Laporan Ilmiah Flowchart dan Analisis Algoritma RAGMINI 2026



Arsitektur RAGMINI

Retrieval-Augmented Generation (RAG)



Flowchart Algoritma RAGMINI



Knowledge Graph

(Pengetahuan Terstruktur)



RAGMINI 2026

Analisis Algoritma RAGMINI









KASMUI

LAPORAN ANALISIS ILMIAH

FLOWCHART DAN ANALISIS ARSITEKTUR ALGORITMA RAGMINI

https://ragmini.kasmui.cloud/

Audit implementasi sistem, Retrieval-Augmented Generation, fallback AI, komputasi astronomi, verifikasi bukti, pertumbuhan pengetahuan, dan rancangan evaluasi

Berdasarkan audit kode sumber PHP | 12 November 2026

Abstrak

Laporan ini menyajikan rekonstruksi arsitektur dan analisis algoritmik RAGMINI berdasarkan pemeriksaan kode sumber PHP terlampir. RAGMINI dikaji sebagai sistem hibrida yang mengombinasikan perutean deterministik, pencarian dokumen dengan BM25, pemeringkatan semantik dengan Gemini, sintesis berbasis konteks, pemeriksaan kualitas jawaban, fallback menuju AI Fikih dan Google AI Mode melalui SerpAPI, penyimpanan arsip, pertumbuhan basis pengetahuan bersyarat, serta integrasi komputasi astronomi SOLUNA dan KHGT. Delapan flowchart, tabel konfigurasi, analisis risiko, pseudocode pengembangan, dan rancangan pengujian disertakan secara utuh dalam dokumen ini. Analisis membedakan bukti statis dari kode dengan keberhasilan runtime yang belum dibuktikan. Temuan utama menunjukkan pendekatan evidence-aware RAG dengan mekanisme penapisan sumber dan penambahan pengetahuan, disertai kelemahan pada verifikasi ulang sumber sekunder, pencabutan persetujuan, skalabilitas indeks leksikal, dan ketergantungan terhadap backend eksternal.

Kata kunci: RAGMINI; Retrieval-Augmented Generation; BM25; Gemini; AI Fikih; SerpAPI; SOLUNA; KHGT; evaluasi jawaban; pertumbuhan pengetahuan.

1. Pendahuluan

RAGMINI telah berkembang dari chatbot RAG sederhana menjadi aplikasi hibrida yang mendukung pencarian dokumen, penalaran model bahasa, rute jawaban terstruktur, komputasi numerik, dan penyimpanan pengetahuan. Oleh karena itu, arsitekturnya tidak cukup direpresentasikan oleh urutan linear “database → AI Fikih → Google”. Diperlukan pemetaan yang memperlihatkan tahapan permintaan, pengambilan bukti, pemeriksaan keluaran, dan mekanisme fallback secara eksplisit.

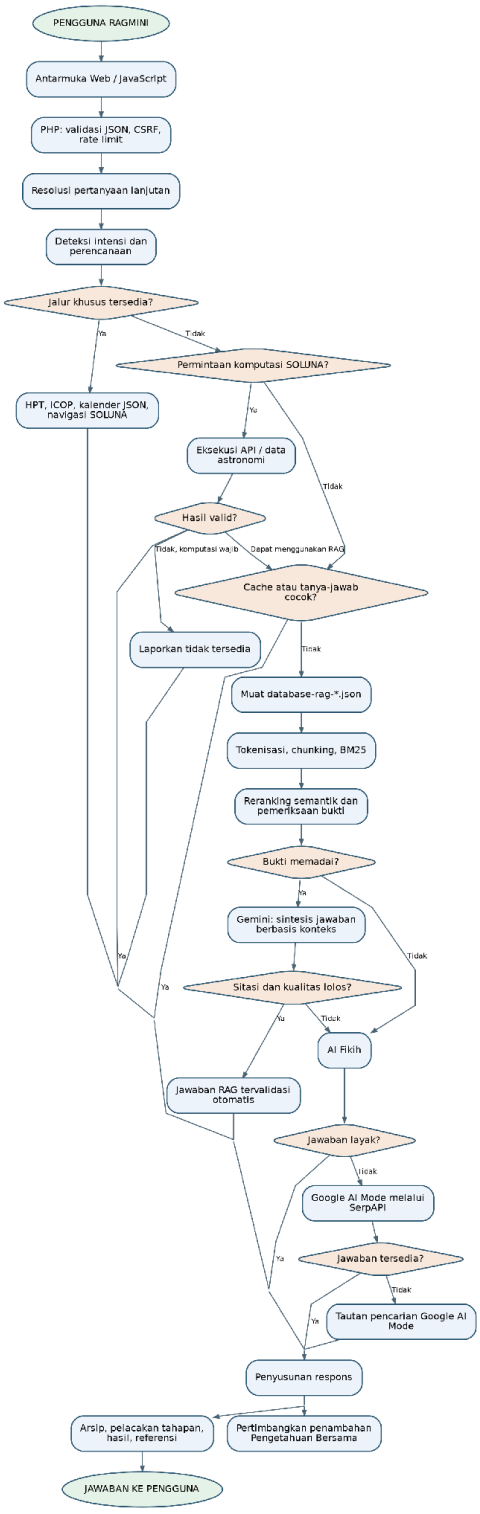
Secara konseptual RAGMINI dapat diklasifikasikan sebagai hybrid, evidence-aware, multi-stage RAG system with knowledge augmentation and domain-specific computational routing. Klasifikasi ini merupakan sintesis analitis atas kode, bukan klaim keberhasilan benchmark. Keabsahan kinerja tiap cabang pada server produksi memerlukan pengujian tersendiri.

2. Metode Audit Kode

Kajian ini menggunakan audit statis terhadap satu berkas PHP utama RAGMINI yang disampaikan pengguna. Pemeriksaan difokuskan pada konstanta konfigurasi, fungsi retrieval dan pemeringkatan, cabang keputusan permintaan utama, pembacaan database, penyimpanan pengetahuan, integrasi layanan eksternal, fungsi komputasi, mekanisme keamanan, serta trace proses. Flowchart disusun dengan mempertahankan hubungan percabangan yang dijelaskan dalam audit sebelumnya. Seluruh diagram dipresentasikan sebagai gambar dan sumber Mermaid dicantumkan pada lampiran agar bentuk serta logikanya dapat direproduksi.

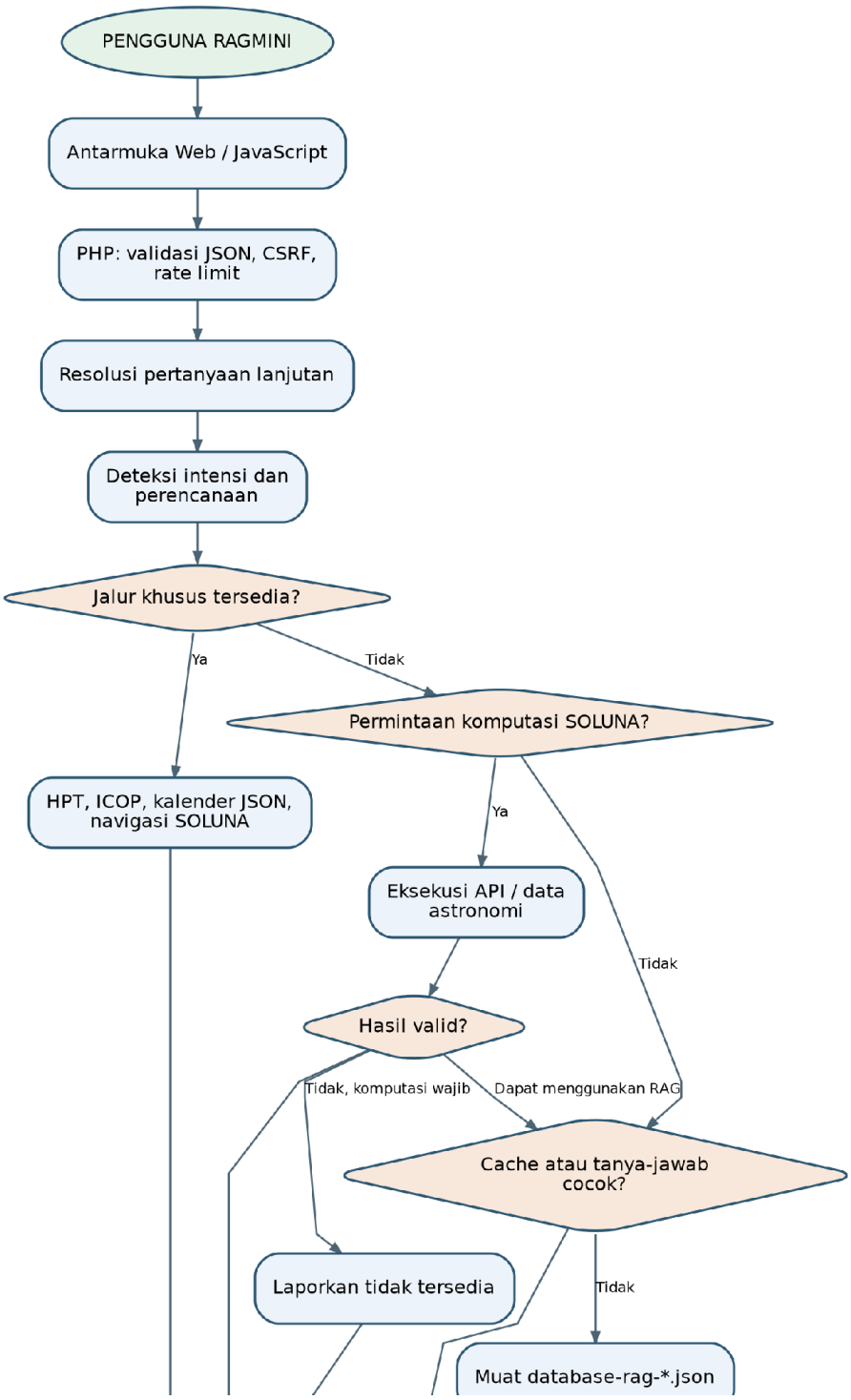
Sumber utama: berkas “Teks yang ditempel (1)(20261111-221032).txt”, memuat kode PHP dan antarmuka JavaScript/HTML RAGMINI. Uraian berkaitan dengan modul terpisah dan keberhasilan jaringan ditandai sebagai batas verifikasi. Dokumen ini tidak menyatakan telah melaksanakan uji API atau evaluasi terhadap server produksi.

3. Flowchart Arsitektur Utama RAGMINI

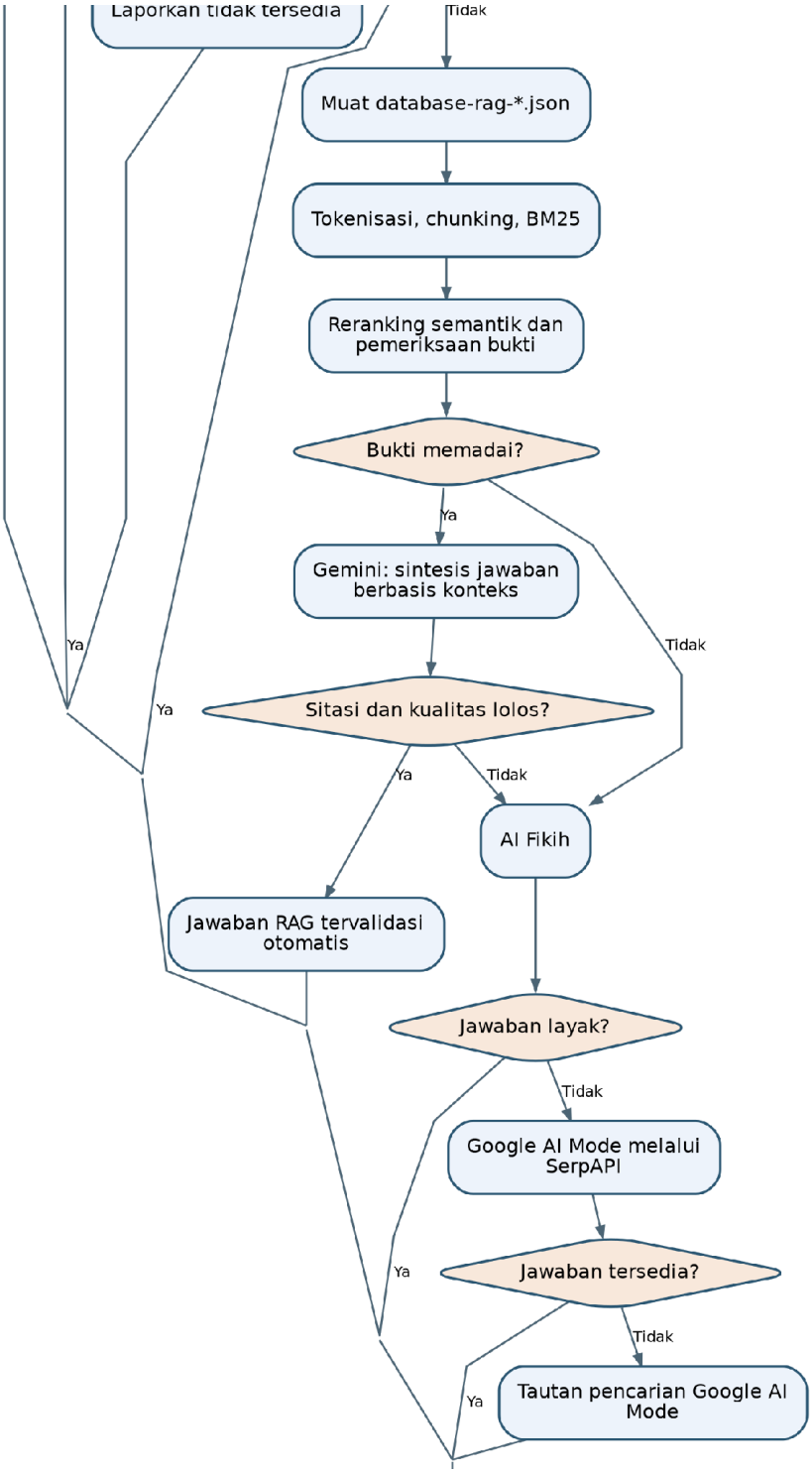


Gambar 1. Alur utama pemrosesan pertanyaan RAGMINI.

Catatan: cabang mencakup jalur utama ask; penambahan Pengetahuan Bersama bersifat bersyarat.



Detail Gambar 1, bagian 1/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).



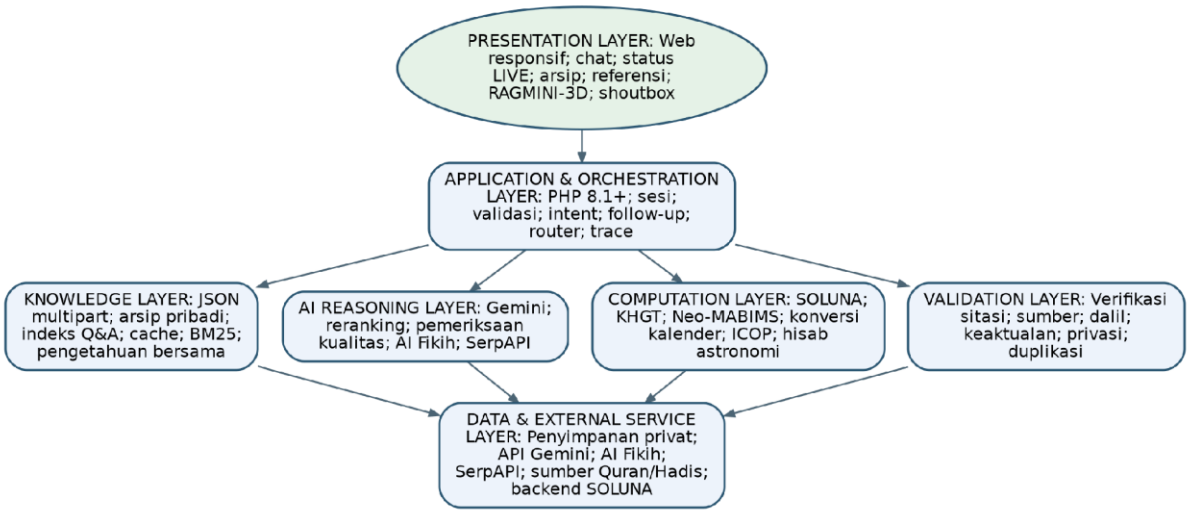
Detail Gambar 1, bagian 2/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).

Alur utama menunjukkan prinsip pemrosesan bertingkat dengan penghentian dini (early exit). Pertanyaan yang dapat dijawab dengan data eksplisit yang relevan tidak harus melewati BM25 atau pemanggilan generatif. Tahapan utama meliputi validasi permintaan, resolusi pertanyaan lanjutan, deteksi intensi, rute khusus atau komputasi, cache Q&A, retrieval, pemeriksaan bukti, generasi jawaban, evaluasi, fallback, serta arsip dan trace.

1. Deterministic routing: pertanyaan tertentu diarahkan ke data kalender, rujukan terkurasi, navigasi, atau mesin komputasi.

2. Retrieval-first: pertanyaan pengetahuan umum diprioritaskan melalui database internal.
3. Evidence-aware generation: sitasi, sumber, dan kualitas keluaran harus melalui pemeriksaan.
4. Multi-stage fallback: AI Fikih dan Google AI Mode menjadi penopang ketika jawaban lokal gagal.
5. Conditional knowledge growth: penambahan pengetahuan hanya setelah memenuhi syarat kualitas, privasi, dan persetujuan.

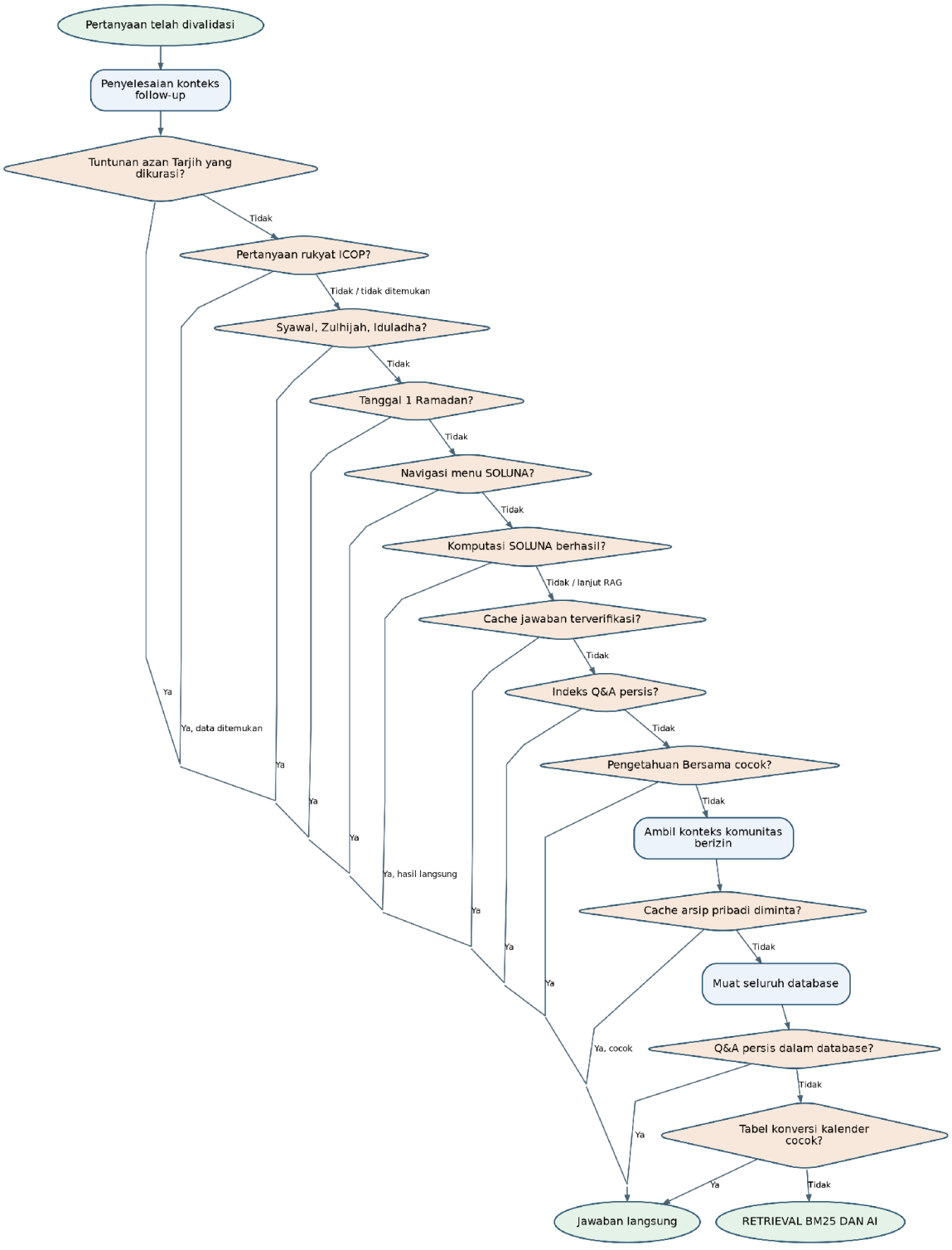
4. Arsitektur Komponen dan Lapisan Sistem



Gambar 2. Lapisan presentasi, orkestrasi, pengetahuan, penalaran, komputasi, validasi, serta layanan eksternal.

Secara konseptual terdapat presentation layer; application and orchestration layer; knowledge layer; AI reasoning layer; computation layer; validation layer; serta data and external service layer. Meskipun lapisan tersebut dapat dibedakan secara fungsional, sejumlah proses masih berada dalam file PHP utama. Pemisahan konseptual penting karena pencarian dokumen, generasi teks, dan hasil numerik harus melalui aturan validasi yang tidak sama.

5. Prioritas Jalur Pemrosesan dan Early Exit



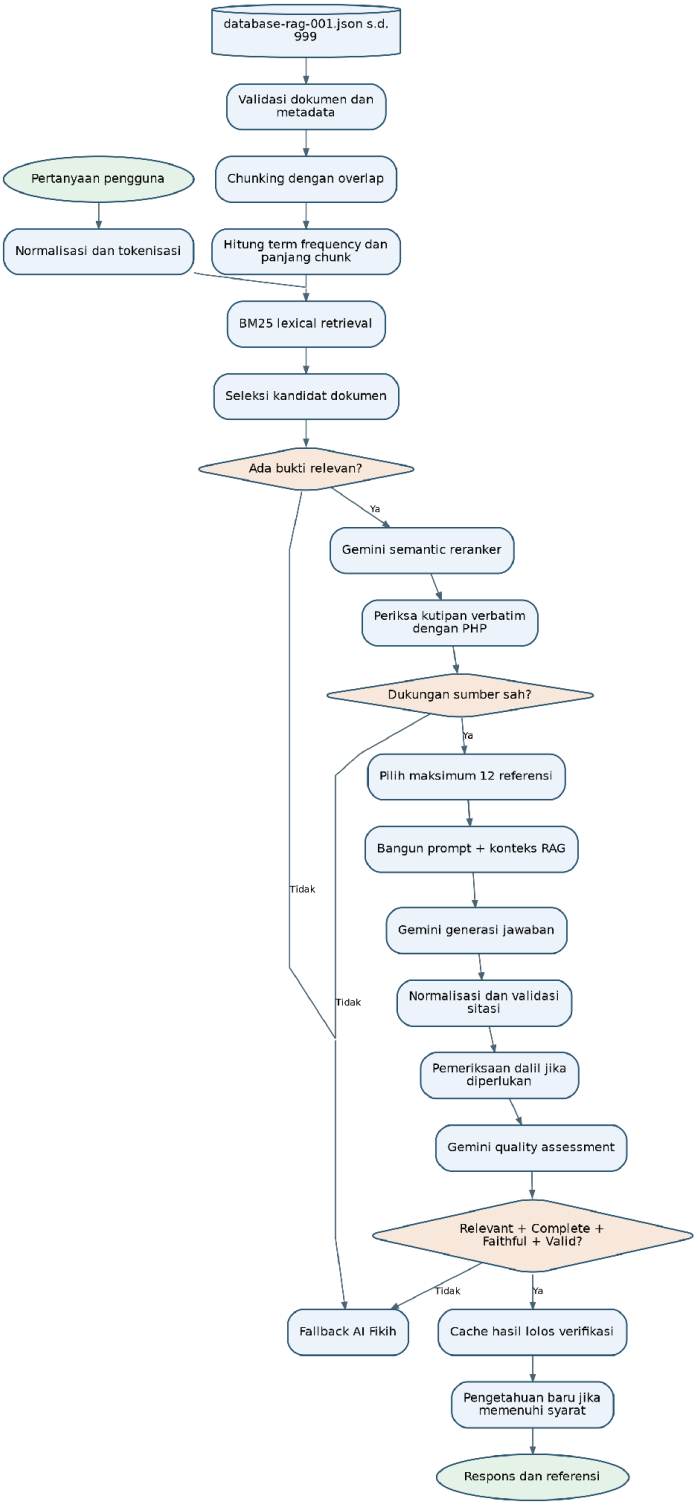
Gambar 3. Urutan prioritas jalur cepat sebelum retrieval BM25.

Jalur cepat mendahulukan data khusus yang sudah disiapkan atau rujukan yang secara eksplisit cocok. Karena itu, jawaban penentuan awal bulan yang telah tersedia dalam JSON tidak wajib dihitung kembali oleh Gemini. Jalur cepat tetap memerlukan pengendalian tipe sumber, masa berlaku data, dan pembatasan agar jawaban sekunder tidak dianggap sebagai dokumen primer.

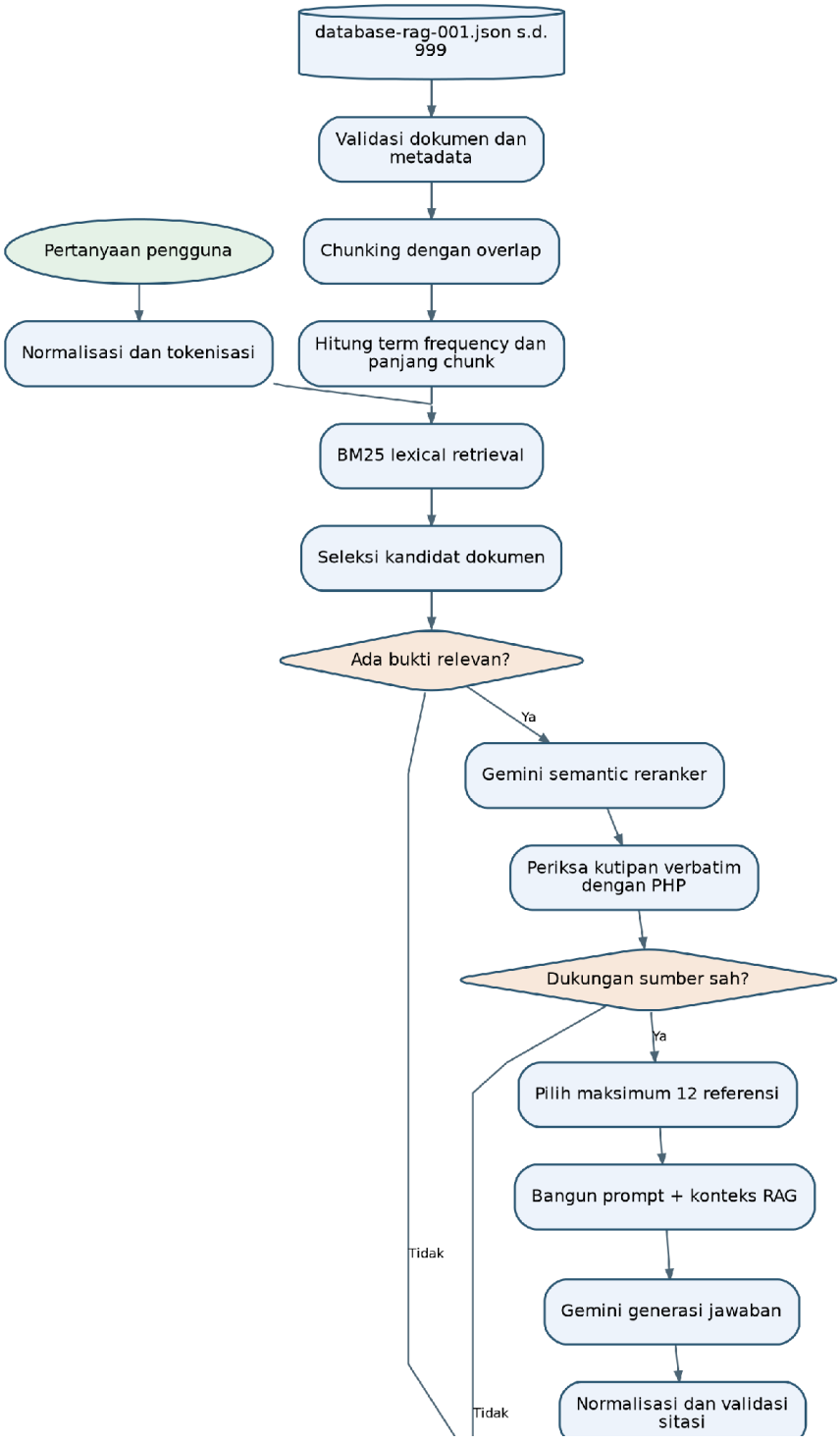
Tabel 1. Mekanisme prioritas dan jenis keluaran.

Jalur	Algoritma	Keluaran
Tuntunan khusus Tarjih	Pencocokan intensi dan jawaban terkurasi	Jawaban HPT dan rujukan terkait
ICOP	Deteksi permintaan dan data ICOP/ringkasan lokal	Hasil rukyat berdasar data tersedia
Ramadan, Syawal, Zulhijah	Pembacaan JSON kalender lokal	Tanggal dan informasi kalender
Navigasi SOLUNA	Pencocokan menu dan tautan	URL modul relevan
Komputasi SOLUNA	Eksekusi API atau arsip numerik	Hasil astronomi sesuai metode
Verified cache	Pencocokan pertanyaan dan tanda tangan database	Jawaban sebelumnya yang lolos pemeriksaan
Indexed Q&A	Kunci pertanyaan ternormalisasi	Jawaban langsung dari Q&A
Pengetahuan Bersama	Exact match atau parafrasa ketat	Jawaban bersama tersaring
Arsip pribadi	Pencocokan jika cache arsip diminta	Jawaban historis pengguna
Database RAG	BM25 dan reranking semantik	Jawaban dokumen jika lolos verifikasi

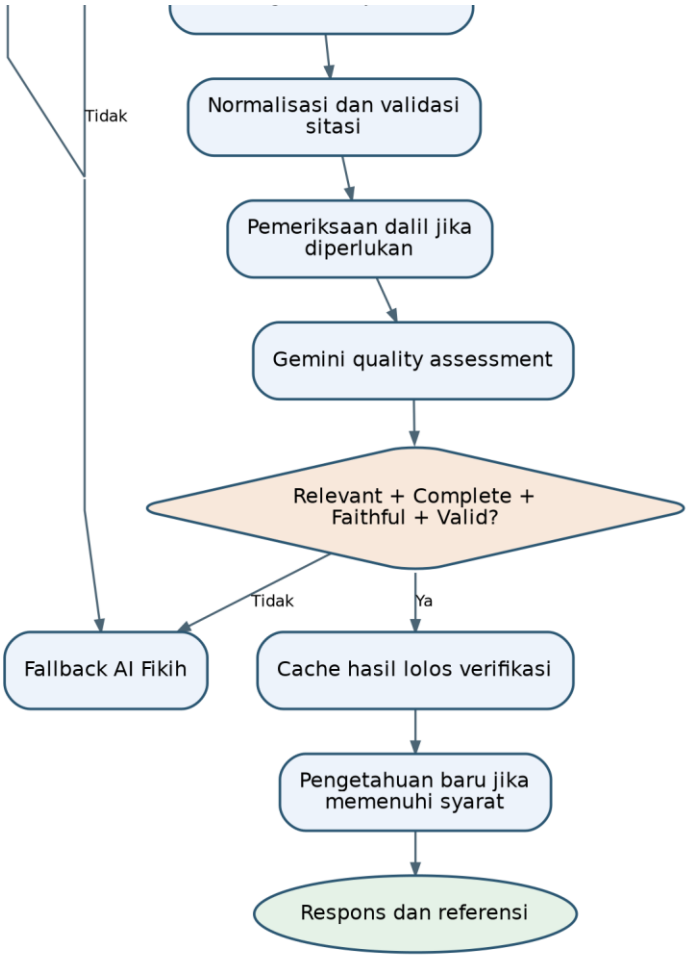
6. Algoritma Inti Retrieval-Augmented Generation



Gambar 4. Tahapan ingestion, preprocessing, BM25, reranking, sintesis, dan quality gate.



Detail Gambar 4, bagian 1/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).



Detail Gambar 4, bagian 2/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).

6.1 Ingestion dan pemuatan dokumen

Fungsi ragLoad() memindai database-rag-001.json hingga database-rag-999.json yang tersedia, bersama berkas legacy database-rag.json jika ada. Pemeriksaan mencakup keberadaan file, izin baca, ukuran, validitas JSON, struktur version dan documents, kelengkapan id/title/content/source, metadata tambahan seperti tags dan jenis sumber, serta konflik ID. Konflik data identik atau berbeda ditangani dengan aturan deduplikasi/validasi. File bagian yang hilang dapat tercatat sebagai peringatan sehingga bagian yang masih valid dapat dibaca.

- Batas file multipart: 1.000.000 byte per bagian.
- Batas file utama lama: 2 MiB.
- Bagian bernomor 001–999; urutan numerik dibaca secara dinamis.
- Metadata dapat mencakup source_type, source_version, published_at, verified_at, valid_until, dan superseded.

6.2 Preprocessing, normalisasi, dan chunking

Fungsi ragTokens() memproses huruf kecil, variasi bahasa Arab, penghapusan harakat untuk pencarian, normalisasi sebagian ejaan Indonesia, dan stopword. Contohnya shalat/sholat/salat menjadi token salat, sedangkan fiqh/fiqih/fikih menjadi istilah yang sama. Fungsi ragChunks()

selanjutnya membagi teks menjadi jendela kata dan karakter dengan overlap agar konteks tidak terputus sepenuhnya di batas potongan.

Tabel 2. Parameter chunking yang disebut dalam audit.

Parameter	Nilai pada audit	Implikasi
Jendela kata	170 kata	Ukuran chunk berbasis kata
Pergeseran kata	140 kata	Overlap 30 kata
Jendela karakter	1.600 karakter	Alternatif pemotongan berbasis karakter
Pergeseran karakter	1.400 karakter	Overlap 200 karakter
Batas chunk sementara	24.000	Batas jumlah fragmen dalam satu konteks pemrosesan

Overlap mendukung pencarian lintas batas fragmen, tetapi pemotongan berdasarkan panjang belum tentu menjaga satu kesatuan ayat, hadis, tabel, paragraf, atau penjelasan metodologis.

6.3 Pemeringkatan leksikal BM25

RAGMINI menggunakan varian BM25 dengan memperhitungkan frekuensi token pada setiap chunk, document frequency, panjang relatif chunk, dan panjang rata-rata kumpulan. Nilai konfigurasi yang diamati adalah k1 = 1,2 dan b = 0,75; jumlah referensi akhir top-k = 12. Candidate retrieval pada konfigurasi yang dibahas dapat mencapai 24 sebelum penyaringan lebih lanjut. Seleksi membatasi maksimal dua fragmen berbeda per dokumen dan menggabungkannya sebagai satu referensi, sekaligus mempertimbangkan persentase cakupan istilah.

BM25 efektif untuk terminologi spesifik pada dokumen fikih dan astronomi, namun dapat kehilangan sinonimi atau parafrasa yang tidak memuat token yang sama. Karena itu, reranking semantik dan/atau retrieval vektor dapat meningkatkan cakupan semantik.

6.4 Semantic reranking dan pemeriksaan kutipan

Fungsi ragSupportedSources() meminta Gemini menilai kandidat sumber yang diambil BM25. Pemeringkat perlu menghasilkan keputusan terstruktur yang menyatakan dukungan dan kutipan sumber. PHP kemudian mengecek bahwa kutipan verbatim benar-benar terdapat dalam fragmen. Mekanisme ini menahan sebagian halusinasi berupa kutipan yang tidak terdapat pada bahan rujukan, tetapi belum menjamin tafsir terhadap kutipan telah tepat.

```
{
 "supported": true,
 "evidence": [
  {"label": "R1", "quote": "Kutipan yang benar-benar terdapat dalam dokumen"}
 ]
}
```

6.5 Pembentukan prompt dan jawaban

Fungsi ragPayload() menyusun konteks dan instruksi untuk ragGemini(). Model diberi konteks sumber terpilih, pola sitasi [R1], [R2], dan seterusnya, serta instruksi untuk menghindari klaim tidak terbukti. Untuk kajian keislaman, prompt memuat pendekatan Bayani, Burhani, Irfani, dan Maqashid Syariah. Pengutipan keagamaan juga dapat melalui pemeriksaan teks Arab, terjemahan, dan referensi bernomor.

Tabel 3. Konfigurasi model dan generasi.

Konfigurasi	Nilai yang tercantum
Model utama	gemini-3.8-flash
Model cadangan	gemini-3.5-flash-lite
Temperatur generasi RAG	0,2
Maksimum keluaran generasi	4.000 token
Mode strict	Aktif

Nama model merupakan nilai yang terdapat dalam kode dan dapat ditimpa melalui environment server. Nilai konfigurasi tidak membuktikan endpoint kedua model tersedia atau aktif pada akun API produksi.

6.6 Validasi jawaban

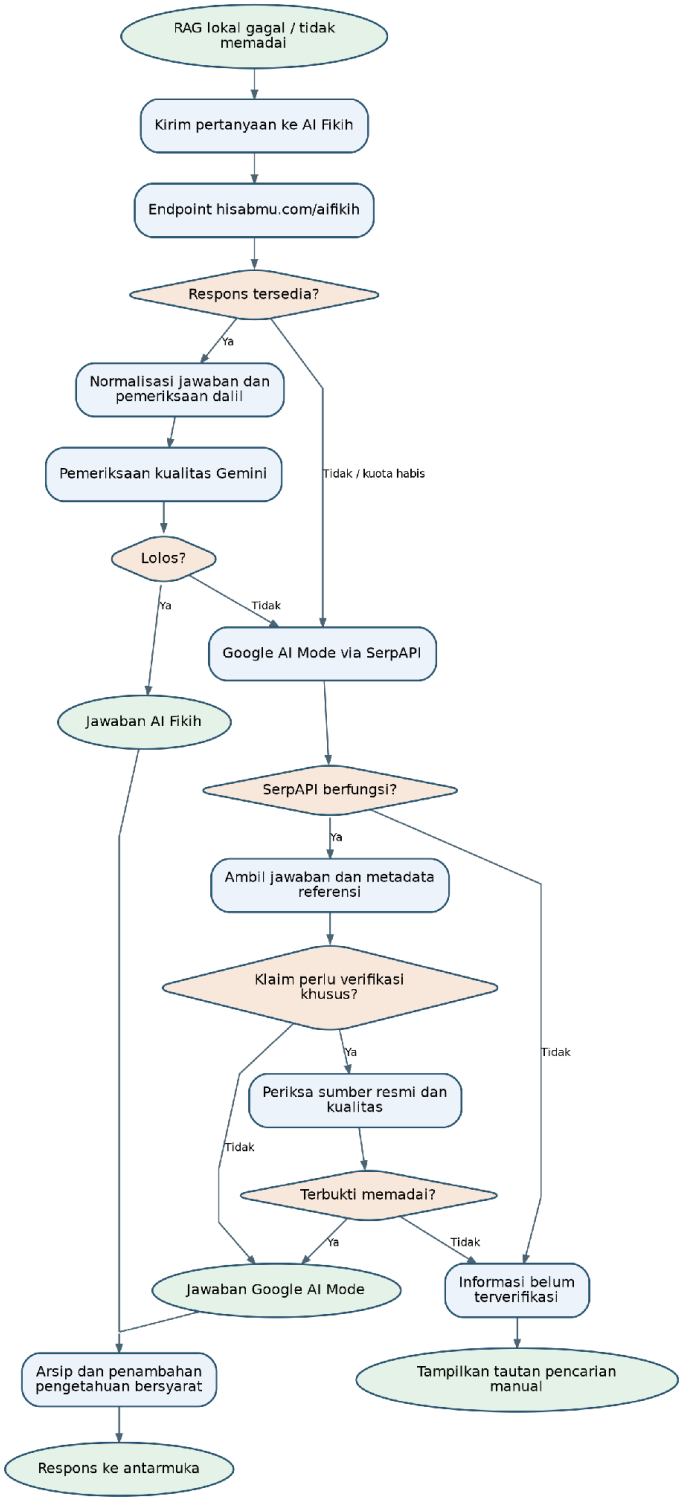
Fungsi ragAssessAnswer() menguji relevant, complete, faithful, dan unsupported_absence. Syarat penerimaan jawaban RAG lokal mengharuskan dukungan yang memadai, jawaban relevan serta lengkap, dan tidak menyatakan ketiadaan keputusan atau dokumen tanpa dasar. Kegagalan validasi membawa proses menuju fallback, bukan menampilkan jawaban lokal sebagai fakta yang telah terverifikasi.

Tabel 4. Elemen quality gate.

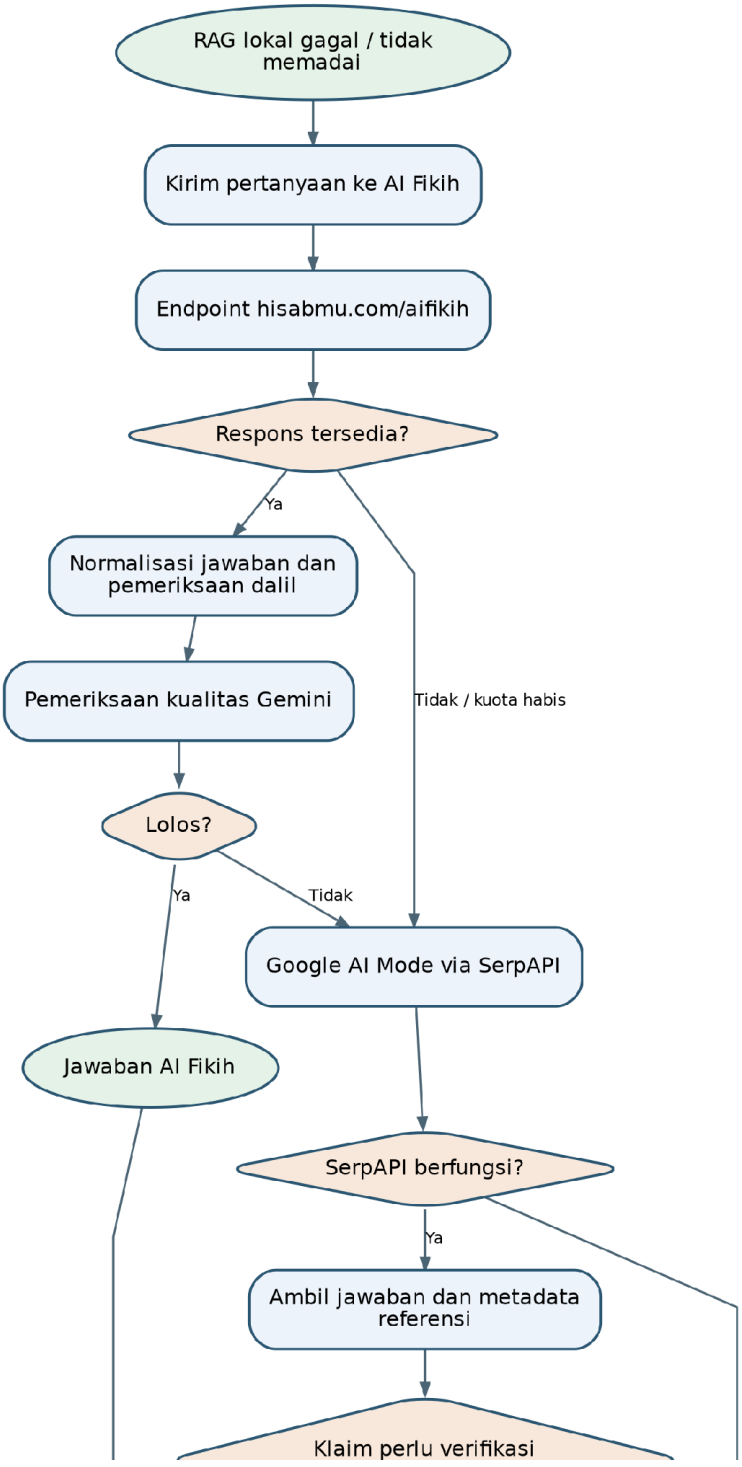
Indikator	Pertanyaan evaluasi
relevant	Apakah jawaban sesuai dengan pertanyaan?
complete	Apakah semua bagian pertanyaan dijawab?
faithful	Apakah pernyataan didukung sumber?
unsupported_absence	Apakah klaim bahwa sesuatu tidak ada dibuat tanpa bukti cukup?

Evaluasi otomatis berbasis Gemini tetap mempunyai kemungkinan salah. Kesalahan dapat berupa salah memahami konteks ayat/hadis, ringkasan yang melampaui bukti, dan bias akibat memakai keluarga model yang sama untuk generator dan evaluator.

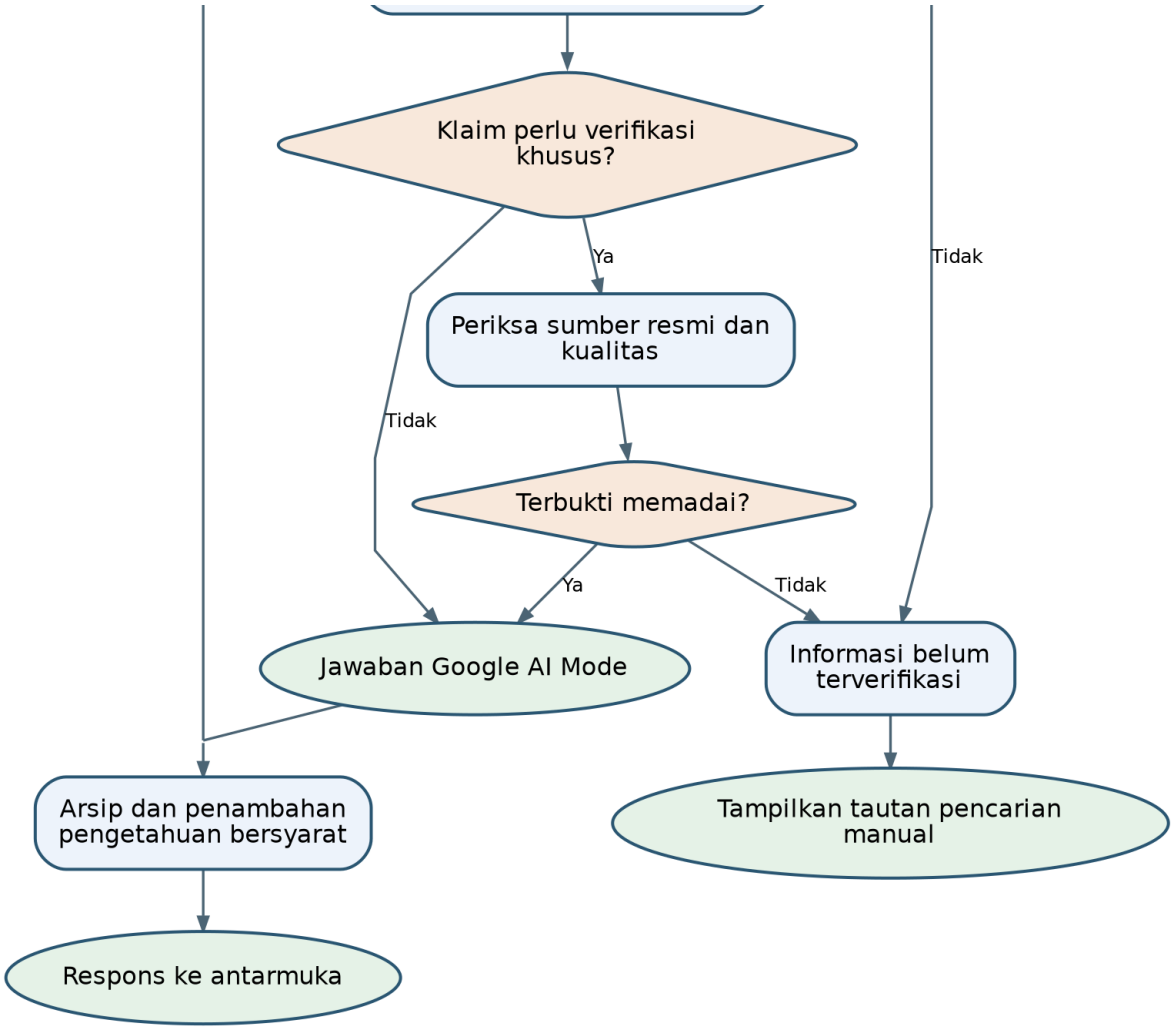
7. Fallback AI Fikih dan Google AI Mode



Gambar 5. Fallback AI Fikih, penilaian kualitas, SerpAPI Google AI Mode, dan tautan manual.



Detail Gambar 5, bagian 1/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).



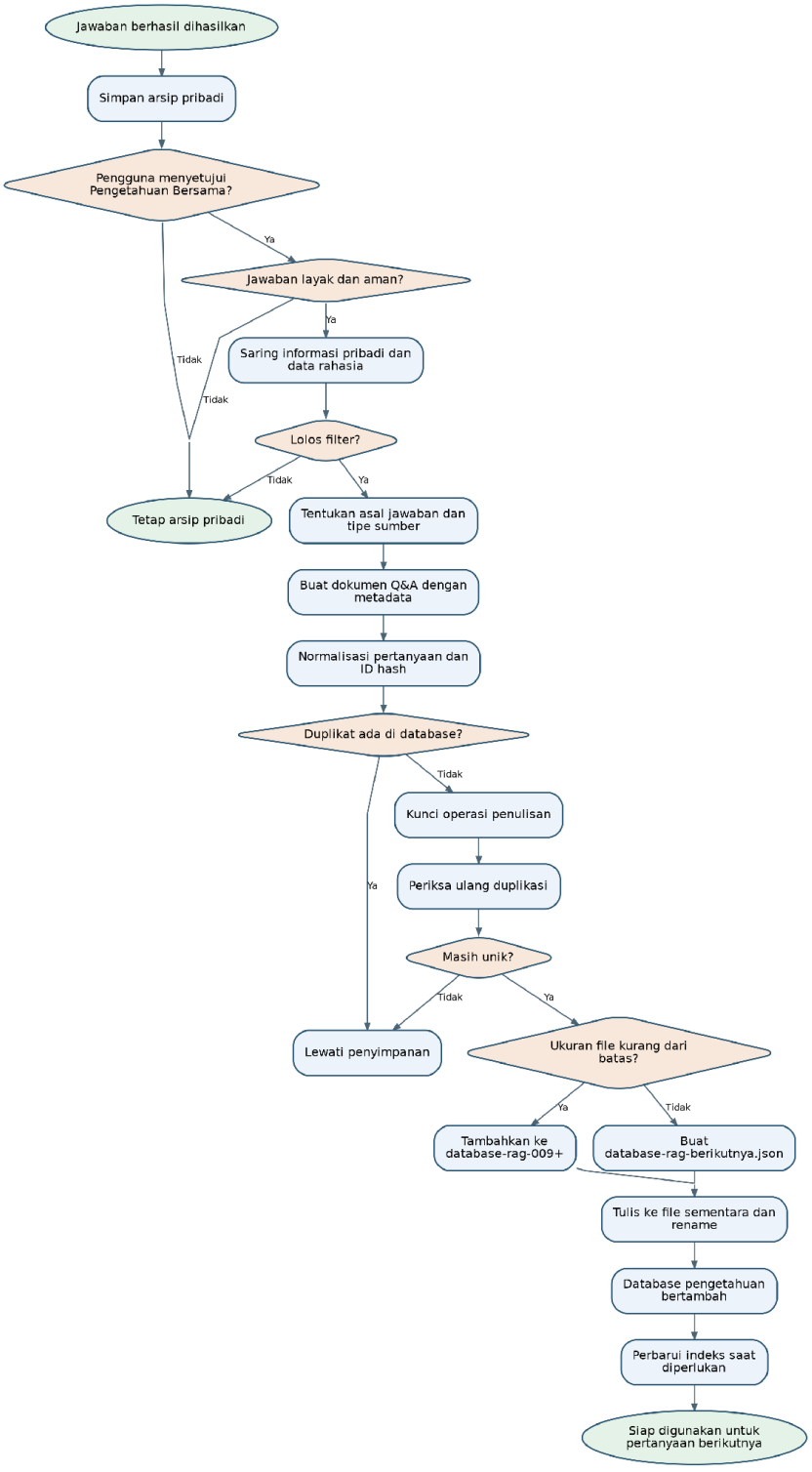
Detail Gambar 5, bagian 2/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).

Fallback dipicu bukan hanya saat database kosong, tetapi juga ketika dukungan sumber tidak cukup, kualitas keluaran gagal, sitasi tidak sah, atau API tidak tersedia. Tahap pertama memanggil AI Fikih pada layanan hisabmu.com/aifikih/index.php lalu menormalisasi jawaban serta memeriksa relevansi dan kelengkapan. Bila gagal, sistem mencoba SerpAPI dengan engine google_ai_mode untuk memperoleh jawaban daring dan metadata rujukan.

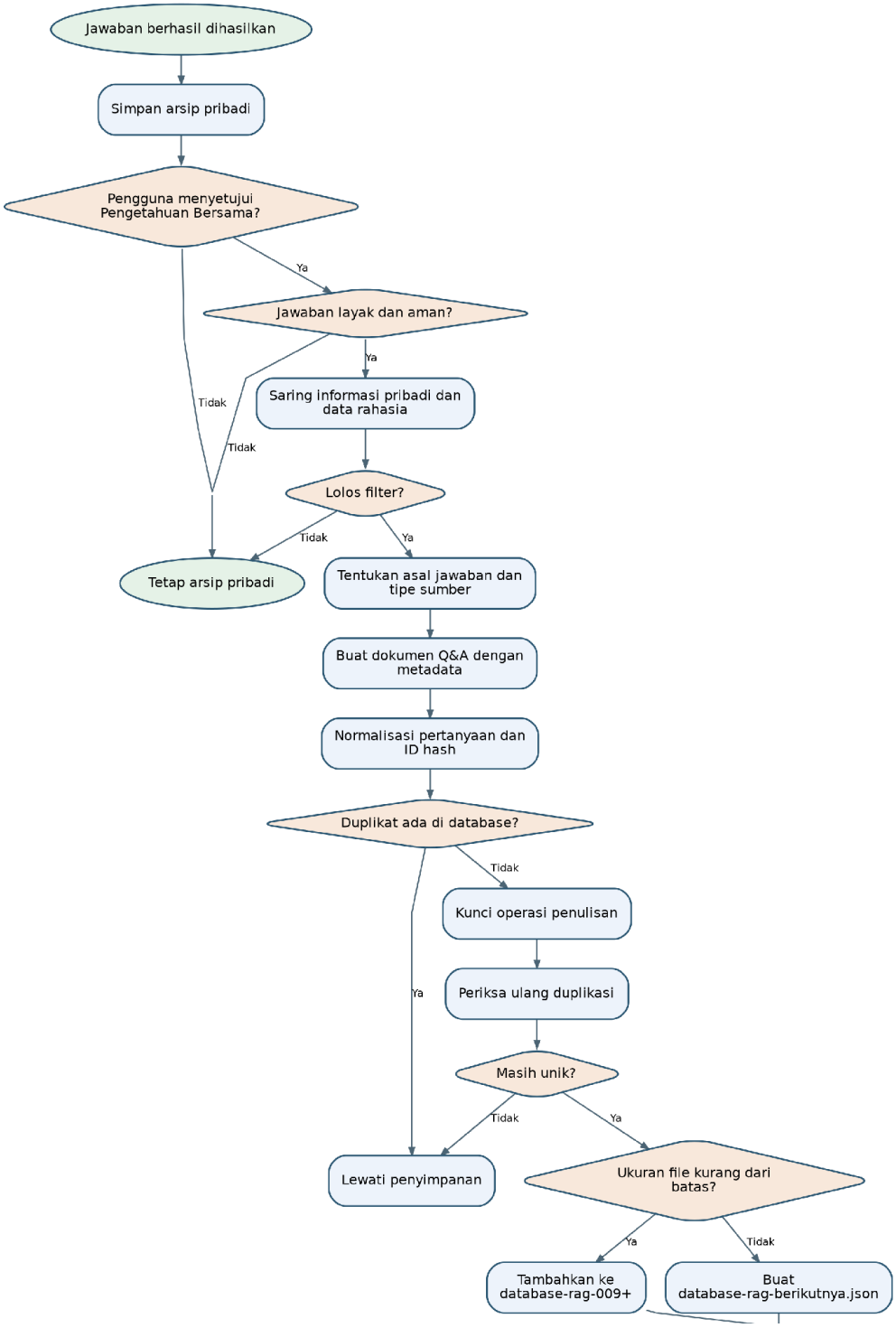
Untuk klaim tertentu, terutama yang mengatasnamakan institusi atau Majelis Tarjih, terdapat pemeriksaan sumber resmi. Standar pemeriksaan antarjalur belum seluruhnya identik: jawaban dari SerpAPI pada beberapa pertanyaan umum tidak mengalami gate yang sama dengan RAG lokal. Jika SerpAPI tidak tersedia, aplikasi dapat menampilkan tautan pencarian manual.

Terdapat fungsi jawaban Gemini dengan Google Search grounding, tetapi pada rute ask utama yang dibahas fungsi tersebut bukan tahap otomatis aktif. Urutan operasional utamanya adalah Database RAG → AI Fikih → Google AI Mode melalui SerpAPI → tautan manual.

8. Pembelajaran Pengetahuan dan Pertumbuhan Database



Gambar 6. Alur penambahan pengetahuan bersama berbasis persetujuan, validasi, dan deduplikasi.



Detail Gambar 6, bagian 1/1 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).

RAGMINI melaksanakan knowledge-base augmentation, bukan pelatihan ulang bobot parameter model Gemini. Mekanismenya mengubah sebagian jawaban yang lolos penapisan menjadi dokumen Q&A dengan metadata, memeriksa duplikasi, lalu menyimpan pada bagian dinamis database-rag-009.json dan seterusnya. Arsip percakapan pribadi tetap merupakan objek yang berbeda dari pengetahuan bersama.

Tabel 5. Perlakuan berdasarkan asal jawaban.

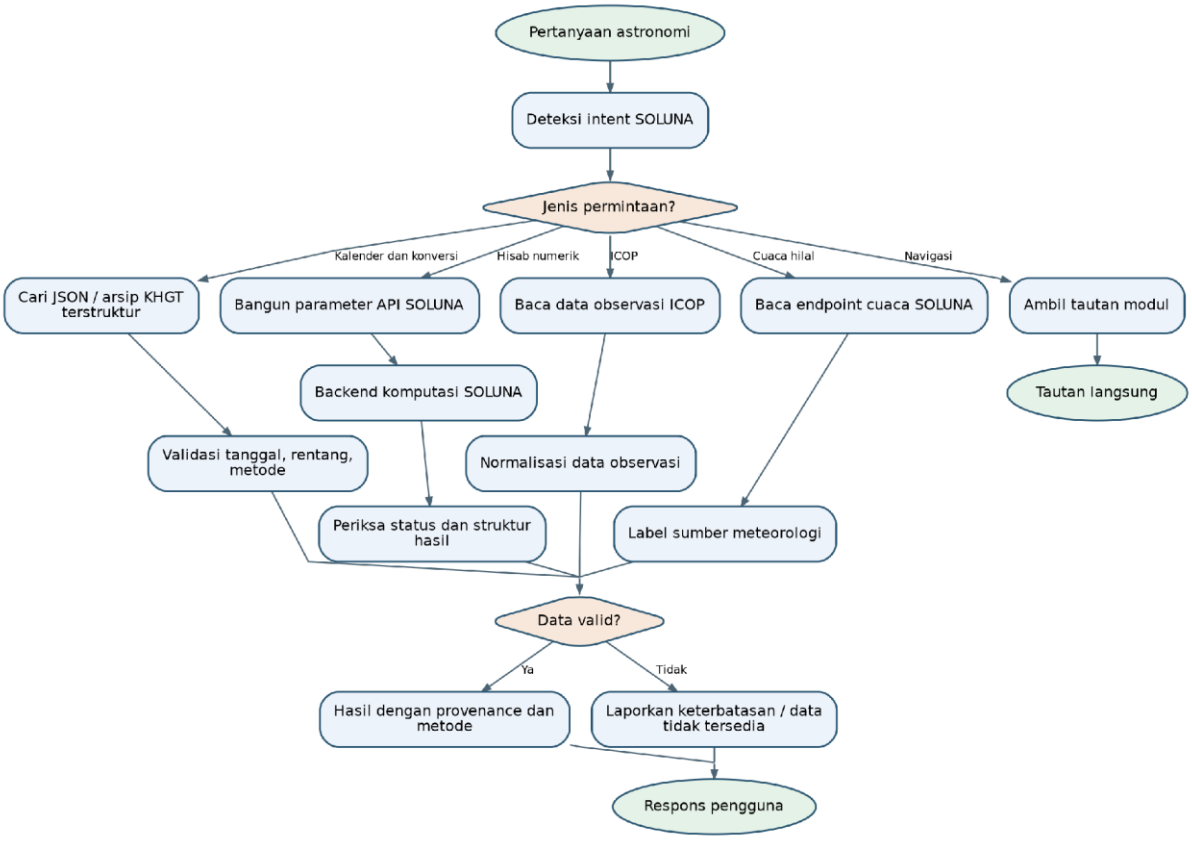
Asal jawaban	Perlakuan dalam knowledge growth
RAG lokal dan Gemini	Dapat disimpan bila didukung bukti primer serta lulus penapisan.
AI Fikih	Dapat disimpan sebagai ringkasan sekunder setelah syarat terpenuhi.
Google AI Mode via SerpAPI	Dapat disimpan sebagai pengetahuan sekunder; bukan dokumen primer.
Arsip pengguna	Dapat disinkronkan bila disaring dan ada persetujuan.
Gagal atau hanya tautan	Tidak layak menjadi dokumen pengetahuan otomatis.

Untuk ringkasan tertentu didukung sumber primer, kode menandai masa berlaku contoh selama 90 hari. Dokumen eksternal juga diberikan metadata asal dan masa berlaku, dengan aturan lebih singkat apabila memerlukan pembaruan fakta. Pemeriksaan kunci pertanyaan mencegah duplikasi terhadap database kurasi maupun bagian otomatis. Penulisan memakai file lock dan file sementara yang diganti melalui rename, sehingga penulisan serentak lebih terkendali. Bila ukuran bagian mendekati 990.000 byte, dokumen baru disimpan di bagian berikutnya, dengan batas eksplisit sampai bagian 999.

Tabel 6. Ruang lingkup pembelajaran sistem.

Bentuk pembelajaran	Status
Menyimpan tanya-jawab baru	Diimplementasikan dengan syarat.
Memanfaatkan jawaban tersimpan sebagai konteks atau jawaban cepat	Diimplementasikan.
Melatih ulang bobot model dari percakapan	Tidak teramati pada kode.

9. Integrasi Astronomi SOLUNA, KHGT, dan ICOP



Gambar 7. Rute permintaan astronomi menuju arsip lokal, backend SOLUNA, ICOP, dan data meteorologi.

RAGMINI mengenali kebutuhan konversi kalender KHGT ke Masehi dan sebaliknya, kalender bulanan, posisi benda langit, fenomena hilal, ephemeris, gerhana, kiblat, jadwal salat, observasi rukyat ICOP, meteorologi pengamatan, serta kemungkinan perbandingan KHGT dan Neo-MABIMS sesuai ketersediaan layanan. Kode dapat memakai arsip kalender terstruktur dan membangun parameter API. Dalam konversi kalender tertentu, tersedia pendekatan pencarian biner terhadap arsip berurutan.

Pemilahan asal angka merupakan syarat ilmiah. Nilai numerik tidak boleh direkayasa oleh model bahasa ketika backend tidak tersedia. Demikian pula, data KHGT tidak boleh diam-diam digunakan untuk menjawab permintaan Neo-MABIMS. Data hisab, kriteria penetapan kalender, data meteorologi, serta hasil rukyat empiris harus diberi provenance dan metode masing-masing.

Keterbatasan verifikasi: keberadaan rute API pada file utama tidak membuktikan ketepatan algoritma astronomi pada modul terpisah ragmini-soluna-client.php, ragmini-khgt-fallback.php, maupun backend SOLUNA. Diperlukan uji integrasi dengan data acuan.

10. Keamanan, Privasi, dan Keandalan Sistem

Tabel 7. Mekanisme perlindungan dan penilaian risiko.

Komponen	Implementasi	Penilaian
API key	Environment dan file rahasia server	Lebih aman daripada hardcode publik.
CSRF	Token untuk request POST	Mengurangi serangan lintas situs.
Rate limiting	Interval sekitar 3 detik per sesi	Tidak memadai sendirian untuk serangan terdistribusi.
Autentikasi	Hash password dan batas percobaan	Fondasi proteksi akun.
Arsip	Direktori .rag-private dengan pembatasan akses	Memisahkan arsip dari dokumen publik.
Penulisan data	File lock serta penulisan atomik	Mengurangi risiko konflik penulisan.
Validasi jawaban	Sumber, sitasi, kualitas	Mengurangi jawaban tanpa dasar.
Pengetahuan bersama	Persetujuan dan filter privasi	Memerlukan pengujian lebih jauh.
Status LIVE	Token acak dan snapshot tahapan	Tahap proses tanpa isi penalaran privat.
Batas waktu	Anggaran hingga 240 detik pada jalur terukur	Membatasi beberapa operasi berat.

Mekanisme live processing trace mencatat request_id, nama tahap, status, dan durasi dalam milidetik; rekaman ini mendukung dashboard latency dan observabilitas tanpa harus mengungkap pertanyaan, kunci API, atau isi penalaran internal.

11. Temuan Audit, Risiko, dan Kelemahan

Tabel 8. Evaluasi implementasi; bukan hasil benchmark kuantitatif.

Aspek	Kekuatan	Keterbatasan
Retrieval	BM25, normalisasi, overlap	Belum menggunakan indeks vektor semantik.
Seleksi sumber	Reranking Gemini dan kutipan verbatim	Penafsiran konteks tetap dapat keliru.
Generasi	Prompt dengan konteks/sitasi	Bergantung pada Gemini.
Evaluasi	Relevansi, kelengkapan, faithfulness	Kesalahan evaluator dapat berkorelasi.
Fallback	AI Fikih dan SerpAPI	Standar kualitas antarjalur tidak seragam.
Knowledge growth	Database bertambah sesuai persetujuan	Ringkasan AI berisiko menumpuk.
Komputasi	Memisahkan hasil numerik dari narasi	Bergantung konektor dan server terpisah.
Skalabilitas	Multipart JSON, cache, indeks Q&A	Pemindaian dokumen melalui PHP meningkat seiring korpus.
Keamanan	CSRF, penguncian file, hash password	Filter privasi pola belum menjamin deteksi lengkap.

11.1 Sumber sekunder sebagai jalur cepat

Indeks Q&A berpotensi mengembalikan ringkasan AI yang pernah disimpan tanpa retrieval ulang ke sumber primer. Meskipun asal jawaban dapat ditandai sebagai sekunder, penyimpanan historis tidak memastikan bahwa informasi masih benar atau merupakan keputusan terbaru. Hierarki sumber yang disarankan: (1) dokumen primer terverifikasi; (2) data resmi dan keluaran komputasi dengan provenance jelas; (3) ringkasan RAG yang didukung sumber primer; (4) jawaban AI eksternal belum diverifikasi independen; dan (5) arsip percakapan sebagai konteks.

11.2 Pencabutan persetujuan dan data turunan

Pada jalur pencabutan persetujuan yang diperiksa, belum terlihat mekanisme penghapusan retroaktif semua dokumen turunan yang sebelumnya disalin ke database-rag-009 dan bagian berikutnya. Mekanisme provenance internal, daftar ketergantungan dokumen, serta penghapusan indeks perlu ditambahkan agar pencabutan persetujuan dapat ditegakkan pada semua salinan yang masih aktif.

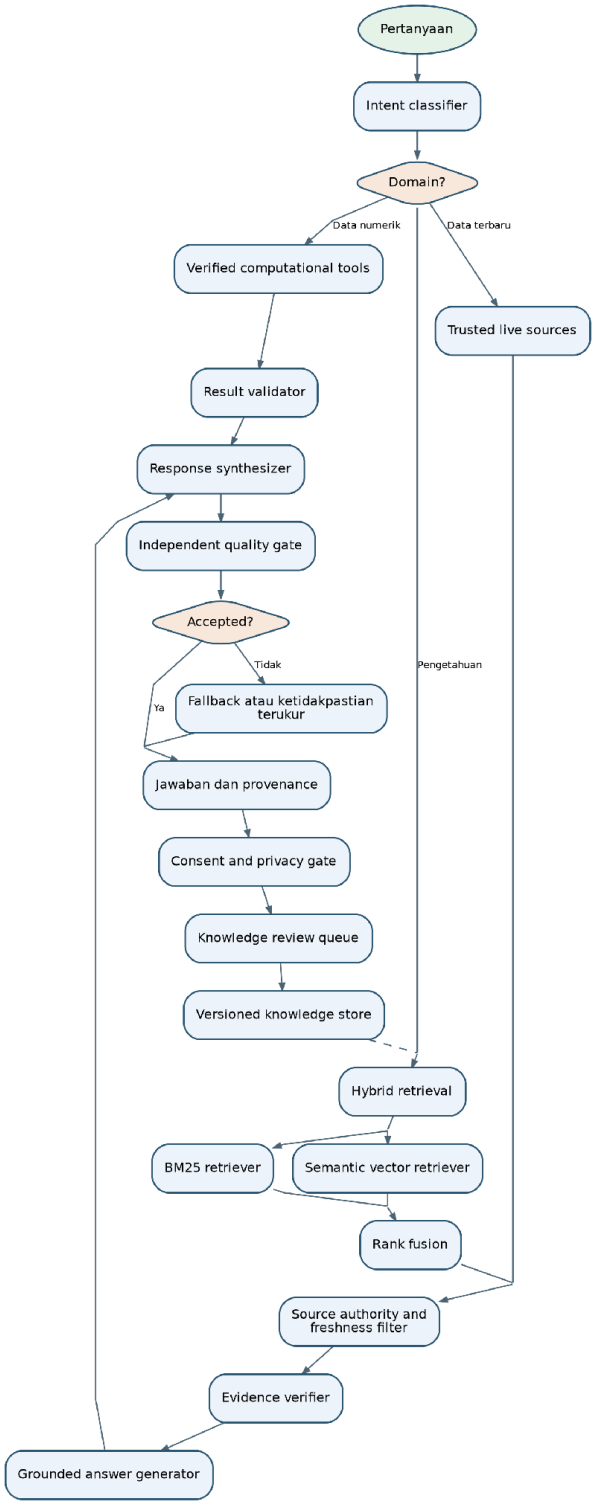
11.3 Skalabilitas dan kompleksitas

Penilaian BM25 dalam kode memindai chunk pada setiap permintaan untuk menghitung frekuensi dan skor. Dengan N chunk dan T istilah kueri, kompleksitas pemindaian secara kasar tumbuh sebanding dengan N × T, di luar biaya pengurutan, komunikasi API, dan evaluasi model. Untuk korpus besar, inverted index dan penyimpanan retrieval persisten dapat mengurangi biaya berulang.

11.4 Status pemeriksaan sintaks

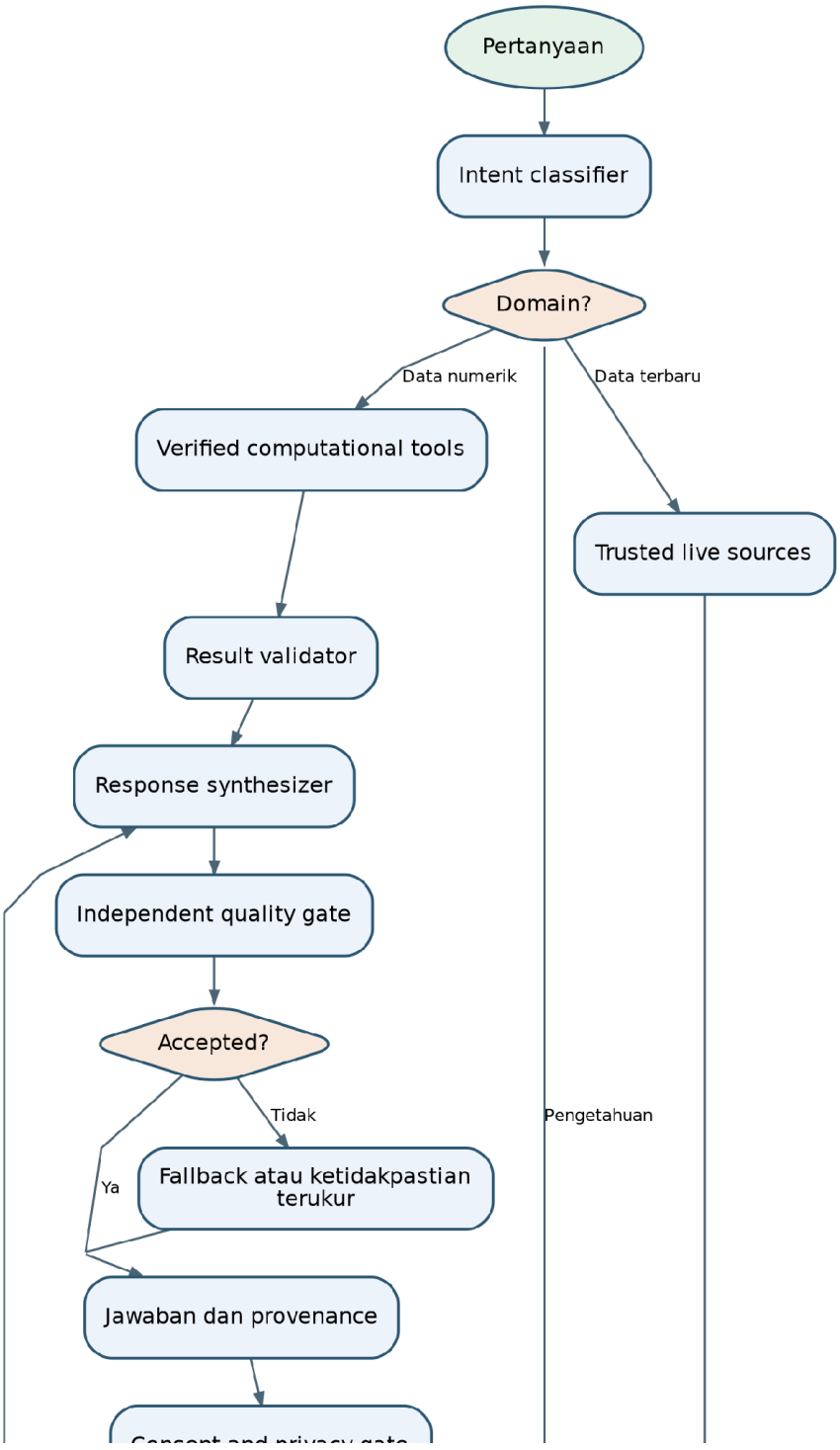
Audit sebelumnya melaporkan bahwa pemeriksaan php -l lulus tanpa kesalahan sintaks. Hasil tersebut adalah pemeriksaan sintaks statis, tidak membuktikan endpoint model tersedia, koneksi API berhasil, integritas database produksi terjaga, atau angka astronomi akurat. Pemindaian awal menyebut sekitar 13.857 baris berkas; angka deklarasi fungsi yang pernah dicatat menggunakan metode penghitungan berbeda (325 termasuk anonim; 274 bernama rag...) dan sebaiknya diperlakukan sebagai catatan pemindaian, bukan ukuran kompleksitas yang telah distandardisasi.

12. Arsitektur Penyempurnaan yang Direkomendasikan

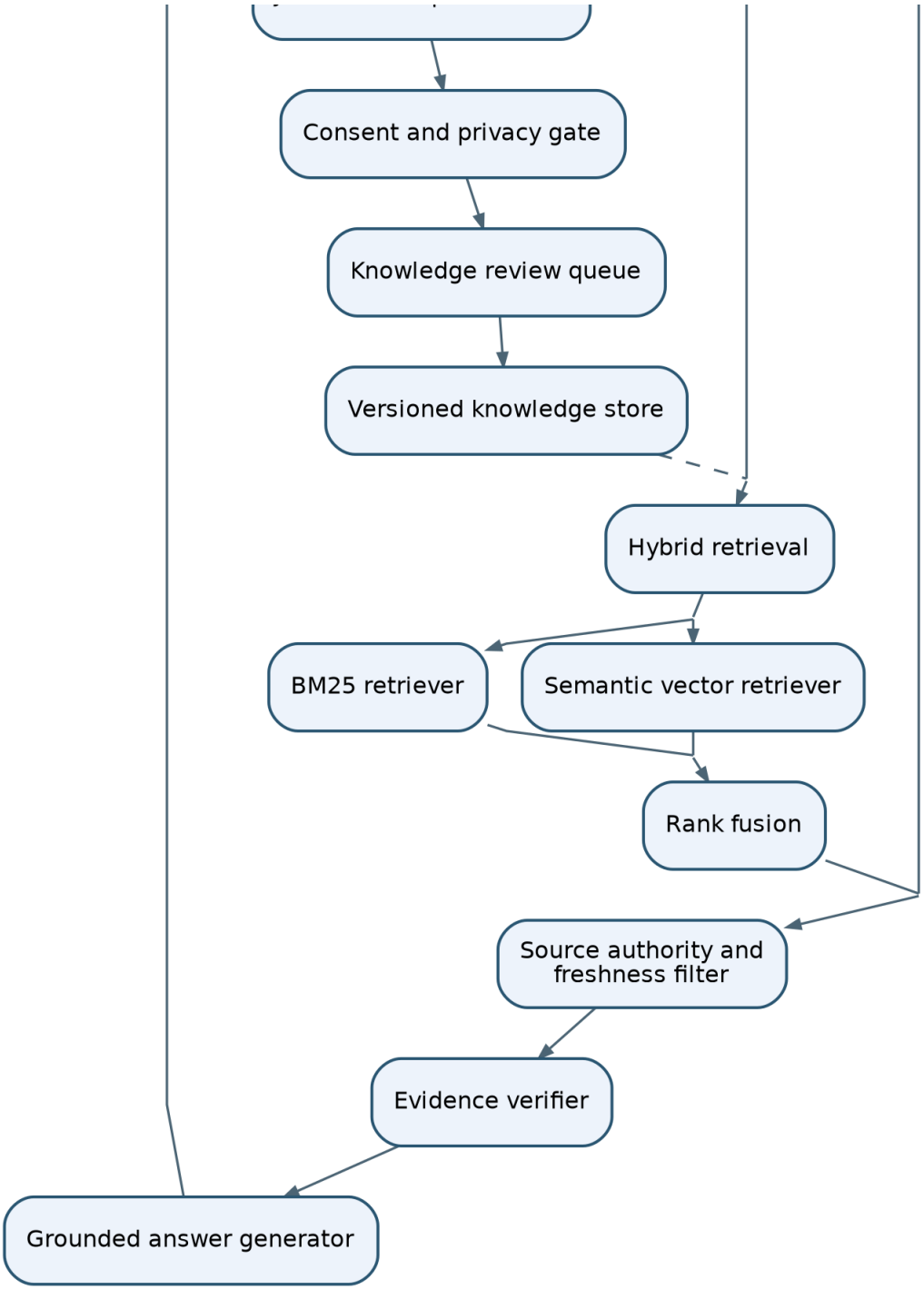


Gambar 8. Rancangan pengembangan Hybrid RAG dengan verifikasi bukti dan tata kelola pengetahuan.

Catatan: komponen vektor, rank fusion, review queue, dan versioned store merupakan usulan desain, bukan klaim implementasi aktif.



Detail Gambar 8, bagian 1/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).



Detail Gambar 8, bagian 2/2 (pembesaran alur diagram; diagram lengkap tetap tercantum di atas).

Pengembangan diarahkan menuju Hybrid RAG + Tool Routing + Evidence Verification + Governed Knowledge Growth. Prinsip utamanya adalah kombinasi retrieval leksikal dan semantik, pemeringkatan berdasarkan otoritas serta keaktualan sumber, gerbang kualitas independen, pelacakan provenance, dan kontrol siklus hidup pengetahuan bersama.

Tabel 9. Rekomendasi prioritas pengembangan.

Prioritas	Perbaikan	Manfaat
P0 — Kritis	Pencabutan persetujuan menghapus dokumen turunan	Perlindungan privasi.

Prioritas	Perbaikan	Manfaat
P0 — Kritis	Validasi ulang Q&A sekunder sebelum jawaban otoritatif	Mencegah penguatan kesalahan.
P1 — Tinggi	Quality gate seragam pada semua fallback	Konsistensi kualitas.
P1 — Tinggi	Indeks BM25 persisten dan retrieval vektor	Recall dan kecepatan.
P1 — Tinggi	Hierarki dokumen primer/ringkasan/arsip	Otoritas rujukan.
P1 — Tinggi	Uji SOLUNA dengan angka acuan	Validitas astronomi.
P2 — Menengah	Modularisasi file PHP	Pemeliharaan dan pengujian.
P2 — Menengah	Dashboard latency dan ketepatan jawaban	Evaluasi empiris.
P2 — Menengah	Versi pengetahuan dan supersession	Mengurangi informasi kedaluwarsa.

13. Pseudocode Algoritma RAGMINI yang Direkomendasikan

Pseudocode berikut mempertahankan tahapan yang sudah ada sambil mengusulkan retrieval vektor, penggabungan peringkat, dan antrean peninjauan pengetahuan. Fungsi yang belum aktif tidak boleh diartikan sebagai fitur yang telah tersedia dalam implementasi sekarang.

```
def answer_ragmini(question, user):
    # 1. Validasi permintaan
    validate_request(question, user)
    effective_question = resolve_followup(question, user)
    intent = classify_intent(effective_question)

    # 2. Jalur domain deterministik
    direct = query_verified_structured_data(intent)
    if direct.is_valid:
        return finalize(direct, user)

    # 3. Komputasi ilmiah
    if intent.requires_computation:
        computed = execute_verified_tool(intent)
        if computed.is_valid:
            return finalize(computed, user)
        return computation_unavailable()

    # 4. Cache yang masih sah
    cached = get_trusted_cache(effective_question)
    if cached and verify_freshness(cached):
        if verify_source_authority(cached):
            return finalize(cached, user)

    # 5. Hybrid retrieval
    documents = load_knowledge_base()
    lexical = bm25_search(effective_question, documents)
    semantic = vector_search(effective_question, documents)
    candidates = rank_fusion(lexical, semantic)

    # 6. Filter provenance dan waktu
    evidence = filter_sources(
        candidates, authority=True, freshness=True, validity=True
    )
    evidence = verify_evidence(effective_question, evidence)

    # 7. Generasi jawaban
    if evidence.is_sufficient:
        answer = generate_grounded_answer(
            question=effective_question, evidence=evidence
        )
        if verify_answer(answer, evidence):
            result = finalize(answer, user)
            if user.consent:
                proposed = sanitize_knowledge(result)
                if proposed.is_safe:
                    queue_knowledge_review(proposed)
            return result

    # 8. Fallback dengan validasi seragam
    for provider in ["ai_fikih", "google_ai_mode"]:
        result = request_provider(provider, effective_question)
        if result.is_available:
            if verify_external_answer(result, intent):
```

```
        return finalize(result, user)

    # 9. Gagal dengan aman
    return insufficient_evidence_response()
```

14. Rancangan Pengujian Empiris

Dokumen audit sumber belum mencakup benchmark penggunaan RAGMINI. Berikut disampaikan rancangan pengujian ilmiah awal berisi 200 pertanyaan yang dibagi atas fikih dan Tarjih, keislaman, KHGT dan falak, verifikasi klaim, pertanyaan lanjutan, informasi mutakhir, serta keamanan/privasi. Angka ini merupakan usulan dataset, bukan hasil pengujian empiris.

Tabel 10. Rancangan stratifikasi dataset evaluasi.

Kategori	Jumlah pertanyaan
Fikih dan Tarjih	50
Akidah dan keislaman	25
KHGT dan falak	40
Verifikasi klaim	30
Pertanyaan lanjutan	25
Data terbaru	20
Privasi dan batas sistem	10
Jumlah	200

Tabel 11. Indikator evaluasi yang direkomendasikan.

Dimensi	Metrik	Tujuan
Retrieval	Recall@12	Apakah bukti relevan ditemukan?
Retrieval	Precision@12	Ketepatan referensi yang dipilih.
Jawaban	Answer relevance	Kesesuaian pertanyaan dan jawaban.
Jawaban	Faithfulness	Dukungan klaim oleh bukti.
Sitasi	Citation accuracy	Kecocokan klaim-sitasi.
Kelengkapan	Answer completeness	Cakupan seluruh subpertanyaan.
Fallback	Fallback success rate	Keberhasilan layanan alternatif.
Privasi	Privacy leakage rate	Deteksi kebocoran informasi sensitif.
Komputasi	Numerical agreement	Kesamaan hasil dengan acuan.
Kinerja	p50/p95 latency	Distribusi waktu respons.
Efisiensi	Token/API cost per answer	Biaya per jawaban.
Pertumbuhan	Acceptance dan duplicate rate	Kualitas penambahan pengetahuan.

Pengujian astronomi wajib dipisahkan dari penilaian kefasihan bahasa. Setiap kasus perlu memuat tanggal, waktu, tempat, skala waktu, sistem koordinat, definisi geosentris atau toposentris, metode hisab, dan kriteria kalender. Variabel yang dapat dibandingkan: ijtimak, tinggi Bulan, elongasi, magrib, moonset, serta tanggal awal bulan menurut metode terkait.

15. Kesimpulan

Hasil pemetaan kode menunjukkan enam kemampuan arsitektural utama: (1) RAG leksikal dengan BM25, normalisasi dan reranking semantik; (2) orkestrasi layanan AI secara bertingkat; (3) evaluasi jawaban berbasis sumber; (4) pertumbuhan basis pengetahuan berdasarkan persetujuan; (5) rute komputasi astronomi melalui SOLUNA, KHGT, dan data rukyat; serta (6) observabilitas proses dan arsip. Secara akademik, RAGMINI dapat dirumuskan sebagai sistem kecerdasan buatan berbasis Hybrid Retrieval-Augmented Generation dengan orkestrasi layanan bertingkat, evaluasi jawaban

berbasis bukti, penambahan basis pengetahuan terkendali, dan integrasi komputasi astronomi serta pengetahuan keislaman.

> *Nilai utama RAGMINI bukan sekadar kemampuan menghasilkan jawaban, melainkan menggabungkan dokumen terkurasi, penalaran model bahasa, layanan eksternal, dan komputasi deterministik ke dalam alur yang dapat diperiksa.*

Pengembangan yang paling mendesak adalah penguatan otoritas dokumen primer, pemeriksaan ulang ringkasan sekunder, mekanisme penghapusan pengetahuan turunan setelah pencabutan persetujuan, penyeragaman quality gate, peningkatan indeks, serta pelaksanaan benchmark empiris. Dengan langkah tersebut, RAGMINI berpotensi menjadi objek penelitian yang terukur dalam pengembangan chatbot ilmiah untuk fikih, hisab-rukyat, dan Kalender Hijriah Global Tunggal.

Daftar Sumber dan Batas Rujukan

1. Kode sumber PHP RAGMINI. “Teks yang ditempel (1)(20261111-221032).txt”, berkas yang diberikan pengguna untuk audit statis (sumber primer analisis implementasi).
2. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33, 9459–9474. (Rujukan konseptual metode RAG).
3. Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. (Rujukan konseptual pemeringkatan BM25).
4. Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. arXiv:1901.04085. (Rujukan konseptual reranking; bukan klaim bahwa kode memakai BERT).

Catatan metodologis: literatur konseptual di atas tidak digunakan untuk mengklaim hasil pengujian RAGMINI. Bukti implementasi berasal dari kode; diagram penyempurnaan dan rancangan benchmark merupakan rekomendasi.

Lampiran A. Sumber Mermaid untuk Seluruh Flowchart

Untuk menjaga reproduktibilitas, setiap diagram juga dicantumkan dalam bentuk kode Mermaid. Gambar pada badan dokumen telah dirender dan diperiksa sebagai elemen visual, bukan hanya teks kode.

Lampiran A.1. Sumber diagram 1

```
flowchart TD
    A(["PENGGUNA RAGMINI"]) --> B["Antarmuka Web / JavaScript"]
    B --> C["PHP: validasi JSON, CSRF, rate limit"]
    C --> D["Resolusi pertanyaan lanjutan"]
    D --> E["Deteksi intensi dan perencanaan"]
    E --> F{"Jalur khusus tersedia?"}
    F -->|"Ya"| G["HPT, ICOP, kalender JSON, navigasi SOLUNA"]
    G --> Z["Penyusunan respons"]
    F -->|"Tidak"| H{"Permintaan komputasi SOLUNA?"}
    H -->|"Ya"| I["Eksekusi API / data astronomi"]
    I --> J{"Hasil valid?"}
    J -->|"Ya"| Z
    J -->|"Tidak, komputasi wajib"| K["Laporkan tidak tersedia"]
    K --> Z
    J -->|"Dapat menggunakan RAG"| L
    H -->|"Tidak"| L{"Cache atau tanya-jawab cocok?"}
    L -->|"Ya"| Z
    L -->|"Tidak"| M["Muat database-rag-*.json"]
    M --> N["Tokenisasi, chunking, BM25"]
    N --> O["Reranking semantik dan pemeriksaan bukti"]
    O --> P{"Bukti memadai?"}
    P -->|"Ya"| Q["Gemini: sintesis jawaban berbasis konteks"]
    Q --> R{"Sitasi dan kualitas lolos?"}
    R -->|"Ya"| S["Jawaban RAG tervalidasi otomatis"]
    S --> Z
    R -->|"Tidak"| T["AI Fikih"]
    P -->|"Tidak"| T
    T --> U{"Jawaban layak?"}
    U -->|"Ya"| Z
    U -->|"Tidak"| V["Google AI Mode melalui SerpAPI"]
    V --> W{"Jawaban tersedia?"}
    W -->|"Ya"| Z
    W -->|"Tidak"| X["Tautan pencarian Google AI Mode"]
    X --> Z
    Z --> Y["Arsip, pelacakan tahapan, hasil, referensi"]
    Y --> AA(["JAWABAN KE PENGGUNA"])
    Z -.-> AB["Pertimbangkan penambahan Pengetahuan Bersama"]
```

Lampiran A.2. Sumber diagram 2

```
flowchart TD
 A(["PRESENTATION LAYER: Web responsif; chat; status LIVE; arsip; referensi; RAGMINI-3D; shoutbox"]) -->
 B["APPLICATION & ORCHESTRATION LAYER: PHP 8.1+; sesi; validasi; intent; follow-up; router; trace"]
 B --> C["KNOWLEDGE LAYER: JSON multipart; arsip pribadi; indeks Q&A; cache; BM25; pengetahuan bersama"]
 B --> D["AI REASONING LAYER: Gemini; reranking; pemeriksaan kualitas; AI Fikih; SerpAPI"]
 B --> E["COMPUTATION LAYER: SOLUNA; KHGT; Neo-MABIMS; konversi kalender; ICOP; hisab astronomi"]
 B --> F["VALIDATION LAYER: Verifikasi sitasi; sumber; dalil; keaktualan; privasi; duplikasi"]
 C --> G["DATA & EXTERNAL SERVICE LAYER: Penyimpanan privat; API Gemini; AI Fikih; SerpAPI; sumber
Quran/Hadis; backend SOLUNA"]
 D --> G
 E --> G
 F --> G
```

Lampiran A.3. Sumber diagram 3

```
flowchart TD
 A(["Pertanyaan telah divalidasi"]) --> B["Penyelesaian konteks follow-up"]
 B --> C{"Tuntunan azan Tarjih yang dikurasi?"}
 C -->|Ya| X(["Jawaban langsung"])
 C -->|Tidak| D{"Pertanyaan rukyat ICOP?"}
```

```
D -->|"Ya, data ditemukan"| X
D -->|"Tidak / tidak ditemukan"| E{"Syawal, Zulhijah, Iduladha?"}
E -->|Ya| X
E -->|Tidak| F{"Tanggal 1 Ramadan?"}
F -->|Ya| X
F -->|Tidak| G{"Navigasi menu SOLUNA?"}
G -->|Ya| X
G -->|Tidak| H{"Komputasi SOLUNA berhasil?"}
H -->|"Ya, hasil langsung"| X
H -->|"Tidak / lanjut RAG"| I{"Cache jawaban terverifikasi?"}
I -->|Ya| X
I -->|Tidak| J{"Indeks Q&A persis?"}
J -->|Ya| X
J -->|Tidak| K{"Pengetahuan Bersama cocok?"}
K -->|Ya| X
K -->|Tidak| L["Ambil konteks komunitas berizin"]
L --> M{"Cache arsip pribadi diminta?"}
M -->|"Ya, cocok"| X
M -->|Tidak| N["Muat seluruh database"]
N --> O{"Q&A persis dalam database?"}
O -->|Ya| X
O -->|Tidak| P{"Tabel konversi kalender cocok?"}
P -->|Ya| X
P -->|Tidak| Q(["RETRIEVAL BM25 DAN AI"])
```

Lampiran A.4. Sumber diagram 4

```
flowchart TD
 A(["Pertanyaan pengguna"]) --> B["Normalisasi dan tokenisasi"]
 D[("database-rag-001.json s.d. 999")] --> C["Validasi dokumen dan metadata"]
 C --> E["Chunking dengan overlap"]
 E --> F["Hitung term frequency dan panjang chunk"]
 B --> G["BM25 lexical retrieval"]
 F --> G
 G --> H["Seleksi kandidat dokumen"]
 H --> I{"Ada bukti relevan?"}
 I -->|Tidak| X["Fallback AI Fikih"]
 I -->|Ya| J["Gemini semantic reranker"]
 J --> K["Periksa kutipan verbatim dengan PHP"]
 K --> L{"Dukungan sumber sah?"}
 L -->|Tidak| X
 L -->|Ya| M["Pilih maksimum 12 referensi"]
 M --> N["Bangun prompt + konteks RAG"]
 N --> O["Gemini generasi jawaban"]
 O --> P["Normalisasi dan validasi sitasi"]
 P --> Q["Pemeriksaan dalil jika diperlukan"]
 Q --> R["Gemini quality assessment"]
 R --> S{"Relevant + Complete + Faithful + Valid?"}
 S -->|Tidak| X
 S -->|Ya| T["Cache hasil lolos verifikasi"]
 T --> U["Pengetahuan baru jika memenuhi syarat"]
 U --> V(["Respons dan referensi"])
```

Lampiran A.5. Sumber diagram 5

```
flowchart TD
 A(["RAG lokal gagal / tidak memadai"]) --> B["Kirim pertanyaan ke AI Fikih"]
 B --> C["Endpoint hisabmu.com/aifikih"]
 C --> D{"Respons tersedia?"}
 D -->|Ya| E["Normalisasi jawaban dan pemeriksaan dalil"]
 E --> F["Pemeriksaan kualitas Gemini"]
 F --> G{"Lolos?"}
 G -->|Ya| H(["Jawaban AI Fikih"])
 G -->|Tidak| I["Google AI Mode via SerpAPI"]
 D -->|"Tidak / kuota habis"| I
 I --> J{"SerpAPI berfungsi?"}
 J -->|Ya| K["Ambil jawaban dan metadata referensi"]
 K --> L{"Klaim perlu verifikasi khusus?"}
 L -->|Ya| M["Periksa sumber resmi dan kualitas"]
 M --> N{"Terbukti memadai?"}
 N -->|Ya| O(["Jawaban Google AI Mode"])
 N -->|Tidak| P["Informasi belum terverifikasi"]
```

```
L -->|Tidak| O
J -->|Tidak| P
P --> Q(["Tampilkan tautan pencarian manual"])
H --> R["Arsip dan penambahan pengetahuan bersyarat"]
O --> R
R --> S(["Respons ke antarmuka"])
```

Lampiran A.6. Sumber diagram 6

```
flowchart TD
 A(["Jawaban berhasil dihasilkan"]) --> B["Simpan arsip pribadi"]
 B --> C{"Pengguna menyetujui Pengetahuan Bersama?"}
 C -->|Tidak| D(["Tetap arsip pribadi"])
 C -->|Ya| E{"Jawaban layak dan aman?"}
 E -->|Tidak| D
 E -->|Ya| F["Saring informasi pribadi dan data rahasia"]
 F --> G{"Lolos filter?"}
 G -->|Tidak| D
 G -->|Ya| H["Tentukan asal jawaban dan tipe sumber"]
 H --> I["Buat dokumen Q&A dengan metadata"]
 I --> J["Normalisasi pertanyaan dan ID hash"]
 J --> K{"Duplikat ada di database?"}
 K -->|Ya| L["Lewati penyimpanan"]
 K -->|Tidak| M["Kunci operasi penulisan"]
 M --> N["Periksa ulang duplikasi"]
 N --> O{"Masih unik?"}
 O -->|Tidak| L
 O -->|Ya| P{"Ukuran file kurang dari batas?"}
 P -->|Ya| Q["Tambahkan ke database-rag-009+"]
 P -->|Tidak| R["Buat database-rag-berikutnya.json"]
 Q --> S["Tulis ke file sementara dan rename"]
 R --> S
 S --> T["Database pengetahuan bertambah"]
 T --> U["Perbarui indeks saat diperlukan"]
 U --> V(["Siap digunakan untuk pertanyaan berikutnya"])
```

Lampiran A.7. Sumber diagram 7

```
flowchart TD
 A(["Pertanyaan astronomi"]) --> B["Deteksi intent SOLUNA"]
 B --> C{"Jenis permintaan?"}
 C -->|"Kalender dan konversi"| D["Cari JSON / arsip KHGT terstruktur"]
 C -->|"Hisab numerik"| E["Bangun parameter API SOLUNA"]
 C -->|ICOP| F["Baca data observasi ICOP"]
 C -->|"Cuaca hilal"| G["Baca endpoint cuaca SOLUNA"]
 C -->|Navigasi| H["Ambil tautan modul"]
 D --> I["Validasi tanggal, rentang, metode"]
 E --> J["Backend komputasi SOLUNA"]
 J --> K["Periksa status dan struktur hasil"]
 F --> L["Normalisasi data observasi"]
 G --> M["Label sumber meteorologi"]
 H --> N(["Tautan langsung"])
 I --> O{"Data valid?"}
 K --> O
 L --> O
 M --> O
 O -->|Ya| P["Hasil dengan provenance dan metode"]
 O -->|Tidak| Q["Laporkan keterbatasan / data tidak tersedia"]
 P --> R(["Respons pengguna"])
 Q --> R
```

Lampiran A.8. Sumber diagram 8

```
flowchart TD
 A(["Pertanyaan"]) --> B["Intent classifier"]
 B --> C{"Domain?"}
 C -->|"Data numerik"| D["Verified computational tools"]
 C -->|Pengetahuan| E["Hybrid retrieval"]
 C -->|"Data terbaru"| F["Trusted live sources"]
```

```
E --> G["BM25 retriever"]
E --> H["Semantic vector retriever"]
G --> I["Rank fusion"]
H --> I
I --> J["Source authority and freshness filter"]
F --> J
J --> K["Evidence verifier"]
K --> L["Grounded answer generator"]
D --> M["Result validator"]
M --> N["Response synthesizer"]
L --> N
N --> O["Independent quality gate"]
O --> P{"Accepted?"}
P -->|Ya| Q["Jawaban dan provenance"]
P -->|Tidak| R["Fallback atau ketidakpastian terukur"]
R --> Q
Q --> S["Consent and privacy gate"]
S --> T["Knowledge review queue"]
T --> U["Versioned knowledge store"]
U -.-> E
```


+ HCl
computation
computing
- KASMUI -
Dosen Kimia, Komputasi, IT, AI UNNES
https://hisabmu.com/
https://kasmui.cloud/
Ketua PCM Gunungpati 2
Anggota Majelis Tabligh PDM Kota Semarang & PWM Jawa Tengah
Tim Pengembang Software KHGT MTT PP Muhammadiyah
Praktisi Ilmu Falak