Jurnal Teknologi Informasi dan Multimedia
p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim

Evaluasi Sistem *Retrieval-Augmented Generation* Berbasis *Low-Code* dalam Meningkatkan Akurasi Konseptual Pembelajaran Sosiologi

Rusmanto [1], Jasrino Maulana Putra [1]*

[1] Program Studi Sistem Informasi, Sekolah Tinggi Teknologi Terpadu Nurul Fikri, Indoneisa

* Korespondensi: jasrinomaulanap@gmail.com

Sitasi: R. Rusmanto, and J. M. Putra, "Evaluasi Sistem *Retrieval-Augmented Generation* Berbasis Low-Code dalam Meningkatkan Akurasi Konseptual Pem-belajaran Sosiologi", *Jurnal Teknologi Informasi Dan Multimedia*, vol. 8, no. 2, pp. 406-416, 2026. https://doi.org/10.35746/jtim.v8i2.1016

Diterima: 28-04-2026
Direvisi: 18-05-2026
Disetujui: 25-05-2026

Abstract: The utilization of Large Language Models (LLMs) in higher education offers significant efficiency, yet it introduces critical risks of information hallucination and conceptual bias, particularly in the discipline of Sociology. This study aims to evaluate the performance of a Retrieval-Augmented Generation (RAG) system based on the low-code platform n8n as a robust solution for hallucination mitigation. The system integrates semantic search using Supabase as a vector database and the Gemini 2.5 Flash model to restrict response generation exclusively to verified academic literature. The research employed an Experimental Single-System Evaluation method with a dual-evaluation approach (quantitative and qualitative) across 50 test instruments. Quantitative testing using the ROUGE-L metric recorded a mean score of 0.354, indicating adequate structural similarity despite variations inherent to the paraphrasing nature of LLMs in analytical tasks. Thematic qualitative analysis of evaluator comments revealed 98.2% positive sentiment, with the dominant theme being "Conceptually Accurate". Ultimately, 100% of the expert panel declared the system suitable for implementation as a reliable supplementary learning medium.

Keywords: Artificial Intelligence; Large Language Model; n8n; Retrieval-Augmented Generation; ROUGE-L

Abstrak: Penggunaan *Large Language Model* (LLM) di pendidikan tinggi menawarkan efisiensi, namun memunculkan risiko halusinasi informasi dan bias konseptual, khususnya pada disiplin ilmu Sosiologi. Penelitian ini bertujuan mengevaluasi kinerja sistem *Retrieval-Augmented Generation* (RAG) berbasis platform *low-code* n8n sebagai solusi mitigasi halusinasi. Sistem ini dibangun dengan mengintegrasikan pencarian semantik menggunakan Supabase sebagai *vector database* dan model Gemini 2.5 Flash untuk membatasi ruang generasi jawaban hanya pada literatur akademik terverifikasi. Penelitian ini menggunakan metode *Experimental Single-System Evaluation* dengan pendekatan evaluasi ganda (kuantitatif dan kualitatif) terhadap 50 instrumen pertanyaan pengujian. Hasil pengujian kuantitatif menggunakan metrik ROUGE-L mencatatkan nilai rata-rata 0,354, yang menunjukkan kesamaan struktural memadai meskipun bervariasi akibat sifat parafrastik LLM pada soal analitis. Hasil analisis kualitatif tematik terhadap komentar evaluator menunjukkan 98,2% sentimen positif dengan tema dominan "Akurat secara konsep". Berdasarkan evaluasi holistik, 100% panel ahli menyatakan sistem ini layak diimplementasikan sebagai media pembelajaran pendukung.

Kata kunci: Kecerdasan Buatan; Large Language Model; n8n; Retrieval-Augmented Generation; ROUGE-L

1. Pendahuluan

Perkembangan *Large Language Models* (LLM) dalam dua tahun terakhir menunjukkan peningkatan yang sangat pesat dalam ekosistem pendidikan tinggi. LLM mulai dimanfaatkan untuk penyusunan materi perkuliahan, penilaian awal tugas, penyederhanaan literatur akademik, hingga pembuatan soal evaluasi berbasis teks. Laporan EDUCAUSE *AI Landscape Study* tahun 2024 menunjukkan bahwa meningkatnya penggunaan AI oleh mahasiswa menjadi pendorong utama perencanaan strategis di lebih dari 70% institusi pendidikan tinggi [1]. Sejalan dengan temuan tersebut, OECD menekankan pentingnya perhatian pada dampak AI terhadap sistem pendidikan dan pemetaan kapabilitas AI terhadap keterampilan manusia, yang mendorong urgensi literasi dan tata kelola di lingkungan akademis [2]. Penelitian terdahulu juga membuktikan bahwa LLM mampu meningkatkan kualitas umpan balik adaptif, efektivitas interaksi pembelajaran daring, serta efisiensi layanan akademik [3,4]. Perkembangan ini menegaskan bahwa LLM telah menjadi komponen penting dalam transformasi digital pendidikan tinggi.

Meskipun menawarkan potensi efisiensi yang signifikan, adopsi teknologi ini terhambat oleh tantangan epistemologis, khususnya terkait keandalan dan kebenaran pengetahuan yang dihasilkan model. Dalam disiplin ilmu yang sarat interpretasi teoretis seperti Sosiologi, risiko ini tidak hanya terbatas pada kesalahan faktual, tetapi juga distorsi konsep yang dapat diperburuk oleh bias dalam data pelatihan model. Penelitian sebelumnya menyoroti bahwa model generatif rentan menghasilkan informasi yang tidak akurat atau tidak berdasar, yang dikenal sebagai *hallucination* [5]. Fenomena ini dapat dipicu oleh keberadaan *noise* dalam proses penarikan informasi dan representasi data [6], serta diperkuat oleh bias bawaan model [7]. Beberapa pendekatan mitigasi telah diusulkan, seperti *Retrieval-Augmented Generation* (RAG) dan teknik verifikasi jawaban bertahap [8,9]. Namun, mayoritas kajian berfokus pada implementasi arsitektur berbasis kode yang kompleks [10], sementara evaluasi efektivitas RAG berbasis platform *low-code* yang lebih praktis untuk diadopsi pendidik masih sangat terbatas [11].

Beberapa studi terdahulu telah secara konsisten mengonfirmasi keunggulan arsitektur RAG dalam meningkatkan reliabilitas sistem AI di pendidikan tinggi. Penelitian oleh Swacha dan Gracel [12] menunjukkan bahwa penggunaan *chatbot* berbasis RAG mampu memberikan respons yang lebih transparan dan dapat ditelusuri sumbernya dibandingkan model generatif murni. Sejalan dengan temuan tersebut, Thüs et al. [13] membuktikan bahwa sistem RAG efektif dalam meningkatkan keterlibatan mahasiswa terhadap literatur ilmiah melalui penyediaan konteks akademik yang presisi. Integrasi mekanisme *retrieval* ini menjadi kunci utama untuk memastikan bahwa teknologi AI tidak hanya berfungsi sebagai asisten produktivitas, tetapi juga sebagai media pembelajaran yang akurat secara faktual.

Meskipun efektivitas RAG telah terbukti di berbagai lini akademik, tantangan besar muncul pada domain pengetahuan sosial seperti Sosiologi yang memiliki tingkat interpretasi teoretis tinggi. Studi oleh Abror dan Rousyati [7] mengingatkan adanya risiko bias sosial dan distorsi makna yang dapat direproduksi oleh LLM jika tidak dikendalikan melalui basis pengetahuan yang terverifikasi secara ketat. Di sisi lain, mayoritas implementasi RAG saat ini masih menuntut keahlian teknis pemrograman yang tinggi, sehingga menciptakan batasan bagi dosen yang ingin mengadopsinya secara mandiri.

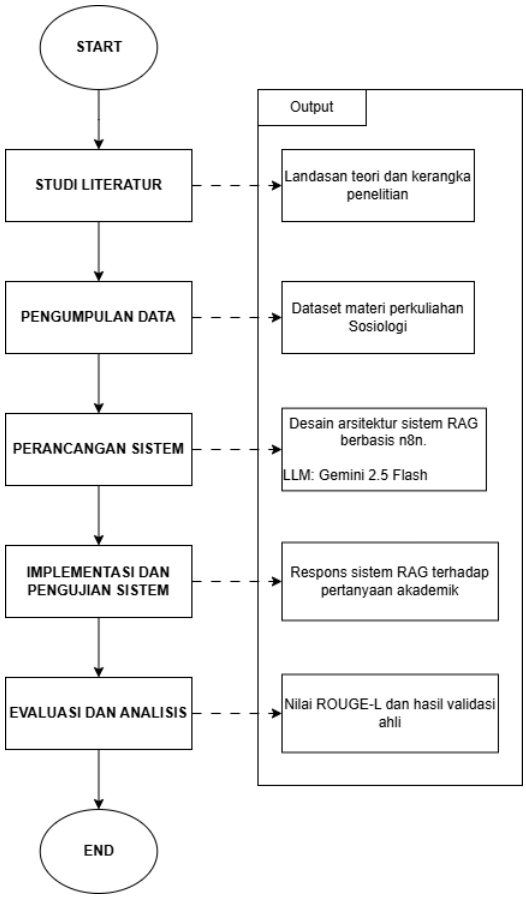
Kesenjangan literatur tersebut menimbulkan persoalan praktis, terutama ketika LLM digunakan sebagai sumber rujukan akademik utama oleh mahasiswa. Tanpa mekanisme *retrieval* yang terverifikasi, penggunaan LLM berisiko tinggi menimbulkan miskonsepsi sosial [14,15]. Berdasarkan celah tersebut, rumusan masalah penelitian ini adalah: Bagaimana efektivitas kinerja sistem *Retrieval-Augmented Generation* (RAG) berbasis platform *low-code* dalam memastikan *Large Language Model* (LLM) menghasilkan jawaban yang akurat, relevan, dan bebas dari bias konseptual pada mata kuliah Sosiologi? Oleh

karena itu, diperlukan evaluasi terstruktur terhadap sistem RAG [16] yang memanfaatkan basis pengetahuan tertutup untuk memastikan bahwa keluaran model selaras dengan konteks akademik dan prinsip pedagogis [12].

Penelitian ini bertujuan untuk mengevaluasi kinerja sistem RAG berbasis platform *low-code* n8n sebagai solusi mitigasi halusinasi untuk mata kuliah Sosiologi. Secara operasional, sistem ini mengintegrasikan Supabase sebagai *vector database* dan model Gemini 2.5 Flash untuk membatasi generasi jawaban hanya pada literatur yang divalidasi [17,18]. Pendekatan eksperimental digunakan dengan mengevaluasi keluaran sistem secara kuantitatif melalui metrik ROUGE-L dan secara kualitatif melalui validasi panel ahli [19,20]. Melalui evaluasi ganda ini, hasil penelitian menunjukkan bahwa sistem mampu mencapai kesamaan struktural yang memadai dalam merespons pertanyaan analitis, sekaligus meraih sentimen positif yang dominan terkait akurasi konsep teoretis. Secara keseluruhan, integrasi RAG *low-code* ini terbukti secara holistik layak diimplementasikan sebagai media pembelajaran pendukung yang andal dan aman dari bias halusinasi.

2. Bahan dan Metode

Penelitian ini dilakukan secara sistematis melalui lima tahapan utama guna menjamin akurasi dan validitas evaluasi kinerja sistem *Retrieval-Augmented Generation* (RAG) dalam domain ilmu sosial. Tahapan penelitian diawali dengan studi literatur untuk membangun landasan teori, diikuti dengan pengumpulan dataset materi perkuliahan, perancangan arsitektur sistem berbasis *low-code*, implementasi pengujian, hingga berakhir pada analisis evaluasi ganda. Seluruh rangkaian prosedur tersebut divisualisasikan dalam alur penelitian pada Gambar 1.

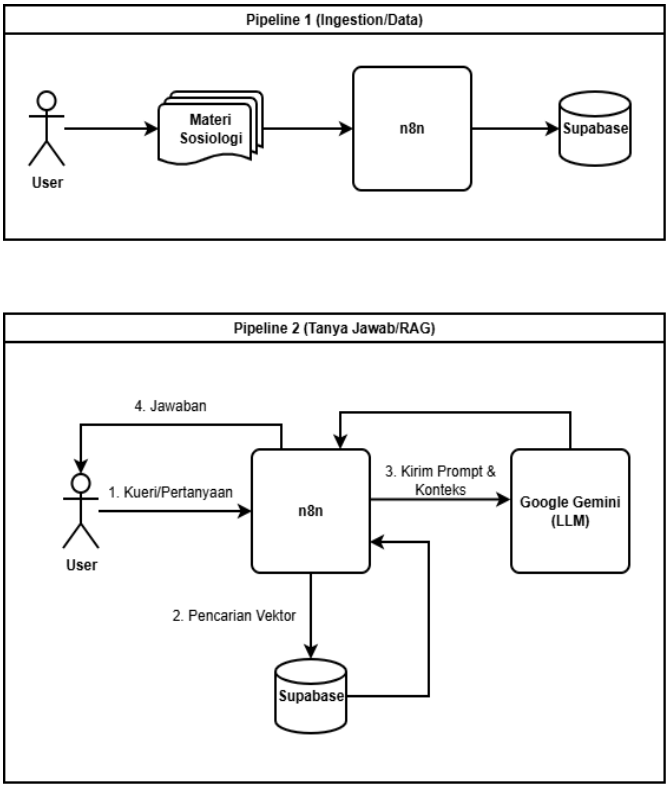


Gambar 1. Tahapan Alur Penelitian

Berdasarkan tahapan yang telah dipetakan, langkah krusial pertama adalah penyediaan bahan penelitian berupa sumber data primer (*knowledge-base*). Dataset yang digunakan mencakup dokumen akademik resmi mata kuliah Sosiologi pada salah satu Perguruan Tinggi Negeri Badan Hukum (PTNBH) di Indonesia, yang meliputi naskah pokok bahasan, modul utama, dan artikel studi kasus pendukung. Guna mematuhi kaidah etik penelitian, identitas institusi, kode mata kuliah, serta nama responden disamarkan untuk menjaga privasi informan dan kerahasiaan data internal. Dokumen-dokumen digital tersebut berfungsi sebagai basis pengetahuan terverifikasi yang menjamin kredibilitas informasi sebelum diintegrasikan ke dalam sistem melalui tahap *preprocessing*.

Setelah basis pengetahuan siap, tahap selanjutnya adalah merealisasikan perancangan sistem dengan memanfaatkan platform *low-code* n8n sebagai lingkungan orkestrasi utama. Pemilihan n8n didasarkan pada fungsinya sebagai pengatur alur kerja atau *workflow orchestrator*, berbeda dengan antarmuka AI generatif siap pakai seperti ChatGPT atau Microsoft Copilot. Penggunaan layanan AI publik tersebut memiliki keterbatasan dalam kontrol privasi data dan sulit diisolasi secara eksklusif untuk hanya membaca dokumen internal institusi, kecuali menggunakan versi *enterprise* yang berbiaya tinggi. Melalui n8n, arsitektur aliran data dapat dikonfigurasi secara independen dengan keamanan yang terjaga. Untuk mendukung konfigurasi independen tersebut, seluruh eksekusi alur kerja sistem dan pemrosesan data dalam eksperimen ini dijalankan menggunakan perangkat keras komputasi dengan spesifikasi prosesor Intel Core i5-11400H, GPU Nvidia Geforce RTX 3050, serta RAM berkapasitas 16 GB.

Arsitektur sistem ini secara operasional terbagi menjadi dua *pipeline* utama, yaitu *pipeline ingestion* data dan *pipeline* tanya-jawab yang diilustrasikan pada Gambar 2. Dalam *pipeline ingestion*, dokumen materi diekstraksi dan dipotong menjadi segmen teks (*chunking*) berukuran 1000 karakter dengan *overlap* 200 karakter guna menjaga kesinambungan konteks. Segmen-segmen tersebut kemudian dikonversi menjadi vektor numerik melalui model *Embeddings Google Gemini* dan disimpan secara terindeks dalam *vector database* Supabase.



Gambar 2. Desain Arsitektur Sistem RAG Berbasis n8n

Integrasi teknis tersebut memungkinkan *pipeline* kedua mengelola interaksi semantik secara *real-time*. Ketika pengguna memasukkan kueri, sistem melakukan pencarian vektor menggunakan metrik *cosine similarity* untuk mengambil potongan referensi paling relevan. Informasi tersebut kemudian dirakit menjadi *prompt* terstruktur yang dikirimkan ke model Gemini 2.5 Flash melalui integrasi API guna menghasilkan jawaban yang akurat dan terhindar dari risiko halusinasi. Pemisahan arsitektur ini memastikan beban komputasi pemrosesan data tidak mengganggu responsivitas sistem saat melayani pengguna.

Tabel 1. Komposisi Instrumen Pengujian

Tipe Soal	Jumlah Soal	Persentase	Justifikasi Berdasarkan Sumber
Definisi (D)	10	20%	Menguji kemampuan dasar RAG dalam mengambil definisi eksplisit.
Teori (T)	20	40%	Menguji pemahaman terhadap kerangka konseptual yang lebih luas.
Analisis Kasus (A)	20	40%	Menguji penerapan teori pada konteks spesifik.
Total	50	100%	

Sebagai tahap akhir dari alur penelitian, kinerja sistem diuji secara empiris menggunakan instrumen yang terdiri dari 50 butir pertanyaan yang dirancang secara proporsional berdasarkan tingkat kognitif sosiologis (Tabel 1). Setiap respons yang dihasilkan sistem direkam dan dievaluasi secara kuantitatif menggunakan metrik ROUGE-L untuk mengukur kesamaan struktural terhadap jawaban acuan (*ground truth*). Metrik ini bekerja dengan mengidentifikasi subsekuens terpanjang yang sama (*Longest Common Subsequence* atau LCS) antara jawaban sistem dan jawaban acuan. Secara matematis, nilai *recall* (R_{LCS}) dan *precision* (P_{LCS}) dihitung berdasarkan rasio panjang LCS terhadap panjang teks referensi (m) dan panjang teks sistem (n) melalui Persamaan (1) dan (2):

$$R_{LCS} = \frac{LCS(\mathrm{X}, \mathrm{Y})}{m} \tag{1}$$

$$P_{LCS} = \frac{LCS(\mathrm{X}, \mathrm{Y})}{n} \tag{2}$$

Berdasarkan nilai tersebut, skor akhir ditentukan melalui perhitungan *F-measure* (F_{LCS}) yang mengintegrasikan kedua komponen tersebut dengan parameter penyeimbang β sebagaimana ditunjukkan pada Persamaan (3):

$$F_{LCS} = \frac{(1+\beta^2)\, R_{LCS}\, P_{LCS}}{R_{LCS} + \beta^2\, P_{LCS}} \tag{3}$$

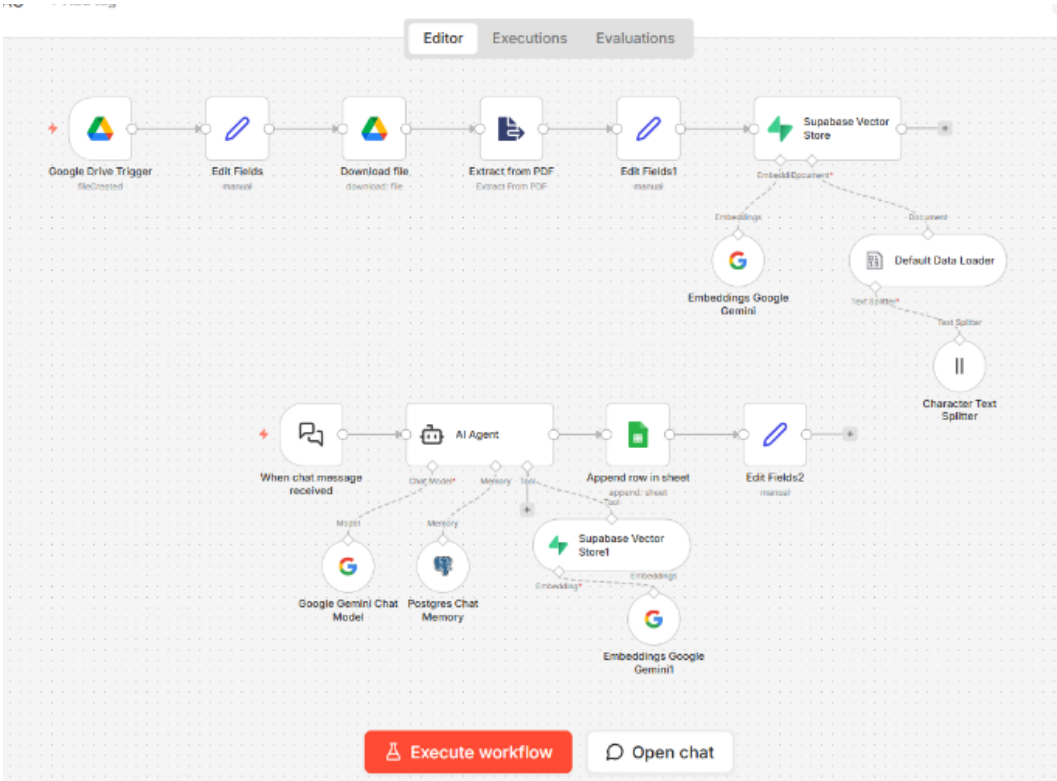
Dalam penelitian ini, nilai β diatur untuk menghasilkan nilai *F-score* sebagai indikator utama performa sistem. Melengkapi pengukuran otomatis tersebut, sistem juga melalui proses validasi ahli yang melibatkan pakar sosiologi dari institusi terkait guna menilai akurasi faktual serta kesesuaian pedagogis melalui penilaian skala Likert dan umpan balik naratif terbuka. Analisis tematik kemudian diterapkan untuk mengelompokkan data naratif menjadi kategori sentimen dan tema evaluasi yang spesifik. Rangkaian prosedur komprehensif ini bertujuan untuk menentukan kelayakan sistem sebagai media pembelajaran pendukung yang andal di pendidikan tinggi.

3. Hasil dan Pembahasan

Bagian ini menyajikan hasil implementasi operasional serta temuan empiris dari pengujian kinerja sistem RAG yang dikembangkan untuk mata kuliah Sosiologi. Evaluasi komprehensif dilakukan secara komplementer melalui dua pendekatan utama, yaitu pengujian kuantitatif menggunakan metrik linguistik komputasional ROUGE-L dan analisis kualitatif tematik berdasarkan penilaian panel ahli. Seluruh rincian data, visualisasi arsitektur, dan hasil tabulasi dirangkai secara sistematis guna memetakan efektivitas serta kelayakan implementasi arsitektur *low-code* ini dalam memitigasi risiko halusinasi informasi pada domain ilmu sosial.

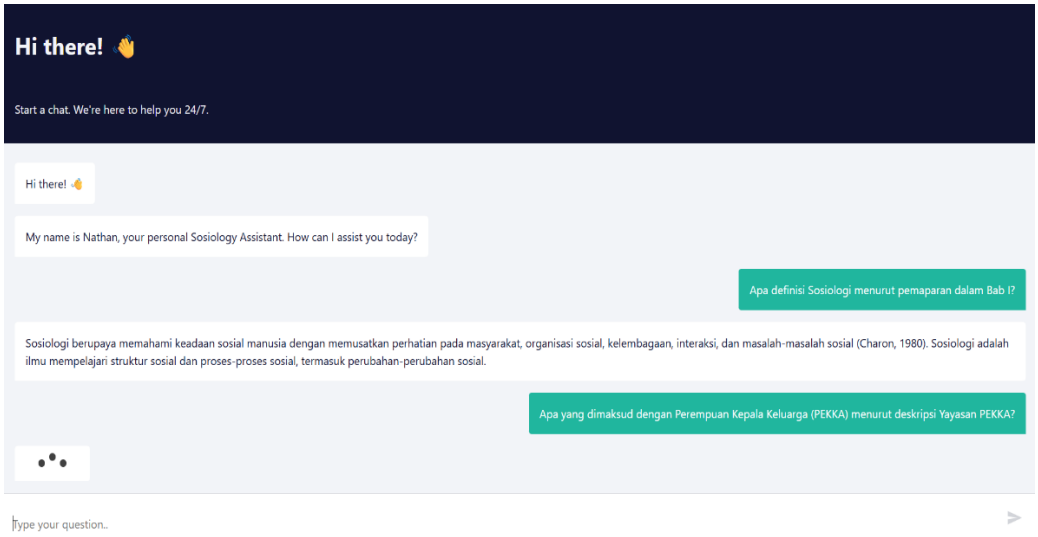
3.1. Implementasi Operasional Sistem dan Antarmuka Pengguna

Arsitektur sistem *Retrieval-Augmented Generation* (RAG) yang digunakan dalam penelitian ini dibangun menggunakan platform *low-code* n8n sebagai orkestrator alur kerja utama yang mengintegrasikan berbagai komponen kecerdasan buatan. Berdasarkan Gambar 3, arsitektur sistem yang telah berhasil diimplementasikan secara operasional terbagi menjadi dua alur utama yang berjalan beriringan, yakni alur pengelolaan basis pengetahuan (*knowledge ingestion*) dan alur pemrosesan pertanyaan (*question answering*).



Gambar 3. Arsitektur Operasional Sistem RAG pada Platform n8n

Pada alur pengelolaan basis pengetahuan, dokumen materi perkuliahan sosiologi diekstraksi dari format PDF menjadi teks. Teks tersebut kemudian dipecah menjadi beberapa segmen (*text splitting*), dikonversi menjadi representasi vektor (*embedding*), dan disimpan secara terstruktur ke dalam pangkalan data vektor Supabase. Sementara itu, pada alur pemrosesan pertanyaan, agen kecerdasan buatan memproses kueri pengguna dengan melakukan pencarian semantik ke dalam pangkalan data vektor tersebut. Hasil pencarian berupa potongan dokumen relevan kemudian dijadikan konteks rujukan bagi model bahasa generatif untuk merumuskan jawaban akhir. Melalui arsitektur ini, keluaran sistem telah berhasil dirangkai untuk mencegah model bahasa merumuskan jawaban di luar konteks akademik yang telah diverifikasi.



Gambar 4. Tampilan Antarmuka Interaktif Sistem RAG

Sebagai hasil akhir dari integrasi kedua alur kerja tersebut, sistem menyediakan antarmuka pengguna (*user interface*) interaktif berupa ruang obrolan (*chat*) yang dapat diakses secara langsung. Sebagaimana ditampilkan pada Gambar 4, antarmuka ini memfasilitasi interaksi tanya-jawab di mana pengguna dapat mengajukan kueri spesifik terkait materi sosiologi, dan sistem akan langsung memberikan respons naratif yang diformulasikan dari dokumen acuan. Ketersediaan antarmuka ini menjadi bukti bahwa rancangan arsitektur *back-end* pada platform n8n telah berhasil diwujudkan menjadi sistem aplikasi fungsional yang siap digunakan untuk tahapan pengujian selanjutnya.

3.2. Evaluasi Kuantitatif Kesamaan Tekstual

Pengujian kuantitatif dilakukan untuk mengukur tingkat kemiripan leksikal dan struktural antara teks jawaban yang dihasilkan oleh sistem RAG terhadap dokumen jawaban acuan (*ground truth*). Pengukuran otomatis ini dieksekusi secara komputasional menggunakan metrik ROUGE-L terhadap 50 butir instrumen pertanyaan sosiologi yang telah diujikan. Ringkasan statistik deskriptif dari hasil kalkulasi algoritma tersebut disajikan secara terkonsolidasi bersama skor validasi ahli pada Tabel 2.

Tabel 2. Rekapitulasi Hasil Pengujian Kuantitatif

Statistik	Skor ROUGE-L	Skor Validasi Ahli (Skala 1-4)
Nilai Minimum	0,053	1,50
Nilai Maksimum	1,000	4,00
Rata-rata (*Mean*)	0,354	3,32
Nilai Tengah (*Median*)	0,298	3,50
Standar Deviasi	0,230	0,75

Berdasarkan hasil kalkulasi komputasional yang terekam pada Tabel 2, pengujian ROUGE-L mencatatkan nilai rata-rata (*mean*) sebesar 0,354. Perolehan angka yang melampaui ambang batas minimal keberhasilan 0,30 ini mengindikasikan bahwa teks keluaran sistem RAG memiliki tingkat kesamaan struktural yang memadai terhadap jawaban acuan (*ground truth*). Melengkapi evaluasi otomatis tersebut, kuantifikasi penilaian kepakaran melalui skala Likert juga menunjukkan tren kinerja yang sejalan. Sistem meraih skor rata-rata sebesar 3,32 dari skala 4,00 (dengan median 3,50), yang secara empiris mengafirmasi bahwa substansi keilmuan yang diformulasikan oleh model sangat valid dan memuaskan standar akademik dosen pengampu. Meskipun demikian, tingginya standar deviasi pada metrik ROUGE-L (0,230) dan skor validasi ahli (0,75)

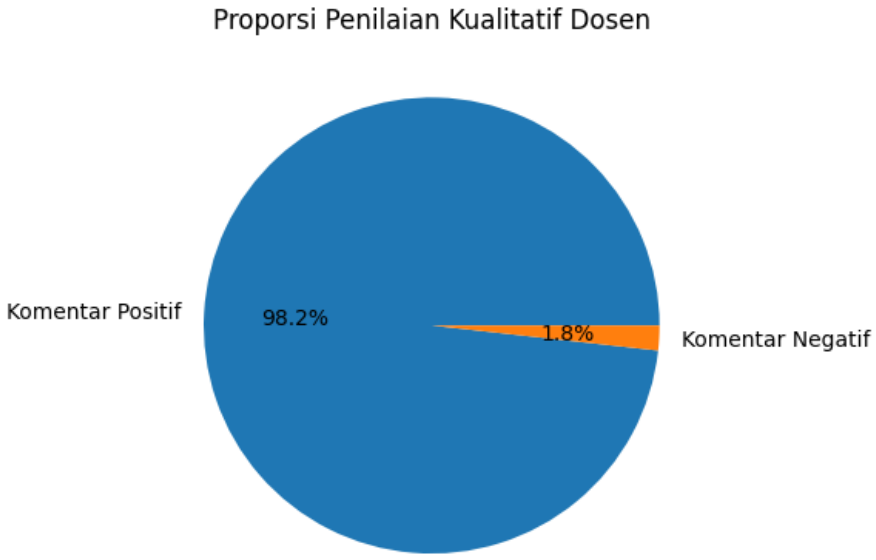
memperlihatkan adanya variasi fluktuasi kinerja yang dipengaruhi oleh tingkat kompleksitas kognitif masing-masing tipe soal.

3.3. Evaluasi Kualitatif dan Analisis Tematik

Guna mendalami kualitas respons sistem melampaui data angka, evaluasi kualitatif dilakukan melalui analisis tematik terhadap umpan balik naratif (komentar terbuka) yang diberikan oleh evaluator. Analisis ini bertujuan untuk memetakan kekuatan serta kelemahan pedagogis keluaran sistem secara deskriptif.

Hasil pemetaan menunjukkan bahwa tema "Akurat secara konsep" merupakan ulasan yang paling mendominasi penilaian pakar, dengan rekam jejak kemunculan mencapai 55 kali. Sebagai representasi deskriptif dari tema dominan ini, beberapa catatan *verbatim* (kutipan langsung) dari evaluator meliputi pernyataan persetujuan mutlak seperti *"Jawaban sudah benar"*, *"Tepat dan detil"*, serta *"Sudah tepat, definisi sosiologi dijawab dengan [baik]"*. Kehadiran ragam komentar ini mengafirmasi bahwa keluaran teks dari sistem RAG secara harfiah telah berhasil memenuhi standar validitas materi keilmuan.

Selain tema dominan tersebut, tercatat pula kemunculan beberapa tema sekunder yang bersifat konstruktif. Tema "Perlu pengayaan" muncul sebanyak 4 kali, merepresentasikan ulasan pakar yang menilai jawaban sistem sudah tepat secara definisi namun menyarankan adanya elaborasi lebih lanjut (misalnya pada kutipan: *"Jawaban sudah benar, tinggal kedalamannya..."*). Sementara itu, tema evaluasi lainnya seperti "Koheren/jelas" dan ulasan korektif berupa "Kurang konteks" masing-masing hanya muncul sebanyak 1 kali dalam keseluruhan tabulasi data evaluasi. Pendalaman kualitatif ini dikonsolidasikan dan divisualisasikan proporsinya pada Gambar 5.



Gambar 5. Sentimen Penilaian Kualitatif Evaluator

Berdasarkan visualisasi pada Gambar 5, pengelompokan keseluruhan komentar terbuka memperlihatkan bahwa sentimen positif memegang persentase mayoritas yang sangat mutlak, yakni mencapai 98,2%. Sebaliknya, proporsi komentar negatif hanya mencakup porsi yang sangat kecil, yakni sebesar 1,8%, yang secara spesifik merujuk pada saran penyempurnaan kedalaman analitis. Secara keseluruhan, penelusuran kualitatif ini membuktikan bahwa sistem RAG berhasil membatasi ruang narasi model generatif untuk beroperasi secara patuh di dalam koridor literatur referensi, sehingga efektif dalam memitigasi risiko fabrikasi atau halusinasi informasi.

3.4. Sintesis Evaluasi dan Analisis Komparatif

Sintesis dan interpretasi mendalam terhadap temuan empiris menunjukkan adanya fenomena ketidaksejajaran (*mismatch*) yang signifikan antara evaluasi linguistik komputasional dan interpretasi kepakaran manusia. Untuk memahami perbandingan ini secara objektif, perlu diperhatikan bahwa kedua metrik beroperasi pada skala yang berbeda: metrik otomatis ROUGE-L beroperasi pada rentang 0–1 (di mana 1 menunjukkan kesamaan tekstual mutlak), sedangkan skor kualitatif menggunakan skala Likert 1–4 (di mana 4 adalah skor kesempurnaan materi). Sebagaimana dirangkum dalam Tabel 3, tingkat kesamaan leksikal cenderung berbanding terbalik dengan kompleksitas kognitif soal yang diujikan.

Tabel 3. Perbandingan Skor Kualitatif dan ROUGE-L Berdasarkan Aspek Pertanyaan

Aspek Pertanyaan	ROUGE-L	Skor Kualitatif
Definisi (D)	0,569	3,70
Teori (T)	0,438	3,55
Analisis Kasus (A)	0,164	2,90

Pada pertanyaan bertipe Definisi, sistem meraih skor kualitatif 3,70 (sangat mendekati batas maksimum 4,00) dan ROUGE-L sebesar 0,569. Dalam konteks evaluasi *Natural Language Processing* (NLP), skor ROUGE-L di atas 0,500 untuk tugas generatif sudah dikategorikan sangat tinggi karena menunjukkan tumpang tindih kata kunci yang masif, atau setara dengan tingkat akurasi tekstual yang sangat baik. Namun, pada tipe Analisis Kasus, skor ROUGE-L menurun drastis menjadi 0,164 sementara skor validasi ahli masih bertahan di angka 2,90 yang tergolong sangat layak.

Hal ini membuktikan bahwa metrik ROUGE-L yang berbasis tumpang tindih urutan kata memiliki keterbatasan dalam menilai jawaban parafrastik dan dinamis. Meskipun algoritma memberikan skor rendah karena perbedaan susunan kata (hanya sekitar 16% kesamaan kata demi kata), pakar manusia tetap memvalidasi bahwa jawaban tersebut sangat koheren secara konsep keilmuan. Perbedaan ini terjadi karena LLM merumuskan penjelasan analisis menggunakan struktur kalimat kreatif yang berbeda dari jawaban acuan, namun secara substansi tetap akurat. Hal ini menegaskan bahwa penilaian kualitatif tetap menjadi standar emas dalam mengukur keandalan AI di sektor pendidikan untuk menjembatani kakuannya metrik komputasional.

Secara operasional, tingginya tingkat validitas konseptual yang diakui oleh ahli tidak lepas dari efisiensi konfigurasi *node* visual pada platform n8n. Platform ini bertindak sebagai orkestrator yang menjembatani lalu lintas data secara linier tanpa celah kebocoran informasi. Pada alur pemrosesan pertanyaan, *node AI Agent* dikonfigurasi sebagai pusat kendali utama yang secara modular terhubung dengan *node Google Gemini Chat Model* sebagai mesin berpikir, *node Postgres Chat Memory* untuk mempertahankan konteks percakapan historis, serta *node Supabase Vector Store* yang berfungsi sebagai alat penarik informasi referensi resmi. Keterpaduan antarkomponen visual ini memotong rantai instruksi yang rumit, sehingga waktu respons sistem tetap terjaga optimal saat melayani interaksi pengguna.

Keunggulan praktis dari arsitektur ini tercermin secara langsung pada mitigasi halusinasi saat pengguna memberikan kueri. Sebagai contoh, ketika pengguna memasukkan pertanyaan analitis, sistem tidak langsung melemparkannya secara bebas ke LLM. Sebaliknya, kueri tersebut ditangkap oleh n8n untuk mengekstrak kata kunci semantik, mencari potongan teks terindeks di Supabase, dan merakitnya kembali ke dalam cetakan instruksi (*prompt template*) yang kaku sebelum diserahkan kepada Gemini 2.5 Flash. Melalui skenario operasional kontrol ganda ini, jawaban akhir yang dibangkitkan oleh kecerdasan buatan terbukti bersih dari fabrikasi fakta, sebagaimana direpresentasikan pada contoh keluaran antarmuka sistem di Gambar 4.

Keberhasilan mitigasi halusinasi tersebut didorong secara fundamental oleh arsitektur RAG yang secara ketat membatasi ruang narasi model hanya pada koridor literatur sosiologi yang telah divalidasi. Dominasi tema umpan balik "Akurat secara konsep" (98,2% sentimen positif) menjadi bukti kuat bahwa integrasi pencarian semantik ini berhasil memaksa model untuk mematuhi konteks akademik. Temuan ini secara konsisten mengonfirmasi sekaligus memperluas kajian Swacha dan Gracel [12] mengenai keunggulan arsitektur berbasis *retrieval* dalam menekan bias informasi, khususnya pada domain ilmu sosial yang rentan terhadap distorsi interpretasi bahasa. Selain itu, riset ini memperkaya metodologi evaluasi yang diusulkan Thüs et al. [13] dengan menyandingkan metrik otomatis dan justifikasi kepakaran secara holistik.

Dari perspektif adopsi teknologi, efektivitas sistem yang dibangun melalui platform *low-code* n8n menjadi afirmasi empiris yang kuat terhadap argumentasi Nakhod [11] mengenai demokratisasi AI. Fakta bahwa keandalan faktual yang mumpuni dapat dicapai melalui manipulasi antarmuka visual membuktikan bahwa platform *low-code* bukan sekadar solusi kompromi, melainkan arsitektur kompetitif yang meruntuhkan hambatan teknis bagi pendidik. Berdasarkan evaluasi holistik penutup, 100% panel ahli sepakat menyatakan bahwa sistem ini layak diimplementasikan sebagai asisten akademik (*co-pilot*) pendukung yang tersedia 24 jam bagi mahasiswa. Meskipun demikian, kelayakan ini diiringi dengan catatan revisi minor, mengingat masih ditemukannya keterbatasan sistem saat merespons pertanyaan lapangan yang sangat kompleks. Hal ini menjadi temuan penting yang mengindikasikan adanya keterbatasan pada rekayasa instruksi (*prompt engineering*) yang digunakan saat ini, sehingga model belum mampu melakukan elaborasi mendalam secara otomatis pada fenomena sosiologi terapan.

4. Kesimpulan

Implementasi arsitektur *Retrieval-Augmented Generation* (RAG) berbasis platform *low-code* n8n terbukti menjadi solusi efektif dalam memastikan *Large Language Model* (LLM) menghasilkan jawaban yang akurat, relevan, dan bebas dari bias konseptual pada mata kuliah Sosiologi. Melalui integrasi pencarian semantik dan pangkalan data vektor Supabase, sistem berhasil mengarahkan LLM untuk membatasi ruang generasinya hanya pada literatur akademik yang telah tervalidasi. Keberhasilan ini dikonfirmasi secara kualitatif oleh panel ahli dengan tingkat kepuasan mencapai 98,2% pada tema "Akurat secara konsep", yang menegaskan bahwa arsitektur RAG mampu memitigasi risiko halusinasi dan bias data turunan pada domain ilmu sosial di pendidikan tinggi.

Meskipun secara operasional telah terbukti efektif, terdapat sejumlah keterbatasan yang perlu dicatat sebagai landasan reflektif. Penggunaan metrik kuantitatif seperti ROUGE-L memiliki keterbatasan dalam menangkap kualitas jawaban yang bersifat analitis dan parafrastik, sehingga penilaian pakar tetap menjadi standar emas dalam validasi konseptual. Selain itu, pengujian yang berfokus pada model Gemini 2.5 Flash dan modul sosiologi dasar menunjukkan bahwa kapabilitas sistem dalam merespons pertanyaan interdisipliner atau studi kasus lapangan yang kompleks masih memerlukan penyempurnaan. Hal ini membuktikan bahwa meskipun teknologi RAG sangat ideal sebagai instrumen penguatan materi dasar, peran dosen sebagai pemegang otoritas validasi konseptual tetap mutlak diperlukan dalam ekosistem akademik.

Sebagai langkah pengembangan di masa mendatang, diperlukan perluasan basis pengetahuan dokumen dengan memasukkan literatur empiris atau jurnal studi kasus kontemporer guna memperkuat kapabilitas analitis sistem pada fenomena sosiologi terapan. Di samping itu, diperlukan optimasi spesifik pada aspek rekayasa instruksi agar sistem mampu menyertakan rujukan teori atau referensi pakar sosiologi secara eksplisit dalam setiap jawaban yang dihasilkan. Akhirnya, disarankan untuk melakukan studi komparatif menggunakan berbagai model bahasa besar lainnya guna memetakan tingkat

efisiensi dan fleksibilitas arsitektur RAG pada berbagai lingkungan komputasi yang berbeda.

Ucapan Terima Kasih: Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Sekolah Tinggi Teknologi Terpadu Nurul Fikri (STT-NF) atas dukungan fasilitas, lingkungan akademik, serta bimbingan yang berharga selama proses penelitian ini berlangsung. Penghargaan khusus juga kami sampaikan kepada salah satu pihak Perguruan Tinggi Negeri Badan Hukum (PTNBH) di Indonesia yang telah memberikan izin akses dan penggunaan dokumen materi mata kuliah Sosiologi sebagai basis data utama dalam pengujian sistem ini. Dukungan administratif maupun teknis dari berbagai pihak tersebut sangat berkontribusi terhadap kelancaran dan keberhasilan penyelesaian naskah penelitian ini.

Referensi

[1] J. Robert, "2024 EDUCAUSE AI Landscape Study, EDUCAUSE, 2024. Available: https://library.educause.edu/resources/2024/2/2024-educause-ai-landscape-study. [Accessed: Nov. 19, 2025].

[2] OECD, "Artificial Intelligence and Education and Skills," OECD, 2024. Available: https://www.oecd.org/en/topics/sub-issues/artificial-intelligence-and-education-and-skills.html. [Accessed: Nov. 19, 2025].

[3] H. Khandakar, S. A. Fazal, K. F. Afnan, and K. K. Hasan, "Implications of artificial intelligence chatbot models in higher education," *IAES Int. J. Artif. Intell.*, vol. 13, no. 4, pp. 3808–3813, 2024, https://doi.org/10.11591/ijai.v13.i4.pp3808-3813.

[4] I. Gligorea, M. Cioca, R. Oancea, A. T. Gorski, H. Gorski, and P. Tudorache, "Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review," *Educ. Sci.*, vol. 13, no. 12, 2023, https://doi.org/10.3390/educsci13121216.

[5] Y. Zhang *et al.*, "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *Comput. Linguist.*, pp. 1–46, 2025, https://doi.org/10.1162/coli.a.16.

[6] F F. Cuconasu *et al.*, "The Power of Noise: Redefining Retrieval for RAG Systems," *SIGIR 2024 - Proc. 47th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 719–729, 2024, https://doi.org/10.1145/3626772.3657834.

[7] D. Abror and Rousyati, "Etika Dan Bias Dalam Llm: Tanggung Jawab Sosial Atas Kecerdasan Buatan Generatif," *J. Unitek*, vol. 18, no. 1, pp. 69–75, 2025, https://doi.org/10.52072/unitek.v18i1.1386.

[8] H. W. Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint*, 2023,: https://doi.org/10.48550/arXiv.2312.10997.

[9] S. Dhuliawala *et al.*, "Chain-of-Verification Reduces Hallucination in Large Language Models," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 3563–3578, 2024, https://doi.org/10.18653/v1/2024.findings-acl.212.

[10] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. Lee, "Retrieval-augmented generation for educational application : A systematic survey," *Comput. Educ. Artif. Intell.*, vol. 8, no. May, p. 100417, 2025, https://doi.org/10.1016/j.caeai.2025.100417.

[11] O. Nakhod, "Using Retrieval-Augmented Generation to Elevate Low-Code Developer Skills," *Artif. Intell.*, 2023, https://doi.org/10.15407/jai2023.03.126.

[12] J. Swacha and M. Gracel, "Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications," *Appl. Sci.*, vol. 15, no. 8, 2025, https://doi.org/10.3390/app15084234.

[13] D. Thüs, S. Malone, and R. Brünken, "Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature," *Front. Psychol.*, vol. 15, no. October, pp. 1–23, 2024, https://doi.org/10.3389/fpsyg.2024.1474892.

[14] S. Dakshit, "Faculty Perspectives on the Potential of RAG in Computer Science Higher Education," *Proc. 25th Annu. Conf. Inf. Technol. Educ. SIGITE 2024*, pp. 19–24, 2024, https://doi.org/10.1145/3686852.3686864.

[15] E. Tyndall, C. Gayheart, A. Some, J. Genz, T. Wagner, and B. Langhals, "Impact of retrieval augmented generation and large language model complexity on undergraduate exams created and taken by AI agents," *Data Policy*, vol. 7, 2025, https://doi.org/10.1017/dap.2025.10024.

[16] Q. Huang, C. Lv, L. Lu, and S. Tu, "Evaluating the Quality of AI-Generated Digital Educational Resources for University Teaching and Learning," pp. 1–18, 2025, https://doi.org/10.3390/systems13030174.

[17] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74-81, https://aclanthology.org/W04-1013/.

[18] A. Janakiraman and B. Ghoraani, "An Empirical Comparison of Text Summarization: A Multi-Dimensional Evaluation of Large Language Models," *arXiv preprint*, 2025, http://arxiv.org/abs/2504.04534.

[19] S. Es, J. James, L. Espinosa-anke, and S. Schockaert, "Ragas: Automated Evaluation of Retrieval Augmented Generation," *arXiv preprint*, 2023, https://arxiv.org/abs/2309.15217.

[20] H. Li *et al.*, "LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods," *arXiv preprint*, 2024, https://arxiv.org/abs/2412.05579.