POLITEKNIK NEGERI BALIKPAPAN
PUSAT PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
Jurnal JTT (Jurnal Teknologi Terpadu)
Jl Soekarno Hatta Km 08 Balikpapan, website http://p3m.poltekba.ac.id,
E-mail:p3m@poltekba.ac.id

SURAT KETERANGAN DITERIMA
ACCEPTANCE LETTER

Berdasarkan hasil evaluasi reviewer dan rapat koordinasi tim editor jurnal JTT (Jurnal Teknologi Terpadu) Politeknik Negeri Balikpapan, artikel anda berikut:

Judul	:	Pengembangan Sistem Tanya Jawab Penyakit Pernapasan Menggunakan Model Phi-4 dengan Pendekatan *Retrieval-Augmented Generation*
Penulis	:	Salsabila Widya, Hendri Ahmadian, Nurrizqa
Korespondensi Penulis	:	saalbillawidya@gmail.com
Korespondensi Editor	:	richkihardi@gmail.com
Jumlah Halaman	:	10 (Sepuluh) Lembar

Dinyatakan DITERIMA pada penerbitan Jurnal JTT Volume 14 Nomor 2 Edisi Oktober 2026, untuk itu artikel tersebut diatas tidak diperkenankan untuk dipublikasikan pada media publikasi yang lain. Atas kerjasamanya diucapkan terimakasih.

Ketua Redaksi JTT Poltekba

Dr. Ir. Randis, S.T., M.T
NIDN. 0024108601

Kepala Pusat Penelitian dan Pengabdian kepada Masyarakat

Hadiyanto, S.T., M.Eng
NIDN. 1108078001

Received :	Accepted:	Published :

Pengembangan Sistem Tanya Jawab Penyakit Pernapasan Menggunakan Model Phi-4 dengan Pendekatan *Retrieval-Augmented Generation*

Salsabila Widya[1*], Hendri Ahmadian[2], Nurrizqa[3]

[1,2,3] Teknologi Informasi, Sains dan Teknologi, Universitas Islam Negeri Ar-Raniry Banda Aceh, Banda Aceh, 23111, Indonesia*

[*]Email: saalbillawidya@gmail.com

Abstract

Large Language Model (LLM)-based question answering systems are capable of generating relevant responses; however, they still have the potential to produce information that is inconsistent with the underlying knowledge source. The Retrieval-Augmented Generation (RAG) approach addresses this limitation by retrieving relevant context before generating responses. This study aims to implement the Phi-4 model with the RAG approach for a respiratory disease question answering system and to evaluate the effect of the fine-tuning process on system performance. Fine-tuning was performed using the Low-Rank Adaptation (LoRA) method, while the RAG implementation employed the BAAI/bge-m3 embedding model and the ChromaDB vector database. The evaluation was conducted using the RAGAS framework with the Context Precision, Context Recall, Faithfulness, and Answer Relevancy metrics on 50 question-answer pairs, and the performance of the fine-tuned Phi-4 model was compared with that of the base Phi-4 model. The results show that the fine-tuned Phi-4 model achieved Context Precision, Context Recall, Faithfulness, and Answer Relevancy scores of 1.0000, 1.0000, 0.9855, and 0.6874, respectively. Compared with the base model, the fine-tuning process improved the Faithfulness and Answer Relevancy scores by 0.0776 and 0.0791, respectively, while the Context Precision and Context Recall scores remained at 1.0000. These results indicate that the fine-tuning process improves answer quality without affecting retrieval performance.

Keywords : Retrieval-Augmented Generation, Phi-4, Fine-tuning, Question Answering System, Respiratory Diseases

Abstrak

Sistem tanya jawab berbasis Large Language Model (LLM) memiliki kemampuan menghasilkan jawaban yang relevan, namun masih berpotensi menghasilkan informasi yang tidak sesuai dengan sumber pengetahuan. Pendekatan Retrieval-Augmented Generation (RAG) dapat mengatasi permasalahan tersebut dengan memanfaatkan proses retrieval untuk mengambil konteks yang relevan sebelum jawaban dihasilkan. Penelitian ini bertujuan mengimplementasikan model Phi-4 dengan pendekatan RAG pada sistem tanya jawab penyakit pernapasan serta mengevaluasi pengaruh proses fine-tuning terhadap kinerja sistem. Proses fine-tuning dilakukan menggunakan metode Low-Rank Adaptation (LoRA), sedangkan implementasi RAG menggunakan model embedding BAAI/bge-m3 dan basis data vektor ChromaDB. Evaluasi dilakukan menggunakan framework RAGAS dengan metrik Context Precision, Context Recall, Faithfulness, dan Answer Relevancy pada 50 sampel pasangan pertanyaan dan jawaban, serta membandingkan model Phi-4 hasil fine-tuning dengan base model Phi-4. Hasil penelitian menunjukkan bahwa model Phi-4 hasil fine-tuning memperoleh nilai Context Precision, Context Recall, Faithfulness, dan Answer Relevancy masing-masing sebesar 1,0000, 1,0000, 0,9855, dan 0,6874. Dibandingkan dengan base model, proses fine-tuning meningkatkan nilai Faithfulness sebesar 0,0776 dan Answer Relevancy sebesar 0,0791, sedangkan nilai Context Precision dan Context Recall tetap sebesar 1,0000. Hasil tersebut menunjukkan bahwa proses fine-tuning meningkatkan kualitas jawaban tanpa memengaruhi kinerja proses retrieval.

Kata Kunci : Retrieval-Augmented Generation, Phi-4, Fine-tuning, Sistem Tanya Jawab, Penyakit Pernapasan

1. Pendahuluan

Perkembangan teknologi informasi telah mendorong pemanfaatan *Artificial Intelligence* (AI) dalam berbagai bidang, termasuk kesehatan. Salah satu perkembangan AI yang banyak dimanfaatkan adalah *Large Language Models* (LLM), yaitu model bahasa yang mampu memahami dan menghasilkan teks melalui pemrosesan bahasa alami. Kemampuan tersebut memungkinkan penerapan sistem tanya jawab yang dapat menyajikan informasi kesehatan secara interaktif dan responsif sehingga berpotensi membantu masyarakat memperoleh informasi awal mengenai kondisi kesehatan [1]. Dalam bidang kesehatan, sistem tanya jawab berbasis LLM dimanfaatkan untuk menyediakan informasi mengenai penyakit pernapasan sebelum pengguna melakukan konsultasi langsung dengan tenaga kesehatan [2]. Selain itu, penyakit pernapasan memiliki prevalensi yang cukup tinggi di masyarakat serta dapat memengaruhi kualitas hidup penderitanya, sehingga diperlukan penyediaan informasi kesehatan yang mudah diakses oleh masyarakat [3].

Meskipun memberikan kemudahan dalam penyediaan informasi, sistem tanya jawab berbasis LLM masih menghadapi permasalahan terkait keakuratan jawaban yang dihasilkan. LLM menghasilkan respons berdasarkan pola bahasa yang dipelajari selama proses pelatihan tanpa selalu merujuk pada sumber informasi, sehingga rentan menghasilkan *hallucination*, yaitu jawaban yang terdengar meyakinkan tetapi tidak sepenuhnya sesuai dengan fakta [4]. Penelitian menunjukkan bahwa LLM dapat menghasilkan jawaban yang tidak akurat dengan tingkat kesalahan berkisar antara 50% hingga 82%, bahkan ketika menggunakan model dengan performa yang lebih baik, *hallucination* masih ditemukan pada sekitar 23% hingga 53% jawaban yang dihasilkan [5]. Kondisi tersebut menunjukkan bahwa kualitas jawaban pada sistem tanya jawab berbasis LLM masih perlu ditingkatkan, terutama pada domain kesehatan yang memerlukan informasi yang akurat.

Untuk mengatasi permasalahan tersebut, berbagai penelitian telah menerapkan pendekatan *Retrieval-Augmented Generation* (RAG) pada model LLM, di antaranya GPT-3.5-Turbo [6], Solar Mini Chat [7], GPT-OSS-20B [8], LLaMA 3.1 8B [9], [10], serta Mistral 7B [11]. Hasil penelitian tersebut menunjukkan bahwa pendekatan RAG mampu meningkatkan relevansi dan kualitas jawaban dengan memanfaatkan informasi dari basis pengetahuan eksternal. Namun, peningkatan tersebut tidak selalu konsisten karena efektivitas RAG bergantung pada model yang digunakan serta kualitas dan relevansi informasi yang berhasil diperoleh [12]. Pada penelitian-penelitian tersebut, RAG umumnya diterapkan pada *base model* tanpa proses *fine-tuning* domain spesifik, sehingga belum banyak penelitian yang membandingkan kinerja model generatif sebelum dan sesudah *fine-tuning* dalam sistem RAG. Selain itu, dibandingkan dengan model seperti LLaMA 3.1 8B [9], [10] atau Mistral 7B [11] yang telah digunakan pada penelitian RAG sebelumnya, penerapan dan evaluasi Phi-4 pada sistem tanya jawab kesehatan berbahasa Indonesia, khususnya pada domain penyakit pernapasan, masih terbatas. Karena itu, efektivitas penerapan RAG pada model Phi-4 dalam meningkatkan kualitas jawaban serta mengurangi *hallucination* masih perlu diteliti lebih lanjut [13].

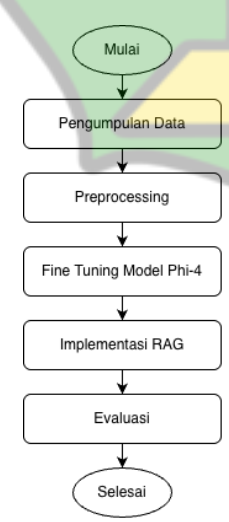
Berdasarkan kondisi tersebut, penelitian ini menerapkan model Phi-4 dengan pendekatan RAG pada sistem tanya jawab penyakit pernapasan. Phi-4 merupakan model bahasa yang dirancang untuk meningkatkan kemampuan penalaran sehingga mampu menghasilkan jawaban yang lebih terstruktur dan relevan [14]. Sementara itu, pendekatan RAG digunakan untuk mengatasi keterbatasan LLM yang bergantung pada pengetahuan yang tersimpan dalam parameter model, dengan menggabungkan kemampuan generatif LLM dan proses pengambilan informasi dari basis pengetahuan eksternal, sehingga memungkinkan model menghasilkan respons berdasarkan informasi yang diperoleh [15]. Pada penelitian ini, proses *fine-tuning* model

Phi-4 dilakukan menggunakan *Low-Rank Adaptation* (LoRA) pada domain penyakit pernapasan. Model hasil *fine-tuning* kemudian digunakan sebagai model generatif dalam sistem tanya jawab berbasis RAG, sehingga memungkinkan penelitian ini membandingkan kinerja model sebelum dan sesudah adaptasi pada domain penyakit pernapasan.

Penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi sistem tanya jawab penyakit pernapasan berbasis model Phi-4 dengan pendekatan RAG. Kinerja sistem dinilai berdasarkan kualitas jawaban yang dihasilkan serta kemampuan sistem dalam mengurangi *hallucination* [16]. Hasil penelitian diharapkan dapat memberikan gambaran mengenai pengaruh *fine-tuning* terhadap kinerja sistem RAG, serta menjadi referensi bagi pengembangan sistem informasi kesehatan berbasis LLM.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional untuk mengimplementasikan dan mengevaluasi sistem tanya jawab penyakit pernapasan berbasis model Phi-4 dengan pendekatan RAG. Pendekatan kuantitatif dipilih karena memungkinkan pengukuran kinerja sistem secara objektif melalui metrik evaluasi untuk menilai kemampuan sistem dalam menghasilkan jawaban yang relevan dan sesuai dengan basis pengetahuan.



Gambar 1. Tahapan Penelitian

Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1. Penelitian diawali dengan pengumpulan data penyakit pernapasan sebagai basis pengetahuan utama. Selanjutnya, dilakukan *preprocessing* untuk menyiapkan data sebelum digunakan pada proses *fine-tuning*. Tahap berikutnya adalah *fine-tuning* model Phi-4 menggunakan metode LoRA untuk menyesuaikan model pada domain penyakit pernapasan. Model hasil *fine-tuning* kemudian diimplementasikan pada sistem tanya jawab berbasis RAG untuk menghasilkan jawaban berdasarkan konteks yang diperoleh dari basis pengetahuan. Tahap terakhir adalah evaluasi untuk mengukur kinerja sistem, baik dari sisi kemampuan *retrieval* maupun kualitas jawaban yang dihasilkan.

2.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Hugging Face, yaitu agufsamudra/alodokter-qna. Dataset tersebut merupakan hasil *scraping* diskusi publik pada situs Alodokter. Dataset berisi 288.106 baris data yang terdiri dari kolom *title*, *question*, *answer*, *doctor_name*, *tag*, dan *url*, yang berisi pasangan pertanyaan dan jawaban terkait informasi kesehatan.

2.2. Preprocessing

Proses *preprocessing* dilakukan untuk memastikan data penyakit pernapasan yang digunakan bersih, konsisten, dan siap digunakan pada proses *fine-tuning* model. Tahapan *preprocessing* yang dilakukan adalah sebagai berikut.

a. Penyaringan data: data difilter berdasarkan *tag* penyakit pernapasan, kemudian data yang tidak termasuk dalam domain penyakit pernapasan dihapus, sehingga hanya data yang relevan yang digunakan.
b. Pembersihan dan normalisasi teks: mencakup penghapusan karakter yang tidak relevan dan simbol khusus, penghapusan sapaan pada teks, serta normalisasi spasi, tanda baca, dan kapitalisasi.
c. Pengelompokan kategori: setiap data dikelompokkan ke dalam kategori, yaitu

umum, *bayi*, *anak*, *lansia*, dan *hamil* berdasarkan informasi yang terdapat pada *question* dan *answer*, yang digunakan sebagai *metadata* dalam proses *retrieval*.

d. Penyesuaian kolom: kolom *title* digunakan sebagai *question*, sedangkan kolom *question* dan *doctor_name* pada dataset dihapus karena menggunakan bahasa informal pengguna dan nama dokter yang tidak diperlukan dalam proses pelatihan model.

Dataset hasil *preprocessing* terdiri dari 18.724 pasangan data pertanyaan dan jawaban penyakit pernapasan. Pada dataset tersebut, tag *vaksin-covid-19-1*, *tuberkulosis* dan *sars-cov-2* merupakan tag dengan jumlah data yang lebih dominan dibandingkan tag lainnya. Dataset hasil *preprocessing* kemudian dibagi menjadi data *training* dan *testing* dengan perbandingan 90:10, dilakukan secara *stratified* berdasarkan *tag* penyakit agar proporsi setiap kategori tetap seimbang. Data *training* selanjutnya dibagi menjadi data *training* dan *validation* dengan perbandingan 90:10 secara acak. Data *training* digunakan untuk proses *fine-tuning* model Phi-4 menggunakan metode *Low-Rank Adaptation* (LoRA), data *validation* digunakan untuk memantau kinerja model selama proses *fine-tuning*, sedangkan data *testing* dipisahkan dari proses pelatihan dan digunakan untuk evaluasi kinerja sistem RAG.

Tabel 1. Contoh Data Hasil Preprocessing

Komponen	Isi Dataset
Question	Bersin terus menerus hidung gatal mata berair penyakit apa?
Answer	Sering bersin-bersin hingga mata kemerahan, berair, dan hidung gatal terutama di pagi hari besar kemungkinan menandakan Anda mengalami *rhinitis* alergi....
Tag	rhinitis-alergi
URL	https://www.alodokter.com/komunitas/topic/bersin-terus-menerus-hidung-gatal-mata-berair-penyakit-apa-
Kategori	umum

Tabel 1 menampilkan contoh data hasil *preprocessing* yang digunakan dalam penelitian. Dataset hasil *preprocessing* terdiri atas kolom *question*, *answer*, *tag*, *url*, dan *kategori*. Kolom *question* berasal dari kolom *title* pada dataset dan berisi pertanyaan, sedangkan kolom *answer* berisi jawaban yang sesuai dengan setiap pertanyaan. Kolom *tag* digunakan pada tahap *preprocessing* untuk proses *filtering* data penyakit pernapasan, kolom *url* digunakan sebagai sumber referensi dokumen, sedangkan kolom *kategori* digunakan sebagai *metadata* untuk mendukung proses *retrieval*. Dataset hasil *preprocessing* tersebut selanjutnya digunakan sebagai dataset pelatihan pada proses *fine-tuning* model Phi-4 untuk menyesuaikan kemampuan model terhadap domain penyakit pernapasan.

2.3. Fine-tuning Model Phi-4

Proses *fine-tuning* dilakukan untuk menyesuaikan kemampuan model Phi-4 dalam memahami dan menjawab pertanyaan pada domain penyakit pernapasan. *Fine-tuning* dilakukan menggunakan *framework* Unsloth dengan *Supervised Fine-Tuning Trainer* (SFTTrainer) dari *library* TRL. Metode *Parameter-Efficient Fine-Tuning* (PEFT) berbasis LoRA digunakan agar proses *fine-tuning* dapat dilakukan secara efisien tanpa memperbarui seluruh parameter.

Tabel 2. Konfigurasi Base Model

Parameter	Nilai
Model	unsloth/phi-4-unsloth-bnb-4bit
Jumlah Parameter	14B
Quantization	4-bit
Maximum Sequence Length	2048 token

Tabel 2 menunjukkan konfigurasi *base model* yang digunakan dalam penelitian ini. Model yang digunakan adalah unsloth/phi-4-unsloth-bnb-4bit dengan kuantisasi 4-bit dan jumlah parameter sebesar 14 miliar. Selain itu, *maximum sequence length* ditetapkan sebesar

2048 token untuk membatasi panjang teks yang dapat diproses oleh model.

Tabel 3. Konfigurasi LoRA

Parameter	Nilai
LoRA Rank (r)	16
LoRA Alpha	16
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Loss Computation	Response Only

Tabel 3 menampilkan konfigurasi LoRA yang digunakan pada proses *fine-tuning*. Nilai *LoRA rank* (r) dan *LoRA alpha* masing-masing ditetapkan sebesar 16, sedangkan *target modules* mencakup modul yang berperan dalam proses *attention* dan *feed-forward* model. Selain itu, perhitungan *loss* hanya dilakukan pada bagian jawaban (*response only*), sehingga model berfokus mempelajari pola jawaban yang relevan dengan domain penyakit pernapasan tanpa memperhitungkan *loss* pada bagian pertanyaan.

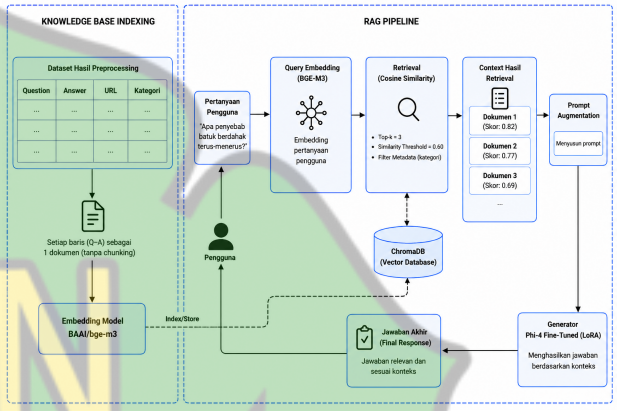
Tabel 4. Hyperparameter Training

Parameter	Nilai
Epoch	2
Batch Size per Device	2
Gradient Accumulation Steps	4
Learning Rate	2×10^{-4}
LR Scheduler	Cosine
Warmup Ratio	0.03
Optimizer	AdamW 8-bit
Weight Decay	0.01

Tabel 4 menunjukkan *hyperparameter* yang digunakan selama proses *training*. Model dilatih selama 2 *epoch* dengan *batch size* sebesar 2 per *device* dan *gradient accumulation steps* sebesar 4, sehingga menghasilkan *effective batch size* sebesar 8. *Learning rate* ditetapkan sebesar 2×10^{-4} menggunakan *cosine scheduler* dan *warmup ratio* sebesar 0.03, sementara optimisasi dilakukan menggunakan AdamW 8-bit dengan *weight decay* sebesar 0.01.

2.4. Implementasi RAG

Implementasi sistem tanya jawab penyakit pernapasan dilakukan menggunakan model Phi-4 hasil *fine-tuning* dengan pendekatan RAG, sehingga sistem mampu menghasilkan jawaban berdasarkan konteks yang relevan dari basis pengetahuan.



Gambar 2. Arsitektur RAG

Tahapan implementasi RAG pada sistem tanya jawab penyakit pernapasan dapat dilihat pada Gambar 2. Proses diawali dengan mengubah setiap pasangan *question* dan *answer* pada dataset hasil *preprocessing* menjadi representasi vektor menggunakan model *embedding* BAAI/bge-m3. Setiap pasangan *question* dan *answer* direpresentasikan sebagai satu dokumen, kemudian disimpan ke dalam basis data vektor ChromaDB menggunakan metrik *Cosine Similarity* dengan *metadata* berupa kategori dan URL.

Setelah proses penyimpanan selesai, sistem menerima pertanyaan dari pengguna dan mengategorikan pertanyaan menjadi umum, bayi, anak, lansia, dan hamil. Kategori yang diperoleh digunakan sebagai *metadata filter* pada proses *retrieval*. Selanjutnya, pertanyaan pengguna diubah menjadi representasi vektor menggunakan model *embedding* BAAI/bge-m3, kemudian dilakukan pencarian pada ChromaDB berdasarkan nilai *Cosine Similarity*.

Pada proses *retrieval*, sistem mengambil tiga dokumen dengan *similarity score* tertinggi (*top-k*=3) menggunakan batas minimum *similarity score* sebesar 0,60. Dari ketiga dokumen tersebut, hanya satu dokumen dengan *similarity score* tertinggi yang digunakan sebagai konteks (k=1) untuk proses generasi. Konteks tersebut kemudian digabungkan dengan pertanyaan pengguna melalui proses *prompt augmentation* sebelum diberikan kepada model Phi-4 hasil *fine-tuning*. Model kemudian menghasilkan jawaban berdasarkan konteks yang tersedia. Selanjutnya, dilakukan pemeriksaan terhadap keluaran untuk mendeteksi jawaban yang berulang sebelum jawaban akhir dikembalikan kepada pengguna.

2.5. Evaluasi

Evaluasi dilakukan untuk mengukur kinerja sistem tanya jawab berbasis RAG dalam mengambil konteks yang relevan serta menghasilkan jawaban yang sesuai. Proses evaluasi menggunakan *framework Retrieval-Augmented Generation Assessment* (RAGAS) yang menyediakan metrik untuk mengevaluasi komponen *retrieval* dan *generation*. Evaluasi menggunakan GPT-5.4-mini sebagai *LLM judge* untuk menilai kualitas jawaban berdasarkan metrik evaluasi yang digunakan [17].

2.5.1. Evaluasi Retrieval

Evaluasi *retrieval* bertujuan untuk menilai kemampuan sistem dalam mengambil konteks yang relevan dari basis pengetahuan. Metrik yang digunakan terdiri atas *Context Precision* dan *Context Recall*.

a. Context Precision

Context Precision digunakan untuk mengukur kemampuan *retriever* dalam menempatkan konteks yang relevan pada peringkat teratas hasil *retrieval* sehingga berguna untuk menjawab pertanyaan, dengan membandingkan setiap konteks terhadap referensi.

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K}(\text{Precision@}k \times v_k)}{\text{Jumlah konteks relevan pada top-}K}$$

dengan:

$$\text{Precision@}k = \frac{\text{True Positives@}k}{\text{True Positives@}k + \text{False Positives@}k}$$

di mana:
K: jumlah konteks teratas yang di-*retrieval*.
v_k: indikator relevansi konteks pada peringkat ke-k (0 atau 1).

b. Context Recall

Context Recall digunakan untuk mengukur sejauh mana informasi yang terdapat pada referensi berhasil tercakup oleh konteks hasil *retrieval*.

$$\text{Context Recall} = \frac{\text{Jumlah klaim pada referensi yang didukung konteks}}{\text{Total jumlah klaim pada referensi}}$$

2.5.2. Evaluasi Generation

Evaluasi *generation* bertujuan untuk menilai kualitas jawaban yang dihasilkan model berdasarkan konteks hasil *retrieval*. Metrik yang digunakan terdiri atas *Faithfulness* dan *Answer Relevancy*.

a. Faithfulness

Faithfulness digunakan untuk mengukur sejauh mana informasi pada jawaban yang dihasilkan model didukung oleh konteks hasil *retrieval*.

$$\text{Faithfulness} = \frac{\text{Jumlah klaim pada jawaban yang didukung konteks}}{\text{Total jumlah klaim pada jawaban}}$$

b. Answer Relevancy

Answer Relevancy digunakan untuk mengukur sejauh mana jawaban yang dihasilkan model relevan dengan pertanyaan pengguna.

$$\text{Answer Relevancy} = \frac{1}{N}\sum_{i=1}^{N} \text{cosine similarity}\left(E_{g_i}, E_o\right)$$

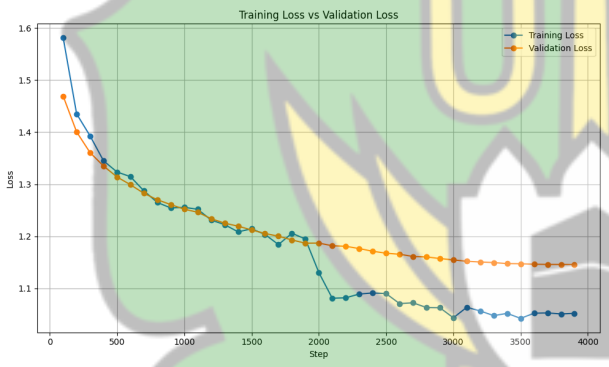
di mana:
E_{g_i}: embedding pertanyaan ke-i hasil generasi
E_o: embedding pertanyaan pengguna
N: jumlah pertanyaan hasil generasi

3. Hasil Penelitian

Penelitian ini menghasilkan sistem tanya jawab penyakit pernapasan berbasis model Phi-4 dengan pendekatan RAG. Hasil penelitian mencakup hasil *fine-tuning* model Phi-4 serta evaluasi kinerja sistem tanya jawab menggunakan *framework* RAGAS. Evaluasi kinerja dilakukan dengan membandingkan model Phi-4 hasil *fine-tuning* dengan pendekatan RAG dan *base model* Phi-4 dengan pendekatan RAG.

3.1. Hasil Fine-Tuning

Proses *fine-tuning* dilakukan untuk menyesuaikan kemampuan model Phi-4 dalam memahami dan menghasilkan jawaban pada domain penyakit pernapasan. Hasil *fine-tuning* dievaluasi berdasarkan perubahan nilai *training loss* selama proses pelatihan.



Gambar 3. Grafik Training Loss

Gambar 3 memperlihatkan perubahan nilai *training loss* dan *validation loss* selama proses *fine-tuning* model Phi-4. Nilai *training loss* mengalami penurunan dari 1,5812 pada awal pelatihan menjadi 1,0522 pada akhir pelatihan, sedangkan *validation loss* menurun dari 1,4691 menjadi 1,1460. Pada *step* 1.700 hingga 2.100, nilai *training loss* mengalami penurunan yang cukup besar, yang menunjukkan adanya penyesuaian model terhadap pola pertanyaan dan jawaban pada data pelatihan. Setelah *step* 2.000, kedua kurva mulai melandai dan menunjukkan perubahan yang relatif kecil hingga akhir pelatihan. Kondisi ini sesuai dengan penggunaan *cosine learning rate scheduler* yang menurunkan *learning rate* secara bertahap, sehingga pembaruan parameter model menjadi lebih kecil mendekati akhir *epoch*. *Validation loss* juga menunjukkan penurunan yang konsisten meskipun berada sedikit di atas *training loss*, yang menunjukkan bahwa model tetap memberikan kinerja yang stabil pada data *validation*. Selain itu, tidak terlihat peningkatan *validation loss* yang berkelanjutan ketika *training loss* terus menurun, sehingga tidak terdapat indikasi *overfitting* yang berarti selama proses *fine-tuning*.

3.2. Evaluasi Sistem RAG

Kinerja sistem tanya jawab penyakit pernapasan berbasis model Phi-4 hasil *fine-tuning* dengan pendekatan RAG dievaluasi menggunakan *framework* RAGAS dengan empat metrik, yaitu *Context Precision*, *Context Recall*, *Faithfulness*, dan *Answer Relevancy*, dengan GPT-5.4-mini sebagai *LLM judge*. Evaluasi dilakukan terhadap 50 sampel pasangan pertanyaan dan jawaban yang dipilih secara acak dari data *testing*. Sampel dibatasi pada data dengan satu *tag* untuk memastikan setiap sampel memiliki satu kategori penyakit atau gejala yang relevan secara spesifik, kemudian dikelompokkan menjadi *tag* penyakit dan *tag* gejala dengan jumlah sampel yang dialokasikan berdasarkan jumlah data pada masing-masing *tag*. Hasilnya, sampel evaluasi mencakup 30 *tag* penyakit dan 13 *tag* gejala. Sebagai pembanding, evaluasi juga dilakukan pada sistem tanya jawab berbasis RAG yang menggunakan *base model* Phi-4 tanpa *fine-tuning* untuk mengetahui sejauh mana proses *fine-tuning* berpengaruh terhadap peningkatan kinerja sistem.

Tabel 5. Hasil Evaluasi Fine-tuned Phi-4 dengan RAG

Metrik	Rata-rata
Context Precision	1.0000
Context Recall	1.0000
Faithfulness	0.9855
Answer Relevancy	0.6874

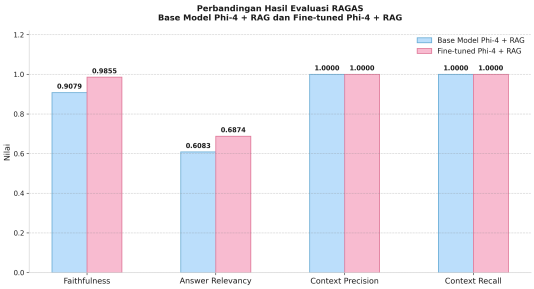
Tabel 5 menampilkan hasil evaluasi sistem tanya jawab menggunakan model Phi-4 hasil *fine-tuning* dengan pendekatan RAG.

Nilai *Context Precision* sebesar 1,0000 memperlihatkan bahwa sistem mampu mengambil konteks yang relevan dan menempatkannya pada peringkat teratas hasil *retrieval*. Nilai *Context Recall* sebesar 1,0000 memperlihatkan bahwa seluruh informasi yang terdapat pada referensi berhasil tercakup oleh konteks hasil *retrieval*. Nilai *Faithfulness* sebesar 0,9855 menunjukkan bahwa sebagian besar informasi pada jawaban yang dihasilkan model didukung oleh konteks hasil *retrieval*. Sementara itu, nilai *Answer Relevancy* sebesar 0,6874 menunjukkan bahwa jawaban yang dihasilkan cukup relevan dengan pertanyaan pengguna, namun relevansinya masih dapat ditingkatkan.

Tabel 6. Hasil Evaluasi Base Model Phi-4 dengan RAG

Metrik	Rata-rata
Context Precision	1.0000
Context Recall	1.0000
Faithfulness	0.9079
Answer Relevancy	0.6083

Tabel 6 menunjukkan hasil evaluasi sistem tanya jawab menggunakan *base model* Phi-4 dengan pendekatan RAG. Nilai *Context Precision* dan *Context Recall* masing-masing mencapai 1,0000, sama seperti pada model Phi-4 hasil *fine-tuning*, yang berarti bahwa proses *retrieval* pada kedua sistem memiliki kinerja yang sama. Perbedaan terlihat pada metrik *Faithfulness* dan *Answer Relevancy* yang masing-masing memperoleh nilai 0,9079 dan 0,6083. Nilai tersebut lebih rendah dibandingkan model Phi-4 hasil *fine-tuning*, sehingga dapat disimpulkan bahwa proses *fine-tuning* memberikan peningkatan terhadap kualitas jawaban yang dihasilkan.



Gambar 4. Perbandingan Hasil Evaluasi

Gambar 4 memperlihatkan perbandingan hasil evaluasi antara *base model* Phi-4 dengan pendekatan RAG dan model Phi-4 hasil *fine-tuning* dengan pendekatan RAG. Pada metrik *Context Precision* dan *Context Recall*, keduanya menghasilkan nilai yang sama, yaitu 1,0000, yang berarti proses *retrieval* memiliki kinerja yang konsisten pada keduanya karena menggunakan konfigurasi yang identik, sehingga proses *fine-tuning* tidak memengaruhi kemampuan sistem dalam mengambil konteks yang relevan. Perbedaan terlihat pada metrik *Faithfulness* dan *Answer Relevancy*. Nilai *Faithfulness* meningkat dari 0,9079 menjadi 0,9855 dengan selisih 0,0776, sedangkan nilai *Answer Relevancy* meningkat dari 0,6083 menjadi 0,6874 dengan selisih 0,0791.

Peningkatan pada kedua metrik tersebut menandakan bahwa proses *fine-tuning* meningkatkan kemampuan model dalam menghasilkan jawaban yang lebih konsisten dengan konteks hasil *retrieval* serta lebih relevan dengan pertanyaan pengguna. Hasil tersebut menunjukkan bahwa peningkatan kinerja sistem lebih dipengaruhi kemampuan model generatif yang semakin baik setelah proses *fine-tuning*, bukan dari perubahan pada mekanisme *retrieval*. Dengan demikian, *fine-tuning* memungkinkan model memanfaatkan konteks hasil *retrieval* secara lebih efektif sehingga jawaban yang dihasilkan menjadi lebih sesuai dengan informasi pada basis pengetahuan.

Meskipun demikian, evaluasi pada penelitian ini sepenuhnya bersifat otomatis menggunakan *framework* RAGAS dengan LLM sebagai *judge*, sehingga hanya menilai konsistensi tekstual antara jawaban dan konteks hasil *retrieval*, bukan kebenaran medis dari jawaban yang dihasilkan. Validasi oleh ahli kesehatan belum dilakukan pada penelitian ini, sehingga akurasi klinis dari jawaban sistem belum dapat dipastikan sepenuhnya.

4. Kesimpulan

Implementasi sistem tanya jawab penyakit pernapasan menggunakan model Phi-4 dengan pendekatan RAG mampu

menghasilkan jawaban berdasarkan konteks yang diperoleh dari basis pengetahuan. Proses *fine-tuning* menggunakan metode LoRA menghasilkan penurunan nilai *training loss* dari sekitar 1,64 menjadi 1,01, yang berarti model mampu mempelajari pola pada dataset penyakit pernapasan secara stabil.

Hasil evaluasi menggunakan *framework* RAGAS menunjukkan bahwa model Phi-4 hasil *fine-tuning* dengan pendekatan RAG memperoleh nilai *Context Precision*, *Context Recall*, *Faithfulness*, dan *Answer Relevancy* masing-masing sebesar 1,0000, 1,0000, 0,9855, dan 0,6874. Dibandingkan dengan *base model* Phi-4 yang menggunakan pendekatan RAG, proses *fine-tuning* meningkatkan nilai *Faithfulness* sebesar 0,0776 dan *Answer Relevancy* sebesar 0,0791, sedangkan nilai *Context Precision* dan *Context Recall* tetap sama, yaitu 1,0000. Hal ini berarti bahwa proses *fine-tuning* meningkatkan kualitas jawaban tanpa memengaruhi kinerja proses *retrieval*.

Hasil penelitian ini menunjukkan bahwa kombinasi *fine-tuning* menggunakan LoRA dengan pendekatan RAG dapat meningkatkan kualitas sistem tanya jawab kesehatan berbasis LLM. Peningkatan kualitas jawaban setelah *fine-tuning* menunjukkan bahwa adaptasi model dapat mendukung penyajian informasi penyakit pernapasan yang lebih sesuai dengan konteks hasil *retrieval*, sehingga berpotensi mendukung penyediaan informasi kesehatan awal bagi masyarakat.

5. Saran

Berdasarkan hasil penelitian ini, penelitian selanjutnya dapat menggunakan dataset yang lebih beragam serta mencakup lebih banyak domain penyakit, membandingkan penggunaan model *embedding* maupun model LLM yang berbeda, dan mempertimbangkan penggunaan beberapa dokumen hasil *retrieval* sebagai konteks masukan. Penelitian selanjutnya juga disarankan melibatkan validasi oleh ahli kesehatan untuk menilai akurasi klinis dan keamanan informasi yang dihasilkan sistem.

6. Daftar Pustaka

[1] N. Rachmat and D. P. Kesuma, "Implementasi Large Language Models Gemini Pada Pengembangan Aplikasi Chatbot Berbasis Android," *Jurnal Ilmu Komputer (JUIK)*, vol. 4, no. 1, pp. 40–52, 2024.

[2] L. Athota, V. K. Shukla, N. Pandey, and A. Rana, "Chatbot for Healthcare System Using Artificial Intelligence," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 619–622. doi: 10.1109/ICRITO48877.2020.9197833.

[3] A. Calderaro, M. Buttrini, B. Farina, S. Montecchini, F. De Conto, and C. Chezzi, "Respiratory tract infections and laboratory diagnostic methods: a review with a focus on syndromic panel-based assays," *Microorganisms*, vol. 10, no. 9, p. 1856, 2022, doi: 10.3390/microorganisms10091856.

[4] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards Mitigating Hallucination in Large Language Models via Self-Reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1827–1843.

[5] M. Omar *et al.*, "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support," *Communications Medicine*, vol. 5, no. 1, p. 330, 2025, doi: 10.1038/s43856-025-01021-3.

[6] Q. Zhou *et al.*, "GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation," *Front. Med. (Lausanne).*, vol. 11, p. 1392555, 2024, doi: 10.3389/fmed.2024.1392555.

[7] N. Son, I. Kang, I. Kim, K. Lee, S. Nam, and D. Lee, "Development and Evaluation of a Retrieval-Augmented Generation-Based Electronic Medical Record Chatbot System," *Healthc. Inform. Res.*, vol. 31, no. 3, pp. 218–225, 2025, doi: 10.4258/hir.2025.31.3.218.

[8] M. A. Abdurrazzaq, E. L. Tjiong, A. Fasya, M. Hiu, and J. Tanuwidjaya, "An Indonesian Chatbot for Disease Diagnosis Using Retrieval-Augmented Generation," *INOVTEK Polbeng-Seri Informatika*, vol. 10, no. 3, pp. 1877–1887, 2025.

[9] I. L. Kharisma, M. S. Hidayat, Somantri, and Kamdan, "Implementasi Retrieval Augmented Generation dalam Sistem Chatbot Dermatologi Berbasis Website," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 11, no. 3, pp. 448–462, 2025, doi: 10.28932/jutisi.v11i3.12258.

[10] G. Samudra, A. T. Zy, and Ermanto, "Implementasi Retrieval Augmented Generation (RAG) Dalam Perancangan Chatbot Kesehatan Pencernaan," *JSAI: Journal Scientific and Applied Informatics*, vol. 8, no. 1, pp. 181–188, 2025, doi: 10.36085.

[11] L. O. M. Y. Prayitno, A. Nurfadilah, S. B. Saudi, W. D. Tsunami, and A. M. Sajiah, "Conversational Agent for Medical Question-Answering Using RAG and LLM," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 4, no. 3, pp. 1894–1899, 2025.

[12] Y. Chen *et al.*, "Exploring the Role of Knowledge Graph-Based RAG in Japanese Medical Question Answering with Small-Scale LLMs," *arXiv preprint arXiv:2504.10982*, pp. 1–10, 2025.

[13] O. K. Gargari and G. Habibi, "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," *Digit. Health*, vol. 11, pp. 1–7, 2025, doi: 10.1177/20552076251337177.

[14] M. Abdin *et al.*, "Phi-4 Technical Report," *arXiv preprint arXiv:2412.08905*, pp. 1–36, 2024.

[15] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

[16] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023, doi: 10.1145/3571730.

[17] S. Es, J. James, L. E. Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.