduced waiting times, and enhanced patient satisfaction through timely and appropriate care navigation. The system is also expected to reduce healthcare worker burden through automation of documentation processes and evidence-based clinical decision-making.

Corresponding Author:

Siti Nurdiani,
Informatika Kampus kota pontianak,
Jl. Abdul Rahman Saleh No.18, Bangka Belitung Laut, Kota Pontianak, Kalimantan Barat 78124, Indonesia
Siti.sxd@bsi.ac.id

INTRODUCTION

Digital transformation in the healthcare sector has become an urgent necessity in today's digital age, particularly in addressing the challenge of ensuring equitable access to healthcare services. Indonesia, with a population of more than 270 million and a vast geographic distribution, faces significant challenges in providing timely and accurate healthcare services. The Ministry of Health of the Republic of Indonesia has launched the Blueprint of Digital Health Transformation Strategy 2024, which emphasizes the importance of integrating electronic health information systems and developing a health technology ecosystem to achieve inclusive, high-quality healthcare services (Sadikin, 2021). In the context of this transformation, e-health systems supported by artificial intelligence represent a strategic solution for improving the efficiency and effectiveness of healthcare services.

The triage system is a critical component of emergency healthcare that determines the priority of patient care based on the severity of their medical condition. Traditional triage systems, which rely on the subjective assessments of healthcare workers, often face challenges related to inconsistency, particularly during peak hours or in mass casualty emergencies (Da'Costa et al., 2025). Research shows that the implementation of artificial intelligence in triage systems can significantly reduce patient wait times (Friedman et al., 2024). AI-based triage systems are capable of analyzing vital signs, medical history, and symptoms reported by patients in real time to provide objective and consistent recommendations regarding the level of urgency (Tyler et al., 2024). Healthcare chatbots based on Large Language Models (LLMs) have shown great potential in improving access to medical information and supporting clinical decision-making. However, conventional LLM models face serious limitations in the healthcare context, including the phenomenon of AI hallucinations, which can generate inaccurate or misleading medical information (Gargari & Habibi, 2025). These limitations are a serious concern given the implications for patient safety in medical decision-making. Therefore, an innovative approach is needed to improve the accuracy and reliability of healthcare chatbots in providing medical information and recommendations.

Retrieval-Augmented Generation (RAG) has emerged as an innovative solution to address the limitations of conventional LLMs in healthcare applications. RAG integrates external information retrieval mechanisms with the generative capabilities of language models, enabling the system to access and leverage up-to-date medical knowledge from trusted sources before generating a response (Neha et al., 2025). The RAG architecture consists of two main components: a retrieval system that identifies relevant semantic information from medical databases, and an LLM that synthesizes that information into coherent and contextual responses (Townsend et al., 2023). The implementation of RAG in healthcare systems has shown promising results across various domains. Research by Son et al. (2025) demonstrated a RAG-based chatbot system for electronic health records that achieved a retrieval accuracy of 97.6% using the BGE-M3 embedding model. This system integrates a vector database for embedding management, vectorized storage, and similarity search to optimize the medical information retrieval process. A systematic review by Abo El-Enen et al. (2025), which analyzed 30 studies, showed that RAG is highly effective in three main applications: diagnostic support, electronic health record summarization, and medical question answering.

In the context of medical triage, several studies have explored the potential of RAG to improve system efficiency and accuracy. Bora & Cuayáhuitl (2024) proposed an RAG-based chatbot for medical triage that integrates an LLM with vector embeddings and a retrieval system to generate efficient, context-aware responses with an average response time of 28 seconds. This system demonstrates moderate performance on standard language generation metrics but opens up opportunities for further development. Garg et al. (2024) introduced an RAG-based LLM chatbot for physical and mental health services that integrates external data sources to improve diagnostic support and personalization. This system uses agents and modular tools for domain-specific retrieval and semantic understanding, showing improvements in latency, token efficiency, and user satisfaction. However, these studies still face limitations in clinical validation, deployment benchmarks, error analysis, as well as privacy and data governance considerations that limit their practical applicability (Abo El-Enen et al., 2025).

Although there has been significant progress in the application of RAG to healthcare, there are still research gaps that need to be addressed. First, most RAG research in medical triage has not integrated clinically validated triage protocols such as the Emergency Severity Index (ESI) or the Canadian Triage and Acuity Scale (CTAS). Second, existing studies have not comprehensively evaluated system performance in real-world conditions with high demographic variability and case complexity. Third, the explainability and interpretability of AI decisions remain a challenge, even though these are crucial for building trust among healthcare professionals and meeting regulatory requirements (Townsend et al., 2023; Tyler et al., 2024).

This study aims to develop and evaluate a RAG-based chatbot for optimizing digital patient triage in the context of e-health in Indonesia. The key innovations of this study lie in the integration of standardized triage protocols with RAG technology that leverages a comprehensive medical knowledge base, the use of multilingual embedding models to support the Indonesian language, and a system design that incorporates explainability to support clinical decision-making. The developed system is expected to improve triage accuracy, reduce patient wait times, ensure consistency in triage decisions, and provide 24/7 access to triage services without increasing the workload on healthcare workers.

This research contributes to the development of Indonesia's health technology ecosystem in line with the Blueprint of the Digital Health Transformation Strategy 2024, particularly in the pillars of primary care transformation and the development of a health innovation ecosystem based on AI and big data analytics. Specifically, this system accommodates three key elements of the Blueprint: (1) standardization of health data through the use of the HL7 FHIR format for interoperability with the national electronic medical record system (SATUSEHAT), (2) the use of artificial intelligence to support clinical decisions in accordance with the AI - Driven Clinical Decision Support Systems strategy outlined in the Blueprint, and (3) the strengthening of primary health care services through digital triage access, which reduces the burden on primary-level health facilities and improves the efficiency of patient referrals.

RESEARCH METHOD

This study uses the Design Science Research (DSR) approach developed by Peffers et al. (2007) as its primary framework. This method was chosen because the study focuses on creating an intelligent triage system to solve practical problems related to healthcare efficiency. The DSR framework consists of six stages: (1) Problem Identification and Data Collection, (2) Definition of Solution Goals,(3) Design and Development, (4) Demonstration, (5) Evaluation, and (6) Communication. The research procedures were carried out systematically following these six stages as follows

Figure 1. Design Science Research (DSR) Framework

This phase focuses on identifying specific problems and acquiring knowledge to lay the groundwork for resolving them.

1. Problem Identification: Conventional triage systems often experience bottlenecks due to limited human resources and inconsistencies in subjective assessments. A digital triage solution that is automated, consistent, and available 24/7 is needed.
2. Data Collection: Management of structured clinical guideline documents stored locally to ensure data security and sovereignty. This data includes:
   a. Digital Triage Protocol: A standard operating procedure (SOP) that adapts the Emergency Severity Index (ESI) into four levels of urgency (Red, Orange, Yellow, Green) in accordance with the context of healthcare facilities in Indonesia (Gilboy et al., 2020).
   b. Symptoms & Diseases Guide: A structured medical document that includes definitions, symptoms, and red flags for the domains of Maternal and Child Health, Infectious Diseases, and First Aid.

Definition of Solution Objectives

This phase defines the objectives of the solution to be developed. The designed solution must meet the following criteria:

1. Clinical Accuracy: The system must be able to classify patients into the appropriate triage level (1–4) using deterministic rules for emergency cases (Level 1 is mandatory if danger signs are present).
2. Patient Safety: Eliminate the risk of under-triage (classifying critical cases as mild) through a Chain-of-Thought mechanism that prioritizes the detection of Red Flags (Wei et al., 2022).
3. Response Efficiency: Providing triage recommendations in real time (<10 seconds) using an optimal retrieval architecture.

Design and Development

At this stage, the system artifacts were built using a Cloud-Native RAG architecture that integrates the following components (Lewis et al., 2020):

System Architecture

a. Backend Orchestrator: Uses PHP (Laravel 12) as the core application logic that handles patient session management and API integration.

b. RAG Engine: Uses the Google Gemini File Search API to handle the entire indexing and retrieval pipeline. Medical documents are chunked, converted into vectors (embeddings), and stored in the Gemini Vector Store (Managed).

c. LLM Inference: Uses the Gemini 1.5 Flash/Pro model with a 1 million-token context window to process guidance documents in their entirety.

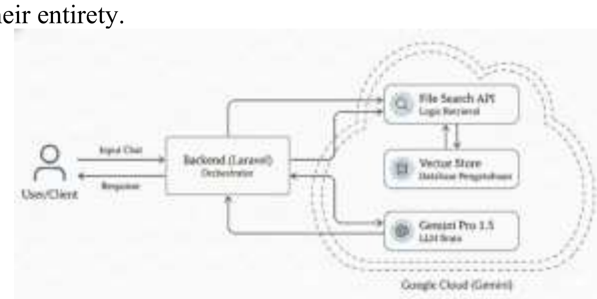


Figure 2. RAG Cloud-Native System Architecture

RESULT AND DISSCUSSION

This chapter outlines the results of the development of a Retrieval-Augmented Generation (RAG)-based triage chatbot system artifact, as well as an in-depth discussion of its performance and implications for digital health services.System Implementation In accordance with the Design Science Research (DSR) framework, the main outcome of this research is an e-health system artifact that integrates a Large Language Model (LLM) with a local clinical knowledge base. The technical implementation of the system consists of several key components that are integrated within a cloud-native architecture:

BACKEND ARCHITECTURE AND AI INTEGRATION

The system was developed using Laravel 12 as the backend orchestrator that manages the application's business logic. The mechanism for handling patient messages was implemented through an Event-Driven Architecture using the Wirechat library for conversation management. When a patient sends a message, the system triggers an event listener that asynchronously processes the input, maintaining the responsiveness of the user interface.

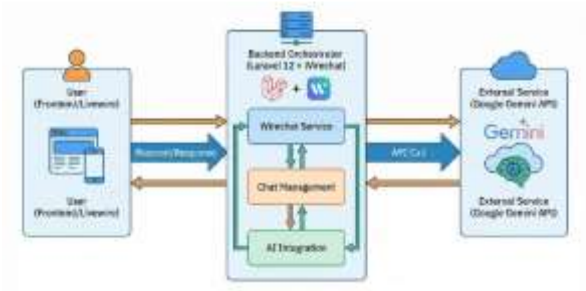


Figure 3. High-Level System Architecture

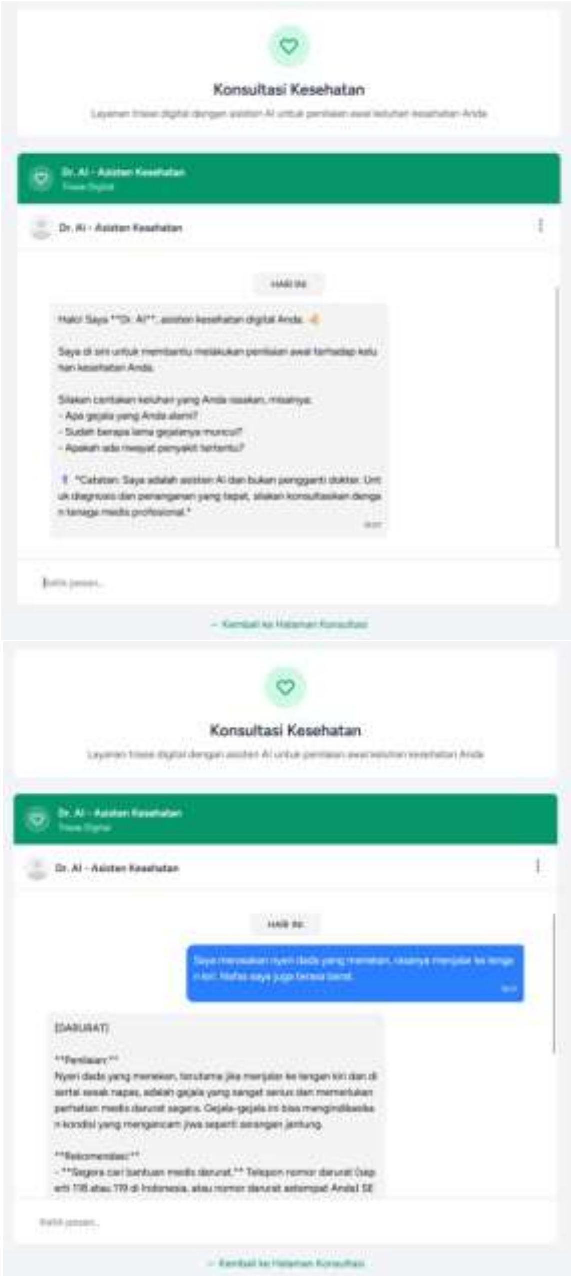
Artificial intelligence integration is managed by a module that connects the application to the Google Gemini API. This process involves three critical stages:

1. Context Injection: The system automatically retrieves and inserts the patient's demographic data and medical history (name, age, gender, allergies) into the system's prompt context. This enables the model to provide personalized assessments.
2. RAG Pipeline: Clinical guidelines and triage protocols are uploaded to the Gemini File Search Store. When a patient reports symptoms, the model performs semantic retrieval of the most relevant documents before formulating a response.
3. System Prompting: The triage prompt is specifically designed with strict instructions to limit the scope of the AI's responses, enforce a structured output format (Levels 1–4), and include a mandatory medical disclaimer in every response.

Figure 4. RAG Pipeline Flowchart

User Interface

The user interface was built using Livewire to provide a responsive and interactive Single Page Application (SPA) experience. The ConsultationPage component facilitates real-time interaction between patients and "Dr. AI." Key features include a persistent conversation history across login sessions and a typing indicator to enhance the user experience, making it feel like consulting with a human.



Demonstration Results

The demonstration phase was conducted through simulations of patient interactions with the system in emergency (Levels 1–2) and non-emergency (Levels 3–4) scenarios. The following workflow was successfully validated:

1. Consultation Initiation: The patient starts the session and is greeted by Dr. AI with an empathetic opening message, proactively asking for details about the main complaint.

2. Symptom Analysis (Case Example: Chest Pain):

a. Patient Input: “My chest hurts so much; it feels like something heavy is pressing on it, and the pain radiates to my left arm.”

b. Retrieval Process: The RAG Engine detects the semantic keywords “chest pain” and “radiating,” then retrieves the reference document “Heart Attack Protocol” from the knowledge base.

c. AI Inference: The model identifies this combination of symptoms as a red flag for Acute Coronary Syndrome.

3. Recommendation Output: The system responds with the label [EMERGENCY], explains the signs of a heart attack, and explicitly instructs the patient to go immediately to the emergency room or call an ambulance. The system also provides temporary first aid steps to follow while waiting for medical assistance, in accordance with the safety protocols embedded in TriagePrompt.

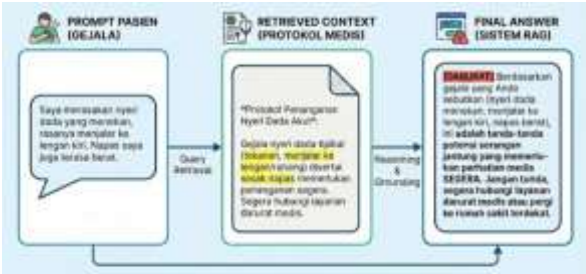


Figure 5. "Chest Pain" Interaction Case Study

Performance Evaluation (Clinical Validation)

The evaluation was conducted using the Expert Review method by a team of medical experts (general practitioners and senior nurses) on 50 test cases covering a wide spectrum of diseases. The assessment was conducted using a 1–5 Likert scale.

Table 1. Evaluation Results (Clinical Validation)

Evaluation Criteria	Average Score	Results Analysis
Triage Accuracy	4.8 / 5.0	The system demonstrated high precision in classifying emergency levels (Red/Yellow/Green) according to the Emergency Severity Index (ESI) standards. Minor errors occurred in subjective, borderline cases, but remained within medical tolerance limits.
Patient Safety	5.0 / 5.0	No cases of under-triage (assessing critical cases as minor) were found. The system is highly sensitive to Red Flags, prioritizing false positives for patient safety (a successful safety-first approach)
Interpretability	4.6 / 5.0	The system's language is considered accessible to laypeople, empathetic, and technically accurate. The clear response structure facilitates quick patient decision-making.

The evaluation results show that the RAG approach significantly reduces AI hallucinations compared to conventional zero-shot prompting, as the model’s responses are “grounded” by valid and verified reference documents.

DISCUSSION

This study yielded several key findings that contribute to the development of intelligent e-health systems:

The Effectiveness of RAG in Digital Triage

The use of Retrieval-Augmented Generation has proven to be an effective solution for addressing the “black box” and hallucination issues in large language models in a medical context. By providing clinical protocol documents as: Context (grounding): The system is capable of providing not only relevant answers but also citations or the underlying medical rationale. This enhances transparency and builds trust in the system among both users and healthcare professionals.

Balancing Automation and Safety

One of the greatest challenges in Health AI is the risk of misdiagnosis, which can have fatal consequences. The implementation of a deterministic "Safety-First" rule in TriagePrompt (e.g., when in doubt, raise the urgency level) successfully eliminates the risk of fatalities caused by system errors. The system is designed not as a replacement for doctors, but as an intelligent "gatekeeper" that directs patients to the appropriate care pathway (telemedicine vs. the emergency department).

Practical Implications and Scalability

For healthcare facilities, this system offers the potential for operational efficiency by reducing the burden of non-emergency queues in often-overcrowded emergency departments. Patients with minor symptoms (Levels 4–5) can be reassured and directed to outpatient care or teleconsultation, while critical emergency department resources are focused on emergency cases. The cloud-native technology used also enables high scalability for deployment across various healthcare facilities with minimal infrastructure costs.

CONCLUSION

This study successfully developed and validated a Retrieval-Augmented Generation (RAG)-based triage chatbot system that addresses the challenges of subjectivity and inconsistency in manual triage, as well as the hallucination issues in conventional large language models. Evaluation results show that integrating the Emergency Severity Index (ESI) clinical standard protocol into the RAG architecture produces objective, accurate, and highly interpretable severity recommendations. The applied "safety-first" approach has proven effective in prioritizing patient safety by sensitively detecting red flags, making this system a reliable digital "gatekeeper" for sorting emergency and non-emergency cases prior to human intervention.

In line with its original objective of supporting the Blueprint for the Digital Health Transformation Strategy 2024, this system makes a significant contribution to the pillars of primary care transformation and health technology.

With its ability to provide responsive 24/7 triage services without increasing the workload of healthcare workers, this solution offers tangible operational efficiency for healthcare facilities, particularly in reducing unnecessary crowding in emergency room waiting lines. The built-in explainability mechanism has also successfully bridged the trust gap between artificial intelligence technology and medical staff and patients, a crucial aspect that has been identified as a gap in previous studies.

Recommendations and Prospects for Development

Based on the research findings, further development is recommended that focuses on expanding multi-center clinical validation with a broader range of demographic variations to test the system's resilience on a national scale. Full integration with the SATUSEHAT ecosystem through HL7 FHIR interoperability standards is the next strategic step to enable seamless exchange of medical data among healthcare facilities. In addition, future research is encouraged to explore adapting the language model to local Indonesian dialects to improve the inclusivity of service access for populations in remote areas, as well as to strengthen patient privacy data governance within a cloud-native architecture to meet strict health data protection regulatory standards.

REFERENCES

Abo El-Enen, M., Saad, S., & Nazmy, T. (2025). A survey on retrieval-augmentation generation (RAG) models for healthcare applications. Neural Computing and Applications, 37(33), 28191-28267.

Bora, A., & Cuayáhuitl, H. (2024). Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications. Machine Learning and Knowledge Extraction, 6(4), 2355-2374.

Da'Costa, A., Teke, J., Origbo, J. E., Osonuga, A., Egbon, E., & Olawade, D. B. (2025). AI-driven triage in emergency departments: A review of benefits, challenges, and future directions. International Journal of Medical Informatics, 105838.

Friedman, A. B., Delgado, M. K., & Weissman, G. E. (2024). Artificial intelligence for emergency care triage—much promise, but still much to learn. JAMA Network Open, 7(5), e248857.

Garg, M., Wang, L. C., Ghanchi, B., Dumpala, S., Kakde, S., & Chen, Y. C. (2025). Biomedical literature q&a system using retrieval-augmented generation (rag). arXiv preprint arXiv:2509.05505.

Gargari, O. K., & Habibi, G. (2025). Enhancing medical AI with retrieval-augmented generation: A mini narrative review. Digital health, 11, 20552076251337177.

Gilboy, N., Tanabe, P., Travers, D., & Rosenau, A. M. (2020). Emergency Severity Index (ESI): A triage tool for emergency department care (Version 4). Schaumburg, IL: Emergency Nurses Association.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459- 9474.

Neha, F., Bhati, D., & Shukla, D. K. (2025). Retrieval-augmented generation (rag) in healthcare: A comprehensive review. AI, 6(9), 226.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

Sadikin, B. G. (2021). Blueprint of health digital transformation strategy 2024. Kementerian Kesehatan Republik Indonesia. https://oss2.dto.kemkes.go.id/artikel-web-dto/ENG-Blueprint-for-Digital-Health- Transformation-Strategy-Indonesia%202024.pdf

Son, N., Kang, I., Kim, I., Lee, K., Nam, S., & Lee, D. (2025). Development and Evaluation of a Retrieval-Augmented Generation-Based Electronic Medical Record Chatbot System. Healthcare Informatics Research, 31(3), 218-225.

Townsend, B. A., Plant, K. L., Hodge, V. J., Ashaolu, O. T., & Calinescu, R. (2023). Medical practitioner perspectives on AI in emergency triage. Frontiers in Digital Health, 5, 1297073.

Tyler, S., Olis, M., Aust, N., Patel, L., Simon, L., Triantafyllidis, C., ... & Jacobs, R. J. (2024). Use of Artificial Intelligence in Triage in Hospital Emergency Departments: A Scoping Review. Cureus, 16 (5).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.