BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

Evaluasi Aplikasi Pembelajaran Berbasis Web Menggunakan Generative Artificial Intelligence dengan Metode ROUGE

Rusmanto[1], Nuranisah[2,*]

[1] Program Studi Sistem Informasi, Sekolah Tinggi Teknologi Terpadu Nurul Fikri, Depok, Indonesia
[2] Program Studi Teknik Informatika, Sekolah Tinggi Teknologi Terpadu Nurul Fikri, Depok, Indonesia
Email: [1]rus@nurulfikri.ac.id, [2,*]nura22097ti@student.nurulfikri.ac.id
Email Penulis Korespondensi: nura22097ti@student.nurulfikri.ac.id

Abstrak−Penelitian ini bertujuan untuk mengevaluasi fungsionalitas serta kualitas jawaban yang dihasilkan oleh aplikasi pembelajaran berbasis web menggunakan *Generative Artificial Intelligence* (GenAI) untuk mata kuliah Pendidikan Pancasila dan Kewarganegaraan (PPKN). Fokus utama dari penelitian ini terletak pada proses evaluasi dari sistem, sementara pengembangan aplikasi hanya dilakukan sebagai sarana untuk menghasilkan data uji. Evaluasi sistem dilakukan melalui dua tahap, yaitu pengujian fungsional menggunakan metode *black-box testing* dan pengukuran kualitas jawaban yang dihasilkan sistem dengan menggunakan metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE). Pengujian *black-box* dilakukan untuk memastikan seluruh fitur inti sistem berjalan sesuai dengan spesifikasi. Hasil pengujian *black-box* menunjukkan tingkat keberhasilan 100% pada seluruh skenario uji yang dilakukan. Selanjutnya, evaluasi kualitas jawaban dilakukan terhadap 50 pasangan data uji yang terdiri dari jawaban GenAI dan teks referensi (*gold standard*) yang disusun oleh dosen PPKN dengan menggunakan metode ROUGE. Hasil evaluasi menunjukkan rata-rata nilai *F1-score* sebesar 97% pada metrik ROUGE-1, ROUGE-2, dan ROUGE-L. Sebanyak 49 dari 50 jawaban termasuk dalam kategori "Sangat Baik" (≥ 0,75), dan 1 jawaban berada pada kategori "Baik". Hasil ini menunjukkan bahwa aplikasi mampu menghasilkan jawaban dengan tingkat kesesuaian tekstual yang sangat tinggi terhadap referensi akademik. Penelitian ini berkontribusi untuk mengisi kekosongan bukti empiris dan menyediakan tolak ukur evaluasi terstandar untuk aplikasi GenAI berbasis web dalam pendidikan, serta menawarkan pendekatan evaluasi yang menggabungkan pengujian fungsional sistem dan pengukuran kualitas jawaban berbasis ROUGE. Namun, evaluasi ini masih terbatas pada aspek linguistik berbasis n-gram dan belum sepenuhnya merepresentasikan kedalaman makna secara semantik.

Kata Kunci: Evaluasi Sistem; Generative Artificial Intelligence; PPKN; ROUGE; Black-Box Testing

Abstract−This study aims to evaluate the functionality and answer quality of a web-based learning application that uses Generative Artificial Intelligence (GenAI) for the Pancasila and Civic Education (PPKN) course. The primary focus of this research lies in the system evaluation process, while the application development was carried out solely as a means of generating test data. The system was evaluated in two stages: functional testing using the black-box testing method and answer quality assessment using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) method. Black-box testing was conducted to ensure that all core system features operated according to specifications. The results of the black-box testing showed a 100% success rate across all test scenarios. Furthermore, answer quality evaluation was performed on 50 test data pairs consisting of GenAI-generated answers and reference texts (gold standards) prepared by PPKN lecturers using the ROUGE method. The evaluation results showed an average F1-score of 97% on the ROUGE-1, ROUGE-2, and ROUGE-L metrics. A total of 49 out of 50 answers were categorized as "Very Good" (≥ 0.75), while 1 answer was categorized as "Good." These findings indicate that the application is capable of generating answers with a very high level of textual similarity to academic references. This study contributes to filling the gap in empirical evidence and provides a standardized evaluation benchmark for web-based GenAI applications in education, while also offering an evaluation approach that integrates system functional testing and ROUGE-based answer quality measurement. However, this evaluation is still limited to linguistic aspects based on n-grams and does not yet fully represent semantic depth.

Keywords: System Evaluation; Generative Artificial Intelligence; PPKN; ROUGE; Black-Box Testing

1. PENDAHULUAN

Era transformasi digital telah menghadirkan revolusi pada berbagai aspek kehidupan, termasuk dalam dunia pendidikan. *Generative Artificial Intelligence* (GenAI) muncul sebagai teknologi yang menjanjikan untuk mengubah paradigma pembelajaran tradisional menjadi lebih interaktif, *personalisasi*, dan adaptif. Teknologi GenAI memiliki kemampuan untuk melakukan ringkasan otomatis, menghasilkan konten tekstual dan menciptakan materi pembelajaran yang relevan secara kontekstual[1]. Dalam konteks pendidikan Indonesia, integrasi teknologi ini menjadi semakin penting seiring dengan diberlakukannya kurikulum merdeka pada pembelajaran yang berpusat pada peserta didik [2].

Mata kuliah Pendidikan Pancasila dan Kewarganegaraan (PPKN) mempunyai peran yang penting dalam pembentukan karakter dan kesadaran kewarganegaraan mahasiswa. Namun, tantangan dalam pembelajaran PPKN di era digital semakin kompleks, terutama dalam menghadapi generasi milenial dan generasi yang lahir atau besar di era digital (*digital native*) yang memiliki karakteristik pembelajaran yang berbeda [3]. Transformasi kewarganegaraan dalam era digital menuntut adaptasi materi, metode, dan nilai-nilai kewarganegaraan digital yang harus disesuaikan dengan konteks lokal Indonesia [4]. Penggunaan GenAI dalam pembelajaran PPKN berpotensi memberikan solusi yang inovatif untuk meningkatkan keterlibatan, aksesibilitas materi, dan pengembangan kemampuan berpikir kritis mahasiswa.

Meskipun potensi GenAI dalam pendidikan sangat menjanjikan, implementasinya memerlukan evaluasi yang komprehensif dan terukur. Metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) telah terbukti efektif sebagai alat evaluasi otomatis untuk mengukur kualitas teks yang dihasilkan oleh sistem AI [5], khususnya dalam tugas-tugas seperti *summarization* dan *text generation* [6]. Penggunaan metode ROUGE dalam konteks evaluasi aplikasi GenAI berbasis web untuk pembelajaran PPKN dapat memberikan penilaian objektif terhadap kualitas konten yang dihasilkan,

akurasi informasi, dan relevansi dengan tujuan pembelajaran [7]. Namun, karena sistem aplikasi GenAI berbasis web untuk mata kuliah PPKN ini masih belum tersedia [3], penelitian ini mencakup proses pengembangan aplikasi GenAI berbasis web terlebih dahulu sebagai persyaratan utama untuk menghasilkan data yang dapat dievaluasi menggunakan metode ROUGE dan pengujian fungsional menggunakan *black-box testing*.

Selain metode ROUGE, pengujian menggunakan *black-box testing* juga dipilih untuk melakukan pengujian pada perangkat lunak yang berfokus pada output yang dihasilkan oleh sistem GenAI berdasarkan input yang diberikan, tanpa mengetahui implementasi internal kode atau sistem tersebut [8]. Metode ini memiliki tujuan untuk memverifikasi apakah aplikasi GenAI berbasis web berfungsi sesuai dengan kebutuhan pengguna dan menghasilkan hasil yang diharapkan. *Black-box testing* sering digunakan dalam pengujian sistem berbasis GenAI, karena memungkinkan evaluasi sistem dari perspektif pengguna yang berinteraksi dengan aplikasi tanpa mengakses kode internal [9].

Berdasarkan tinjauan literatur sistematis, terdapat beberapa kesenjangan penelitian yang signifikan dalam evaluasi aplikasi GenAI berbasis web untuk pendidikan menggunakan metode ROUGE, khususnya dalam konteks PPKN. Pertama, protokol evaluasi untuk aplikasi GenAI berbasis web dalam pendidikan masih belum terstandarisasi, dengan berbagai studi menggunakan metrik yang berbeda-beda sehingga menyulitkan replikasi dan perbandingan penelitian [6]. Kedua, masih terbatas korpus dan model yang disesuaikan khusus untuk domain kewarganegaraan Indonesia, sehingga model GenAI umum sering kali menghasilkan konten yang kurang kontekstual [10]. Ketiga, aspek etika dan keselarasan nilai dalam penerapan GenAI untuk pendidikan kewarganegaraan masih memerlukan kajian mendalam. Penelitian lokal menunjukkan perlunya mekanisme pengawasan yang etis agar hasil GenAI selaras dengan nilai-nilai Pancasila dan konteks kebijakan pendidikan Indonesia [11].

Selain itu, penelitian-penelitian terdahulu lebih banyak berfokus pada pemanfaatan GenAI dalam pendidikan secara umum atau implementasi teknis *Retrieval-Augmented Generation* (RAG) pada sistem tanya jawab, sedangkan penelitian ini secara spesifik mengevaluasi kualitas jawaban GenAI pada domain PPKN menggunakan metrik kuantitatif berbasis kemiripan teks dengan acuan *gold standard* dari dosen. Berdasarkan kondisi tersebut, penelitian ini diposisikan untuk mengisi gap tersebut dengan mengimplementasikan aplikasi GenAI berbasis web pada materi PPKN, untuk memverifikasi fungsionalitasnya melalui *black-box testing*, serta melakukan pengukuran kualitas jawaban GenAI secara kuantitatif menggunakan ROUGE-1, ROUGE-2, dan ROUGE-L terhadap jawaban referensi yang telah disusun oleh dosen.

Berdasarkan identifikasi *research gap* di atas, maka terdapat dua rumusan masalah utama yang akan dijawab pada penelitian ini adalah sebagai berikut: 1) Bagaimana mengevaluasi fungsionalitas aplikasi GenAI berbasis web untuk mata kuliah PPKN yang dikembangkan terlebih dahulu dengan menggunakan *black-box testing* sebagai tahap awal sebelum dievaluasi kualitas jawaban yang dihasilkan sistem menggunakan metode ROUGE? 2) Bagaimana mengevaluasi kualitas dan relevansi jawaban yang dihasilkan aplikasi GenAI berbasis web untuk mata kuliah PPKN dengan menggunakan metode ROUGE?

Urgensi penelitian ini didorong oleh beberapa faktor kritis. Pertama, digitalisasi kurikulum dan transformasi pembelajaran di era *Society* 5.0 menuntut inovasi teknologi yang dapat meningkatkan kualitas pendidikan kewarganegaraan [4]. Kedua, generasi mahasiswa saat ini memiliki karakteristik *digital native* yang memerlukan pendekatan pembelajaran yang lebih interaktif dan *teknologi-enhanced* [3]. Ketiga, perlunya bukti empiris yang kuat untuk mendukung integrasi AI dalam pendidikan, khususnya untuk mata kuliah yang berkaitan dengan pembentukan karakter nilai-nilai kebangsaan [12].

Dari perspektif praktis, evaluasi pengembangan aplikasi GenAI berbasis web menggunakan metode ROUGE untuk mata kuliah PPKN dapat membantu mengatasi variasi akses dan motivasi belajar mahasiswa, terutama dalam konteks pembelajaran *hybrid* atau jarak jauh [3]. Dari sisi ilmiah, penelitian ini berkontribusi untuk mengisi kekosongan bukti empiris dan menyediakan tolak ukur evaluasi yang terstandar untuk aplikasi GenAI berbasis web dalam pendidikan [13]. Sementara dari aspek etis, penelitian ini penting untuk menguji keselarasan nilai model AI terhadap prinsip-prinsip kewarganegaraan lokal guna mencegah dampak negatif pada pembentukan sikap dan karakter mahasiswa [11].

Penelitian ini bertujuan untuk menjawab beberapa hal yang berkaitan dengan rumusan masalah, antara lain: 1) Memahami proses evaluasi fungsional sistem aplikasi GenAI berbasis web untuk mata kuliah PPKN yang telah dikembangkan terlebih dahulu dengan menggunakan metode *black-box testing* sebagai tahap awal sebelum dilakukan evaluasi kualitas jawaban yang dihasilkan sistem menggunakan metode ROUGE. 2) Memahami proses evaluasi kualitas dan relevansi jawaban yang dihasilkan aplikasi GenAI berbasis web untuk pembelajaran mata kuliah PPKN dengan menggunakan metode ROUGE. Adapun kontribusi penelitian ini terletak pada penyediaan evaluasi yang mengintegrasikan pengujian fungsional sistem dan pengukuran kualitas jawaban berbasis metrik kuantitatif pada konteks pembelajaran PPKN, yang masih dibahas secara terbatas pada penelitian sebelumnya.

Untuk mencapai tujuan tersebut, dilakukan evaluasi sistem GenAI berbasis web yang diintegrasikan ke dalam sistem pembelajaran PPKN dengan menggunakan metode ROUGE dan *black-box testing* untuk pengujian fungsional sistem aplikasi GenAI berbasis web menggunakan *Large Language Model* (LLM) dan *Retrieval-Augmented Generation* (RAG). Dengan demikian, penelitian ini menempatkan pengembangan aplikasi sebagai tahapan pendukung, sedangkan fokus utamanya tetap pada evaluasi fungsionalitas sistem dan kualitas jawaban yang dihasilkan.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

2. METODOLOGI PENELITIAN

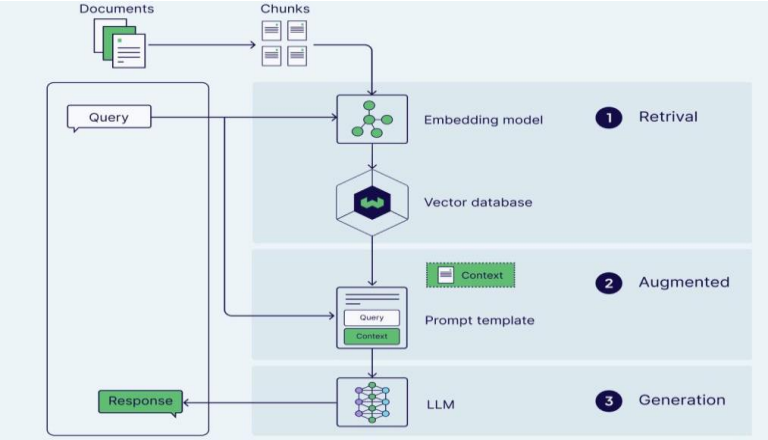
2.1 Kajian Metode yang Digunakan

a. Kecerdasan Artifisial Generatif (GenAI)
Kecerdasan Artifisial Generatif (GenAI) merupakan konsep AI yang mampu menghasilkan teks atau materi pembelajaran secara otomatis. GenAI mendukung proses pembelajaran yang adaptif dan reflektif, karena mahasiswa dapat mengeksplorasi materi PPKN sesuai gaya belajar masing-masing [14].

b. *Large Language Model* (LLM)
Large Language Model (LLM) merupakan komponen yang berperan dalam memahami konteks bahasa alami dan menghasilkan teks pembelajaran yang relevan secara semantik. LLM dilatih menggunakan data besar untuk menyesuaikan hasil jawaban sistem AI dengan konteks PPKN yang bernuansa nilai dan moralitas bangsa [15].

c. *Retrieval-Augmented Generation* (RAG)
Retrieval-Augmented Generation (RAG), bertugas untuk mengambil informasi dari data eksternal untuk meningkatkan faktualitas dan validitas materi yang dihasilkan oleh LLM. Integrasi RAG memastikan bahwa sistem AI tidak hanya kreatif tetapi juga faktual dalam konteks akademik [16].

d. Generative Learning Engine
Generative Learning Engine merupakan gabungan dari dua komponen, yaitu LLM dan RAG. Sistem ini bekerja dengan cara melakukan pencarian informasi yang relevan melalui *embedding model* dan *vector database* (retrieval), selanjutnya data yang ditemukan diolah untuk memperkaya konteks dengan menggunakan *prompt template* (*augmentation*) [17]. Setelah itu LLM menghasilkan jawaban berdasarkan konteks yang telah diperluas, proses ini memungkinkan sistem untuk dapat menghasilkan jawaban yang relevan dengan berbasis pada data referensi [17].



Gambar 1. Proses RAG

Gambar 1 tersebut yang mempresentasikan alur kerja dari *Retrieval-Augmented Generation* (RAG) yang terdiri dari tiga tahap utama, yaitu *retrieval, augmentation,* dan *generation*. Proses RAG dimulai ketika sistem menerima pertanyaan (*Query*) dari pengguna dan melakukan *embedding model* untuk mencari informasi yang relevan di dalam *vector database*. Selanjutnya, informasi yang telah ditemukan dipadukan dengan *query* dalam *prompt template* untuk membentuk konteks yang lebih lengkap (*augmentation*). Terakhir, *Large Language Model* (LLM) memproses konteks tersebut untuk menghasilkan jawaban yang akurat dan relevan berdasarkan data referensi.

e. Pengujian dengan *Black-box Testing*
Pengujian menggunakan *black-box testing* sangat cocok untuk pengujian perangkat lunak aplikasi berbasis web, karena pengujian ini berfokus pada fungsi dan respons sistem terhadap input pengguna. Sistem yang diuji dengan metode ini tidak memerlukan akses kode sumber internal, yang mempermudah proses pengujian dan verifikasi fungsionalitas sistem secara efisien [9].

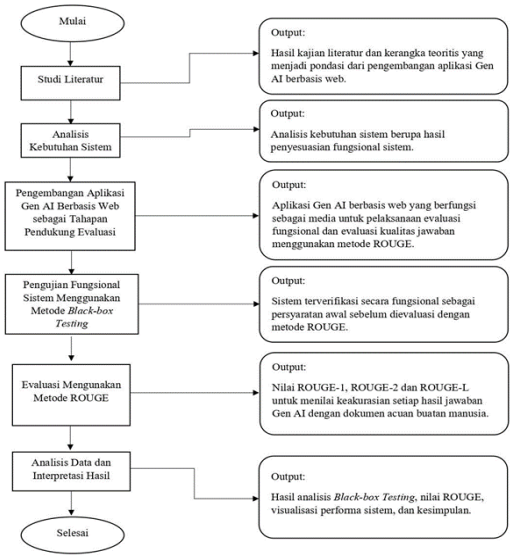
f. Penilaian Kualitas Hasil dengan ROUGE
Bagian ini menunjukkan proses evaluasi teks menggunakan metode ROUGE yang dilakukan secara manual oleh peneliti. Evaluasi ini meliputi perbandingan n-gram overlap, relevansi semantik, dan kelengkapan isi teks acuan dosen PPKN. Pengukuran manual ini memungkinkan penilaian lebih mendalam terhadap konteks penilaian Pancasila dan moralitas yang mungkin tidak dapat terukur secara otomatis [6].

2.2 Tahapan Penelitian

Penelitian ini menggunakan metode kuantitatif dengan fokus utama, yaitu evaluasi kualitas dan relevansi jawaban yang dihasilkan aplikasi GenAI berbasis web menggunakan metode ROUGE (Recall-Oriented Understudy for Gisting Evaluation), sedangkan pengembangan aplikasi GenAI berbasis web hanya dilakukan secara minimal sebagai sarana untuk menghasilkan data uji [5]. Metode kuantitatif digunakan untuk menghasilkan data objektif dan terukur yang dapat dianalisis secara statistik, sesuai dengan praktik evaluasi performa sistem pembelajaran digital di Indonesia [18]. Aplikasi

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

GenAI yang dikembangkan berfungsi sebagai chatbot pembelajaran yang mampu memberikan materi dan jawaban otomatis terkait mata kuliah PPKN dengan memanfaatkan Large Language Model (LLM) dan Retrieval-Augmented Generation (RAG) [15]. Tahapan penelitian disusun secara berurutan untuk memastikan bahwa proses evaluasi dilakukan secara valid, dimulai dari studi literatur hingga analisis data dan interpretasi hasil.



Gambar 2. Tahapan Penelitian

Tahapan penelitian ditunjukkan pada Gambar 2, dimulai dari studi literatur, analisis kebutuhan, pengembangan aplikasi, pengujian *black-box*, evaluasi ROUGE hingga analisis dan interpretasi hasil. Adapun penjelasan pada tiap tahapan sebagai berikut:

a. Studi Literatur
Tahap awal pada penelitian ini adalah studi literatur, pada tahap ini peneliti mengumpulkan konsep dan teori dari penelitian terdahulu yang relevan dengan evaluasi serta pengembangan aplikasi pembelajaran berbasis web menggunakan GenAI. Peneliti juga melakukan kajian terhadap literatur yang membahas LLM, RAG, serta metode ROUGE. Hasil dari kajian ini digunakan untuk membangun landasan teori dan memastikan penelitian memiliki kebaruan yang jelas. Menurut studi oleh Astuti dan Baysha [19], menekankan bahwa studi literatur penting dilakukan untuk menjamin relevansi aplikasi pembelajaran berbasis AI dengan teori pendidikan yang valid.

b. Analisis Kebutuhan
Tahapan selanjutnya adalah melakukan identifikasi kebutuhan fungsional dan non-fungsional sistem yang akan dikembangkan. Peneliti menentukan karakteristik pengguna, jenis data teks yang digunakan untuk evaluasi, serta fitur utama aplikasi seperti antarmuka chatbot dan integrasi AI. Analisis kebutuhan ini dilakukan dengan tujuan untuk memastikan sistem yang dikembangkan relevan dengan kebutuhan pembelajaran dan konteks PPKN. Menurut penelitian oleh Fajriati [18], menegaskan bahwa analisis kebutuhan merupakan langkah fundamental agar rancangan teknologi pembelajaran sesuai dengan konteks dan kebutuhan pengguna.

c. Pengembangan Aplikasi GenAI Berbasis Web sebagai Tahapan Pendukung Evaluasi
Tahapan ini bertujuan untuk membangun aplikasi GenAI berbasis web yang akan digunakan sebagai media dalam proses evaluasi sistem. Aplikasi dikembangkan menggunakan React.js sebagai frontend dan Node.js sebagai backend, serta diintegrasikan dengan API LLM dan RAG. Namun, meskipun aplikasi ini benar-benar dibangun dan berfungsi sebagai chatbot pembelajaran PPKN, pengembangannya bukan merupakan fokus utama pada penelitian ini, melainkan hanya berupa tahapan pendukung untuk menyediakan platform yang dapat dievaluasi fungsionalitas dan kualitas jawaban yang dihasilkannya. Integrasi LLM dengan RAG terbukti efektif dalam meningkatkan relevansi dan akurasi hasil keluaran teks pada sistem AI [20].

d. Pengujian Fungsional sistem
Tahap ini dilakukan pengujian yang bertujuan untuk memastikan seluruh fungsi dari sistem berjalan sesuai ekspektasi dari pengguna dengan menggunakan metode *black-box testing* [21], peneliti memberikan input tertentu dan memeriksa apakah sistem menghasilkan output yang sesuai tanpa memeriksa kode program. Tahap ini meliputi pengujian fungsional antarmuka pengguna, konektivitas API, dan konsistensi respons AI terhadap pertanyaan yang diinput pengguna. Pengujian fungsional ini penting dilakukan untuk memverifikasi stabilitas sistem sebelum masuk ke tahap evaluasi performa [9].

e. Evaluasi Menggunakan Metode ROUGE
Tahap ini merupakan tahap evaluatif untuk menilai kualitas dari jawaban yang dihasilkan sistem GenAI menggunakan metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) [5]. Evaluasi dilakukan dengan membandingkan jawaban hasil GenAI dengan referensi yang disusun oleh manusia. Metode ROUGE dipilih karena

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

mampu mengukur kemiripan linguistik antara jawaban hasil GenAI dan referensi manusia berdasarkan pada jumlah kata atau urutan kalimat yang sama. Indikator yang digunakan dalam metode ini meliputi ROUGE-1, ROUGE-2, dan ROUGE-L, yang masing-masing indikator berfungsi untuk mengukur kemiripan kata tunggal (*unigram*), pasangan kata (*bigram*), dan urutan kata terpanjang (*Longest Common Subsequence*) [22]. Penelitian oleh Halimah, Agustian dan Ramadhani menunjukkan bahwa kombinasi dari ketiga indikator ROUGE ini efektif untuk menilai kemiripan jawaban hasil GenAI dengan referensi yang disusun oleh manusia [23].
Nilai ROUGE dihitung menggunakan rumus sebagai berikut:

1. ROUGE-1
   Mengukur kesamaan teks hasil dan teks referensi dengan menghitung jumlah kata tunggal yang sama atau unigram.

$$\text{ROUGE} - 1\ precision = \frac{jumlah\ unigram\ kata\ sama}{jumlah\ total\ unigram\ dalam\ hasil} \quad (1)$$

$$\text{ROUGE} - 1\ recall = \frac{jumlah\ unigram\ kata\ sama}{jumlah\ total\ unigram\ dalam\ referensi} \quad (2)$$

$$\text{ROUGE} - 1\ f1\text{-}score = 2\text{x}\frac{precision\ \times recall}{precision\ + recall} \quad (3)$$

2. ROUGE-2
   Mengukur kesamaan teks hasil dan teks referensi dengan menghitung jumlah dua kata berturut-turut yang sama atau *bigram*.

$$\text{ROUGE} - 2\ precision = \frac{jumlah\ bigram\ kata\ sama}{jumlah\ total\ bigram\ dalam\ hasil} \quad (4)$$

$$\text{ROUGE} - 2\ recall = \frac{jumlah\ bigram\ kata\ sama}{jumlah\ total\ bigram\ dalam\ referensi} \quad (5)$$

$$\text{ROUGE} - 2\ f1\text{-}score = 2\text{x}\frac{precision\ \times recall}{precision\ + recall} \quad (6)$$

3. ROUGE-L
   Mengukur kesamaan berdasarkan pada LCS yaitu urutan kata terpanjang yang muncul dalam kedua teks secara berurutan namun tidak harus berdekatan.

$$\text{ROUGE} - \text{L}\ precision = \frac{LCS\ dalam\ hasil}{jumlah\ total\ unigram\ dalam\ hasil} \quad (7)$$

$$\text{ROUGE} - \text{L}\ recall = \frac{LCS\ dalam\ hasil}{jumlah\ total\ unigram\ dalam\ referensi} \quad (8)$$

$$\text{ROUGE} - \text{L}\ f1\text{-}score = 2\text{x}\frac{precision\ \times recall}{precision\ + recall} \quad (9)$$

Proses perhitungan nilai ROUGE dilakukan menggunakan *library* evaluasi teks berbasis Python, yaitu *rouge-score*, yang banyak digunakan dalam penelitian terkait *Natural Language Processing* (NLP). Sebelum dilakukan perhitungan, seluruh teks jawaban GenAI dan teks referensi yang telah disusun oleh dosen PPKN melalui tahap *preprocessing* yang meliputi penghapusan spasi berlebih, serta tokenisasi berbasis *whitespace*. Penelitian ini tidak melakukan penghapusan *stopword* maupun *stemming* untuk menjaga keutuhan struktur kalimat asli pada materi PPKN yang bersifat normatif. Perhitungan dilakukan pada metrik ROUGE-1, ROUGE-2, dan ROUGE-L dengan menggunakan pendekatan *F1-score* sebagai indikator utama karena mempertimbangkan keseimbangan antara *precision* dan *recall*. Seluruh proses evaluasi dilakukan secara konsisten pada 50 pasangan data uji dengan konfigurasi parameter yang sama untuk memastikan *reprodusibilitas* hasil.

f. Analisis Data dan Interpretasi Hasil
   Tahapan ini bertujuan untuk mengolah dan menganalisis hasil evaluasi ROUGE secara deskriptif kuantitatif. Nilai ROUGE yang diperoleh dari setiap jawaban GenAI dibandingkan dan dihitung rata-ratanya untuk menentukan tingkat kemiripan antara jawaban yang dihasilkan GenAI dengan referensi yang disusun oleh manusia. Nilai-nilai tersebut kemudian dikategorikan menjadi empat tingkat kualitas teks: ≥ 0,75 (Sangat Baik), 0,50–0,74 (Baik), 0,25–0,49 (Cukup), dan < 0,25 (Kurang). Kategorisasi nilai ROUGE mengacu pada standar evaluasi teks berbahasa Indonesia dalam penelitian terdahulu, sehingga interval kualitas digunakan sebagai batas yang objektif dalam menilai tingkat kesesuaian jawaban yang dihasilkan GenAI [5].

3. HASIL DAN PEMBAHASAN

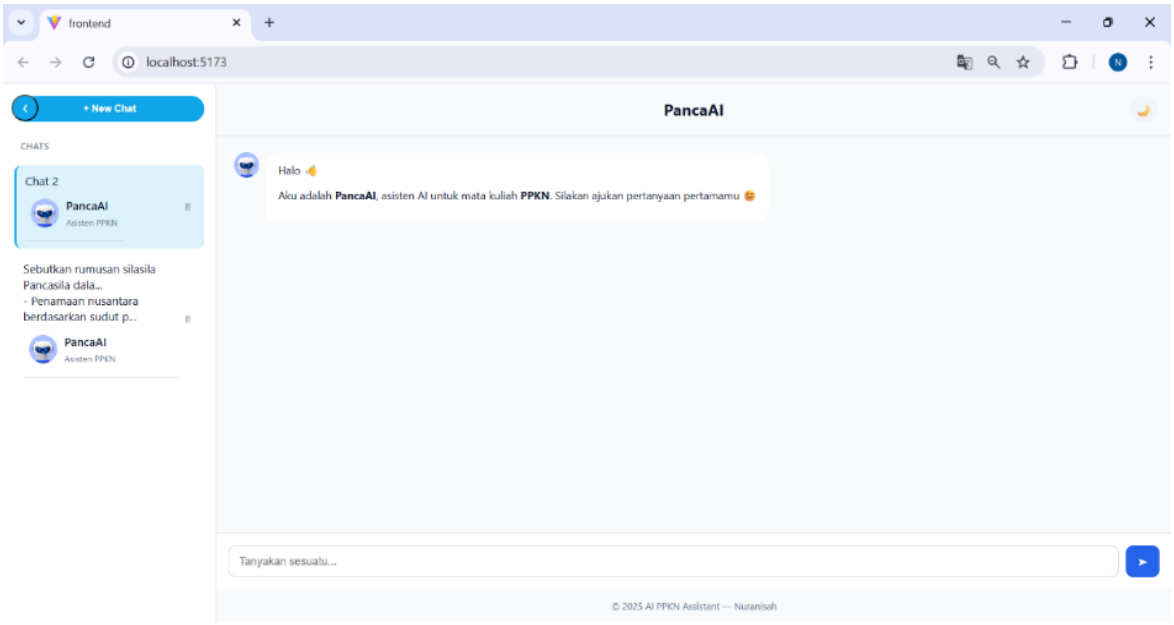
3.1 Implementasi Aplikasi PancaAI Berbasis Web

Aplikasi yang telah dikembangkan diberi nama PancaAI dan berfungsi sebagai sistem pembelajaran berbasis chatbot untuk membantu mahasiswa dalam memahami materi PPKN secara interaktif. Implementasi aplikasi ini bertujuan untuk menghasilkan data uji yang kemudian akan dilakukan pengujian fungsionalitas sistem dengan menggunakan metode *black-box testing* untuk memastikan seluruh sistem berjalan sesuai rancangan, dan penilaian kualitas jawaban yang

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

dihasilkan menggunakan metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) untuk mengukur akurasi serta relevansi jawaban sistem dengan materi referensi.

Antarmuka utama aplikasi PancaAI ditampilkan dalam bentuk halaman chatbot berbasis web. Pada tampilan awal, pengguna akan disambut dengan pesan pengenalan sistem yang menyatakan bahwa PancaAI merupakan asisten AI untuk mata kuliah PPKN. Antarmuka sistem ini terdiri dari beberapa komponen utama, yaitu area navigasi chat di sisi kiri, area percakapan utama di bagian tengah, serta kolom input pertanyaan pada bagian bawah. Menu navigasi chat berfungsi untuk menampilkan riwayat percakapan, serta membuat sesi percakapan baru (*new chat*), serta menghapus riwayat chat. Selain itu, sistem ini juga menyediakan fitur pengaturan tampilan *light mode* dan *dark mode* untuk meningkatkan kenyamanan pengguna dalam berinteraksi dengan sistem.



Gambar 3. Antarmuka Utama Aplikasi PancaAI Berbasis Web

Pada Gambar 3, diperlihatkan tampilan awal sistem ketika pengguna memulai sesi percakapan baru, yang ditandai dengan munculnya pesan inisialisasi dari chatbot. Pesan sapaan tersebut berfungsi sebagai konfirmasi peran sistem kepada pengguna sebagai asisten AI khusus mata kuliah PPKN. Pada bagian bawah, terdapat kolom *text input* dengan label instruksional "Tanyakan sesuatu..." yang bertindak sebagai *entry point* utama bagi pengguna untuk mengirimkan pertanyaan pada sistem.

Proses interaksi pengguna dengan sistem dimulai ketika pengguna mengetikkan pertanyaan pada kolom input yang tersedia. Pertanyaan tersebut kemudian diproses oleh sistem menggunakan mekanisme *Retrieval-Augmented Generation* (RAG). Sistem terlebih dahulu melakukan pencarian konteks materi PPKN yang relevan dari basis data, kemudian *Large Language Model* (LLM) menghasilkan jawaban berdasarkan konteks yang telah diperluas. Jawaban yang dihasilkan sistem ditampilkan secara langsung pada area chat dalam bentuk teks naratif atau poin-poin penjelasan. Sehingga, alur kerja sistem pada penelitian ini dapat dipahami sebagai rangkaian proses input pertanyaan, *retrieval* konteks berbasis RAG, *generation* jawaban oleh LLM, dan penyajian respons kepada pengguna melalui antarmuka web. Uraian dari proses ini penting karena menjadi dasar bagi evaluasi pada dua tahap selanjutnya, yaitu pengujian fungsional sistem dan pengukuran kualitas jawaban yang dihasilkan.

3.2 Hasil Pengujian Fungsional Sistem

Evaluasi terhadap sistem PancaAI dilakukan melalui dua tahap utama, yaitu pengujian fungsional dengan menggunakan metode *black-box testing* dan evaluasi kualitas jawaban sistem menggunakan metode ROUGE. Pendekatan ini dilakukan untuk memastikan bahwa sistem tidak hanya berjalan stabil secara teknis, tetapi juga menghasilkan jawaban yang layak secara akademik. Pengujian fungsional menjadi tahap awal untuk memverifikasi kesesuaian operasional sistem terhadap spesifikasi rancangan, sedangkan evaluasi dengan metode ROUGE digunakan untuk mengukur kualitas tekstual jawaban dibandingkan dengan referensi yang disusun oleh dosen PPKN.

Pengujian fungsional dilakukan terhadap seluruh fitur inti sistem, yang mencakup input pertanyaan, mekanisme *retrieval* berbasis RAG, proses generasi jawaban oleh LLM, serta manajemen sesi percakapan dan pengaturan tampilan antarmuka aplikasi. Rekapitulasi hasil pengujian dapat dilihat pada Tabel 1 berikut:

Tabel 1. Rekapitulasi Hasil Pengujian *Black-box*

No	Fitur yang Diuji	Jumlah Skenario	PASS	FAIL	Tingkat Keberhasilan
1	Input pertanyaan	1	1	0	100%
2	RAG	1	1	0	100%

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

No	Fitur yang Diuji	Jumlah Skenario	PASS	FAIL	Tingkat Keberhasilan
3	Generate jawaban PPKN	1	1	0	100%
4	Generate jawaban non PPKN	1	1	0	100%
5	Riwayat chat	1	1	0	100%
6	New chat	1	1	0	100%
7	Menghapus riwayat chat	1	1	0	100%
8	*Light mode* dan *dark mode*	1	1	0	100%
	Total	8	8	0	100%

Berdasarkan Tabel 1, seluruh skenario pengujian memperoleh status berhasil dengan tingkat keberhasilan 100%. Secara teknis, hasil ini menunjukkan bahwa integrasi antarmuka aplikasi, mekanisme *retrieval* berbasis RAG, pemrosesan oleh LLM, serta manajemen sesi berjalan stabil tanpa ditemukan kesalahan fungsional. Stabilitas sistem ini penting dalam sistem GenAI, karena kegagalan pada salah satu komponen dapat mempengaruhi kualitas respons secara keseluruhan. Namun demikian, secara metodologis perlu ditegaskan bahwa keberhasilan fungsional tidak secara otomatis merepresentasikan kualitas substansi jawaban. Pengujian *black-box* hanya memastikan sistem bekerja sesuai spesifikasi, bukan mengukur ketepatan konseptual atau kedalaman argumentasi. Oleh karena itu, evaluasi lanjutan terhadap kualitas jawaban menjadi tahap yang krusial.

Hasil pengujian ini juga memperjelas bahwa tahapan verifikasi operasional berhasil terpenuhi sebelum dilakukan evaluasi kualitas jawaban yang dihasilkan oleh sistem. Ini artinya, seluruh komponen utama yang menjadi alur kerja sistem, mulai dari input pertanyaan, *retrieval* konteks, *generation* jawaban, hingga pengelolaan sesi dapat digunakan secara konsisten sebagai dasar pengambilan data uji untuk selanjutnya dilakukan evaluasi ROUGE.

Apabila diperhatikan secara lebih rinci, keberhasilan pada fitur input pertanyaan menunjukkan bahwa sistem mampu menerima masukan pengguna tanpa kendala. Keberhasilan pada mekanisme RAG juga menunjukkan bahwa sistem dapat melakukan pencarian konteks materi secara tepat sebelum proses *generation* dilakukan. Kemudian, keberhasilan pada fitur generate jawaban PPKN dan non-PPKN menunjukkan bahwa sistem mampu memberikan respons sesuai dengan ruang lingkup topik pertanyaan yang diberikan. Selain itu, keberhasilan pada fitur riwayat chat, penambahan sesi baru, penghapusan riwayat chat, serta pengaturan tampilan menunjukkan bahwa aspek interaksi pengguna telah berjalan sesuai rancangan. Sehingga, hasil pengujian *black-box* tidak hanya menegaskan kesiapan teknis sistem, tetapi juga memperkuat validitas tahapan evaluasi berikutnya.

3.3 Hasil Evaluasi Kualitas Jawaban Menggunakan ROUGE

Evaluasi kualitas jawaban dilakukan menggunakan metrik *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE), yang meliputi ROUGE-1, ROUGE-2, dan ROUGE-L. Untuk memberikan gambaran sintesis performa sistem secara agregat, ringkasan statistik berupa nilai rata-rata dan rentang skor dijadikan pada Tabel 2 berikut:

Tabel 2. ROUGE *Evaluation Summary*

Metrik	Rata-rata F1	Nilai Minimum	Nilai Maksimum
ROUGE-1	97%	61%	100%
ROUGE-2	97%	58%	100%
ROUGE-L	97%	61%	100%

Berdasarkan Tabel 2, nilai rata-rata pada ketiga metrik menunjukkan tingkat kesamaan tekstual yang sangat tinggi antara jawaban sistem GenAI dan teks referensi yang disusun oleh dosen PPKN, yaitu sekitar 97%. Skor maksimum 100% menunjukkan bahwa pada beberapa pertanyaan, sistem menghasilkan jawaban identik dengan referensi. Hal ini menunjukkan efektivitas mekanisme *retrieval* dalam menyediakan konteks yang sesuai sebelum proses generasi dilakukan. Namun, interpretasi skor tinggi ini harus dilakukan secara kritis. ROUGE-1 mengukur kesamaan *unigram*, sehingga lebih mencerminkan kesesuaian terminologi dibandingkan kemampuan penalaran yang mendalam. Dengan kata lain, skor tinggi pada metrik ini lebih menunjukkan akurasi faktual dan konsistensi terminologi, bukan kompleksitas *reasoning*.

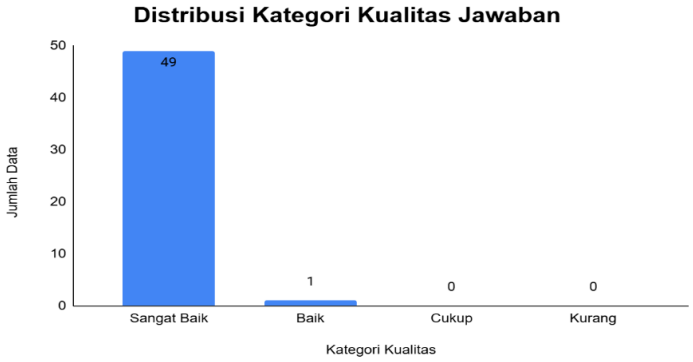
ROUGE-2 dan ROUGE-L memberikan gambaran tambahan mengenai kualitas jawaban yang dihasilkan oleh sistem. ROUGE-2 menunjukkan tingkat kesamaan kata atau frasa, sehingga metrik ini dapat mencerminkan konsistensi sistem dalam menyusun jawaban yang mendekati struktur referensi. Sementara itu, ROUGE-L menunjukkan tingkat kesamaan urutan informasi melalui *longest common subsequence*, sehingga dapat menggambarkan sejauh mana pola penyampaian jawaban sistem berjalan dengan referensi akademik. Dengan begitu, ketiga metrik tersebut dapat saling melengkapi dalam menunjukkan bahwa sistem tidak hanya menghasilkan terminologi yang tepat, tetapi juga menyusun jawaban dengan pola informasi yang relatif konsisten terhadap referensi.

Nilai ROUGE yang telah diperoleh kemudian dikategorikan ke dalam empat tingkat kualitas, yaitu Sangat Baik ($\geq$ 0,75), Baik (0,50–0,74), Cukup (0,25–0,49), dan Kurang ($<$ 0,25), mengacu pada standar evaluasi teks berbahasa Indonesia pada penelitian terdahulu [5]. Hasil kategorisasi menunjukkan bahwa dari 50 data uji yang digunakan, sebanyak 49 data termasuk kategori “Sangat Baik”, 1 data berada pada kategori “Baik”. Sedangkan, untuk kategori “Cukup” dan “Kurang” tidak ditemukan data yang berada pada kategori tersebut.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

Tabel 3. Distribusi Kategori Kualitas Jawaban

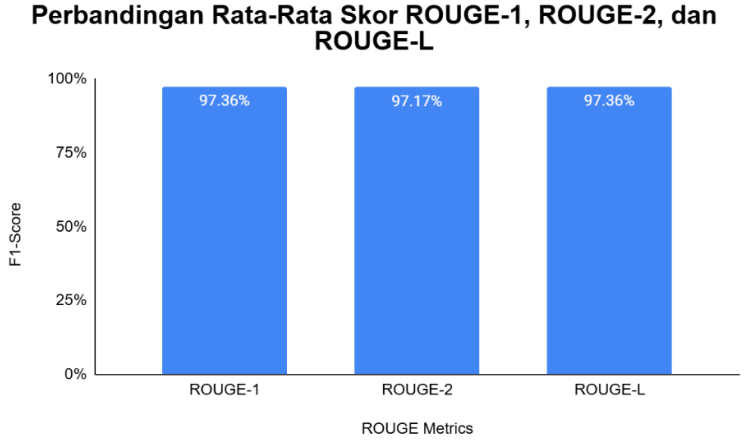
Kategori	Jumlah Data
Sangat Baik (≥ 0,75)	49
Baik (0,50–0,74)	1
Cukup	0
Kurang	0



Gambar 4. Grafik Distribusi Kategori Kualitas Jawaban

Berdasarkan Tabel 3 dan Gambar 4, mayoritas jawaban yang dihasilkan oleh aplikasi PancaAI berada pada kategori "Sangat Baik", yaitu sebanyak 49 dari 50 total data uji yang digunakan. Hanya satu data uji yang termasuk kategori "Baik", sementara tidak ditemukan jawaban yang masuk ke dalam kategori "Cukup" maupun "Kurang". Distribusi ini menunjukkan bahwa secara umum kualitas jawaban yang dihasilkan sistem memiliki tingkat kesesuaian yang sangat tinggi terhadap teks referensi. Dominasi kategori "Sangat Baik" juga mengindikasikan konsistensi performa sistem dalam menghasilkan jawaban yang relevan dan lengkap pada berbagai jenis pertanyaan PPKN. Hasil ini memperkuat temuan bahwa aplikasi PancaAI mampu memberikan respons yang sesuai dengan konteks akademik dan dapat diandalkan sebagai asisten pembelajaran berbasis GenAI dalam mata kuliah PPKN.

Distribusi kategori tersebut juga menunjukkan bahwa performa sistem tidak hanya tinggi pada nilai rata-rata agregat, tetapi juga relatif stabil pada hampir seluruh data uji. Hanya terdapat satu data pada kategori "Baik", yang menunjukkan adanya variasi kecil dalam struktur jawaban, namun tidak menurunkan relevansi utamanya terhadap materi acuan. Dengan demikian, hasil kategorisasi memperkuat interpretasi bahwa sistem memiliki konsistensi keluaran yang baik dalam menjawab pertanyaan faktual, konseptual, dan normatif pada mata kuliah PPKN.



Gambar 5. Perbandingan Rata-rata Skor ROUGE

Pada Gambar 5, diperlihatkan perbandingan rata-rata skor F1 pada tiga metrik evaluasi, yaitu ROUGE-1, ROUGE-2, dan ROUGE-L dengan nilai yang relatif tinggi dan konsisten pada kisaran di atas 97%. ROUGE-1 dan ROUGE-L memiliki nilai yang lebih tinggi dibandingkan ROUGE-2, yang mengindikasikan bahwa sistem GenAI mampu mempertahankan kesesuaian leksikal serta urutan informasi yang selaras dengan referensi, meskipun terdapat sedikit variansi dalam struktur frasa atau pasangan kata. Perbedaan tipis pada ROUGE-2 menunjukkan bahwa adanya fleksibilitas sintaktis dalam proses generasi jawaban oleh model, bukan penurunan kualitas makna. Secara keseluruhan pola ini memperlihatkan bahwa sistem tidak hanya memproduksi terminologi yang relevan, tetapi juga menjaga

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 3, April 2026 | Hal 842-852
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i3.1032

konsistensi struktur informasi, sehingga mendukung temuan bahwa integrasi mekanisme RAG dan LLM mampu menghasilkan jawaban yang stabil dan sesuai dengan konteks materi PPKN.

3.4 Pembahasan

Secara lebih mendalam, tingginya nilai rata-rata ROUGE pada ketiga metrik menunjukkan bahwa sistem GenAI memiliki konsistensi performa dalam menghasilkan jawaban yang sesuai dengan referensi akademik. Hal ini mengindikasikan bahwa mekanisme RAG berfungsi efektif dalam menyediakan konteks yang relevan sebelum proses generasi dilakukan oleh LLM. Dengan memanfaatkan basis data materi PPKN yang telah disesuaikan dengan kurikulum, sistem mampu mengurangi kemungkinan munculnya informasi yang tidak relevan atau menyimpang dari topik. Dengan demikian, hasil ini menunjukkan bahwa integrasi *retrieval* dalam sistem GenAI berkontribusi terhadap peningkatan kesesuaian tekstual dan stabilitas respons.

Namun demikian, evaluasi hasil tidak hanya dapat ditafsirkan sebagai keberhasilan teknis semata. Nilai ROUGE yang tinggi, khususnya pada ROUGE-1 dan ROUGE-L, lebih mencerminkan kesamaan terminologi dan struktur penyampaian dibandingkan kedalaman penalaran konseptual. Dalam konteks pembelajaran PPKN yang tidak hanya menuntut reproduksi informasi, tetapi juga pemahaman nilai dan argumentasi normatif, kesamaan tekstual belum sepenuhnya merepresentasikan kualitas reflektif atau analisis jawaban. Oleh karena itu meskipun secara kuantitatif sistem menunjukkan performa yang sangat baik, evaluasi ini masih terbatas pada dimensi linguistik dan belum mengukur dimensi semantik secara menyeluruh.

Perbedaan kecil antara ROUGE-1, ROUGE-2, dan ROUGE-L juga memberikan gambaran karakteristik sistem. Nilai ROUGE-2 yang sedikit lebih rendah dibandingkan metrik lainnya menunjukkan adanya variasi dalam struktur frasa. Variasi ini dapat diinterpretasikan sebagai fleksibilitas sintaktis model generatif dalam menyusun kalimat tanpa mengubah makna utama. Dengan kata lain, sistem tidak sekedar menyalin teks referensi, melainkan mampu melakukan parafrase dalam batas kesesuaian konteks. Temuan ini mendukung asumsi bahwa model berbasis *transformer* memiliki kemampuan generatif adaptif, meskipun tetap bergantung pada konteks yang disediakan oleh mekanisme *retrieval.*

Jika dibandingkan dengan penelitian terdahulu, hasil penelitian ini sejalan dengan temuan Akter et al. (2022) yang menyatakan bahwa metrik ROUGE efektif dalam mengukur kesamaan teks pada sistem berbasis GenAI [6]. Penelitian ini juga mendukung hasil Hadwirianto et al. (2023) yang menunjukkan bahwa ROUGE pada tugas peringkasan teks (*summarization*) [5], sedangkan penelitian ini mengaplikasikannya pada sistem *Question Answering* dalam domain pembelajaran PPKN. Dengan demikian, penelitian ini memperluas konteks penerapan ROUGE pada evaluasi sistem GenAI berbasis web dalam bidang pendidikan.

Di sisi lain, beberapa penelitian terkait GenAI dalam pendidikan lebih banyak berfokus pada persepsi pengguna atau dampak terhadap motivasi belajar, tanpa melakukan evaluasi kuantitatif terhadap kualitas jawaban sistem. Dalam hal ini, penelitian ini memberikan kontribusi berbeda karena mengintegrasikan pengujian fungsional dan evaluasi kuantitatif berbasis metrik yang terstandarisasi. Pendekatan ini memungkinkan penilaian yang lebih objektif dan dapat direplikasi oleh peneliti lain pada konteks yang serupa.

Meskipun hasil penelitian menunjukkan performa yang sangat baik, terdapat beberapa keterbatasan yang perlu dipertimbangkan dalam interpretasi hasil. Pertama, evaluasi hanya menggunakan metrik berbasis kesamaan n-gram sehingga belum sepenuhnya merepresentasikan kesamaan makna secara semantik. Kedua, penelitian ini belum melakukan perbandingan langsung antara sistem berbasis RAG dan sistem generatif tanpa RAG, sehingga kontribusi spesifik mekanisme *retrieval* terhadap peningkatan kualitas jawaban belum dapat dibuktikan secara eksperimental. Ketiga, jumlah data uji sebanyak 50 pertanyaan meskipun telah mencakup variasi materi faktual, konseptual, dan normatif, masih dapat diperluas untuk meningkatkan generalisasi hasil.

Selain itu, tingginya skor ROUGE yang diperoleh dalam penelitian ini juga perlu ditafsirkan secara hati-hati. Skor yang mendekati kesamaan yang sempurna berpotensi dipengaruhi oleh mekanisme RAG yang mengambil konteks langsung dari basis data yang menjadi referensi. Namun, evaluasi ini lebih mencerminkan tingkat kesesuaian tekstual terhadap materi acuan dibandingkan kemampuan generative model secara independen. Penelitian ini belum membandingkan performa sistem dengan model generative tanpa mekanisme *retrieval*, sehingga kontribusi spesifik RAG terhadap peningkatan skor belum diuji secara eksperimental. Oleh karena itu, penelitian lanjutan disarankan untuk melakukan studi komparatif guna memperkuat validasi utama.

Secara keseluruhan, evaluasi hasil menunjukkan bahwa sistem PancaAI mampu menghasilkan jawaban yang stabil dan konsisten dengan referensi akademik berdasarkan indikator kuantitatif yang digunakan. Integrasi RAG dan LLM terbukti efektif dalam menjaga relevansi dan struktur informasi dalam konteks pembelajaran PPKN. Hasil ini mendukung penelitian terdahulu yang menyatakan bahwa penggabungan mekanisme *retrieval* dan model generatif dapat meningkatkan kualitas respons, sekaligus memberikan kontribusi empiris pada pengembangan sistem GenAI dalam bidang pendidikan kewarganegaraan.

4. KESIMPULAN

Berdasarkan hasil penelitian, evaluasi fungsionalitas aplikasi GenAI berbasis web dilakukan menggunakan metode *black-box testing* sebagai tahap awal sebelum evaluasi kualitas jawaban. Pengujian dilakukan terhadap seluruh fitur inti sistem, meliputi proses input pertanyaan, mekanisme *Retrieval-Augmented Generation* (RAG), proses generasi jawaban oleh

Large Language Model (LLM), serta pengelolaan sesi dan riwayat percakapan. Hasil pengujian menunjukkan tingkat keberhasilan 100% pada seluruh skenario uji, yang mengindikasikan bahwa sistem telah berjalan sesuai dengan spesifikasi rancangan dan stabil secara operasional. Dengan demikian, evaluasi fungsionalitas melalui *black-box testing* terbukti efektif untuk memastikan kesiapan sistem sebelum dilakukan pengujian kualitas jawaban. Selanjutnya, evaluasi kualitas dan relevansi jawaban dilakukan menggunakan metode *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) terhadap 50 pertanyaan yang mencakup materi faktual, konseptual, dan normatif terkait pembelajaran PPKN. Hasil evaluasi menunjukkan rata-rata nilai F1-score sebesar 97% pada metrik ROUGE-1, ROUGE-2, dan ROUGE-L. Nilai tersebut menunjukkan tingkat kesamaan tekstual yang sangat tinggi antara jawaban sistem dan referensi akademik yang digunakan sebagai *gold standard*. Tingginya skor ROUGE-1 dan ROUGE-L mengindikasikan kesesuaian terminologi dan struktur penyampaian informasi, sedangkan perbedaan kecil pada ROUGE-2 menunjukkan adanya variasi sintaktis tanpa mengurangi relevansi makna. Dengan demikian, metode ROUGE dapat digunakan secara efektif untuk mengevaluasi kualitas dan relevansi jawaban sistem GenAI dalam konteks pembelajaran PPKN. Meskipun hasil penelitian menunjukkan performa yang sangat baik, evaluasi yang dilakukan masih terbatas pada pengukuran kesamaan teks berbasis n-gram dan belum sepenuhnya merepresentasikan kedalaman pemahaman konseptual secara semantik. Oleh karena itu, penelitian selanjutnya disarankan untuk mengintegrasikan metrik berbasis kesamaan semantik atau penilaian pakar independen guna memperoleh evaluasi yang lebih komprehensif. Selain itu, penelitian lanjutan direkomendasikan untuk melakukan studi komparatif antara sistem berbasis *Retrieval-Augmented Generation* (RAG) dan sistem generatif tanpa mekanisme *retrieval*. Perbandingan tersebut dapat dilakukan dengan mengukur perbedaan skor ROUGE maupun indikator evaluasi lainnya untuk mengetahui secara lebih objektif kontribusi mekanisme *retrieval* terhadap peningkatan kualitas jawaban. Dengan pendekatan tersebut, validitas empiris mengenai efektivitas integrasi RAG dalam sistem GenAI untuk pembelajaran PPKN dapat diperkuat secara metodologis.

REFERENCES

[1] A. Verma and N. V, “Harnessing Generative AI in Education: From Theory to Real-World Impact,” Preprints, May 30, 2025, doi: 10.20944/preprints202505.1177.v3.

[2] S. Harahap and Z. N. Napitupulu, “Pengaruh Teknologi terhadap Pendidikan di Indonesia: Systematic Literature Review,” Rekognisi: Jurnal Pendidikan dan Kependidikan, vol. 8, no. 2, Dec. 2023. [Online]. Available: https://jurnal.unusu.ac.id/index.php/rekognisi/article/view/162

[3] D. Hermawan, C. Dermawan, and P. Bestari, “Transforming Citizenship Education in the Digital Era: Challenges and Opportunities for the Indonesian Millennial Generation,” Unnes Political Science Journal, vol. 8, no. 1, pp. 30–38, Jun. 2024, doi: 10.15294/upsj.v8i1.5783.

[4] Fatimah and D. A. Nugroho, “Strengthening Digital Citizenship Values in Pancasila dan Civics Learning in the 21st Century,” in Proceedings of the 4th Annual Civic Education Conference (ACEC 2022), 2023, pp. 948–954, doi: 10.2991/978-2-38476-096-1_99.

[5] M. R. Hadwiriantor, F. Hamami, and O. N. Pratiwi, “Extractive Text Summarization terhadap Artikel Berita Indonesia Berbasis Machine Learning,” eProceedings of Engineering, vol. 11, no. 4, Jul. 2024. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/23804

[6] M. Akter, N. Bansal, and S. K. Karmaker, “Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE?,” in Findings of the Association for Computational Linguistics: ACL 2022, May 2022, pp. 1547–1560, doi: 10.18653/v1/2022.findings-acl.122.

[7] I. D. Saputra, N. S. Harahap, S. Agustian, M. Fikry, and L. Oktavia, “Aplikasi Web Question Answering Menggunakan Langchain OpenAI tentang Peraturan Perundang-undangan Bidang Pendidikan,” Journal of Computer System and Informatics (JoSYC), vol. 6, no. 1, pp. 293–304, Nov. 2024, doi: 10.47065/josyc.v6i1.6182.

[8] A. Yuniarti et al., “Aplikasi Konsultasi Psikologi Berbasis Flutter dan ChatGPT Menggunakan Metode Extreme Programming,” Journal of Digital Business and Technology Innovation (DBESTI), vol. 2, no. 1, pp. 14–20, 2025. [Online]. Available: https://journal.nurulfikri.ac.id/index.php/DBESTI

[9] R. Darman, “Peran ChatGPT Sebagai Artificial Intelligence dalam Menyelesaikan Masalah Pertanahan dengan Metode Studi Kasus dan Black Box Testing,” Tunas Agraria, vol. 7, no. 1, pp. 18–46, Jan. 2024, doi: 10.31292/jta.v7i1.256.

[10] J. Zhang and D. Sun, “A Systematic Review of Generative Artificial Intelligence in Education,” in 2025 7th International Conference on Computer Science and Technologies in Education (CSTE), Apr. 2025, pp. 552–556, doi: 10.1109/CSTE64638.2025.11092288.

[11] R. D. Agustin, S. Wiyono, and R. Yamanto, “Analysis of Value Alignment and Ethical Guardianship of Learning with AI in Civic Education,” Jurnal Moral Kemasyarakatan, vol. 9, no. 2, pp. 255–265, Nov. 2024, doi: 10.21067/jmk.v9i2.10650.

[12] Z. Chen and W. Zhou, “Ethical Shifts and Innovative Approaches to Civic Education under Generative Artificial Intelligence,” in Proceedings of the 2024 2nd International Conference on Language, Innovative Education and Cultural Communication (CLEC 2024), 2024, pp. 181–187, doi: 10.2991/978-2-38476-263-7_25.

[13] I. Jurenka et al., “Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach,” arXiv preprint arXiv:2407.12687, Dec. 2025. [Online]. Available: http://arxiv.org/abs/2407.12687

[14] Q. Xia, X. Weng, F. Ouyang, T. J. Lin, and T. K. F. Chiu, “A Scoping Review on How Generative Artificial Intelligence Transforms Assessment in Higher Education,” International Journal of Educational Technology in Higher Education, Dec. 2024, doi: 10.1186/s41239-024-00468-z.

[15] D. Lee et al., “The Impact of Generative AI on Higher Education Learning and Teaching: A Study of Educators’ Perspectives,” Computers and Education: Artificial Intelligence, vol. 6, Jun. 2024, doi: 10.1016/j.caeai.2024.100221.

[16] M. Klesel and H. F. Wittmann, “Retrieval-Augmented Generation (RAG),” Business and Information Systems Engineering, vol. 67, no. 4, pp. 551–561, Aug. 2025, doi: 10.1007/s12599-025-00945-3.

[17] G. D. Albert and A. Voutama, "Pengembangan Chatbot Berbasis PDF Menggunakan Local Retrieval-Augmented Generation (RAG) dan Ollama," Jurnal Informatika dan Teknik Elektro Terapan, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6361.
[18] A. Fajriati, W. Wisroni, and C. Handrianto, "Pemanfaatan Teknologi Artificial Intelligence (AI) dalam Pembelajaran Berbasis Peserta Didik di Era Digital," Wahana Pedagogika, vol. 6, no. 2, pp. 71–85, Dec. 2024.
[19] E. R. P. Astuti and M. H. Baysha, "Evaluasi Efektivitas Sistem Umpan Balik Berbasis AI dalam Meningkatkan Hasil Belajar Mahasiswa," EDUTECH: Jurnal Inovasi Pendidikan Berbantuan Teknologi, vol. 4, no. 3, pp. 122–136, 2024, doi: 10.51878/edutech.v4i3.3142.
[20] A. S. Wiradinata and V. C. Mawardi, "Abstractive Text Summarization Berita Bahasa Indonesia Menggunakan Retrieval-Augmented Generation," Jurnal Ilmu Komputer dan Sistem Informasi, vol. 13, no. 1, 2025, doi: 10.24912/jiksi.v13i1.32861.
[21] Firdaus, I. E. Putra, F. Kesumaningtyas, N. Sahrun, and T. Hadyanto, "Perancangan Sistem Cashless Payment Berbasis Aplikasi Mobile dan Web Menggunakan Teknologi QR Code," Jurnal Sains Informatika Terapan, vol. 4, no. 3, pp. 547–553, Oct. 2025, doi: 10.62357/jsit.v4i3.786.
[22] Z. Idhafi, S. Agustian, F. Yanto, and N. Safaat H, "Peringkas Teks Otomatis pada Artikel Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance," Jurnal CoSciTech (Computer Science and Information Technology), vol. 4, no. 3, pp. 609–618, Dec. 2023, doi: 10.37859/coscitech.v4i3.6311.
[23] Halimah, S. Agustian, and S. Ramadhani, "Peringkasan Teks Otomatis (Automated Text Summarization) pada Artikel Berbahasa Indonesia Menggunakan Algoritma LexRank," Jurnal CoSciTech (Computer Science and Information Technology), vol. 3, no. 3, pp. 371–381, Dec. 2022, doi: 10.37859/coscitech.v3i3.4300.