Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

Implementasi Fine Tuning Menggunakan Metode QLoRA Pada Sistem Tanya Jawab Hadits

Muhammad Azmi[1], Fitra Kurnia[2], Nazruddin Safaat Harahap[3], Muhammad Irsyad[4], Pizaini[5], Arif Marsal[6]
[1,2,3,4,5] Teknik Informatika, Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
[6] Sistem Informasi, Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau
[1]12250115675@students.uin-suska.ac.id , [2]fitra.k@uin-suska.ac.id , [3]nazruddin.safaat@uin-suska.ac.id*, [4]irsyadtech@uin-suska.ac.id, [5]pizaini@uin-suska.ac.id, [6]arif.marsal@uin-suska.ac.id

Abstract

Islamic hadith is the second primary source of Islamic law, yet Large Language Models applied to this domain suffer from factual hallucination and misattributed citations. This study develops a hadith question-answering system on Qwen 2.5 7B Instruct through a three-stage pipeline: (1) Supervised Fine tuning with QLoRA on 988 instruction-response pairs curated from 1,730 raw samples via Instruction Following Difficulty Scoring; (2) Direct Preference Optimization Iteration 1 using an off-policy strategy that induces behavioral regression due to distribution shift; and (3) Direct Preference Optimization Iteration 2 applying a fully on-policy strategy, generating chosen (T=0.1) and rejected (T=0.9) responses from the same model, yielding 509 validated pairs. The Hybrid Retrieval-Augmented Generation component indexes 65,811 hadith records from 11 books in Qdrant Cloud, combining BGE-M3 dense embeddings with BM25 sparse retrieval via Reciprocal Rank Fusion. Evaluation with RAGAS v0.2.6 (GPT-4o as judge) and BERTScore (xlm-roberta-base) shows Direct Preference Optimization Iteration 2 achieves higher Faithfulness (0.676 vs 0.633), equal Context Precision (1.000), and near-identical BERTScore F1 (0.8621 vs 0.8615). These results confirm that on-policy Direct Preference Optimization produces more stable behavioral alignment for domain-specific language models.

Keywords: BERTScore, DPO, Hybrid RAG, QLoRA, RAGAS

Abstrak

Hadits Islam merupakan sumber hukum kedua dalam ajaran Islam, namun penerapan Large Language Model (LLM) pada domain ini menghadapi masalah halusinasi faktual dan kesalahan atribusi sitasi. Penelitian ini mengembangkan sistem tanya jawab hadits berbasis Qwen 2.5 7B Instruct melalui tiga tahap. Pertama, Supervised Fine tuning (SFT) dengan Quantized Low-Rank Adaptation (QLoRA) pada 988 pasangan instruksi-respons yang dipilih dari 1.730 data mentah menggunakan Instruction Following Difficulty (IFD) Scoring berbasis rasio perplexity pada rentang P20–P80. Kedua, Direct Preference Optimization (DPO) Iterasi 1 dengan strategi off-policy menyebabkan regresi perilaku model akibat perbedaan distribusi data. Ketiga, DPO Iterasi 2 dengan strategi on-policy penuh menghasilkan respons chosen (T=0,1) dan rejected (T=0,9) dari model SFT yang sama, menghasilkan 509 pasangan valid. Komponen Hybrid Retrieval-Augmented Generation (RAG) mengindeks 65.811 hadits dari 11 kitab di Qdrant Cloud menggunakan BGE-M3 dan BM25 dengan Reciprocal Rank Fusion. Evaluasi RAGAS v0.2.6 dengan hakim GPT-4o dan BERTScore berbasis xlm-roberta-base menunjukkan DPO Iterasi 2 memperoleh Faithfulness lebih tinggi (0,676 vs 0,633), Context Precision sempurna (1,000) pada kedua model, dan BERTScore F1 yang setara (0,8621 vs 0,8615). Temuan ini mengonfirmasi bahwa strategi on-policy DPO menghasilkan keselarasan perilaku yang lebih stabil untuk model bahasa domain-spesifik.

Kata kunci: BERTScore, DPO, Hybrid RAG, QLoRA, RAGAS

1. Pendahuluan

Hadits merupakan sumber hukum Islam kedua setelah Al-Qur'an yang memuat seluruh perkataan, perbuatan, dan ketetapan Nabi Muhammad SAW yang diriwayatkan sebagai pedoman hukum dan etika umat Muslim [1]. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) melaporkan bahwa tingkat penetrasi internet di Indonesia telah mencapai 79,5% pada tahun 2024 dengan jumlah pengguna lebih dari 221 juta jiwa [2], sehingga mendorong pergeseran masif perilaku pencarian informasi keagamaan dari media cetak ke platform digital [3]. Pergeseran ini disertai risiko penyebaran informasi hadits yang tidak terverifikasi karena masyarakat awam masih bergantung pada mesin pencari umum yang menghasilkan teks tanpa sanad yang jelas.

Penerapan LLM (model bahasa berbasis kecerdasan buatan) pada domain hadits Islam menghadapi tiga tantangan utama. Pertama, fenomena halusinasi di mana model menghasilkan informasi yang terdengar meyakinkan namun tidak akurat termasuk nomor hadits yang tidak valid, yang berpotensi menimbulkan kekeliruan dalam pemahaman hukum Islam [4]. Kedua, keterbatasan pemahaman terminologi teknis seperti sanad (rantai perawi) dan matan (isi hadits) pada model bahasa umum yang belum disesuaikan untuk domain ini [5]. Ketiga, ketidakmampuan model bahasa umum untuk secara konsisten menolak pertanyaan di luar domain hadits sekaligus menyertakan sitasi yang

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

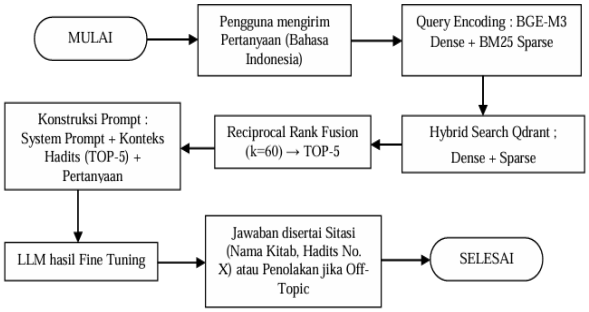
terverifikasi dalam setiap jawaban kondisi yang jika dibiarkan dapat menyebarkan informasi keagamaan yang tidak akurat kepada masyarakat awam. Penelitian sebelumnya yang mengandalkan antarmuka pemrograman aplikasi (API) LLM komersial tanpa fine tuning hanya mencapai akurasi 84,29% dan masih gagal menangani kueri yang kompleks [6].

Dua pendekatan yang terbukti efektif mengatasi keterbatasan ini adalah fine tuning domain dan Retrieval-Augmented Generation (RAG). Fine tuning menggeser distribusi probabilitas model agar sesuai dengan terminologi domain target [7], sementara RAG memastikan setiap jawaban berlandaskan dokumen referensi yang dapat diverifikasi melalui mekanisme retrieval dari basis pengetahuan eksternal [8]. Kombinasi kedua pendekatan menghasilkan sinergi yang lebih kuat karena model yang telah di-fine-tune lebih mampu memanfaatkan konteks dokumen Arab dari RAG dibandingkan model dasar yang belum mengenal struktur hadits [9].

Penelitian ini mempresentasikan pipeline terintegrasi yang menggabungkan fine tuning bertahap dengan Hybrid RAG untuk sistem tanya jawab hadits. Kontribusi utama penelitian ini adalah: (1) implementasi pipeline Supervised Fine Tuning (SFT) dengan Instruction Following Difficulty Scoring [10] sebagai mekanisme seleksi data otomatis; (2) analisis empiris komparatif off-policy Direct Preference Optimization versus on-policy Direct Preference Optimization yang mengungkap mekanisme regresi behavioral akibat distribution shift [11]; dan (3) kerangka evaluasi berlapis menggunakan RAGAS [12] dan BERTScore [13] untuk menilai kualitas sistem dari dua dimensi yang berbeda.

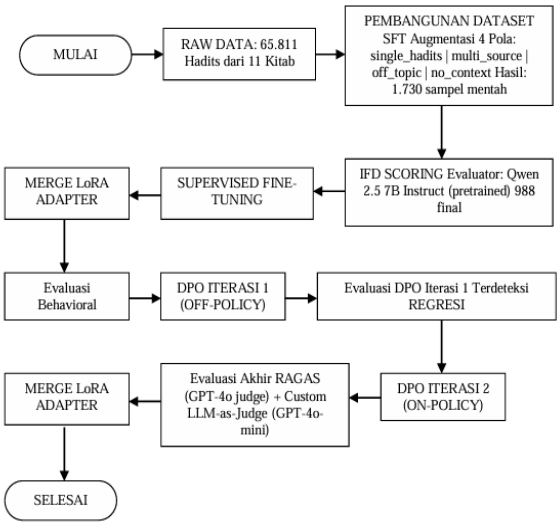
2. Metode Penelitian

Penelitian ini merupakan penelitian rancang bangun yang mengadopsi metode Prototyping. Metode ini dipilih karena pengembangan sistem berbasis Large Language Model (LLM) bersifat sangat eksploratif, di mana konfigurasi optimal hiperparameter, strategi pembangunan data, dan arsitektur pipeline tidak dapat ditentukan secara teoritis tanpa eksperimen berulang [14]. Seluruh proses pelatihan dilaksanakan pada platform Google Colaboratory dengan Graphics Processing Unit (GPU) NVIDIA A100 SXM4 berkapasitas Video Random Access Memory (VRAM) 80 GB. Inferensi model dilakukan secara lokal menggunakan Ollama sebagai engine serving model dalam format GGUF, karena pendekatan local deployment menjamin privasi data pengguna dan kontrol teknis penuh terhadap perilaku model, sekaligus menghindari perubahan yang tidak terdokumentasi akibat pembaruan layanan antarmuka pemrograman aplikasi (API) pihak ketiga [15]. Arsitektur sistem dan alur pelatihan dijelaskan pada Gambar 1 dan Gambar 2.



Gambar 1. Flowchart Sistem

Alur sistem dimulai saat pengguna mengirim pertanyaan Bahasa Indonesia yang dikodekan secara paralel oleh BGE-M3 menjadi *dense vector* 1.024 dimensi (semantik) dan *sparse vector* BM25 (leksikal). Keduanya digunakan dalam *Hybrid Search* pada Qdrant Cloud yang mengindeks 65.811 hadits dari 11 kitab melalui HNSW dan *inverted index* secara simultan. Hasil kedua jalur digabungkan via *Reciprocal Rank Fusion* (k=60) untuk menghasilkan TOP-5 hadits dengan relevansi tertinggi. Kelima hadits beserta metadatanya disusun bersama *system prompt* dan pertanyaan pengguna menjadi satu blok prompt yang dikirim ke model DPO2 (GGUF Q5_K_M via Ollama), yang kemudian menghasilkan jawaban bersitasi format (Nama Kitab, Hadits No. X) atau penolakan sopan apabila pertanyaan berada di luar domain hadits



Gambar 2. Flowchart Pipeline Pelatihan

Pipeline pelatihan dimulai dari 65.811 hadits mentah dari 11 kitab. Data kemudian dibangun menjadi dataset Supervised Fine tuning (SFT) melalui augmentasi empat pola interaksi yang menghasilkan 1.730 sampel mentah. Sampel disaring menggunakan IFD Scoring untuk menghasilkan dataset SFT final berkualitas tinggi. Setelah pelatihan SFT dengan QLoRA selesai dan adapter LoRA digabungkan dengan model dasar, dilakukan evaluasi behavioral untuk memverifikasi stabilitas distribusi model sebelum memasuki tahap

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

JUPTIK

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

Direct Preference Optimization (DPO). DPO Iterasi 1 dengan strategi off-policy dieksekusi, namun evaluasi behavioral pasca-training mendeteksi regresi, sehingga dilakukan DPO Iterasi 2 dengan strategi on-policy. Model akhir dievaluasi menggunakan RAGAS dan BERTScore.

2.1. Dataset Korpus Hadits dan Infrastruktur RAG

Basis pengetahuan sistem dibangun dari 65.811 riwayat hadits yang bersumber dari 11 kitab otoritatif yang disimpan dalam database MySQL terstruktur. Dataset ini mencakup seluruh Kutubut Tis'ah (sembilan kitab utama) yang diakui secara global oleh para sarjana hadits [16], ditambah Riyadhus Shalihin sebagai kitab panduan etika harian yang paling banyak dirujuk, dan Musnad Syafi'i sebagai kitab hadits hukum fiqih. Distribusi lengkap korpus disajikan pada Tabel 1.

Tabel 1. Distribusi Korpus Hadits dari 11 Kitab

No	Nama Kitab	Hadits
1	Shahih Bukhari	7.008
2	Shahih Muslim	5.362
3	Musnad Ahmad	26.363
4	Muwatho Malik	1.594
5	Musnad Syafi'i	1.800
6	Sunan Nasa'i	5.662
7	Sunan Abu Daud	4.590
8	Sunan Tirmidzi	3.891
9	Sunan Ibnu Majah	4.285
10	Sunan Darimi	3.367
11	Riyadhus Shalihin	1.889
	TOTAL	65.811

Seluruh 65.811 hadits diindeks ke Qdrant Cloud menggunakan named vectors ganda: dense vector berdimensi 1.024 dari model BGE-M3 [17] untuk pencarian semantik, dan sparse vector berbasis BM25 untuk pencocokan leksikal eksak. Qdrant dipilih karena kemampuan native hybrid search, dan efisiensi memori melalui kuantisasi vektor [18].

2.2. Pembangunan dan Seleksi Dataset SFT

Dataset SFT dibangun melalui pipeline otomatis yang mengambil hadits dari MySQL dan menghasilkan pasangan instruksi-respons dalam empat pola yang merepresentasikan variasi kondisi penggunaan sistem RAG di dunia nyata [4]: (1) single_hadits, yaitu satu hadits relevan dalam konteks dan model menjawab dengan sitasi wajib; (2) multi_source, yaitu beberapa hadits dalam konteks dengan jawaban enumeratif; (3) off_topic, yaitu pertanyaan di luar domain hadits dan model menolak dengan sopan; (4) no_context, yaitu pertanyaan valid namun konteks tidak relevan dan model mengakui keterbatasan tanpa mengarang sitasi. Variasi keempat pola ini penting karena model yang hanya dilatih pada satu kondisi akan gagal merespons kondisi lainnya secara tepat [19]. Dari pipeline ini dihasilkan 1.730 sampel mentah yang kemudian harus diseleksi kualitasnya.

Seleksi kualitas dilakukan menggunakan IFD Scoring [10]. Metode ini dipilih karena pendekatan kurasi manual pada ribuan sampel tidak praktis dan rentan terhadap inkonsistensi, sementara seleksi acak tidak mempertimbangkan nilai informatif setiap contoh bagi model. IFD Scoring bekerja secara otomatis dengan mengukur manfaat setiap contoh instruksi bagi proses pembelajaran model melalui rasio perplexity respons tanpa instruksi terhadap perplexity bersyarat dengan instruksi :

$$IFD(x,y) = \frac{PPL_{(y)}}{PPL_{(y|x)}} \quad (1)$$

Nilai IFD yang tinggi menandakan bahwa instruksi x memberikan informasi yang bermakna bagi model untuk menghasilkan respons y, sehingga contoh tersebut bernilai tinggi untuk pelatihan. Sebaliknya, nilai IFD yang sangat rendah menandakan contoh yang sudah mudah diprediksi model tanpa instruksi sehingga tidak memberikan sinyal pembelajaran yang berarti [10]. Li et al. [10] membuktikan bahwa seleksi berbasis IFD dengan 10% data terbaik menghasilkan performa yang setara atau lebih baik dibandingkan pelatihan dengan keseluruhan data. Strategi filter menggunakan rentang persentil ke-20 (P20) hingga persentil ke-80 (P80) dipilih untuk membuang sampel di kedua ujung ekstrem distribusi: sampel di bawah P20 terlalu trivial, sedangkan sampel di atas P80 terlalu sulit atau mengandung noise yang dapat merusak konvergensi pelatihan [10]. Evaluasi skor IFD dilakukan menggunakan model Qwen 2.5 7B Instruct pretrained sebagai evaluator melalui dua forward pass per sampel. Statistik IFD dan distribusi dataset final disajikan pada Tabel 2 dan Tabel 3.

Tabel 2. Statistik Skor IFD Dataset SFT

Statistik	Nilai
Total diproses	1.730 (valid: 1.533)
Tidak valid (inf/NaN)	197
IFD minimum / maksimum	1,5345 / 77,8747
IFD rata-rata (mean)	5,8851
IFD median	5,0567
Persentil ke-20 (P20)	3,7604
Persentil ke-80 (P80)	7,4959
Lolos filter P20–P80	919
Dataset SFT final (setelah QC)	988 contoh

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

Proses seleksi dataset SFT menggunakan IFD Scoring diawali dengan 1.730 sampel mentah, namun hanya 1.533 yang menghasilkan nilai IFD valid sebanyak 197 sampel menghasilkan nilai tak terhingga (inf) atau tidak terdefinisi (NaN) akibat kegagalan komputasi perplexity dan langsung dieliminasi. Dari 1.533 sampel valid, distribusi skor IFD menunjukkan rentang yang sangat lebar (1,5345 hingga 77,8747) dengan rata-rata 5,8851 dan median 5,0567, mengindikasikan adanya sampel trivial di ujung bawah dan sampel mengandung noise di ujung atas. Oleh karena itu, diterapkan filter persentil P20–P80 yang mempertahankan sampel dengan IFD di antara 3,7604 dan 7,4959, menghasilkan 919 sampel. Angka ini kemudian bertambah menjadi 988 setelah melalui *quality check* berbasis aturan yang mencakup standarisasi *system prompt* pada 88 sampel yang sebelumnya menggunakan versi *system prompt* yang lebih pendek (1.392 karakter) sehingga distandarisasi ke versi kanonik 2.350 karakter, dan sampel-sampel tersebut dinyatakan valid untuk dimasukkan kembali ke dataset final.

Tabel 3. Distribusi Dataset SFT Final

Pola	Jumlah	Proporsi
single_hadits	705	78,3%
off_topic	109	12,1%
no_context	76	8,4%
multi_source	10	1,1%
Total	988	100%

Distribusi yang didominasi pola single_hadits (78,3%) mencerminkan kondisi penggunaan sistem RAG yang paling umum, selaras dengan temuan Zhou et al. [18] bahwa kualitas pola dominan lebih menentukan performa model dibandingkan keseimbangan distribusi yang dipaksakan secara artifisial. Pola off_topic (12,1%) dan no_context (8,4%) dipertahankan dalam proporsi yang memadai karena kemampuan menolak pertanyaan di luar domain merupakan aspek kritis keselamatan sistem LLM: model yang tidak dilatih secara eksplisit untuk menolak akan cenderung menjawab pertanyaan apapun meskipun di luar kapasitasnya, yang pada domain hadits dapat mengakibatkan penyebaran informasi keagamaan yang tidak akurat [19].

2.3. Supervised Fine tuning (SFT) dengan QLoRA

SFT merupakan tahap yang diperlukan sebelum alignment karena model bahasa besar yang hanya menjalani pre-training umum belum mampu mengikuti instruksi secara konsisten, apalagi dalam domain spesifik seperti hadits Islam yang memiliki terminologi teknis tersendiri [7]. Tanpa SFT, model cenderung merespons dengan gaya bahasa bebas tanpa memperhatikan format sitasi, batasan domain, maupun struktur respons yang diinginkan. Dalam konteks penelitian ini, SFT digunakan untuk menanamkan tiga perilaku kritis yang tidak dimiliki model dasar: (1) penyertaan sitasi hadits dalam format baku, (2) penolakan pertanyaan di luar domain, dan (3) penyajian isi hadits tanpa menyimpulkan hukum fiqih secara mandiri.

SFT dengan QLoRA dipilih atas dua pertimbangan utama. Pertama, efisiensi memori: fine tuning penuh pada model 7 miliar parameter membutuhkan memori GPU lebih dari 100 GB yang tidak tersedia pada infrastruktur penelitian ini. QLoRA mengatasi hal ini dengan cara membekukan bobot model asli dan hanya melatih sepasang matriks kecil (adapter) yang disisipkan ke setiap lapisan transformer, sehingga kebutuhan memori turun drastis menjadi sekitar 15 GB tanpa mengorbankan performa secara signifikan [20]. Kedua, pendekatan ini terbukti mempertahankan pengetahuan umum model pretrained sekaligus mengadaptasikannya pada domain spesifik [7],

Adapter LoRA [21] ditambahkan pada tujuh modul proyeksi model Qwen 2.5 7B dengan rank r=64 dan alpha α=128. Nilai rank yang relatif besar dipilih karena adaptasi domain hadits memerlukan kapasitas representasi lebih tinggi dibandingkan tugas adaptasi gaya bahasa sederhana [7]. Pelatihan dilakukan selama 3 epoch dengan effective batch size 16 dan learning rate 2×10^{-4}. Setelah selesai, adapter digabungkan dengan model dasar menghasilkan checkpoint sft_merged, kemudian dikonversi ke format GGUF untuk dijalankan secara lokal melalui Ollama dengan kuantisasi Q5_K_M yang memberikan keseimbangan terbaik antara kualitas output dan kebutuhan memori [20].

2.4. Evaluasi Behavioral Sebelum DPO

Sebelum memasuki tahap DPO, evaluasi behavioral terhadap model SFT dilakukan sebagai prasyarat yang tidak dapat dilewati. Alasan utamanya adalah bahwa dalam kerangka DPO, model SFT berfungsi sebagai model referensi (πref) yang menentukan titik awal dan batas ruang optimasi [11]. Rafailov et al. [11] menegaskan bahwa kualitas model referensi secara langsung menentukan efektivitas DPO: model referensi yang belum stabil secara behavioral akan menghasilkan distribusi referensi yang tidak bermakna, sehingga gradien DPO tidak dapat bekerja dengan arah yang tepat.

Evaluasi behavioral mencakup pengujian manual terhadap tiga aspek yang menjadi target alignment: (1) konsistensi penyertaan sitasi dalam format yang benar untuk pertanyaan on-topic; (2) kemampuan menolak pertanyaan di luar domain hadits tanpa memberikan jawaban yang dikarang; dan (3) kehati-hatian dalam merespons pertanyaan hukum fiqih tanpa menyimpulkan hukum secara mandiri. Hanya setelah model SFT dinyatakan stabil pada ketiga aspek tersebut, proses pembangunan dataset DPO dimulai

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

menggunakan model SFT sebagai generator respons chosen dan rejected, karena data preferensi yang valid hanya dapat dihasilkan dari model yang distribusinya sudah stabil dan representatif [22].

2.5. Direct Preference Optimization (DPO) Iterasi 1: Strategi Off-Policy dan Alasan Perubahan Strategi

DPO dipilih sebagai metode alignment karena bekerja dengan langsung mengajari model mana respons yang disukai (chosen) dan mana yang tidak disukai (rejected), tanpa memerlukan sistem penilaian terpisah yang mahal secara komputasi seperti pada pendekatan Reinforcement Learning From Human Feedback (RLHF) konvensional [11]. Loss function DPO dirumuskan sebagai :

$$\mathcal{L}DPO = -\log\sigma\left(\beta\left[log\frac{\pi_\theta(ychosen|x)}{\pi_{ref}(ychosen|x} - log\frac{\pi_\theta(yrejected|x)}{\pi_{ref}(yrejected|x}\right]\right) \quad (2)$$

di mana yc adalah respons chosen (disukai), yr adalah respons rejected (tidak disukai), β adalah hiperparameter yang mengontrol seberapa jauh model diizinkan menyimpang dari model referensi πref, dan σ adalah fungsi sigmoid [11].

Dataset DPO Iterasi 1 terdiri dari 496 pasangan valid yang tersebar dalam lima kategori target. Kategori terbesar adalah target1_citation dengan 178 pasangan (35,9%) yang bertujuan memperkuat konsistensi format sitasi hadits. Kategori target2_offtopic mencakup 100 pasangan (20,2%) untuk melatih penolakan pertanyaan di luar domain, diikuti normal_single sebanyak 98 pasangan (19,8%) sebagai reinforcement respons hadits tunggal, target3_noanalogy 80 pasangan (16,1%) untuk menghilangkan analogi modern yang tidak relevan, dan normal_multi 40 pasangan (8,1%) untuk respons multi-hadits. Respons chosen untuk kategori off-topic dan no-analogy menggunakan template penolakan standar, sementara kategori citation dan normal menggunakan respons yang dihasilkan dari instruksi GPT-4o via API OpenRouter. Namun, strategi ini mengandung risiko inheren berupa distribution shift: ketika data preferensi dihasilkan oleh model dengan distribusi berbeda dari model referensi (sft_merged), gradien DPO dapat mendorong model keluar dari distribusi yang telah dibangun selama SFT, sehingga merusak kemampuan yang sudah ada [11]. Risiko ini telah diidentifikasi secara teoritis oleh Rafailov et al. [11], dan Guo et al. [22] mengonfirmasi secara empiris bahwa off-policy DPO secara konsisten menghasilkan alignment yang tidak stabil dibandingkan on-policy DPO. Berdasarkan itulah, evaluasi behavioral pasca-training DPO Iterasi 1 dirancang sebagai mekanisme deteksi dini terhadap regresi, sebelum model digunakan atau dikembangkan lebih lanjut.

2.6. DPO Iterasi 2: Strategi On-Policy

DPO Iterasi 2 dirancang menggunakan strategi on-policy penuh untuk mengatasi akar masalah distribution shift yang teridentifikasi pada DPO Iterasi 1. Pada strategi on-policy, respons chosen (yc) dan respons rejected (yr) keduanya dihasilkan oleh model yang sama yang akan dioptimasi, yaitu model SFT [22]. Dengan cara ini, distribusi data preferensi konsisten dengan distribusi model referensi, sehingga gradien DPO bekerja dalam ruang distribusi yang sama dan tidak mendorong model melampaui batas kemampuan yang sudah dibangun. Chen et al. [23] secara empiris membuktikan bahwa pendekatan on-policy menghasilkan alignment yang jauh lebih stabil dibandingkan off-policy pada model bahasa domain-spesifik.

Respons chosen dihasilkan dengan temperature T=0,1 untuk mendapatkan respons yang mendekati deterministik dan berkualitas tinggi. Temperature rendah dipilih karena menghasilkan respons dengan sitasi yang konsisten dan penolakan yang tepat, yang menjadi target perilaku system [23]. Respons rejected dihasilkan dengan temperature T=0,9 dikombinasikan dengan instruksi yang sengaja memancing perilaku tidak diinginkan, mengikuti pendekatan self-play yang menghasilkan sinyal preferensi lebih informatif dibandingkan template manual [22]. Model dasar yang digunakan adalah sft_merged (bukan checkpoint DPO Iterasi 1) untuk menghindari akumulasi regresi. Nilai β=0,2 dipilih lebih tinggi dari DPO Iterasi 1 (β=0,1) agar model lebih konservatif terhadap model referensi dan regresi lebih terkendali [11]. Pipeline menghasilkan 509 pasangan valid yang diprioritaskan berdasarkan kebutuhan: target2_offtopic 144 pasangan (28,3%) sebagai prioritas utama pemulihan penolakan, target1_citation 142 pasangan (27,9%), target3_noanalogy 93 pasangan (18,3%), normal_single 93 pasangan (18,3%), dan normal_multi 37 pasangan (7,3%).

2.7. Arsitektur Hybrid Retrieval Augmented Generation

Sistem RAG mengadopsi arsitektur Advanced RAG [8] dengan Hybrid Retrieval yang menggabungkan dense retrieval berbasis BGE-M3 [17] dan sparse retrieval BM25. BGE-M3 dipilih karena kemampuan multi-functionality dalam satu model yang mendukung lebih dari 100 bahasa termasuk Arab dan Indonesia [17]. Dense retrieval unggul pada pencarian semantik berdasarkan tema hadits, sedangkan BM25 unggul pada pencarian eksak seperti nama perawi spesifik atau nomor hadits tertentu. Penggabungan kedua metode dilakukan melalui Reciprocal Rank Fusion (RRF) [24]:

$$RRF(d) = \sum_{i=1}^{m}\left[\frac{1}{k + rank_i(d)}\right] \quad (3)$$

dengan k=60 sebagai konstanta default yang mencegah dokumen berperingkat tinggi mendominasi dan TOP-K ditetapkan pada lima hadits per kueri. System prompt

(2.350 karakter) berfungsi sebagai lapisan kendali yang menetapkan: (1) identitas sebagai asisten referensi hadits bukan mufti; (2) kewajiban sitasi dalam format (Nama Kitab, Hadits No. X) hanya dari konteks yang diberikan; (3) larangan keras mengarang nomor hadits; dan (4) penolakan eksplisit jika konteks tidak relevan.

2.8. Konfigurasi Evaluasi

Evaluasi dilakukan menggunakan dua framework komplementer pada 19 pertanyaan uji yang mencakup dua kategori: *hadits* (15 pertanyaan langsung) dan *parafrase* (4 pertanyaan berformat ulang untuk menguji robustness). Kedua model Base dan DPO2 menerima input identik berupa system prompt, konteks retrieval TOP-5 dari Qdrant Cloud, dan pertanyaan yang sama, sehingga selisih hasil sepenuhnya mencerminkan dampak pipeline fine tuning.

Framework pertama adalah RAGAS v0.2.6 [12] dengan model hakim GPT-4o via OpenRouter dan embedding *text-embedding-3-small*, yang mengukur empat dimensi kualitas pipeline RAG secara otomatis: Faithfulness, Answer Relevancy, Context Precision, dan Context Recall. *Reference string* (ground truth) untuk RAGAS dihasilkan oleh GPT-4o berdasarkan aturan perilaku model.

Framework kedua adalah BERTScore [13] F1 menggunakan model *xlm-roberta-base*, yang mengukur kemiripan semantik antara jawaban model dan *gold answer* berbasis referensi. Gold answer disusun secara manual dan divalidasi oleh validator ahli di bidang studi hadits dan fiqih, untuk memastikan akurasi konten hadits. Tidak seperti RAGAS yang menggunakan hakim LLM, BERTScore menghasilkan skor deterministik berbasis representasi token yang tidak terpengaruh oleh bias hakim terhadap format jawaban tertentu.

3. Hasil dan Pembahasan

Evaluasi dilakukan secara komparatif antara model DPO Iterasi 2 (selanjutnya DPO2) dan model Qwen 2.5 7B Instruct original (selanjutnya Base) menggunakan 19 pertanyaan uji yang mencakup dua kategori: hadits (H01-H15) dan parafrase (P01, P02, P04, P05). Kedua model menerima konteks retrieval yang identik sehingga perbedaan hasil semata-mata mencerminkan dampak pipeline fine tuning.

3.1. Hasil IFD Scoring dan Pembentukan Dataset SFT

Proses IFD Scoring berhasil mereduksi 1.730 sampel mentah menjadi 988 contoh berkualitas tinggi. Dari total 1.730 sampel, sebanyak 197 sampel (11,4%) menghasilkan nilai skor tidak valid akibat kegagalan komputasi perplexity dan langsung dieliminasi. Dari 1.533 sampel valid yang tersisa, filter persentil P20-P80 menghasilkan 919 sampel. Angka ini kemudian bertambah menjadi 988 setelah proses quality check dan standarisasi system prompt. Rentang skor IFD yang sangat lebar antara 1,53 hingga 77,87 menunjukkan variasi tingkat kesulitan yang signifikan pada data mentah, sehingga seleksi berbasis persentil sangat diperlukan untuk membuang sampel yang terlalu trivial maupun yang terlalu sulit bagi model evaluator.

Distribusi akhir dataset yang didominasi pola single_hadits (78,3%) mencerminkan kondisi penggunaan sistem RAG paling umum. Pola off_topic (12,1%) dan no_context (8,4%) yang dipertahankan dalam proporsi yang cukup terbukti efektif menanamkan kemampuan penolakan pertanyaan di luar domain pada tahap SFT.

3.2. Hasil Supervised Fine tuning

Tahap SFT menghasilkan pergeseran behavioral yang paling signifikan dibandingkan tahap-tahap berikutnya. Model SFT menunjukkan peningkatan Behavior Accuracy dari 0,933 (Base) menjadi 1,000 dan peningkatan kemampuan penolakan pertanyaan di luar domain yang dramatis dari level baseline 0,500 menjadi 1,000. Peningkatan ini mengonfirmasi bahwa dataset SFT yang diseleksi melalui IFD Scoring efektif menanamkan pola perilaku yang diinginkan, khususnya kemampuan menolak pertanyaan di luar domain hadits yang tidak dimiliki model dasar.

Keberhasilan tahap SFT juga tercermin dari peningkatan kemampuan penilaian hukum fiqih (LLM-as-Judge Hukum) dari 0,650 pada model Base menjadi 0,750 pada model SFT. Model SFT berhasil memahami bahwa pertanyaan tentang hukum fiqih harus dijawab dengan menyajikan isi hadits yang relevan tanpa berani menyimpulkan hukum secara mandiri.

3.3. Direct Preference Optimization Iterasi 1 : Regresi Off-Policy

DPO Iterasi 1 dengan strategi off-policy mengakibatkan regresi behavioral yang signifikan. Kemampuan penolakan pertanyaan di luar domain yang telah dibangun pada tahap SFT (1,000) mengalami penurunan kembali ke level baseline. Fenomena ini sesuai dengan prediksi teoritis Rafailov et al. [11] bahwa DPO sangat sensitif terhadap ketidaksesuaian distribusi antara model referensi dan sumber data preferensi. Ketika respons chosen dihasilkan oleh GPT-4o dengan distribusi yang berbeda dari model SFT, gradien DPO mendorong model keluar dari distribusi yang telah dibangun selama SFT. Guo et al. [23] mengonfirmasi secara empiris bahwa off-policy DPO secara konsisten menghasilkan alignment yang tidak stabil untuk model bahasa domain-spesifik.

Secara sederhana Pada tahap ini, sistem "mengajari" model menggunakan contoh jawaban ideal yang dibuat oleh AI lain (GPT-4o), bukan oleh model itu sendiri. Ibarat melatih seorang koki dengan resep dari koki berbeda yang memiliki gaya masak yang sangat berbeda alih-alih semakin mahir, sang koki justru

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

JUPTIK

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

menjadi bingung dan melupakan sebagian keahlian yang sudah ia kuasai sebelumnya.Temuan ini menjadi dasar keputusan untuk mengubah strategi pada DPO Iterasi 2.

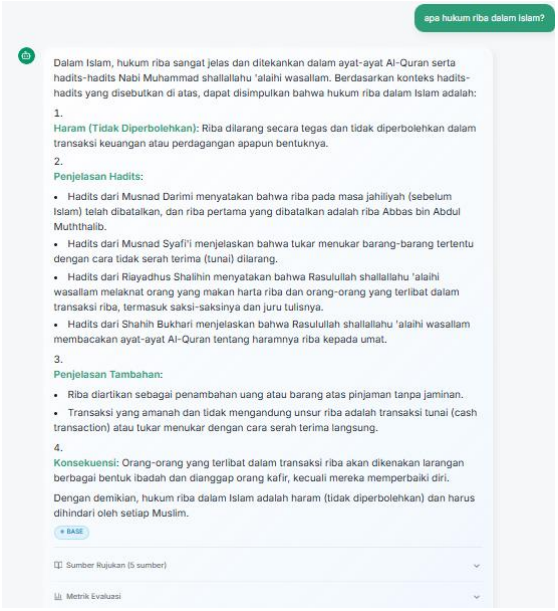
3.4. Direct Preference Optimization Iterasi 2: Pemulihan On-Policy

DPO Iterasi 2 dengan strategi on-policy berhasil memulihkan stabilitas alignment. Dengan menggunakan respons chosen dan rejected yang keduanya dihasilkan oleh model SFT sendiri pada temperature berbeda, konsistensi distribusi antara data preferensi dan model yang dioptimasi terjaga. Kemampuan penilaian hukum fiqih meningkat secara signifikan dari 0,750 (SFT) menjadi 1,000 pada DPO2, yang berarti model berhasil menyajikan isi hadits untuk pertanyaan hukum tanpa berani menyimpulkan hukum secara mandiri.

Secara sederhana berbeda dari iterasi sebelumnya, kali ini model dilatih menggunakan contoh jawaban yang dihasilkan oleh dirinya sendiri satu versi jawaban yang hati-hati dan berkualitas tinggi (chosen) serta satu versi jawaban yang sengaja dibuat kurang tepat (rejected). Dengan cara ini, model belajar membedakan mana yang baik dan mana yang buruk dari pengalamannya sendiri, bukan dari standar orang lain yang belum tentu sesuai. Hasilnya, kemampuan yang sempat hilang pada iterasi 1 berhasil dipulihkan bahkan ditingkatkan.

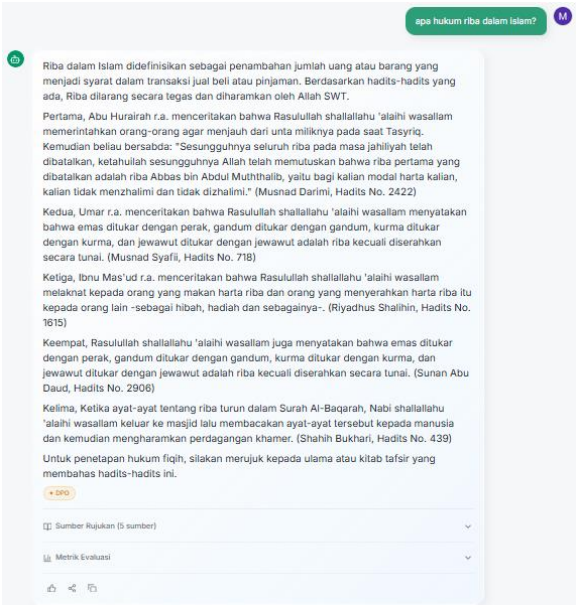
3.5 Contoh Perbandingan Kualitatif Jawaban Base vs DPO iterasi 2.

Untuk memperjelas perbedaan dampak fine tuning secara kualitatif, berikut disajikan perbandingan jawaban kedua model terhadap pertanyaan yang sama menggunakan konteks retrieval yang identik. Jawaban Model Base :



Gambar 3. Jawaban Model Base (sebelum fine tuning)

Jawaban Model DPO iterasi 2:



Gambar 4. Jawaban DPO iterasi 2 (sesudah fine tuning)

Analisis Perbandingan :

Dari perbandingan di atas terlihat empat perbedaan utama yang langsung mencerminkan dampak pipeline fine tuning :

Pertama, Kehati-hatian dalam penetapan hukum fiqih. Model Base menyimpulkan hukum secara mandiri di poin ke-4 dengan menyebutkan konsekuensi berupa "larangan berbagai bentuk ibadah dan dianggap orang kafir" sebuah penyimpulan yang tidak ditemukan dalam hadits mana pun yang ada dalam konteks retrieval dan berpotensi sangat menyesatkan. Model DPO2 justru mengakhiri jawabannya dengan mengarahkan pengguna kepada ulama yang berwenang, sesuai dengan target perilaku sebagai asisten referensi, bukan sebagai mufti.

Kedua, Kesetiaan terhadap konteks (Faithfulness). Model DPO2 secara konsisten menyajikan isi hadits berdasarkan konteks yang diberikan, dengan mencantumkan nama kitab dan nomor hadits untuk setiap pernyataan (misalnya, *Musnad Darimi, Hadits No. 2422*). Model Base menyusun jawaban dalam poin-poin tematik yang menggabungkan isi hadits dengan penyimpulan tambahan yang tidak berdasar pada konteks yang diberikan. Perbedaan inilah yang secara langsung menjelaskan selisih skor Faithfulness antara kedua model dalam evaluasi RAGAS (0,676 vs 0,633).

Ketiga, Halusinasi faktual. Pernyataan pada poin ke-4 jawaban Base bahwa pelaku riba "dianggap orang kafir" adalah contoh nyata halusinasi faktual dalam domain keislaman model menghasilkan pernyataan hukum yang terdengar otoritatif namun tidak ditemukan dalam hadits mana pun yang ada dalam konteks, dan secara substansi merupakan generalisasi

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

yang tidak tepat. Model DPO2 tidak menghasilkan klaim semacam ini.

Keempat, Format sitasi. Model DPO2 secara konsisten menuliskan sumber dalam format *(Nama Kitab, Hadits No. X)* untuk setiap hadits yang disebutkan. Model Base hanya menyebutkan nama kitab secara naratif tanpa nomor hadits yang spesifik di beberapa bagian. Konsistensi format sitasi ini adalah salah satu perilaku kritis yang berhasil ditanamkan melalui tahap SFT dan diperkuat melalui DPO Iterasi 2.

3.6. Golden Answer

Evaluasi kuantitatif pada penelitian ini menggunakan 19 pertanyaan uji yang terdiri atas dua kategori, yaitu pertanyaan hadits (H01–H15) dan pertanyaan parafrase (P01, P02, P04, P05). Setiap pertanyaan dilengkapi dengan golden answer atau jawaban acuan beserta daftar sitasi hadits yang diharapkan, yang telah divalidasi oleh ahli hadits. Berikut adalah Tabel dari golden answer untuk 19 pertanyaan uji tersebut :

Tabel 4. Golden Answer

ID	Pertanyaan Uji	Sitasi yang Diharapkan
H01	Apa keutamaan shalat berjamaah?	RS 1064, RS 1065, RS 1069, RS 1070
H02	Bagaimana adab makan yang benar menurut hadits?	RS 750, RS 751
H03	Apa keutamaan menuntut ilmu dalam Islam?	RS 476, RS 1387, RS 1388
H04	Apa balasan bagi orang yang berbakti kepada kedua orang tua?	RS 312, RS 317, RS 321, SM 4624
H05	Bagaimana cara menjaga lisan agar tidak terjerumus dosa?	RS 1520, RS 1522
H06	Bagaimana sikap seorang muslim ketika ditimpa musibah?	RS 37, SIM 1588
H07	Apa adab seorang muslim ketika bertamu dan menerima tamu?	RS 706, MA 16549, SM 3255, SAD 3256, SAD 3258
H08	Apa hukum minuman yang memabukkan menurut hadits?	MA 14351, SN 5512, SN 5592, SAD 3195
H09	Apa keutamaan bersedekah dalam Islam?	MA 7100, SIM 239
H10	Mengapa niat itu penting dalam setiap amal ibadah?	SB 4682, RS 7
H11	Apa keutamaan menyambung tali silaturahmi?	SM 4630, ST 1825, MA 23200
H12	Apa keutamaan berwudhu dengan sempurna sebelum shalat?	RS 1024, RS 1026, RS 1027, RS 1028
H13	Apa keutamaan mengerjakan shalat malam atau tahajud?	RS 1167, RS 1175, SIM 1077
H14	Bagaimana hukum berbohong dan dampaknya menurut hadits?	SAD 4275,SD 2599, MA 3899
H15	Bagaimana anjuran berbuat baik kepada tetangga?	RS 304, RS 311, MA 23126, MA 9999
P01	Bagaimana keutamaan mengerjakan shalat secara berjamaah di masjid?	RS 1064, RS 1065, RS 1071
P02	Mengapa kita diperintahkan untuk berbakti kepada orang tua?	RS 312, RS 317, SM 4624, SIM 3651
P04	Bagaimana pandangan Islam tentang menggambar makhluk bernyawa?	SM 3945, RS 1681
P05	Apa hukum mendengarkan lagu-lagu zaman sekarang?	MA 23879, SM 1479, SIM 1888

3.6. Hasil Evaluasi RAGAS

Evaluasi RAGAS menggunakan 19 pertanyaan uji dengan pipeline hybrid retrieval dari Qdrant Cloud. Hakim GPT-4o menilai empat dimensi kualitas system RAG secara otomatis [12]. Hasil perbandingan disajikan pada Tabel 5.

Tabel 5. Hasil Evaluasi RAGAS: Base vs DPO2

Metrik RAGAS	Base	DPO2	Δ	Unggul
Ragas - Faithfulness	0,633	0,676	+0,043	DPO
Ragas - Context Precision	1,000	1,000	0,000	Seimbang
Ragas - Context Recall	0,783	0,781	-0,002	Base
Ragas - Answer Relevancy	0,720	0,726	+0,006	DPO

DPO2 menunjukkan keunggulan pada Faithfulness (+0,043), yang merupakan metrik terpenting untuk sistem tanya jawab hadits karena mengukur seberapa setia jawaban model terhadap konteks hadits yang diberikan [12]. Nilai Context Precision yang sempurna (1,000) pada kedua model mengonfirmasi bahwa sistem hybrid retrieval BGE-M3 dengan BM25 dan RRF bekerja dengan sangat baik dalam menempatkan dokumen hadits paling relevan di posisi teratas. Context Recall yang hampir identik (0,783 vs 0,781) menunjukkan bahwa pipeline fine tuning tidak merusak kemampuan model dalam mengekstraksi informasi dari konteks retrieval.

Peningkatan Faithfulness pada DPO2 secara konsisten terlihat pada beberapa pertanyaan spesifik. Pertanyaan H01 (keutamaan shalat berjamaah) menunjukkan peningkatan dari 0,714 pada Base menjadi 1,000 pada DPO2. Pertanyaan H04 (berbakti kepada orang tua) menunjukkan peningkatan paling drastis dari 0,000 pada Base menjadi 0,636 pada DPO2, yang berarti

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

model Base sama sekali tidak menggunakan konteks retrieval untuk pertanyaan tersebut. Pertanyaan H14 (hukum berbohong) meningkat dari 0,455 menjadi 1,000 dan pertanyaan P02 (berbakti kepada orang tua, parafrase) meningkat dari 0,375 menjadi 0,667.

Nilai Answer Relevancy yang seimbang (0,720 vs 0,726) menunjukkan bahwa kedua model menjawab pertanyaan dengan tingkat relevansi yang setara. Perlu dicatat bahwa metrik ini diukur melalui rekayasa balik pertanyaan dari jawaban menggunakan kemiripan semantik [12], sehingga respons penolakan yang tepat pada pertanyaan di luar domain dapat menghasilkan skor yang lebih rendah secara mekanis meskipun perilaku penolakan tersebut adalah respons yang benar.

3.6. Hasil Evaluasi BERTScore

BERTScore [13] mengukur kemiripan semantik jawaban model terhadap gold answer yang telah divalidasi ahli menggunakan model xlm-roberta-base. Hasil perbandingan disajikan pada Tabel 6.

Tabel 6. Hasil BERTScore F1: Base vs DPO2 vs Gold Answer

BertScore	Base	DPO2	Δ	Unggul
Precision	0,8446	0,8463	+0,0017	DPO
Recall	0,8807	0,8774	0,0033	Base
F1	0,8621	0,8615	-0,0006	Seimbang

Hasil BERTScore yang hampir identik pada ketiga komponen mengonfirmasi dua temuan penting. Pertama, proses fine tuning bertahap tidak menyebabkan catastrophic forgetting pada model DPO2 tetap menghasilkan konten yang secara semantik setara dengan jawaban ideal manusia meskipun telah melewati SFT dan dua iterasi DPO. Kedua, analisis terhadap komponen Precision dan Recall mengungkapkan pola yang informatif: DPO2 memperoleh Precision rata-rata lebih tinggi (0,8463 vs 0,8446), mengindikasikan bahwa token-token dalam jawabannya lebih presisi secara semantik terhadap gold answer konsisten dengan efek DPO yang menekan elaborasi di luar konteks. Sebaliknya, Base memperoleh Recall rata-rata lebih tinggi (0,8807 vs 0,8774), yang dapat dijelaskan oleh kecenderungan model dasar menghasilkan jawaban yang lebih panjang sehingga lebih banyak elemen gold answer yang tercakup. Keseimbangan antara Precision dan Recall menghasilkan F1 yang praktis identik (selisih 0,0006), membuktikan bahwa perbedaan Faithfulness antara kedua model dalam evaluasi RAGAS (0,633 vs 0,676) bukan disebabkan oleh perbedaan kualitas konten secara keseluruhan, melainkan oleh perbedaan seberapa ketat model mengikuti konteks retrieval yang diberikan.

3.7. Ringkasan Pergeseran Pipeline

Pergeseran metrik pada setiap tahap pipeline memberikan gambaran menyeluruh tentang dampak masing-masing tahap fine tuning. Pada tahap SFT, terjadi peningkatan paling konsisten dan signifikan: Behavior Accuracy naik dari 0,933 menjadi 1,000, kemampuan penolakan off-topic melonjak dari 0,500 menjadi 1,000, dan penilaian hukum fiqih meningkat dari 0,650 menjadi 0,750. Tahap ini mengonfirmasi bahwa seleksi data berbasis IFD Scoring efektif menanamkan pola perilaku yang diinginkan.

Pada tahap DPO Iterasi 1 dengan strategi off-policy, terjadi regresi behavioral yang terdokumentasi, khususnya pada kemampuan penolakan off-topic yang kembali ke level baseline. Ini merupakan bukti empiris distribution shift yang diprediksi oleh Rafailov et al. [11]. Pada tahap DPO Iterasi 2 dengan strategi on-policy, metrik kritis dipulihkan: Context Precision dan Context Recall bertahan pada level SFT, penilaian hukum fiqih mencapai nilai tertinggi (1,000), dan Faithfulness meningkat menjadi 0,676 melampaui model Base (0,633). BERTScore F1 yang hampir setara (0,8621 vs 0,8615) pada kedua model mengonfirmasi tidak ada degradasi kualitas konten akibat proses alignment.

3.8. Analisis Signifikansi dan Perbandingan dengan Penilitian Terhadulu

Peningkatan Faithfulness sebesar 0,043 (dari 0,633 menjadi 0,676) yang dicapai DPO2 perlu diinterpretasikan dalam konteks signifikansinya, bukan sekadar selisih angka absolut. Dalam domain hadits Islam, setiap peningkatan Faithfulness memiliki implikasi yang jauh melampaui metrik teknis semata. Faithfulness mengukur seberapa ketat jawaban model berlandaskan pada konteks retrieval yang diberikan; dengan kata lain, peningkatan sebesar 6,8% ini berarti model DPO2 secara statistis lebih jarang menghasilkan pernyataan hukum atau konten hadits yang tidak memiliki pijakan dalam sumber referensi yang terverifikasi. Dalam konteks penyebaran informasi keagamaan, perbedaan ini bersifat kritis: satu pernyataan yang dikarang model dapat menyesatkan pembaca awam mengenai hukum Islam yang bersumber dari hadits yang tidak valid. Hasil kualitatif pada Subbab 3.5 memperlihatkan secara konkret bagaimana model Base menghasilkan klaim hukum fiqih yang tidak ditemukan dalam hadits manapun di dalam konteks retrieval, sementara DPO2 secara konsisten menolak membuat simpulan hukum mandiri. Peningkatan Faithfulness yang terukur ini, oleh karenanya, merupakan cerminan langsung dari berkurangnya risiko halusinasi faktual pada domain yang sensitif.

Dari sudut pandang akademik, temuan penelitian ini memberikan kontribusi empiris yang relevan pada tiga level. Pertama, pada level metodologi seleksi data, hasil IFD Scoring yang berhasil mengompresi 1.730 sampel mentah menjadi 988 sampel berkualitas tinggi dengan tetap mencapai Behavior Accuracy 100%

Jurnal Pengembangan Teknologi Informasi dan Komunikasi

2026 Vol. 4 No.1 Hal: 160-170 e-ISSN: 2987-7016

mengonfirmasi secara empiris klaim Li et al. [10] bahwa seleksi berbasis perplexity ratio mampu menggantikan keseluruhan dataset tanpa penurunan performa bermakna, namun kini dalam konteks domain yang lebih spesifik dan sensitif dibandingkan eksperimen orisinal mereka. Kedua, pada level strategi alignment, penemuan bahwa off-policy DPO menyebabkan regresi behavioral yang dapat terdeteksi (Subbab 3.3) memberikan bukti empiris tambahan bagi prediksi teoritis Rafailov et al. [11] dan konfirmasi eksperimental Guo et al. [23] dalam skenario model bahasa domain-spesifik berbahasa Indonesia/Arab, yang belum pernah dieksplorasi dalam penelitian sebelumnya. Ketiga, pada level arsitektur sistem, pencapaian Context Precision sempurna (1,000) pada kedua model menunjukkan bahwa kombinasi BGE-M3 dengan BM25 melalui RRF mampu memberikan presisi retrieval yang sangat tinggi untuk teks Arab-Indonesia, memperkuat temuan Chen et al. [17] tentang kemampuan multilingual BGE-M3 pada bahasa-bahasa di luar dataset evaluasi orisinalnya.

Untuk memposisikan kontribusi penelitian ini secara lebih jelas, Tabel 7 menyajikan perbandingan sistematis dengan penelitian-penelitian terdahulu yang memiliki topik serupa.

Tabel 7. Perbandingan dengan Penelitian Terdahulu

Penelitian	Domain	Evaluasi	Capaian / Keterbatasan
Sutiyo [6] (2024)	Tanya jawab Tafsir Al-Jalalain (bahasa Indonesia)	Akurasi manual 84,29%	Bergantung API pihak ketiga; kueri kompleks gagal; tidak mengukur Faithfulness
Herwanza et al. [9] (2024)	Chatbot hadits via Telegram (9 kitab hadits)	Evaluasi subjektif pengguna; tidak ada metrik RAGAS/BERTScore	Tidak mengatasi halusinasi secara sistematis; tidak ada alignment khusus domain; evaluasi tidak terstandar
Alnefaie et al. [5] (2023)	Tanya jawab Al-Quran dengan GPT-4 (bahasa Arab)	Akurasi manual; tidak terstandar	Mengidentifikasi keterbatasan LLM umum pada teks Islam; tidak ada solusi fine tuning; tidak ada Faithfulness retrieval
Penelitian ini	Tanya jawab hadits Islam (65.811 hadits, 11 kitab, bahasa Indonesia-Arab)	RAGAS v0.2.6 + BERTScore F1	Faithfulness 0,676; Context Precision 1,000; Behavior Accuracy 100%; deployment lokal privat

Dari perbandingan pada Tabel 7, posisi kontribusi penelitian ini menjadi lebih jelas. Penelitian-penelitian terdahulu yang menggunakan domain Islam umumnya mengandalkan API LLM komersial tanpa fine tuning, sehingga tidak dapat mengatasi masalah halusinasi faktual secara sistematis dan evaluasinya tidak terstandar. Penelitian ini menjadi yang pertama menerapkan pipeline alignment iteratif berbasis on-policy DPO secara khusus pada domain hadits Islam dengan teks bilingual Indonesia-Arab, menggunakan kerangka evaluasi berlapis yang terstandar (RAGAS dan BERTScore), dan mendeploy model secara lokal untuk menjamin privasi data. Perbedaan mendasar ini menempatkan penelitian ini tidak sekadar sebagai peningkatan inkremental, melainkan sebagai proof-of-concept pipeline alignment yang dapat direplikasi dan dikembangkan untuk domain-domain teks keagamaan lain yang membutuhkan keandalan sitasi tinggi

4. Kesimpulan

Penelitian ini berhasil merancang, mengimplementasikan, dan mengevaluasi sistem tanya jawab hadits Islam berbasis pipeline fine tuning bertahap dan Hybrid RAG. Beberapa kesimpulan utama dapat ditarik dari penelitian ini.

Pertama, seleksi data berbasis IFD Scoring terbukti efektif dalam mereduksi dataset dari 1.730 menjadi 988 contoh berkualitas tinggi. Model SFT yang dihasilkan mencapai Behavior Accuracy 100% dan kemampuan penolakan off-topic yang melonjak signifikan, mengonfirmasi bahwa kualitas data lebih menentukan daripada kuantitas dalam SFT untuk domain spesifik.

Kedua, DPO Iterasi 1 dengan strategi off-policy menyebabkan regresi behavioral yang terdokumentasi secara empiris, mengonfirmasi prediksi teoritis tentang distribution shift. Temuan ini memberikan bukti konkret bahwa strategi off-policy tidak cocok untuk skenario di mana model yang di-fine-tune dan sumber data preferensi memiliki distribusi yang berbeda secara signifikan.

Ketiga, DPO Iterasi 2 dengan strategi on-policy memulihkan dan meningkatkan metrik kritis. Faithfulness meningkat menjadi 0,676 dibandingkan 0,633 pada model Base, Context Precision sempurna (1,000) pada kedua model, dan BERTScore F1 yang hampir identik (0,8621 vs 0,8615) mengonfirmasi tidak ada degradasi kualitas konten. Strategi on-policy terbukti menghasilkan alignment behavioral yang lebih stabil untuk LLM domain-spesifik.

Untuk penelitian selanjutnya direkomendasikan : (1) ekspansi iterasi on-policy DPO untuk memulihkan kemampuan penolakan off-topic ke level SFT; (2) eksplorasi metode alignment alternatif seperti Identity Preference Optimization (IPO) atau Kahneman-Tversky Optimization (KTO) sebagai pembanding.

Daftar Rujukan

[1] M. Fikri and U. Hasanah, "Unsur-Unsur Hadis dan Asbabul Wurud Hadis dalam Studi Ilmu Hadits," *ojs.uma*, vol. 1, no. 2, pp. 120–128, 2023, https://doi.org/10.31289/aij.v1i2.10180

[2] APJII, "APJII Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang," *APJII*, 2024. https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang (accessed Mar. 07, 2026).

[3] M. Wildan, S. imam A. Pratama, and D. Sugiarto, "Gen Z Muslims, Social Contestation, and Digital Citizenship in Indonesia," *Tribakti*, vol. 36, no. 1, pp. 165–182, 2025, https://doi.org/10.33367/tribakti.v36i1.6421

[4] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models : A Survey," *arxiv*, vol. 5, pp. 1–21, 2024, doi: 10.48550/arXiv.2312.10997.

[5] S. Alnefaie, E. Atwell, and M. A. Alsalka, "Is GPT-4 a Good Islamic Expert for Answering Quran Questions ?," in *Research Gate*, 2023, no. Rocling, pp. 124–133, doi: Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023).

[6] F. Sutiyo, "Implementasi Question Answering Berbasis Chatbot Telegram Pada Tafsir," *Implementasi Question Answering Berbasis Chatbot Telegram Pada Tafsir Al-Jalalain Menggunakan Langchain dan LLM*, vol. 4, no. 5, pp. 2464–2472, 2024, doi: 10.30865/klik.v4i5.1784.

[7] Z. Han, C. Gao, J. Liu, J. J. Zhang, and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models : A Comprehensive Survey," *arxiv*, pp. 1–25, 2024, doi: 10.48550/arXiv.2403.14608.

[8] W. Fan, S. Wang, H. Li, and D. Yin, "A Survey on RAG Meeting LLMs : Towards Retrieval-Augmented Large Language Models," *in Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '24)*, pp. 6491–6501, 2024, https://doi.org/10.1145/3637528.3671470

[9] N. A. M. Herwanza, N. S. Harahap, F. Yanto, and F. Insani, "Penerapan Langchain Retriever Dengan Model Chat Openai Dalam Pengembangan Sistem Chatbot Hadis Berbasis Telegram," *JTIM: Jurnal Teknologi Informasi dan Multimedia*, vol. 6, no. 1, pp. 70–83, 2024, doi: https://doi.org/10.35746/jtim.v6i1.514.

[10] M. Li *et al.*, "From Quantity to Quality : Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning," *NAACL*, vol. 1, pp. 7602–7635, 2024, https://doi.org/10.18653/v1/2024.naacl-long.421

[11] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization : Your Language Model is Secretly a Reward Model," *arxiv*, vol. 3, no. NeurIPS, 2024, doi: https://doi.org/10.48550/arXiv.2305.18290.

[12] S. Es, J. James, L. Espinosa-anke, and S. Schockaert, "RAGAS : Automated Evaluation of Retrieval Augmented Generation," *in Proc. 18th Conf. European Chapter of the Association for Computational Linguistics (EACL 2024): System Demonstrations*, pp. 150–158, 2024, doi: 10.18653/v1/2024.eacl-demo.16.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," *in Proc. Int. Conf. Learning Representations (ICLR 2020)*, pp. 1–43, 2020, doi: 10.48550/arXiv.1904.09675.

[14] E. Bjarnason, F. Lang, and A. Mjöberg, "An empirically based model of software prototyping : a mapping study and a multi - case study," *Empirical Software Engineering*, vol. 8, no. 5, 2023, https://doi.org/10.1007/s10664-023-10331-w

[15] R. K. Pradhana, D. P. Seerapu, G. Routhu, S. K. Manda, and G. R. Jami, "Locally Deployed NLP System for Secure Document Summarization and Context-Aware Question Answering Using LLMs and Vector Embeddings," *International Journal on Science and Technology (IJSAT)*, vol. 16, no. 2, pp. 1–12, 2025, https://doi.org/10.71097/IJSAT.v16.i2.4275

[16] L. S. A. Burhani, "Perkembangan Digitalisasi Hadis : Analisis Ensiklopedia Hadits 9 Imam Karya Lidwa Pusaka," *Jurnal Hukum Syariah*, vol. 4, pp. 23–37, 2021, https://doi.org/10.32506/johs.v4i1-03

[17] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," *in Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024, https://doi.org/10.18653/v1/2024.findings-acl.137

[18] S. Ockerman, U. States, S. Ockerman, R. Underwood, N. Chia, and K. Chard, "Exploring Distributed Vector Databases Performance on HPC Platforms : A Study with Qdrant Exploring Distributed Vector Databases Performance on HPC Platforms : A Study with Qdrant," *in Proc. HPC Asia 2025*, no. December, 2025, https://doi.org/10.1145/3731599.3767404

[19] C. Zhou *et al.*, "LIMA : Less Is More for Alignment," *in Advances in Neural Information Processing Systems (NeurIPS 2023)*, vol. 1, pp. 1–15, 2023, doi: 10.48550/arXiv.2305.11206.

[20] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA : Efficient Finetuning of Quantized LLMs," *n Advances in Neural Information Processing Systems (NeurIPS 2023)*, vol. 36, pp. 10088–10115, 2023, https://doi.org/10.52202/075280-0441

[21] E. Hu *et al.*, "LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS," *arxiv*, vol. 2, pp. 1–26, 2021, doi: https://doi.org/10.48550/arXiv.2106.09685.

[22] Z. Chen, Y. Deng, H. Yuan, K. Ji, and Q. Gu, "Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models," *arxiv*, vol. 3, 2022, doi: 10.48550/arXiv.2401.01335.

[23] D. IGuo *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," *arxiv*, vol. 2, 2026, doi: https://doi.org/10.48550/arXiv.2501.12948.

[24] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended RAG : Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers," *arxiv*, 2024, https://doi.org/10.1109/MIPR62202.2024.00031