

PENGEMBANGAN CHATBOT RETRIEVAL-AUGMENTED GENERATION (RAG) UNTUK LAYANAN INFORMASI AKADEMIK DAN PEMERINTAHAN

Teguh Pribadi^{1*}, Arif Noor Iman²

Program Studi Teknik Informatika¹, Program Studi Sistem Informasi²

STMIK DCI¹, Universitas Ekuitas Indonesia²

Jl. Sutisna Senjaya No.158-A, Cikalang, Kec. Tawang, Kab. Tasikmalaya¹

Jl. Khp Hasan Mustopa No.31, Neglasari, Kec. Cibeunying Kaler, Kota Bandung, Jawa Barat 40124²

teguh.2022210063@gmail.com¹, arif.noor@ekuitas.ac.id²

Abstrak

Layanan akademik dan pemerintahan menghadapi kebutuhan informasi yang semakin tinggi, berulang, dan sering kali tersebar di berbagai dokumen resmi, portal web, surat edaran, regulasi, dan standar operasional prosedur. Chatbot konvensional berbasis aturan dapat menjawab pertanyaan sederhana, tetapi kurang adaptif terhadap variasi bahasa alami, perubahan kebijakan, serta kebutuhan jawaban yang membutuhkan bukti sumber. Artikel ini menyajikan perancangan framework chatbot berbasis Retrieval-Augmented Generation (RAG) untuk meningkatkan akurasi, keterlacakan, dan keamanan layanan informasi pada institusi pendidikan tinggi maupun instansi pemerintahan. Kerangka yang diusulkan mencakup tahap akuisisi dokumen, praproses, segmentasi teks, pembentukan *embedding*, penyimpanan pada *vector database*, *retrieval*, *reranking*, pembangkitan jawaban oleh *large language model*, pemberian sitasi sumber, *guardrail* keamanan, dan evaluasi berlapis. Metode penelitian menggunakan pendekatan *design science research* melalui analisis kebutuhan, perancangan arsitektur, penyusunan pipeline RAG, rancangan pengujian, serta pemetaan risiko dan mitigasi. Evaluasi yang disarankan meliputi *precision@k*, *recall@k*, *mean reciprocal rank*, *context precision*, *context recall*, *faithfulness*, *answer relevancy*, *answer correctness*, *response latency*, dan *System Usability Scale*. Hasil perancangan menunjukkan bahwa RAG chatbot berpotensi menjadi lapisan layanan informasi yang lebih akurat dan dapat diaudit dibanding chatbot generik, khususnya ketika sistem diwajibkan menjawab berdasarkan dokumen resmi yang

mutakhir. Kontribusi utama artikel ini adalah kerangka arsitektur dan evaluasi yang dapat dijadikan *blueprint* awal untuk implementasi RAG chatbot pada layanan akademik dan layanan publik digital. Penelitian selanjutnya diharapkan dapat melanjutkan *blueprint* ini ke tahap pengembangan prototipe, implementasi pada lingkungan layanan yang sebenarnya, serta pengujian empiris untuk menilai kinerja, keamanan, dan penerimaan pengguna.

Kata kunci :

Chatbot, RAG, Large Language Model, Vector Database.

Abstract

Academic and government services face an increasing demand for information that is repetitive and often dispersed across official documents, web portals, circular letters, regulations, and standard operating procedures. Conventional rule-based chatbots are able to answer simple questions, but they are less adaptive to variations in natural language, policy changes, and the need for answers supported by source evidence. This article presents the design of a Retrieval-Augmented Generation (RAG)-based chatbot framework to improve the accuracy, traceability, and security of information services in higher education institutions and government agencies. The proposed framework includes document acquisition, preprocessing, text segmentation, embedding generation, storage in a vector database, retrieval, reranking, answer generation by a large language model, source citation, security guardrails,

and multilayered evaluation. The research method adopts a design science research approach through needs analysis, architecture design, RAG pipeline development, testing design, and risk mapping with mitigation strategies. The recommended evaluation metrics include precision@k, recall@k, mean reciprocal rank, context precision, context recall, faithfulness, answer relevancy, answer correctness, response latency, and the System Usability Scale. The design results indicate that a RAG chatbot has the potential to serve as a more accurate and auditable information service layer than a generic chatbot, particularly when the system is required to generate answers based on up-to-date official documents. The main contribution of this article is an architectural and evaluation framework that can serve as an initial blueprint for implementing RAG chatbots in academic services and digital public services. Future research is expected to extend this blueprint to the prototype development stage, implement it in real-world service environments, and conduct empirical testing to evaluate its performance, security, and user acceptance.

Keywords:

Chatbot, RAG, Large Language Model, Vector Database.

I. PENDAHULUAN

Transformasi digital pada perguruan tinggi dan pemerintahan mendorong organisasi untuk menyediakan layanan informasi yang cepat, konsisten, dan tersedia secara berkelanjutan. Mahasiswa, calon mahasiswa, dosen, pegawai, dan masyarakat sering mengajukan pertanyaan yang berulang, misalnya terkait jadwal akademik, persyaratan administrasi, beasiswa, layanan surat, perizinan, alur pengaduan, atau prosedur layanan publik. Pada praktiknya, informasi tersebut tersebar di berbagai halaman web, dokumen PDF, surat edaran, SOP, peraturan internal, dan basis data layanan sehingga pencarian manual memerlukan waktu dan dapat menghasilkan interpretasi yang tidak seragam.

Chatbot menjadi salah satu solusi karena menyediakan antarmuka percakapan berbasis bahasa alami. Akan tetapi, chatbot berbasis aturan memiliki keterbatasan pada pola pertanyaan yang tidak eksplisit, sedangkan chatbot generatif murni berisiko menghasilkan jawaban yang tampak meyakinkan

tetapi tidak bersumber pada dokumen resmi. Dalam konteks layanan akademik dan pemerintahan, kesalahan informasi tidak hanya menurunkan kepuasan pengguna, tetapi juga dapat menimbulkan risiko administratif, hukum, reputasi, dan pelanggaran perlindungan data.

Retrieval-Augmented Generation (RAG) menawarkan pendekatan yang lebih sesuai untuk domain layanan resmi karena menggabungkan kemampuan retrieval terhadap basis pengetahuan eksternal dengan kemampuan *large language model* (LLM) dalam memahami pertanyaan dan menyusun jawaban natural. RAG pertama kali dipopulerkan sebagai arsitektur yang menggabungkan memori parametrik model bahasa dengan memori non-parametrik berupa korpus eksternal. Pendekatan ini relevan untuk layanan yang informasinya sering berubah karena basis pengetahuan dapat diperbarui tanpa harus melatih ulang model dari awal.

Pada layanan akademik, RAG chatbot dapat diarahkan untuk menjawab pertanyaan berdasarkan katalog akademik, pedoman kurikulum, kalender akademik, FAQ fakultas, prosedur administrasi, pedoman MBKM, dan dokumen layanan mahasiswa. Pada layanan pemerintahan, arsitektur yang sama dapat dipakai untuk membantu akses informasi terkait standar pelayanan, regulasi, persyaratan administrasi, kanal pengaduan, dan program bantuan publik. Di Indonesia, kebutuhan ini sejalan dengan arah Sistem Pemerintahan Berbasis Elektronik (SPBE) yang menekankan tata kelola layanan digital terpadu, sekaligus wajib memperhatikan prinsip perlindungan data pribadi.

Studi terdahulu memperlihatkan bahwa penerapan RAG pada pendidikan umumnya masih berfokus pada satu institusi atau satu jenis layanan. Neupane et al. (2024) mengembangkan chatbot untuk sumber daya universitas dan mengevaluasinya dengan RAGAS serta System Usability Scale; Shanthakumar et al. (2025) berfokus pada layanan university open day; sedangkan Khasanova Zafar kizi dan Suh (2025) membandingkan pipeline RAG untuk penerimaan mahasiswa internasional. Survei Swacha dan Gracel (2025) juga menunjukkan bahwa aplikasi RAG pendidikan telah berkembang untuk beragam tujuan, tetapi implementasinya cenderung spesifik pada konteks dan kebutuhan tertentu.

Pada ranah pemerintahan, Xu et al. (2024) membatasi penerapan RAG pada akses informasi

anggaran pemerintah kota, sedangkan Kusuma et al. (2024) mengembangkan Trisurya untuk meningkatkan akses layanan publik multibahasa di Indonesia. Literatur tersebut membuktikan kelayakan RAG, tetapi studi pendidikan dan pemerintahan masih banyak berjalan sebagai implementasi vertikal yang terpisah. Research gap yang diangkat artikel ini adalah masih terbatasnya kerangka lintas domain untuk layanan informasi resmi yang secara bersamaan mengintegrasikan versioning dokumen, metadata hak akses, sitasi sumber, eskalasi manusia, audit log, serta evaluasi retrieval, generation, usability, operasional, dan governance. Artikel ini mengisi celah tersebut melalui blueprint konseptual yang dapat diadaptasi pada perguruan tinggi dan instansi pemerintahan.

Masalah utama yang dibahas dalam artikel ini adalah bagaimana merancang RAG chatbot yang tidak hanya mampu menjawab secara natural, tetapi juga mampu membatasi jawaban pada dokumen resmi, menampilkan rujukan sumber, menolak pertanyaan di luar cakupan, dan menyediakan mekanisme audit. Dengan demikian, tujuan jurnal ini adalah menyusun kerangka pengembangan RAG chatbot untuk layanan akademik dan pemerintahan, menjelaskan arsitektur teknis, menyusun pipeline data, menentukan skenario evaluasi, serta memetakan risiko keamanan dan tata kelola.

Kontribusi jurnal ini meliputi tiga hal. Pertama, jurnal ini menyajikan rancangan arsitektur RAG chatbot yang dapat diadaptasi pada lingkungan perguruan tinggi maupun instansi publik. Kedua, jurnal ini merumuskan pipeline operasional dari dokumen mentah hingga jawaban bersitasi. Ketiga, jurnal ini menyusun rancangan evaluasi multidimensi yang mencakup kualitas retrieval, kualitas pembangkitan jawaban, pengalaman pengguna, kinerja sistem, serta kepatuhan terhadap tata kelola data. Secara ilmiah, kebaruan artikel ini bukan pada pengenalan komponen RAG secara terpisah, melainkan pada integrasi lintas domain antara manajemen pengetahuan resmi, kontrol keamanan dan tata kelola, keterlacakan jawaban, serta evaluasi berlapis dalam satu blueprint yang dapat diaudit.

II. METODE PENELITIAN

Artikel ini menggunakan pendekatan *design science research* karena fokus utamanya adalah menghasilkan rancangan konseptual berupa kerangka arsitektur, pipeline, dan model evaluasi sistem.

Tahapan penelitian terdiri atas identifikasi masalah, studi literatur, analisis kebutuhan layanan, perancangan arsitektur, perancangan pipeline RAG, perancangan skenario pengujian, serta pemetaan risiko dan mitigasi. Pendekatan ini sesuai untuk penelitian Sistem Informasi dan Teknik Informatika yang bertujuan membangun rancangan solusi terhadap masalah organisasi. Dengan demikian, luaran penelitian dibatasi pada rancangan framework dan mockup interface sebagai tahap awal dari konsep yang dirancang, belum sampai prototipe aplikasi yang telah di implementasikan.

Objek layanan yang dimodelkan adalah layanan informasi akademik dan layanan informasi pemerintahan. Pada layanan akademik, contoh sumber data meliputi kalender akademik, buku pedoman akademik, kurikulum, SOP administrasi, panduan beasiswa, pedoman tugas akhir, dan FAQ fakultas. Pada layanan pemerintahan, contoh sumber data meliputi standar pelayanan, persyaratan administrasi, regulasi, jadwal layanan, kanal pengaduan, dan panduan program publik. Seluruh dokumen yang dimasukkan ke basis pengetahuan harus bersifat resmi, memiliki versi, tanggal berlaku, unit pemilik, dan status validitas.

Rancangan sistem tidak mengasumsikan bahwa LLM memiliki kewenangan untuk membuat keputusan administratif. Chatbot hanya berperan sebagai lapisan informasi awal yang menjawab berdasarkan sumber resmi. Untuk layanan yang membutuhkan keputusan, verifikasi, atau pemrosesan data pribadi, sistem harus mengarahkan pengguna ke petugas atau sistem transaksi resmi. Pembatasan ini penting agar chatbot tidak berubah menjadi pengambil keputusan otomatis tanpa dasar hukum, otorisasi, dan akuntabilitas yang jelas.

Tabel 1. Sumber Data Dan Metadata Basis Pengetahuan

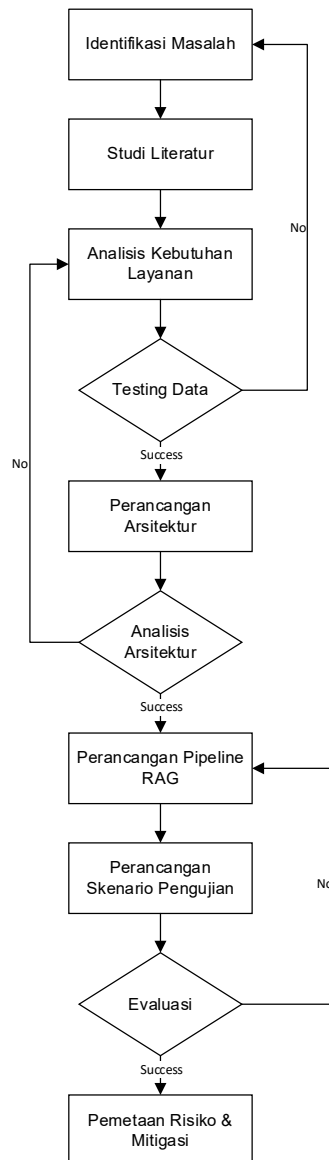
Jenis Dokumen	Contoh Isi	Metadata Wajib
Dokumen akademik	Pedoman kurikulum, kalender, SOP	Unit, tahun, versi, tanggal berlaku
Dokumen layanan publik	Standar pelayanan, syarat administrasi	Instansi, dasar hukum, status, kanal resmi
FAQ resmi	Pertanyaan berulang dan	Topik, pemilik, tanggal revisi

Regulasi	jawaban baku Peraturan, surat edaran, kebijakan	Nomor, tahun, hierarki, masa berlaku
Konten portal	Halaman web layanan	URL, waktu crawl, checksum
Dokumen akademik	Pedoman kurikulum, kalender, SOP	Unit, tahun, versi, tanggal berlaku

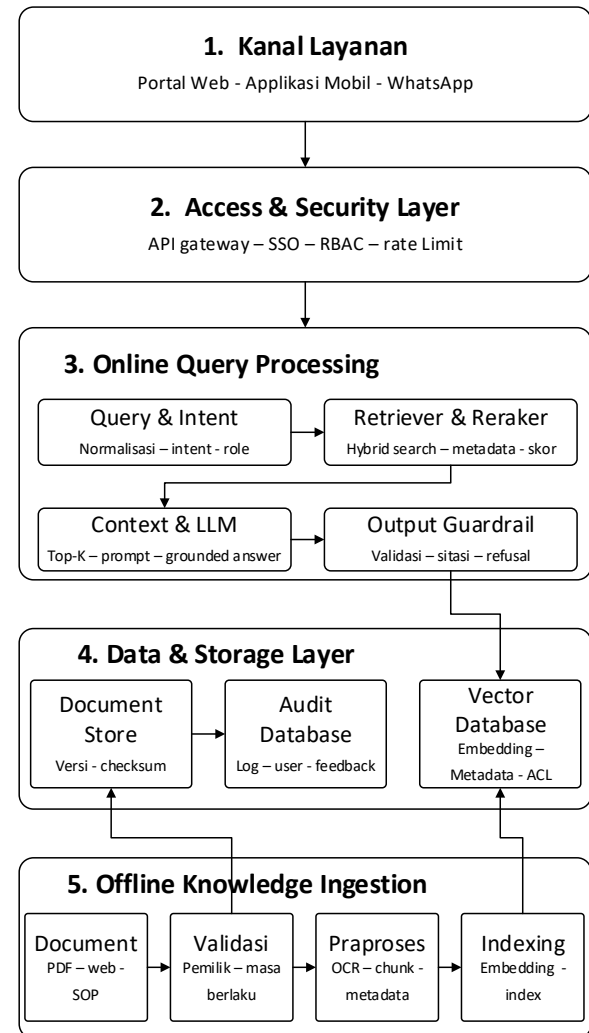
III. ANALISIS DAN PERANCANGAN

III.1 Rancangan Framework RAG Chatbot

Framework yang diusulkan menyatukan jalur pengelolaan pengetahuan offline, pemrosesan pertanyaan online, penyimpanan data, keamanan akses, dan tata kelola dalam lima lapisan yang saling terhubung. Rancangan ini ditempatkan sebagai blueprint konseptual yang dapat diadaptasi pada layanan akademik dan pemerintahan.



Gambar 1. Flowchart Metode Penelitian



Gambar 2. Rancangan framework RAG Chatbot Flowchart Metode Penelitian

Lapisan kanal layanan menyediakan interaksi pengguna. Lapisan access and security menangani autentikasi, otorisasi, pembatasan permintaan, dan pencatatan identitas sesi. Lapisan online query

processing menjalankan normalisasi pertanyaan, retrieval, reranking, penyusunan konteks, pembangkitan jawaban, validasi, sitasi, serta refusal ketika bukti tidak memadai.

Lapisan data and storage menyimpan dokumen, embedding, metadata, audit log, dan feedback. Jalur offline knowledge ingestion memastikan dokumen telah divalidasi, memiliki versi, masa berlaku, unit pemilik, serta hak akses sebelum diindeks. Privacy-by-design, security-by-design, provenance, monitoring, dan human escalation menjadi kontrol lintas lapisan.

III.2 Alur Pipeline RAG

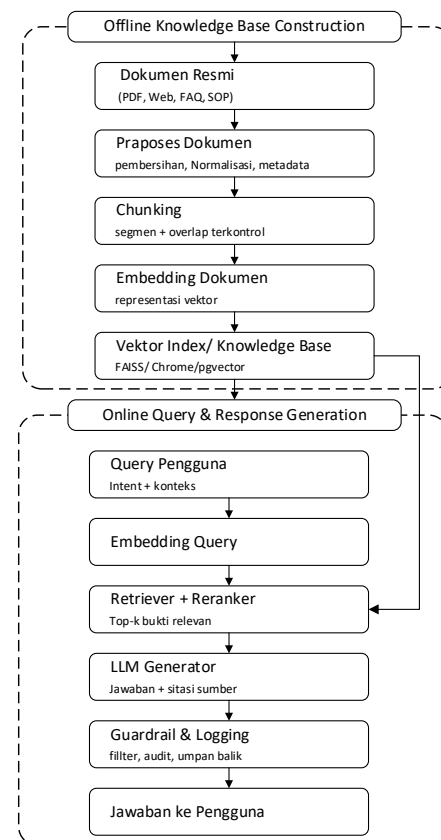
Pipeline RAG yang diusulkan dimulai dari pengumpulan dokumen resmi. Dokumen kemudian dibersihkan dari elemen *noise*, seperti *header* berulang, nomor halaman, navigasi web, dan teks duplikat. Selanjutnya, teks disegmentasi ke dalam unit-unit teks yang distandarisasi dan disertai mekanisme *overlap* untuk mempertahankan kesinambungan konteks antarsegmen. Setiap segmen teks diberi metadata agar jawaban dapat ditelusuri kembali ke sumbernya. *Embedding* dibuat untuk setiap segmen teks dan disimpan pada vector database. Ketika pengguna mengajukan pertanyaan, sistem mengubah pertanyaan menjadi *embedding*, mencari segmen teks yang relevan, melakukan *reranking*, dan menyusun prompt yang berisi pertanyaan, konteks, instruksi, serta aturan penolakan apabila konteks tidak memadai.

Jawaban yang dihasilkan harus memuat ringkasan informasi, rujukan sumber, dan batasan jawaban. Untuk pertanyaan yang memerlukan data pribadi atau tindakan administratif, chatbot tidak boleh meminta data sensitif secara bebas, tetapi mengarahkan pengguna ke kanal layanan resmi yang telah diautentikasi. Jika skor retrieval rendah atau dokumen yang ditemukan tidak relevan, sistem harus memberikan jawaban penolakan yang informatif, misalnya menyatakan bahwa informasi tidak ditemukan pada basis pengetahuan resmi.

III.3 Arsitektur Logis

Arsitektur sistem terdiri atas lapisan antarmuka, API gateway, layanan autentikasi, chat service, retriever, generator, guardrail, vector database, relational database, dan ingestion service. Lapisan antarmuka dapat berupa portal web kampus, aplikasi

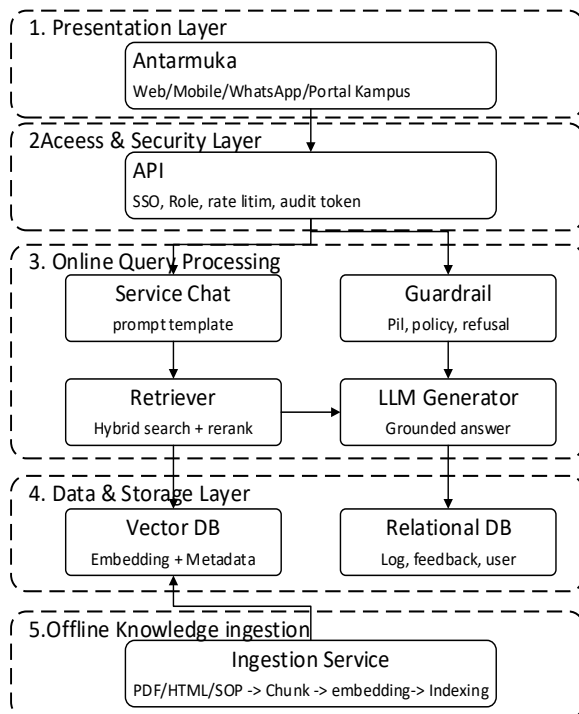
mobile, sistem layanan publik, atau kanal percakapan seperti WhatsApp. API gateway menangani autentikasi, pembatasan permintaan, logging, dan pemisahan hak akses. Retriever menjalankan pencarian semantik atau hybrid search, sedangkan generator menyusun jawaban berdasarkan konteks yang diberikan.



Gambar 3. Flowchart pipeline RAG chatbot untuk layanan informasi resmi

Kontrak komponen menetapkan bahwa retriever mengembalikan segmen beserta metadata dan skor, reranker mengurutkan bukti berdasarkan relevansi serta validitas sumber, generator hanya menyusun jawaban dari konteks terpilih, dan output guardrail memvalidasi faithfulness, data pribadi, sitasi, refusal, serta kebutuhan eskalasi. Document store, vector database, dan relational database mempertahankan versioning, provenance, kontrol akses, audit log, dan feedback.

Komponen *guardrail* menjadi penting karena domain layanan resmi membutuhkan pembatasan. *Guardrail* dapat memeriksa apakah pertanyaan terkait layanan yang tersedia, apakah konteks yang diambil cukup, apakah jawaban mengandung data pribadi, serta apakah terdapat indikasi *prompt injection*. Dalam implementasi yang lebih matang, prompt template harus menginstruksikan model agar tidak menjawab di luar dokumen resmi, tidak membuat prosedur baru, tidak menyimpulkan kebijakan yang tidak disebutkan dokumen, dan selalu menampilkan sumber rujukan.



Gambar 4. Arsitektur logis RAG chatbot untuk layanan akademik dan pemerintahan

III.4 Arsitektur Logis

Algorithm 1. RAG Chatbot untuk Layanan Informasi Resmi

PROGRAM Sistem_Tanya_Jawab_Grounded

```

=====
Nama Program : Sistem_Tanya_Jawab_Grounded
Deskripsi    : Memproses pertanyaan pengguna dengan
               metode pencarian dokumen berbasis
               vector database dan

```

```

menghasilkan jawaban yang dilengkapi
sumber.
Input        : Pertanyaan pengguna (q) dan profil
               akses (u)
Output       : Jawaban dengan sumber (a) atau pesan
               penolakan
=====
=====

```

```

q           : string
q_normal    : string
u           : ProfilPengguna
intent      : string
e_q         : Vector
D_k         : ListDokumen
D_ranked    : ListDokumen
konteks     : ListDokumen
prompt      : string
a_awal     : string
a           : string
sumber      : ListSumber
statusAkses : boolean
hasilValidasi : boolean
skorTertinggi : real

```

AMBANG_RELEVANSI : real ← 0.70

```

BEGIN
  INPUT (q)
  INPUT (u)

  {1. Memvalidasi identitas dan hak akses
  pengguna}
  statusAkses ← ValidasiIdentitasDanAkses(u)

  IF statusAkses = FALSE THEN
    a ← "Permintaan ditolak: identitas atau hak
    akses tidak valid."
    SimpanLogAudit(u, q, a)
    OUTPUT(a)
    STOP
  ENDIF

  {2. Menormalisasi pertanyaan dan mendeteksi
  intent layanan}
  q_normal ← Normalisasi(q)
  intent ← DeteksiIntent(q_normal)

  IF intent = "TIDAK_DIDUKUNG" THEN
    a ← "Permintaan ditolak: intent tidak
    didukung oleh layanan."
    SimpanLogAudit(u, q_normal, a)
    OUTPUT(a)
    STOP
  ENDIF

  {3. Membuat embedding dari pertanyaan}
  e_q ← BuatEmbedding(q_normal)

  {4. Mengambil kandidat dokumen dari vector
  database}
  D_k ← AmbilKandidatDokumen(e_q)

  IF D_k = KOSONG THEN
    a ← "Permintaan ditolak: tidak ditemukan
    dokumen yang relevan."
    SimpanLogAudit(u, q_normal, a)
    OUTPUT(a)
    STOP
  ENDIF

  {5. Melakukan reranking berdasarkan relevansi
  dan metadata}
  D_ranked ← Reranking(D_k, q_normal)
  skorTertinggi ← AmbilSkorTertinggi(D_ranked)

```

```

{6. Memeriksa skor relevansi dokumen}
IF skorTertinggi < AMBANG_RELEVANSI THEN
  a ← "Permintaan ditolak: tingkat relevansi
sumber terlalu rendah."
  SimpanLogAudit(u, q_normal, a)
  OUTPUT(a)
  STOP
ENDIF

{7. Menyusun prompt untuk LLM}
konteks ← AmbilKonteksTerbaik(D_ranked)
prompt ← SusunPrompt(
  "Jawab berdasarkan konteks yang
diberikan",
  konteks,
  q_normal
)

{8. Memanggil LLM untuk menghasilkan jawaban
grounded}
a_awal ← PanggilLLM(prompt)

{9. Memvalidasi jawaban dan melampirkan sumber}
hasilValidasi ← ValidasiOutput(a_awal, konteks)

IF hasilValidasi = FALSE THEN
  a ← "Jawaban ditolak karena tidak sesuai
dengan sumber yang tersedia."
  SimpanLogAudit(u, q_normal, a)
  OUTPUT(a)
  STOP
ENDIF

sumber ← EkstrakSumber(konteks)
a ← LampirkanSumber(a_awal, sumber)

{Menyimpan aktivitas ke dalam log audit}
SimpanLogAudit(u, q_normal, intent, sumber, a)

OUTPUT(a)

END

```

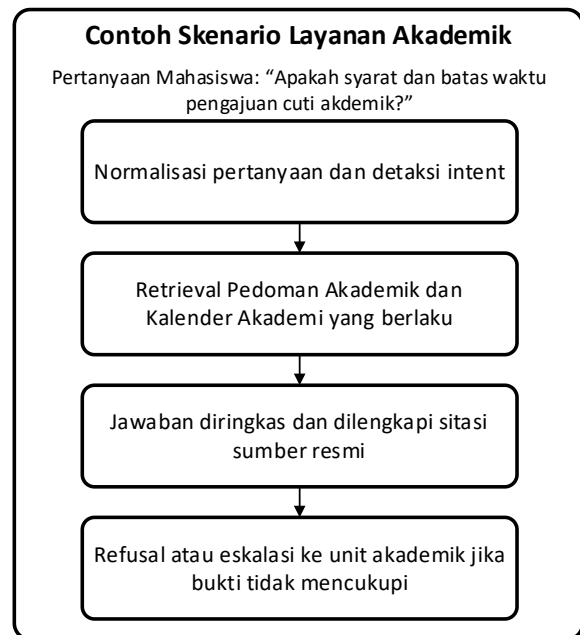
Tabel 2. Kebutuhan Fungsional Dan Non-Fungsional Sistem

Kategori	Kebutuhan	Keterangan
Fungsional	Tanya jawab berbasis dokumen	Jawaban hanya berasal dari sumber resmi
Fungsional	Sitasi sumber	Menampilkan judul, unit, dan tautan dokumen
Fungsional	Penolakan terkontrol	Menolak jika konteks tidak cukup
Fungsional	Feedback pengguna	Like/dislike, koreksi, eskalasi
Non-fungsional	Keamanan	Autentikasi, otorisasi, audit log
Non-fungsional	Kinerja	Respons cepat pada jam layanan tinggi
Non-fungsional	Kepatuhan	Sejalan dengan PDP, SPBE, dan kebijakan institusi
Fungsional	Tanya jawab	Jawaban hanya berasal

	berbasis dokumen	dari sumber resmi
Fungsional	Sitasi sumber	Menampilkan judul, unit, dan tautan dokumen

III.5 Ilustrasi Implementasi Konseptual Dan Skenario Penggunaan

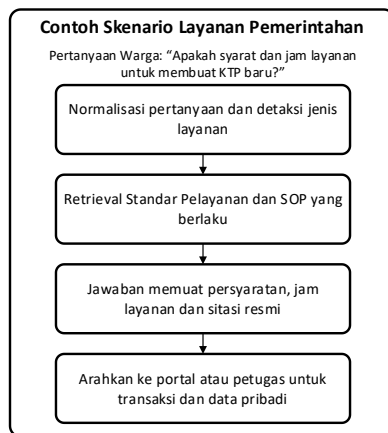
Untuk memperjelas penerapan rancangan, dua contoh berikut disusun sebagai ilustrasi implementasi konseptual. Contoh tersebut bukan hasil implementasi atau pengujian empiris, tetapi memetakan aliran pertanyaan, sumber yang diambil, bentuk jawaban, serta kondisi penolakan atau eskalasi.



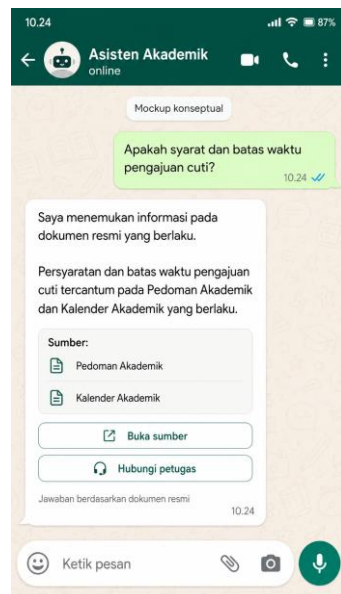
Gambar 5. Ilustrasi Skenario Penggunaan RAG Chatbot Pada Layanan Akademik

Pada skenario layanan akademik, mahasiswa mengajukan pertanyaan: "Apa syarat dan batas waktu pengajuan cuti akademik?" Sistem menormalisasi pertanyaan, mendeteksi intent cuti akademik, lalu mengambil segmen dari Pedoman Akademik dan Kalender Akademik yang metadata versinya masih berlaku. Reranker memilih bukti yang paling relevan. LLM kemudian menyusun ringkasan persyaratan dan batas waktu hanya dari konteks terpilih serta menampilkan judul dokumen, unit pemilik, dan tautan sumber. Jika dokumen berbeda versi atau bukti tidak memadai, chatbot tidak menyimpulkan sendiri dan mengarahkan mahasiswa ke unit akademik.

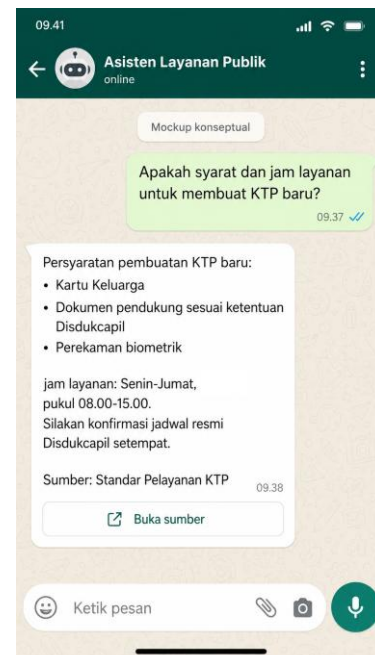
Pada skenario layanan pemerintahan, warga mengajukan pertanyaan: "Apa persyaratan dan jam layanan untuk pengajuan dokumen administrasi?" Sistem mengambil Standar Pelayanan dan SOP dari instansi yang sesuai, memeriksa status berlaku, lalu menyajikan daftar persyaratan, jam layanan, kanal resmi, dan sitasi sumber. Jika pengguna meminta pengajuan transaksi, pemeriksaan status permohonan, atau pemrosesan data pribadi, chatbot mengarahkan pengguna ke portal yang telah diautentikasi atau kepada petugas. Chatbot tetap berfungsi sebagai pemberi informasi dan tidak mengambil keputusan administratif.



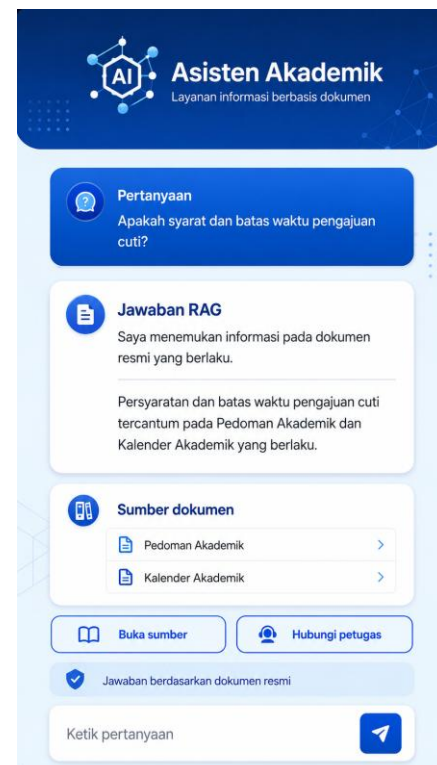
Gambar 6. Ilustrasi Skenario Penggunaan RAG Chatbot Pada Layanan Pemerintah



Gambar 7. Ilustrasi Penggunaan RAG Chatbot Pada Layanan Akademik pada WhatsApp



Gambar 8. Ilustrasi Penggunaan RAG Chatbot Pada Layanan Pemerintah pada WhatsApp



Gambar 9. Ilustrasi Penggunaan RAG Chatbot Pada Layanan Akademik pada Aplikasi Akademis



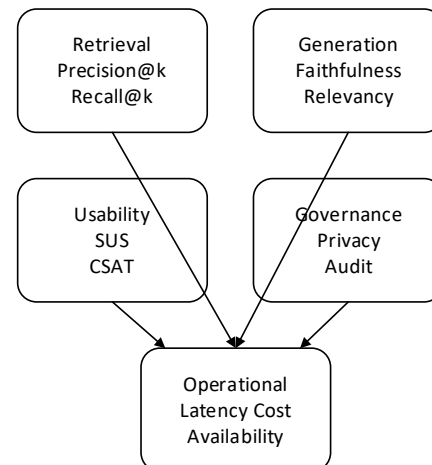
Gambar 10. Ilustrasi Penggunaan RAG Chatbot Pada Layanan Pemerintah pada Apps Pelayanan Publik

Kedua skenario tersebut mengoperasionalkan framework melalui jalur offline untuk validasi dan pengindeksan dokumen, serta jalur online untuk retrieval, reranking, pembangkitan jawaban, guardrail, sitasi, refusal, dan eskalasi. Setiap skenario juga dapat diturunkan menjadi kasus uji yang memuat dokumen relevan, jawaban ideal, keputusan penolakan, kelengkapan sitasi, dan catatan audit. Dengan demikian, ilustrasi ini memperjelas penerapan rancangan tanpa menyatakan bahwa sistem telah diimplementasikan atau diuji secara empiris.

III.6 Valuasi Dan Skenario Pengujian

Evaluasi RAG chatbot harus dilakukan secara berlapis karena kualitas sistem tidak dapat diwakili oleh satu metrik. Retrieval perlu diuji untuk memastikan dokumen relevan muncul pada posisi teratas. Generation perlu diuji untuk memastikan jawaban setia pada konteks. Usability perlu diuji agar sistem mudah digunakan oleh mahasiswa, pegawai, atau masyarakat. Selain itu, aspek operasional seperti response latency, *availability*, biaya inferensi, serta keamanan data perlu diukur karena sistem akan digunakan dalam layanan nyata.

Dataset evaluasi dapat dibuat dari daftar pertanyaan yang sering diajukan, simulasi skenario layanan, dan pertanyaan adversarial. Pertanyaan dikategorikan menjadi *factual*, *procedural*, *multi-hop*, *out-of-scope*, *policy-sensitive*, dan *privacy-sensitive*. Untuk setiap pertanyaan, evaluator manusia menyusun jawaban ideal dan sumber dokumen yang benar. Selanjutnya, sistem diuji dengan membandingkan dokumen yang diambil, kualitas jawaban, dan kesesuaian terhadap kebijakan layanan.



Gambar 11. Dimensi Evaluasi RAG Chatbot

Skenario pengujian perlu memuat pertanyaan yang mudah serta pertanyaan yang berisiko. Pertanyaan mudah misalnya jadwal registrasi, persyaratan surat aktif kuliah, atau jam layanan. Pertanyaan multi-hop memerlukan penggabungan beberapa dokumen, misalnya hubungan antara prasyarat mata kuliah dan ketentuan pengambilan SKS. Pertanyaan *out-of-scope* menguji kemampuan sistem menolak permintaan di luar basis pengetahuan. Pertanyaan *privacy-sensitive* menguji apakah sistem menolak permintaan data pribadi pengguna lain. Pertanyaan *prompt injection* menguji apakah sistem tetap mematuhi instruksi sistem walaupun pengguna mencoba memaksa model mengabaikan dokumen resmi.

Tabel 3. Rancangan Metrik Evaluasi Rag Chatbot

Dimensi	Metrik	Tujuan
Retrieval	Precision@k	Mengukur proporsi konteks relevan pada top-k
Retrieval	Recall@k	Mengukur cakupan bukti relevan yang ditemukan

Retrieval	MRR	Mengukur posisi dokumen benar pertama
Generation	Faithfulness	Mengukur kesetiaan jawaban terhadap konteks
Generation	Answer relevancy	Mengukur kesesuaian jawaban dengan pertanyaan
Generation	Answer correctness	Mengukur kedekatan dengan jawaban ideal
Usability	SUS dan kepuasan pengguna	Mengukur kemudahan dan penerimaan
Operasional	Latency dan availability	Mengukur kelayakan operasional layanan
Governance	Auditability	Mengukur keterlacakan jawaban dan log

III.7 Keamanan Data Dan Tata Kelola

RAG chatbot untuk layanan akademik dan pemerintahan harus dirancang dengan prinsip *privacy-by-design* dan *security-by-design*. Dokumen yang mengandung data pribadi tidak boleh langsung dimasukkan ke basis pengetahuan terbuka tanpa proses klasifikasi dan masking. Metadata hak akses perlu diterapkan pada setiap segmen teks sehingga pengguna hanya menerima konteks yang sesuai dengan perannya. Misalnya, mahasiswa hanya dapat memperoleh informasi umum dan informasi pribadinya sendiri setelah autentikasi, sedangkan petugas tertentu dapat mengakses panduan operasional internal sesuai kewenangan.

Risiko utama dalam RAG chatbot meliputi kebocoran data pribadi, *prompt injection*, *retrieval* terhadap dokumen usang, jawaban tanpa sumber, ketergantungan pengguna terhadap AI, dan bias informasi. Risiko tersebut dapat dikurangi melalui validasi dokumen, *versioning*, *access control*, *content moderation*, *audit log*, *human escalation*, dan *review* berkala. Kerangka NIST AI RMF serta ISO/IEC 23894:2023 dapat digunakan untuk mengatur proses identifikasi, pengukuran, dan mitigasi risiko AI.

Pada konteks Indonesia, sistem perlu memperhatikan prinsip SPBE dan perlindungan data pribadi. SPBE menuntut layanan digital yang terpadu, efektif, efisien, dan akuntabel. Sementara itu, Undang-Undang Pelindungan Data Pribadi menekankan kewajiban pengendali data untuk melindungi keamanan data pribadi dalam pemrosesan.

Oleh karena itu, RAG chatbot tidak boleh dipandang sebagai fitur tambahan semata, melainkan bagian dari tata kelola layanan digital institusi.

Tabel 4. Risiko Dan Mitigasi Pada Rag Chatbot Layanan Resmi

Risiko	Dampak	Mitigasi
Prompt injection	Model mengabaikan instruksi	System prompt kuat, input filtering, output validation
Dokumen usang	Jawaban tidak berlaku	Versioning, status berlaku, review berkala
Kebocoran data pribadi	Pelanggaran privasi	Access control, masking, minimisasi data
Retrieval salah	Jawaban tidak relevan	Reranking, threshold, feedback loop
Jawaban tanpa sumber	Sulit diaudit	Citation wajib dan refusal policy
Bias layanan	Ketidakadilan pengguna	Audit pertanyaan, monitoring, kanal eskalasi

III.8 Posisi Penelitian Dan Kontribusi Ilmiah

Dibandingkan dengan penelitian sebelumnya, yang memberikan bukti implementasi pada satu universitas dan skenario layanan tertentu, kerangka dalam artikel ini menempatkan pengelolaan versi dokumen, metadata kewenangan, sitasi, refusal policy, dan audit log sebagai persyaratan arsitektural sejak tahap ingestion hingga pembangkitan jawaban. Dibandingkan studi Khasanova Zafar kizi dan Suh (2025) yang berfokus pada perbandingan pipeline untuk penerimaan mahasiswa internasional, cakupan artikel ini diperluas pada kumpulan dokumen akademik dan layanan publik yang berbeda tingkat risiko serta otorisasinya.

Pada konteks pemerintahan, kontribusi artikel ini adalah menyatukan kebutuhan yang dapat dipindahkan antar-domain: provenance dan masa berlaku sumber, kontrol akses berbasis metadata, penolakan ketika bukti tidak memadai, eskalasi ke petugas, dan evaluasi multidimensi. Dengan demikian, kontribusinya bersifat integratif dan konseptual, bukan klaim keunggulan empiris; nilai ilmiahnya terletak pada rancangan yang dapat diuji dan diaudit sebagai dasar implementasi lanjutan di lingkungan layanan resmi.

Rancangan RAG chatbot memberikan beberapa manfaat strategis. Pertama, sistem dapat mengurangi beban pertanyaan berulang pada petugas layanan karena pertanyaan umum dijawab secara otomatis. Kedua, pengguna memperoleh akses informasi yang lebih cepat dan natural tanpa harus menelusuri banyak dokumen. Ketiga, organisasi tetap dapat menjaga kontrol terhadap sumber jawaban karena basis pengetahuan berasal dari dokumen resmi. Keempat, jawaban dapat dilengkapi dengan sumber sehingga pengguna dan auditor dapat memverifikasi informasi yang diberikan.

Dibandingkan chatbot berbasis aturan, RAG chatbot lebih fleksibel terhadap variasi bahasa pengguna. Dibandingkan LLM generik, RAG chatbot lebih sesuai untuk domain resmi karena jawabannya dapat dibatasi pada dokumen yang ditentukan. Namun, keunggulan tersebut hanya muncul apabila korpus dikelola dengan baik. RAG tidak otomatis menjamin jawaban benar apabila dokumen tidak lengkap, segmen teks buruk, embedding tidak sesuai bahasa/domain, atau prompt tidak membatasi perilaku model. Dengan demikian, kualitas sistem sangat dipengaruhi oleh manajemen pengetahuan institusi.

Untuk layanan akademik, implementasi awal sebaiknya dimulai dari domain informasi rendah risiko seperti FAQ administrasi, kalender akademik, prosedur umum, dan panduan layanan mahasiswa. Setelah kualitas retrieval dan guardrail terbukti stabil, cakupan dapat diperluas ke layanan yang lebih kompleks. Untuk pemerintahan, implementasi awal dapat diarahkan ke informasi standar pelayanan dan persyaratan administrasi, bukan keputusan substantif atau penilaian kelayakan warga. Pembatasan ini menjaga agar chatbot berfungsi sebagai asisten informasi, bukan pengganti pejabat atau sistem keputusan resmi.

Keterbatasan jurnal ini adalah belum melakukan eksperimen langsung pada korpus institusi tertentu, sehingga belum menyajikan nilai akurasi empiris. Namun, jurnal ini memberikan blueprint yang dapat digunakan oleh peneliti atau pengembang untuk membangun prototipe dan melakukan evaluasi lanjutan. Penelitian berikutnya dapat mengimplementasikan prototipe pada satu institusi, membandingkan retrieval sparse, dense, dan hybrid, menguji model embedding bahasa Indonesia, serta mengukur pengaruh reranking terhadap faithfulness dan latency.

IV. KESIMPULAN

Penelitian ini menghasilkan rancangan framework RAG chatbot untuk layanan informasi akademik dan pemerintahan. Kerangka yang diusulkan mencakup pengelolaan dokumen resmi, pipeline embedding dan retrieval, pembangkitan jawaban berbasis LLM, guardrail keamanan, sitasi sumber, *audit log*, serta evaluasi multidimensi. RAG chatbot dipandang lebih sesuai daripada chatbot rule-based dan LLM generik untuk layanan resmi karena mampu menggabungkan fleksibilitas percakapan dengan grounding pada sumber institusional.

Ilustrasi skenario layanan akademik dan pemerintahan memperlihatkan bagaimana rancangan menghubungkan pertanyaan pengguna dengan sumber resmi, jawaban bersitasi, penolakan terkontrol, dan eskalasi kepada petugas. Ilustrasi tersebut tetap bersifat konseptual sehingga efektivitasnya perlu dibuktikan melalui pengembangan prototipe dan pengujian empiris.

Keberhasilan implementasi bergantung pada kualitas basis pengetahuan, metadata dokumen, strategi segmen teks, model embedding, mekanisme retrieval, prompt template, guardrail, dan tata kelola risiko. Evaluasi harus mencakup kualitas retrieval, kualitas jawaban, pengalaman pengguna, performa operasional, serta kepatuhan terhadap keamanan dan perlindungan data. Dengan rancangan yang tepat, RAG chatbot dapat menjadi solusi layanan digital yang akurat, dapat ditelusuri, dan adaptif terhadap perubahan kebijakan.

REFERENSI

- A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. (2024). "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in Proc. ICLR.
- A. Folstad and P. B. Brandtzaeg. (2017). "Chatbots and the new world of HCI," *Interactions*, vol. 24, no. 4, pp. 38-42.
- A. K. Shanthakumar, F. Fassihi-Tash, A. Lotfi, and J. J. Bird. (2025). "Retrieval augmented large language model chatbots in higher education: A study on university open days," in *Advances in Computational Intelligence Systems*, Springer, pp. 32-44.

- A. R. Kusuma, C. R. Wongso, D. Aldrich, and F. Darari. (2024). "Trisurya: A local language-powered omnichannel chatbot as a solution for the enhancement of e-government quality and accessibility of public services in Indonesia," in Proc. 6th Int. Conf. Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, pp. 601-607, doi: 10.1109/ICIMCIS63449.2024.10956359.
- A. Vaswani et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- C. W. Okonkwo and A. Ade-Ibijola. (2021). "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100033.
- D. Oreski, S. Oreski, and G. Oreski. (2024). "Retrieval augmented generation in large language models for educational chatbot applications," in *CEUR Workshop Proceedings*, vol. 3938.
- E. Adamopoulou and L. Moussiades. (2020). "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, p. 100006.
- F. Petroni et al. (2021). "KILT: A benchmark for knowledge intensive language tasks," in Proc. NAACL-HLT, pp. 2523-2544.
- G. Izacard and E. Grave. (2021). "Leveraging passage retrieval with generative models for open domain question answering," in Proc. EACL, pp. 874-880.
- H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu. (2022). "A survey on retrieval-augmented text generation," arXiv:2202.01110.
- ISO/IEC, ISO/IEC 23894:2023 Information technology - Artificial intelligence - Guidance on risk management, Geneva, Switzerland: ISO.
- J. Dempere, K. Modugu, A. Hesham, and L. K. Ramasamy. (2023). "The impact of ChatGPT on higher education," *Frontiers in Education*, vol. 8.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, pp. 4171-4186.
- J. Johnson, M. Douze, and H. Jegou. (2021). "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535-547, 2021; arXiv:1702.08734.
- J. Swacha. (2025). "Retrieval-augmented generation (RAG) chatbots for education: A systematic review," *Applied Sciences*, vol. 15, no. 8, p. 4234.
- J. Swacha and M. Gracel. (2025). "Retrieval-augmented generation (RAG) chatbots for education: A survey of applications," *Applied Sciences*, vol. 15, no. 8, p. 4234, doi: 10.3390/app15084234.
- J. Wei et al. (2022). "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824-24837.
- J. Xu, J. Wang, J. Leung, and J. Gu. (2024). "GRASP: Municipal budget AI chatbots for enhancing civic engagement," in Proc. 2024 IEEE Int. Conf. Big Data (BigData), Washington, DC, USA, pp. 7438-7442, doi: 10.1109/BigData62323.2024.10825975.
- K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang. (2020). "REALM: Retrieval-augmented language model pre-training," in Proc. ICML, pp. 3929-3938.
- L. Gao, X. Ma, J. Lin, and J. Callan. (2023). "Precise zero-shot dense retrieval without relevance labels," in Proc. ACL, pp. 1762-1777.
- L. Huang et al. (2025). "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Information Systems*.
- M. Khasanova Zafar kizi and Y. Suh. (2025). "Design and performance evaluation of LLM-based RAG pipelines for chatbot services in international student admissions," *Electronics*, vol. 14, no. 15, p. 3095, doi: 10.3390/electronics14153095.

- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1.
- N. Reimers and I. Gurevych. (2019) "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, pp. 3982-3992.
- OECD. (2019). OECD Principles on Artificial Intelligence, Paris, France: OECD.
- Pemerintah Republik Indonesia, Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi.
- P. Lewis et al. (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474.
- P. Liang et al. (2022). "Holistic evaluation of language models," arXiv:2211.09110.
- S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. (2024). "RAGAS: Automated evaluation of retrieval augmented generation," in Proc. EACL Demo, pp. 150-158.
- S. Neupane et al. (2024). "From questions to insightful answers: Building an informed chatbot for university resources," arXiv:2405.08120.
- S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler. (2021). "Are we there yet? A systematic literature review on chatbots in education," Frontiers in Artificial Intelligence, vol. 4.
- S. Yao et al. (2023). "ReAct: Synergizing reasoning and acting in language models," in Proc. ICLR.
- UNESCO. (2023). Guidance for Generative AI in Education and Research. Paris, France: UNESCO.
- United Nations Department of Economic and Social Affairs, United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development. New York, NY, USA: United Nations.
- V. Karpukhin et al. (2020). "Dense passage retrieval for open-domain question answering," in Proc. EMNLP, pp. 6769-6781.
- Y. Gao et al. (2023). "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997.
- Z. Li et al. (2025). "Retrieval-augmented generation for educational application: A survey," Computers and Education: Artificial Intelligence.