BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1415-1423
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1154

Analisis Limitasi Performa Penilaian Esai Otomatis pada Aplikasi ESAO Berdasarkan Metrik BLEU dan ROUGE

Akhmam Fahmi, Nuraini*, Maulana Fakih Latief

Program Studi Teknik Informatika, Sekolah Tinggi Teknologi Terpadu Nurul Fikri, Depok, Indonesia
Email: [1]akhmam.fahmi@nurulfikri.ac.id, [2,*]nura22170ti@student.nurulfikri.ac.id, [3]maulana.latief@nurulfikri.ac.id
Email Penulis Korespondensi: nura22170ti@student.nurulfikri.ac.id

Abstrak−Perkembangan GenAI telah mendorong pemanfaatan teknologi *automated essay scoring* melalui berbagai platform, salah satunya adalah aplikasi ESAO (*Essay Analytic Online*). Meskipun sistem berbasis LLM ini mampu menghasilkan narasi umpan balik penilaian secara otomatis, standarisasi metode evaluasi untuk mengukur keandalan teks tersebut masih menghadapi tantangan besar. Penelitian ini bertujuan untuk menguji kesesuaian metrik *Bilingual Evaluation Understudy* (BLEU) dan *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) sebagai instrumen pengukur performa luar tekstual dari aplikasi ESAO. Metode penelitian dilakukan dengan mengomparasikan teks umpan balik dari ESAO terhadap draf penilaian otentik dosen pada tiga karakteristik materi ujian yang berbeda, yaitu analisis kondisi *dataset*, statistik deskriptif, serta korelasi dan regresi. Hasil pengujian menunjukkan nilai rata-rata metrik BLEU sebesar 0,0522 dan ROUGE sebesar 0,1255. Penelitian ini mengungkap bahwa rendahnya perolehan angka tersebut bukan merepresentasikan kegagalan fungsional dari aplikasi ESAO, melainkan menunjukkan adanya limitasi dan ketidaksesuaian mendasar dari penggunaan metrik leksikal kaku (*word-based metrics*) dalam menilai teks generatif yang dinamis. Metrik BLEU dan ROUGE sangat bergantung pada tumpang-tindih unit kata (*n-gram overlap*) yang kaku, sehingga gagal menangkap kesamaan substansi semantik, konteks penalaran akademik, dan variasi linguistik yang dihasilkan oleh ESAO. Penelitian ini menyimpulkan bahwa metrik evaluasi tradisional seperti BLEU dan ROUGE kurang akurat dan tidak kompatibel untuk dijadikan standar ukur tunggal bagi performa *Generative AI* dalam konteks penilaian pendidikan, sehingga diperlukan transisi menuju metrik evaluasi berbasis kedekatan semantik (*semantic-based metrics*) di masa depan.

Kata Kunci: Automated Essay Scoring; Generative Artificial Intelligenc; BLEU; ROUGE; ESAO

Abstract−The development of GenAI has encouraged the use of automated essay scoring technology through various platforms, one of which is the ESAO (Essay Analytic Online) application. Although this LLM-based system is capable of automatically generating assessment feedback narratives, standardizing evaluation methods to measure the reliability of these texts still faces significant challenges. This study aims to test the suitability of the Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics as instruments to measure the extratextual performance of the ESAO application. The research method was carried out by comparing feedback texts from ESAO with authentic lecturer assessment drafts on three different characteristics of the exam material: dataset condition analysis, descriptive statistics, and correlation and regression. The test results showed an average value of the BLEU metric of 0.0522 and ROUGE of 0.1255. This study revealed that low scores do not represent a functional failure of the ESAO application, but rather indicate fundamental limitations and shortcomings in using rigid lexical metrics (word-based metrics) in assessing dynamic generative texts. The BLEU and ROUGE metrics rely heavily on rigid n-gram overlap, thus failing to capture the semantic similarity, academic reasoning context, and linguistic variation generated by ESAO. This study concludes that traditional evaluation metrics such as BLEU and ROUGE are inaccurate and incompatible as a single benchmark for Generative AI performance in the context of educational assessment, necessitating a transition to semantic-based metrics in the future.

Keywords: Automated Essay Grading; Generative Artificial Intelligence; BLEU; ROUGE; ESAO

1. PENDAHULUAN

Integrasi *Generative Artificial Intelligence* (GenAI) telah memicu transformasi yang sangat signifikan di berbagai sektor global, termasuk dalam pendidikan tinggi yang membawa paradigma baru dalam proses pembelajaran dan penilaian [1]. Pemanfaatan model bahasa besar, *Large Language Models* (LLM) seperti ChatGPT, Gemini dan model lainnya tidak hanya terbatas pada pencarian informasi, selain itu juga mendukung proses pembelajaran, mulai dari membantu mahasiswa memahami materi yang kompleks hingga memberikan umpan balik akademik dengan mudah [2]. Namun, di balik potensi efisiensi dan skalabilitas yang ditawarkan GenAI, muncul tantangan krusial terkait validitas dan akurasi dalam menilai kompetensi kognitif tingkat tinggi, khususnya pada tugas-tugas esai [3]. Penilaian esai secara manual oleh dosen merupakan proses yang memakan waktu dan sumber daya, sehingga mendorong pengembangan sistem *Automated Essay Scoring* (AES) untuk meringankan beban ini [4]. Meskipun demikian, kesenjangan antara kecepatan inovasi teknologi GenAI dan kebutuhan akan validasi ilmiah yang ketat dalam konteks penilaian pendidikan menjadi isu mendesak yang perlu diatasi.

Aplikasi ESAO (Essay Analytic Online) hadir sebagai platform edukasi berbasis GenAI di Indonesia yang dirancang untuk melatih dan menilai kemampuan mahasiswa melalui studi kasus. Aplikasi ini memanfaatkan teknologi LLM untuk memberikan skor awal terhadap elemen-elemen penalaran mahasiswa. Meskipun telah diimplementasikan dalam kegiatan akademis, performa sistem ESAO belum pernah diuji secara empiris melalui perbandingan sistematis dengan penilaian dosen. Tanpa evaluasi yang ilmiah, tingkat akurasi dan reliabilitas dari umpan balik penilai yang dihasilkan oleh aplikasi ini masih menjadi pertanyaan. Penelitian ini bertujuan untuk mengisi kekosongan tersebut dengan melakukan analisis kritis terhadap limitasi performa penilaian esai otomatis pada aplikasi ESAO, khususnya dalam diskrepansi antara penilaian GenAI dan standar penilaian manusia.

Penelitian terkait pemanfaatan LLM dalam asesmen pendidikan telah menunjukkan perkembangan pesat dalam beberapa tahun terakhir. Hal ini didorong oleh Ramesh dan Sanampudi (2021) bahwa sistem penilaian esai otomatis telah

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1415-1423
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1154

dikembangkan menggunakan teknik kecerdasan buatan, tantangan utama tetap terletak pada sulitnya menilai parameter relevansi isi dan koherensi secara akurat dibanding dengan penilaian manual [5]. Dalam aspek teknis penilaian esai, Nur Rokhman *et al.* (2025) menunjukkan bahwa penggunaan metode *Cosine Similarity* yang dikombinasikan dengan pembobotan TF-IDF sangat efektif untuk mengukur kemiripan semantik antara jawaban mahasiswa dengan kunci jawaban dosen, sehingga mampu menekan subjektivitas dan inkonsistensi penilai dosen [6]. Sebagai upaya penguatan metodologi, Ayaan dan Ng (2025) mengusulkan evaluasi hibrida yang mengintegrasikan teknik *Natural Language Processing* (NLP) tradisional seperti kesamaan *Jaccard* dan *cosine*, dengan analisis semantik mendalam menggunakan *Universal Sentence Encoder* guna menangkap esensi argumen secara luas dan mendalam [7]. Penelitian-penelitian ini umumnya berfokus pada pengembangan model AES atau perbandingan metrik tradisional, namun seringkali belum secara eksplisit menyoroti diskrepansi performa yang signifikan antara AI dan penilaian manusia, terutama dalam konteks umpan balik penilaian.

Evaluasi terhadap performa model AI itu sendiri menjadi krusial untuk menjamin kualitas *output*. Penelitian oleh Maulana Nur Rokhim *et al*. (2025) mengevaluasi akurasi dan presisi LLM dalam menghasilkan teks teknis, di mana ditemukan bahwa meskipun LLM memiliki reliabilitas tinggi, tetap diperlukan metrik evaluasi yang ketat agar *output* yang dihasilkan tetap sesuai dengan parameter kualitas yang ditetapkan [8]. Di sisi lain, Fianu et al. (2025) dalam tinjauan sistematisnya tentang model pretrained dalam AES, menunjukkan bahwa meskipun sistem berbasis *transformer* dan LLM seringkali memiliki nilai *Quadratic Weighted Kappa* (QWK) yang lebih tinggi, peningkatan tersebut lebih terkait dengan keselarasan rubrik dan pemodelan spesifik sifat daripada skala model itu sendiri [9]. Xu dan Mahmud (2024) juga melakukan tinjauan sistematis mengenai kompetensi sistem AES dalam skenario pendidikan nyata, menyoroti tantangan implementasi di lapangan [10]. Berbeda dengan penelitian-penelitian sebelumnya yang cenderung berfokus pada peningkatan akurasi model atau perbandingan metrik yang ada, penelitian ini secara spesifik menginvestigasi sejauh mana aplikasi ESAO, sebagai implementasi GenAI, mampu mendekati standar penilaian dosen, dengan fokus pada identifikasi dan analisis diskrepansi yang muncul.

Metode evaluasi yang digunakan dalam penelitian ini adalah metrik *Bilingual Evaluation Understudy* (BLEU) dan *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE). BLEU mengukur presisi teks yang dihasilkan mesin dengan membandingkan tumpang tindih n-gram terhadap teks referensi, sedangkan ROUGE menitikberatkan pada aspek *recall* atau cakupan informasi penting [11]. Meskipun BLEU dan ROUGE merupakan standar yang telah lama digunakan dalam evaluasi teks otomatis, terutama dalam terjemahan mesin dan peringkasan, terdapat metrik lain yang lebih canggih seperti METEOR (*Metric for Evaluation of Translation with Explicit Ordering*) dan BERTScore. METEOR mengatasi beberapa keterbatasan BLEU dengan mempertimbangkan kesamaan semantik, *stemming*, dan *paraphrasing*, sehingga berkorelasi lebih baik dengan penilaian manusia [12]. Sementara itu, BERTScore memanfaatkan *contextual embeddings* dari model BERT untuk mengukur kesamaan semantik antar kalimat, menawarkan evaluasi yang lebih mendalam dibandingkan metrik berbasis n-gram [13].

Namun, dalam penelitian ini, BLEU dan ROUGE tetap digunakan sebagai parameter dasar kuantitatif untuk memetakan diskrepansi awal antara *output* mesin dan teks referensi manusia. Justifikasi penggunaan metrik ini adalah sebagai langkah awal kuantitatif untuk mengidentifikasi pola diskrepansi pada tingkat leksikal dan cakupan informasi, sebelum analisis kualitatif lebih lanjut atau penerapan metrik berbasis semantik yang lebih kompleks. Diskusi akademik mengenai keterbatasan metrik ini dalam menangkap esensi semantik pada umpan balik pedagogis diakui, namun skor yang dihasilkan tetap menjadi indikator valid untuk kesesuaian tekstual awal.

Kebaruan penelitian ini terletak pada analisis mendalam terhadap diskrepansi performa aplikasi ESAO, sebuah platform AES berbasis GenAI di Indonesia, dibandingkan dengan penilaian dosen. Dengan menggunakan kombinasi metrik BLEU dan ROUGE, penelitian ini tidak hanya mengukur tingkat kesamaan, tetapi juga secara kritis menginterpretasikan implikasi dari skor yang rendah terhadap validitas dan reliabilitas sistem. Tujuan utama penelitian ini adalah untuk mengevaluasi secara kuantitatif limitasi performa GenAI pada aplikasi ESAO, memberikan gambaran objektif mengenai keandalan sistem sebagai alat bantu, dan menyajikan kontribusi nyata bagi pengembang aplikasi ESAO dalam kalibrasi model AI mereka. Diharapkan, penelitian ini dapat meningkatkan kepercayaan terhadap penggunaan alat bantu penilaian berbasis AI dengan memastikan bahwa teknologi yang digunakan benar-benar transparan dan objektif dalam mendukung pengembangan kemampuan intelektual mahasiswa, sekaligus menyoroti area-area yang memerlukan perbaikan signifikan.

2. METODOLOGI PENELITIAN

2.1 Kajian Metode Penelitian

Kajian metode penelitian ini memaparkan landasan teoretis dan teknologi yang digunakan untuk menganalisis limitasi performa aplikasi ESAO. Pembahasan dimulai dari konsep kecerdasan buatan generatif sebagai objek utama, serta penggunaan metrik pengolahan bahasa alami sebagai instrumen evaluasi kuantitatif untuk mengidentifikasi bahasa tertentu.

a. *Generative Artificial Intelligence*
   Generative Artificial Intelligence atau GenAI merupakan cabang dari kecerdasan buatan yang berfokus pada penciptaan konten baru berdasarkan pola yang dipelajari dari data yang ada [14]. Dalam penelitian ini, fungsi GenAI dibatasi hanya untuk memberikan skor angka, tanpa kemampuan untuk menghasilkan narasi umpan balik penilaian

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1415-1423
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1154

yang kompleks. Kemampuan sistem ini terbatas pada penilaian kuantitatif yang kaku, sehingga tidak dapat meniru gaya atau substansi penjelasan kualitatif layaknya seorang dosen [15].

b. Large Language Models

Large Language Models (LLM) adalah arsitektur di balik GenAI yang dilatih menggunakan dataset teks berskala masif. Pada aplikasi ESAO, LLM berperan sebagai mesin pengolah kata yang memprediksi setiap kata dalam umpan balik penilaian. Sifat generatif LLM memungkinkan variasi dalam respons, namun hal ini juga berpotensi menimbulkan inkonsistensi dan ketidakakuratan, yang menjadi salah satu limitasi yang dieksplorasi dalam penelitian ini [5].

c. *Natural Language Processing*

Natural Language Processing (NLP) adalah bidang ilmu yang menjembatani interaksi antara bahasa manusia dengan komputer [16]. Dalam penelitian ini, teknik NLP digunakan sebagai metodologi evaluasi untuk mengukur kemiripan semantik dan struktural antara dua buah teks. Pendekatan ini lebih unggul dibandingkan pencocokan kata kunci sederhana karena sistem mampu memahami konteks dan makna esai secara luas untuk menghasilkan akurasi skor angka yang setara dengan penilaian dosen

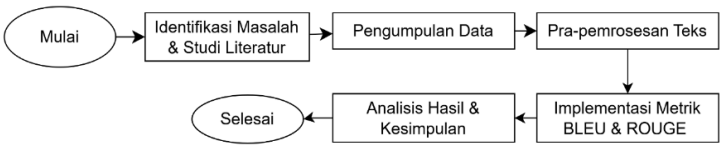
d. *Bilingual Evaluation Understudy*

BLEU adalah metrik evaluasi otomatis dalam NLP yang digunakan untuk mengukur tingkat presisi dari teks yang dihasilkan oleh mesin [17]. Cara kerja BLEU adalah dengan menghitung tumpang tindih unit kata atau *n-gram* antara teks hasil GenAI pada ESAO dengan teks penilaian dosen. Dalam penelitian ini, nilai BLEU digunakan untuk mendeteksi keaslian dan akurasi informasi dalam umpan balik penilaian, serta memastikan tidak adanya halusinasi atau informasi tambahan yang tidak relevan dengan jawaban mahasiswa.

e. *Recall-Oriented Understudy for Gisting Evaluation*

ROUGE merupakan metrik evaluasi yang menitikberatkan pada aspek *recall* atau cakupan informasi [18]. Jika BLEU fokus pada ketepatan kata yang muncul, ROUGE berfungsi untuk mengukur seberapa banyak informasi penting dari kunci jawaban dosen yang berhasil ditangkap dan dituliskan kembali oleh sistem ESAO. Variasi ROUGE-L digunakan untuk melihat urutan kata terpanjang yang sama (*Longest Common Subsequence*), yang sangat efektif dalam mengevaluasi apakah alur logika berpikir kritis mahasiswa telah dinilai dengan benar oleh sistem.

2.2 Tahapan Penelitian



Gambar 1. Tahap penelitian

Berdasarkan Gambar 1, tahapan penelitian ini dirancang untuk menganalisis limitasi performa penilaian esai otomatis
pada aplikasi ESAO. Proses ini melibatkan serangkaian langkah sistematis yang dijelaskan
sebagai berikut:

a. Identifikasi Masalah dan Studi Literatur

Pada tahap ini dilakukan pengkajian terhadap permasalahan utama dalam penelitian, yaitu identifikasi limitasi performa sistem GenAI pada aplikasi ESAO dibandingkan dengan penilaian dosen. Selain itu, dilakukan studi literatur terhadap konsep-konsep yang relevan seperti GenAI, *automated essay scoring*, serta metrik evaluasi berbasis text similarity menggunakan BLEU dan ROUGE. Hasil dari tahap ini adalah perumusan masalah yang jelas dan landasan teoretis yang kuat untuk mendukung penelitian.

b. Pengumpulan Data

Tahap ini bertujuan untuk memperoleh data yang akan digunakan dalam proses evaluasi. Jumlah data yang digunakan dalam penelitian ini masih terbatas karena penelitian difokuskan sebagai studi awal (*preliminary study*) untuk mengevaluasi performa *automated essay scoring* berbasis GenAI pada aplikasi ESAO. Meskipun jumlah soal dan data uji belum besar, data yang digunakan telah mewakili beberapa tipe analisis yang berbeda, yaitu analisis kondisi dataset, statistik deskriptif, serta korelasi dan regresi. Dengan demikian, data tersebut dinilai cukup untuk memberikan gambaran awal mengenai kemampuan sistem dalam melakukan penilaian esai secara otomatis.

c. Pra-pemrosesan Teks

Tahap ini dilakukan untuk menyiapkan data agar berada dalam kondisi yang konsisten dan siap untuk dianalisis. Proses pra-pemrosesan meliputi *case folding* untuk menyeragamkan huruf, penghapusan tanda baca, tokenisasi untuk memecah teks menjadi unit kata, serta pembersihan teks dari tanda baca dan karakter yang tidak relevan. Hasil dari tahap ini adalah data teks yang telah bersih dan siap digunakan dalam proses evaluasi.

d. Implementasi Metrik BLEU dan ROUGE

Tahap ini merupakan proses inti dalam penelitian, yaitu melakukan evaluasi terhadap performa sistem GenAI. Pada tahap ini dilakukan perbandingan antara teks hasil penilaian GenAI dengan teks referensi yang diberikan oleh dosen. Metrik BLEU digunakan untuk mengukur tingkat kesesuaian berdasarkan presisi n-gram, sedangkan metrik ROUGE

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1415-1423
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1154

digunakan untuk mengukur kelengkapan dan relevansi informasi berdasarkan *recall*. Hasil dari tahap ini berupa nilai skor evaluasi yang merepresentasikan performa sistem dalam menyamai penilaian dosen.

e. Analisis Hasil dan Kesimpulan
Tahap terakhir adalah melakukan analisis terhadap hasil evaluasi yang telah diperoleh. Skor BLEU dan ROUGE dianalisis untuk mengetahui tingkat kesesuaian antara penilaian GenAI dan dosen, serta kualitas umpan balik penilaian yang dihasilkan oleh sistem. Hasil analisis ini kemudian digunakan sebagai dasar dalam menyusun kesimpulan penelitian, termasuk implikasi dari limitasi yang ditemukan dan rekomendasi untuk pengembangan lebih lanjut.

2.3 Metode Evaluasi

Metode evaluasi pada penelitian ini menggunakan pendekatan *text similarity* untuk mengukur tingkat kesamaan antara hasil umpan balik penilaian otomatis yang dihasilkan oleh sistem GenAI pada aplikasi ESAO dengan penilaian dosen. Pendekatan ini dipilih karena penelitian berfokus pada limitasi performa sistem dalam mereplikasi penilaian manusia. Metrik evaluasi yang digunakan dalam penelitian ini adalah BLEU (*Bilingual Evaluation Understudy*) dan ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). BLEU digunakan untuk mengukur tingkat kesamaan struktur kalimat dan pemilihan kata berdasarkan presisi n-gram, sedangkan ROUGE digunakan untuk mengukur tingkat kelengkapan dan relevansi informasi berdasarkan *recall* [19]. Penggunaan kedua metrik tersebut dilakukan secara terpisah agar dapat memberikan evaluasi yang lebih komprehensif terhadap performa sistem *automated essay scoring* berbasis GenAI.

Pada proses evaluasi, teks hasil penilaian GenAI digunakan sebagai *candidate text*, sedangkan penilaian dosen digunakan sebagai *reference text*. Sebelum proses perhitungan dilakukan, kedua teks terlebih dahulu melalui tahap pra-pemrosesan untuk memastikan konsistensi data. Selanjutnya dilakukan perhitungan skor BLEU dan ROUGE untuk setiap pasangan data guna mengetahui tingkat kesesuaian antara hasil penilaian sistem dan penilaian dosen. Untuk mengukur tingkat kesamaan struktur teks, penelitian ini menggunakan rumus BLEU sebagai berikut:

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n \log\ p_n) \quad (1)$$

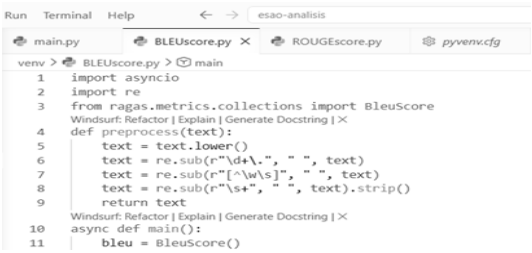
Perhitungan skor BLEU dilakukan untuk mengukur tingkat kesamaan struktur teks antara luaran sistem dengan referensi manusia. Nilai akhir BLEU, yang berada pada rentang 0 hingga 1, diperoleh melalui fungsi eksponensial dari jumlah tertimbang logaritma alami presisi n-gram (pn*pn*). Dalam rumus ini, N*N* menunjukkan jumlah tingkatan n-gram yang dipertimbangkan, sementara wn*wn* merepresentasikan bobot (*weight*) yang diberikan pada setiap tingkatan n-gram tersebut. Selain itu, terdapat komponen *Brevity Penalty* (BP) yang berfungsi sebagai penalti apabila teks hasil penilaian GenAI jauh lebih pendek dibandingkan dengan teks referensi dosen, guna memastikan evaluasi yang lebih adil terhadap panjang teks [20].

Sementara itu, untuk mengukur tingkat cakupan kelengkapan dan relevansi informasi digunakan rumus ROUGE-N sebagai berikut.

$$ROUGE - N = \frac{\sum_n -gram\ \ min(Count\ mahasiswa, Count\ dosen)}{\sum_n -gram\ Count\ dosen} \quad (2)$$

Sementara itu, metrik ROUGE-N digunakan untuk mengevaluasi tingkat cakupan kelengkapan dan relevansi informasi dalam umpan balik yang dihasilkan. Skor ROUGE-N, yang juga memiliki rentang nilai antara 0 hingga 1, dihitung dengan membandingkan jumlah tumpang tindih n-gram minimum antara jawaban mahasiswa (*Count mahasiswa*) dan kunci jawaban dosen (*Count dosen*) terhadap total seluruh n-gram pada dokumen acuan. Penggunaan fungsi minimum (*min*) dalam rumus ini bertujuan untuk membatasi pengaruh tumpang tindih kata yang berulang secara berlebihan, sehingga hasil evaluasi tetap objektif dalam mencerminkan seberapa banyak informasi penting dari dosen yang berhasil ditangkap oleh sistem [21].

Penelitian ini menggunakan bahasa pemrograman Python sebagai *tools* utama dalam proses pengolahan data dan evaluasi teks karena memiliki dukungan yang luas pada bidang *Natural Language Processing* (NLP). Untuk mendukung proses komputasi metrik evaluasi secara terstruktur, penelitian ini memanfaatkan modul evaluasi teks dari *framework* RAGAS (*Retrieval Augmented Generation Assessment*). Meskipun RAGAS umumnya dikenal untuk pengujian sistem berbasis *knowledge-base*, penelitian ini secara spesifik hanya mengisolasi fungsi komputasi metrik tradisional RougeScore dan BleuScore yang tersedia di dalam ekosistemnya untuk mengukur kedekatan linguistik teks. Implementasi pemanggilan fungsi evaluasi berbasis skrip asinkronus Python menggunakan *framework* tersebut disajikan secara ringkas pada tabel berikut.



Gambar 2. Implementasi *Framework* RAGAS

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1415-1423
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1154

Gambar 1 berikut adalah penggunaan *framework* RAGAS yang bertujuan untuk mengintegrasi metrik BLEU dan ROUGE, sehingga mampu memberikan perspektif yang lebih holistik terhadap batasan performa sistem ESAO.

Implementasi sistem ESAO diawali dari proses autentikasi pengguna, pemilihan soal esai, hingga penginputan jawaban mahasiswa. Setelah jawaban dikirimkan, sistem GenAI akan memproses teks untuk menghasilkan skor dan umpan balik penilaian otomatis berdasarkan isi jawaban mahasiswa. Hasil keluaran tersebut kemudian digunakan sebagai *candidate text* dalam proses evaluasi penelitian. Pada tahap implementasi evaluasi, hasil penilaian otomatis dari sistem ESAO dibandingkan dengan penilaian dosen sebagai *reference text*. Sebelum proses evaluasi dilakukan, data teks terlebih dahulu melalui tahap *preprocessing* yang meliputi *case folding*, penghapusan tanda baca, tokenisasi, dan normalisasi spasi untuk memastikan konsistensi data. Selanjutnya dilakukan proses perhitungan skor BLEU dan ROUGE guna mengukur tingkat kesamaan struktur teks serta kelengkapan informasi antara hasil penilaian sistem dan penilaian dosen. Melalui rangkaian prosedur tersebut, performa *automated essay scoring* berbasis GenAI pada aplikasi ESAO dapat dievaluasi secara kuantitatif dan objektif. Adapun perangkat dan teknologi yang digunakan dalam penelitian ini dirangkum pada Tabel berikut.

Tabel 1. *Tools* dan Teknologi Penelitian

Komponen	Teknologi/Metode
Bahasa Pemrograman	Python
Framework Evaluasi	RAGAS
Sistem AES	ESAO
Model AI	*Generative* AI berbasis LLM
Metrik Evaluasi	BLEU dan ROUGE

Tabel 1 menunjukkan bahwa arsitektur teknologi dalam penelitian ini mengombinasikan keunggulan bahasa Python dan model Generative AI berbasis LLM untuk menjalankan sistem ESAO. Keandalan hasil penilaian tersebut kemudian divalidasi menggunakan *framework* RAGAS yang mengintegrasikan metrik BLEU dan ROUGE.

3. HASIL DAN PEMBAHASAN

Hasil penelitian disajikan untuk memetakan tingkat kesesuaian antara umpan balik penilaian otomatis yang dihasilkan oleh sistem GenAI pada aplikasi ESAO dengan narasi penilaian dari dosen. Analisis ini bertujuan untuk mengidentifikasi limitasi performa leksikal dan cakupan informasi dari sistem berbasis LLM dalam skenario penilaian esai analitis.

3.1 Hasil Pengujian Data

Data yang digunakan dalam pengujian pada penelitian ini merupakan jawaban esai dari 50 mahasiswa pada Ujian Tengah Semester (UTS) yang dikerjakan melalui aplikasi ESAO. Data uji terdiri dari tiga soal esai yang menuntut mahasiswa untuk melakukan analisis data, analisis statistik deskriptif, serta analisis hubungan antar variabel. Setiap jawaban mahasiswa kemudian dinilai oleh sistem GenAI pada aplikasi ESAO dan dibandingkan dengan penilaian dosen sebagai acuan (*reference text*). Perbandingan tersebut dilakukan menggunakan metrik BLEU dan ROUGE sebagaimana telah dijelaskan pada BAB III. Adapun tiga soal yang digunakan sebagai data pengujian pada Tabel 2.

Tabel 2. Daftar soal UTS

No	Contoh Soal Implisit	Fokus Analisis
1.	Analisis data tersebut, meliputi ukuran dataset, tipe data dari tiap variabel, keberadaan *missing values*, data duplikat, dan *outlier*, serta berikan kesimpulan apakah data tersebut dapat langsung diolah.	Kualitas dan kesiapan data
2.	Analisis seluruh variabel data dengan statistik deskriptif, uji normalitas distribusi data, serta berikan kesimpulan dari hasil analisis tersebut	Statistik deskriptif dan normalitas
3.	Analisis korelasi, regresi, koefisien determinasi, dan kuadrat residu antara lama belajar dengan hasil skor, serta berikan kesimpulan atas analisis tersebut.	Hubungan antar variabel dan regresi

Berdasarkan Tabel 2, narasi instrumen soal yang disajikan telah diringkas dari dokumen aslinya, Soal pertama berfokus pada kemampuan mahasiswa dalam memahami kondisi awal dataset, seperti kelengkapan data, tipe variabel, dan keberadaan nilai ekstrem. Soal kedua menilai kemampuan mahasiswa dalam melakukan analisis statistik deskriptif dan memahami distribusi data. Sementara itu, soal ketiga mengukur kemampuan mahasiswa dalam menganalisis hubungan antara variabel lama belajar dan hasil skor melalui korelasi, regresi, koefisien determinasi, serta kuadrat residu. Hasil penilaian dari sistem GenAI kemudian dibandingkan dengan penilaian dosen untuk mengetahui tingkat kesesuaian menggunakan metrik BLEU dan ROUGE. Untuk memberikan gambaran mengenai data acuan yang digunakan dalam proses evaluasi, ditampilkan contoh penilaian dosen sebagai *reference text* pada tabel 3.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1415-1423
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1154

Tabel 3. Jawaban Dosen sebagai *Reference Text* Objektif

No	Soal UTS	Penilaian Dosen
1.	Analisis kondisi dataset	Dataset terdiri dari 255 baris dan 6 kolom dengan tipe data mayoritas numerik (*float*). Ditemukan beberapa permasalahan yaitu *missing values* pada semua variabel numerik, data duplikat (5 entri), serta *outlier* pada beberapa variabel. Oleh karena itu, dataset memerlukan proses *preprocessing* (pembersihan data) sebelum analisis lebih lanjut. Secara umum, data menunjukkan variasi yang cukup beragam. Jam belajar memiliki variasi tinggi, jam tidur relatif stabil, kehadiran dan nilai sebelumnya cukup heterogen, sedangkan nilai ujian memiliki variasi sedang dengan mayoritas berada pada kisaran menengah. Hal ini menunjukkan karakteristik mahasiswa yang beragam dalam perilaku belajar dan performa akademik.
2.	Statistik deskriptif dan uji normalitas	Secara umum, data menunjukkan variasi yang cukup beragam. Jam belajar memiliki variasi tinggi, jam tidur relatif stabil, kehadiran dan nilai sebelumnya cukup heterogen, sedangkan nilai ujian memiliki variasi sedang dengan mayoritas berada pada kisaran menengah. Hal ini menunjukkan karakteristik mahasiswa yang beragam dalam perilaku belajar dan performa akademik. Berdasarkan uji Shapiro-Wilk, hanya variabel *exam_score* yang berdistribusi normal, sedangkan variabel lainnya tidak. Ini menunjukkan bahwa sebagian besar data tidak memenuhi asumsi normalitas.
3.	Korelasi dan regresi	Terdapat hubungan yang kuat antara jam belajar dan nilai ujian dengan koefisien korelasi Pearson sebesar 0,78, yang berarti semakin lama waktu belajar, semakin tinggi nilai ujian yang diperoleh. Model regresi menunjukkan bahwa 60% variasi nilai ujian dapat dijelaskan oleh jam belajar (R^2 = 0,60). Persamaan regresi: *Exam Score = 23,62 + 1,63 × Jam Belajar*. Artinya setiap tambahan 1 jam belajar meningkatkan nilai ujian sekitar 1,63 poin. Nilai residual menunjukkan masih terdapat sekitar 40% variasi yang tidak dijelaskan model, sehingga faktor lain selain jam belajar juga mempengaruhi nilai ujian.

Tabel 3 menunjukkan contoh penilaian dosen yang digunakan sebagai acuan dalam proses evaluasi. Penilaian ini menjadi dasar pembanding terhadap hasil penilaian yang dihasilkan oleh sistem GenAI.

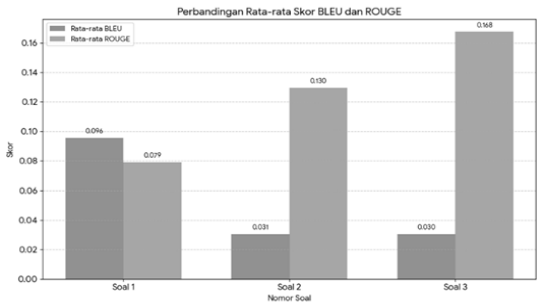
3.2 Analisis Metrik BLEU dan ROUGE

Hasil perhitungan metrik menunjukkan diskrepansi yang signifikan antara luaran sistem GenAI dan penilaian dosen. Tabel 4 merangkum skor yang diperoleh untuk setiap kategori soal.

Tabel 4. Hasil Evaluasi BLEU dan ROUGE

Soal UTS	BLEU *Score*	ROUGE *Score*
Analisis kondisi dataset	0,09575	0,07922
Statistik deskriptif dan uji normalitas	0,03055	0,12960
Korelasi dan regresi	0,03035	0,16770
Rata-rata	0,0522	0,1255

Berdasarkan Tabel 4, nilai BLEU dan ROUGE pada setiap soal menunjukkan tingkat kesesuaian yang berbeda antara penilaian GenAI dan penilaian dosen. Nilai BLEU digunakan untuk melihat kesamaan struktur dan pemilihan kata antara hasil penilaian aplikasi ESAO dan dosen, sedangkan nilai ROUGE digunakan untuk melihat sejauh mana informasi penting dalam penilaian dosen tercakup dalam hasil penilaian aplikasi ESAO. Perbedaan nilai pada setiap soal dapat dipengaruhi oleh tingkat kompleksitas jawaban yang dianalisis. Dengan demikian, hasil pengujian pada tiga soal UTS ini digunakan untuk memperoleh gambaran awal mengenai performa sistem *automated essay scoring* berbasis GenAI pada aplikasi ESAO dalam menilai jawaban esai mahasiswa. Nilai BLEU dan ROUGE yang diperoleh menjadi dasar untuk analisis lebih lanjut pada bagian berikutnya. Untuk mempermudah dalam memahami hasil pengujian, data evaluasi yang telah diperoleh kemudian divisualisasikan dalam bentuk grafik. Visualisasi ini bertujuan untuk memberikan gambaran yang lebih jelas mengenai perbandingan nilai BLEU dan ROUGE pada setiap soal UTS yang digunakan sebagai data pengujian.



Gambar 3. Perbandingan skor BLEU dan ROUGE

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1415-1423
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1154

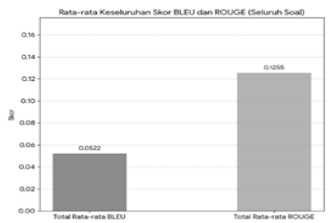
Berdasarkan Gambar 8 terlihat jelas bahwa nilai ROUGE pada setiap soal cenderung lebih tinggi dibandingkan dengan nilai BLEU. Fenomena ini mengindikasikan bahwa sistem GenAI pada aplikasi ESAO lebih mampu mencakup informasi penting dari penilaian dosen dibandingkan dengan mencocokkan struktur atau susunan kata secara presisi. Hal ini sejalan dengan karakteristik model bahasa besar yang cenderung menghasilkan teks yang koheren dan informatif, namun mungkin tidak selalu mereplikasi gaya atau formulasi kalimat yang spesifik dari teks referensi.

Pada Soal 1, ditemukan fenomena unik di mana nilai BLEU 0,09575 mencapai titik tertingginya di antara ketiga soal dan bahkan berhasil mengungguli nilai ROUGE 0,07922. Secara linguistik komputasi, kondisi ini dapat dijelaskan oleh sifat soal pertama yang memiliki batasan token informasi yang sangat kaku. Ketika mahasiswa diminta menganalisis ukuran dataset dan keberadaan data duplikat, baik sistem ESAO maupun dosen secara inheren dipaksa oleh konteks teknis untuk menggunakan kata-kata kunci baku seperti "*missing values*", "*outlier*", "duplikat", "baris", dan "kolom". Keterbatasan variasi sinonim untuk istilah teknis ini menyebabkan tumpang-tindih uni-gram (satu kata) dan bi-gram (dua kata) menjadi sangat tinggi, yang secara langsung meningkatkan skor BLEU. Tingginya presisi leksikal ini mencerminkan kesamaan terminologi yang tidak dapat dihindari dalam domain spesifik ini. Namun, nilai ROUGE untuk Soal 1 menjadi yang terendah. Hal ini disebabkan oleh adanya deskripsi kualitatif tambahan yang sangat panjang dari dosen mengenai manifestasi perilaku belajar mahasiswa di bagian akhir catatan referensinya. Penilaian dosen tidak hanya berfokus pada aspek teknis, tetapi juga meluas ke interpretasi pedagogis. Karena model LLM pada ESAO dirancang untuk fokus mendeteksi parameter teknis dan konten akademik tanpa memperluas narasi ke arah interpretasi psikologi belajar atau observasi perilaku, banyak informasi dari teks referensi dosen yang gagal dicakup (*low recall*) oleh sistem ESAO. Kegagalan ini secara langsung menurunkan skor ROUGE pada variasi soal ini, mengindikasikan bahwa sistem ESAO belum mampu menangkap dimensi kualitatif yang lebih luas dari umpan balik dosen.

Pada Soal 2, struktur performa metrik mengalami pergeseran yang sangat drastis. Nilai BLEU merosot tajam ke angka 0,03055, sementara nilai ROUGE melonjak naik ke angka 0,12960. Penurunan nilai BLEU yang signifikan ini mendeteksi adanya jurang pemisah yang lebar pada aspek susunan sintaksis kalimat (*phrase structure*) dan pilihan kata yang lebih bervariasi. Sebagai contoh, dalam menginterpretasikan kegagalan pemenuhan asumsi normalitas, model GenAI pada ESAO cenderung memproduksi kalimat formal yang kaku seperti: "Data tidak berdistribusi normal karena nilai signifikansi berada di bawah ambang batas". Di sisi lain, teks catatan dosen mungkin menggunakan formulasi yang lebih naratif dan kontekstual, seperti: "Ini menunjukkan bahwa sebagian besar data tidak memenuhi asumsi normalitas". Secara semantik dan esensi konseptual, kedua kalimat tersebut menyampaikan kebenaran materi yang sama. Namun, secara matematis, metrik BLEU menghukum perbedaan susunan kata ini dengan skor mendekati nol karena tidak adanya keselarasan urutan token (*n-gram mismatch*). Hal ini menyoroti sensitivitas BLEU terhadap variasi leksikal dan sintaksis, bahkan ketika makna inti tetap terjaga. Sebaliknya, metrik ROUGE mengalami kenaikan yang signifikan pada Soal 2 karena metrik ini mengukur recall. Mengingat kata-kata kunci utama seperti "normalitas", "distribusi", dan nama uji statistik "Shapiro-Wilk" tetap muncul di kedua dokumen (umpan balik ESAO dan penilaian dosen), ROUGE merekam bahwa sistem ESAO telah berhasil menangkap substansi materi yang diujikan oleh dosen, meskipun dikemas dalam struktur kalimat yang berbeda. Ini menunjukkan bahwa meskipun ESAO mungkin tidak mereplikasi gaya penulisan dosen, ia berhasil mengidentifikasi dan menyertakan konsep-konsep kunci yang relevan dengan topik statistik.

Pada Soal 3, menunjukkan polarisasi metrik yang paling ekstrem, dengan nilai BLEU sebesar 0,03035 dan pencapaian ROUGE tertinggi sebesar 0,16770. Soal korelasi dan regresi menuntut tingkat interpretasi kognitif tertinggi karena melibatkan representasi matematis berupa angka koefisien, nilai determinasi, serta rumus persamaan regresi linear. Rendahnya nilai BLEU pada soal ini disebabkan oleh perbedaan kecil namun krusial dalam cara LLM dan dosen menuliskan rumus dan angka desimal. Bagi metrik pencocokan karakter string murni seperti BLEU, perbedaan kecil pada penulisan karakter, spasi, atau simbol matematika dianggap sebagai kesalahan total (mismatch), sehingga nilai presisinya jatuh secara drastis. Ini menggarisbawahi keterbatasan BLEU dalam mengevaluasi konten yang sangat terstruktur dan sensitif terhadap format. Namun, skor ROUGE untuk Soal 3 justru menjadi yang tertinggi. Hal ini dikarenakan struktur umpan balik GenAI pada ESAO mampu mengikuti alur logika berpikir kaku dari interpretasi regresi. Konsep-konsep kunci seperti korelasi "Pearson", nilai kontribusi variasi ("60%"), serta interpretasi residual ("40%") berhasil diekstrak dan dituliskan kembali oleh sistem ESAO. Karena kuantitas kata kunci konseptual yang berhasil diselamatkan (*recalled*) dari referensi dosen cukup banyak, skor ROUGE naik secara signifikan. Ini membuktikan bahwa sistem memiliki pemahaman konteks hubungan antar variabel yang kuat dan mampu mengidentifikasi elemen-elemen penting dari analisis regresi, meskipun dengan formulasi yang berbeda dari dosen. Tingginya ROUGE menunjukkan kemampuan sistem untuk menangkap esensi informasi, bahkan jika presisi leksikalnya rendah.

3.3 Pembahasan dan Analisis Limitasi



Gambar 4. Rata-rata keseluruhan skor BLEU dan ROUGE

Berdasarkan Gambar 4 hasil evaluasi, diperoleh rata-rata nilai BLEU sebesar 0,0522 dan ROUGE sebesar 0,1255. Secara akademis, skor ini menunjukkan bahwa sistem GenAI pada aplikasi ESAO memiliki limitasi performa yang sangat serius dalam mereplikasi standar penilaian dosen. Pembahasan berikut menganalisis penyebab teknis dan konseptual dari rendahnya skor tersebut tanpa mengabaikan kegagalan sistem yang terdeteksi.

a. Kegagalan Presisi Leksikal (Skor BLEU Rendah)
   Skor BLEU yang hanya mencapai 0,05 mengindikasikan bahwa sistem gagal dalam menyamai struktur kalimat dan pilihan kata yang digunakan oleh dosen. Analisis mendalam menunjukkan bahwa sistem GenAI cenderung memberikan umpan balik penilaian, sementara dosen menggunakan terminologi teknis yang sangat spesifik dan padat (misalnya, penyebutan angka statistik yang presisi). Perbedaan gaya penulisan ini menyebabkan tumpang tindih *n-gram* menjadi sangat minimal, yang secara teknis menjelaskan rendahnya skor BLEU.
b. Keterbatasan Cakupan Informasi (Skor ROUGE Rendah)
   Meskipun skor ROUGE 0,12 sedikit lebih tinggi daripada BLEU, nilai ini tetap menunjukkan bahwa sistem hanya mampu menangkap sekitar 12% informasi penting dari penilaian dosen. Hal ini mengindikasikan adanya masalah pada relevansi konten. Sistem seringkali melewatkan detail krusial seperti nilai koefisien korelasi yang spesifik atau interpretasi mendalam terhadap *outlier*, yang justru menjadi fokus utama dalam penilaian dosen.
c. Limitasi Metrik Evaluasi
   Penelitian ini juga menyoroti bahwa metrik leksikal seperti BLEU dan ROUGE mungkin memiliki keterbatasan dalam mengevaluasi umpan balik penilaian dalam pendidikan yang bersifat semantik. Namun, dalam konteks penelitian ini, rendahnya skor tersebut tetap menjadi indikator valid bahwa sistem belum mampu menghasilkan luaran yang mendekati standar *gold standard* yang ditetapkan.

Meskipun demikian, keberhasilan sistem dalam mencetak skor ROUGE yang lebih tinggi pada soal-soal kompleks membuktikan potensi aplikasi ESAO sebagai instrumen pendukung yang aman dari gejala *AI hallucination*. Sistem terbukti tetap beroperasi dalam korpus akademik yang relevan. Hasil eksperimen ini memberikan rekomendasi kuat bagi pengembangan sistem selanjutnya untuk mengadopsi teknik *Few-Shot Prompting* guna menyelaraskan gaya bahasa sistem dengan gaya penulisan dosen, serta urgensi perluasan alat ukur menggunakan metrik berbasis kedekatan vektor semantik (*embedding-based similarity*) pada pengujian skala besar di masa mendatang.

4. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan mengenai evaluasi performa teks umpan balik pada aplikasi ESAO, penelitian ini berhasil membuktikan adanya limitasi metodologis yang signifikan pada penggunaan metrik BLEU dan ROUGE. Berdasarkan data pengujian, diperoleh rata-rata nilai BLEU sebesar 0,0522 dan ROUGE sebesar 0,1255. Diskusi kritis dalam penelitian ini menegaskan bahwa capaian angka yang rendah tersebut tidak serta merta mencerminkan tidak andalan sistem ESAO dalam menghasilkan narasi penilaian. Fenomena ini merupakan dampak langsung dari karakteristik metrik BLEU dan ROUGE yang bekerja secara leksikal mekanis, di mana sistem hanya menghitung tumpang-tindih *n-gram mismatch* secara harfiah terhadap draf acuan dosen. Kelemahan mendasar dari metrik ini adalah ketidakmampuannya dalam mendeteksi *semantic equivalence* dan penalaran kontekstual. Meskipun aplikasi ESAO mampu memberikan teks umpan balik dengan substansi akademik yang benar, variasi sinonim, perbedaan struktur sintaksis, dan fleksibilitas gaya bahasa generatif di luar teks acuan dosen akan langsung dinilai salah oleh metrik BLEU dan ROUGE. Oleh karena itu, ketidaksesuaian instrumen ukur inilah yang memicu bias hasil evaluasi. Implikasinya terhadap ekosistem akademik adalah perlunya kehati-hatian dalam menafsirkan skor NLP tradisional pada teknologi GenAI agar tidak terjadi salah kesimpulan terhadap performa aplikasi. Sebagai rekomendasi teknis, aplikasi ESAO tetap memiliki potensi besar sebagai alat bantu dosen dalam efisiensi penilaian, namun proses evaluasi performanya di masa mendatang harus beralih dari metrik berbasis kata ke metrik berbasis pemahaman bahasa alami yang lebih lanjut, seperti BERTScore atau *framework* RAGAS. Hal ini diperlukan demi mendapatkan representasi validitas yang objektif terhadap kualitas penalaran kecerdasan buatan.

REFERENCES

[1] E. Shidbringoid, “Proses Adopsi Teknologi Generative Artificial Intelligence Dalam Dunia Pendidikan : Perspektif Teori Difusi Inovasi Adoption Process Of Generative Artificial Intelligence Technology In Education : Diffusion Of Innovation Theory Perspective,” *J. Pendidik. Dan Kebud.*, Vol. 9, No. 1, Pp. 110–133, 2024, Doi: 10.24832/Jpnk.V9i1.4859.
[2] B. A. Dewantara And L. K. Dewi, “Generative Ai Dalam Pembelajaran Mahasiswa: Antara Inovasi Pendidikan Dan Integritas Akademik Keywords: Kata Kunci,” *J. Ilm. Ilmu Pendidik.*, Vol. 8, No. 7, Pp. 8209–8217, 2025, Doi: 10.54371/Jiip.V5i12.1910.
[3] D. Baidoo-Anu And L. O. Ansah, “Education In The Era Of Generative Artificial Intelligence (Ai): Understanding The Potential Benefits Of Chatgpt In Promoting Teaching And Learning,” *J. Ai*, Vol. 7, No. December, Pp. 52–62, 2023, Doi: 10.61969/Jai.1337500.
[4] J. Atkinson And D. Palma, “An Llm-Based Hybrid Approach For Enhanced Automated Essay Scoring,” *Sci. Rep.*, Vol. 15, No. 14551, Pp. 1–9, 2025, Doi: 10.1038/S41598-025-87862-3.
[5] D. Ramesh And S. K. Sanampudi, “An Automated Essay Scoring Systems: A Systematic Literature Review,” *Artif. Intell. Rev.*, Vol. 55, No. 3, Pp. 2495–2527, 2022, Doi: 10.1007/S10462-021-10068-2.
[6] N. Rokhman, P. A. Maulan, And N. A. Wirahuda, “Analisis Penilaian Esai Secara Otomatis Menggunakan Natural Language Processing (Nlp) Dan Cosine Similarity,” *Go Infotech J. Ilm. Stmik Aub*, Vol. 31, No. 1, Pp. 41–52, 2025, Doi:

10.36309/Goi.V31i1.359.

[7] A. Ayaan And K. Ng, "Automated grading using natural language processing and semantic analysis," MethodsX, Vol. 14, P. 103395, 2025, Doi: 10.1016/J.Mex.2025.103395.

[8] A. Info, "Evaluasi Akurasi Dan Presisi Large Language Model (Llm)," *J. Ilm. Inform. With Cc By Nc Licence*, Vol. 10, No. 1, Pp. 48–60, 2025, Doi: 10.35316/Jimi.V10i1.48-60.

[9] E. Fianu, F. Amankwah-Sarfo, P. Ofori, J. K. Amoako, And H. Sumani, "From Traditional Machine Learning Models To Large Language Models: A Systematic Literature Review Of Automated Essay Scoring," *Sn Comput. Sci.*, Vol. 7, No. 5, P. 406, 2026, Doi: 10.1007/S42979-026-05028-Y.

[10] W. Xu, R. Mahmud, And W. A. I. L. A. M. Hoo, "A Systematic Literature Review : Are Automated Essay Scoring Systems Competent In Real-Life Education Scenarios ?," *Ieee Educ. Soc. Sect.*, Vol. 12, No. June, Pp. 77639–77657, 2024, Doi: 10.1109/Access.2024.3399163.

[11] R. Junqueira And V. P. Moreira, "The Inadequacy Of Automatic Evaluation Metrics In Question Answering : A Case-Study In Portuguese," *Proc. 17th Int. Conf. Comput. Process. Port.*, Vol. 1, No. Propor, Pp. 551–561, 2026, [Online]. Available: Https://Aclanthology.Org/2026.Propor-1.54.Pdf

[12] M. S. Maksum, T. Arifin, R. Rohidin, M. Azril, B. Prasetya, And I. Fardian, "Optimalisasi Algoritma Terjemahan Bahasa Dengan Model Transformer: Pendekatan Statistical Machine Learning," *Infotech J.*, Vol. 10, No. 2, Pp. 282–287, 2024, Doi: 10.31949/Infotech.V10i2.11132.

[13] Y. Yuniati, K. M. Fitria, Melvi, S. Purwiyanti, E. Nasrullah, And M. A. Muhammad, "Analisis Performa Ekstraksi Konten Gpt-3 Dengan Matrik Bertscore Dan Rouge," *J. Teknol. Inf. Dan Ilmu Komput.*, Vol. 11, No. 6, Pp. 1273–1280, 2024, Doi: 10.25126/Jtiik.2024118088.

[14] L. Banh And G. Strobel, "Generative Artificial Intelligence," *Electron. Mark.*, Vol. 33, No. 1, Pp. 1–17, 2024, Doi: 10.1007/S12525-023-00680-1.

[15] B. Arslan *Et Al.*, "Opportunities And Challenges Of Using Generative Ai To Personalize Educational Assessment," *Front. Artif. Intell.*, Vol. 7, No. Perspective, Pp. 1–8, 2024, Doi: 10.3389/Frai.2024.1460651.

[16] S. M. S. Mohammadabadi, B. C. Kara, C. Eyupoglu, C. Uzay, M. S. Tosun, And O. Karakuss, "A Survey Of Large Language Models : Evolution , Architectures , Adaptation , Benchmarking , Applications , Challenges , And Societal Implications," *Electronics*, Vol. 14, No. 18, Pp. 1–31, 2025, Doi: 10.3390/Electronics14183580.

[17] E. Reiter, "A Structured Review Of The Validity Of Bleu," *Comput. Linguist.*, Vol. 44, No. 3, Pp. 393–401, 2025, Doi: 10.1162/Coli_A_00322.

[18] M. Barbella And G. Tortora, "Rouge Metric Evaluation For Text Summarization Techniques," *Ssrn Electron. J.*, Pp. 1–31, 2022, Doi: 10.2139/Ssrn.4120317.

[19] D. Li *Et Al.*, "From Generation To Judgment : Opportunities And Challenges Of Llm-As-A-Judge," *Emnlp*, Vol. Proceeding, Pp. 2758–2792, 2025, Doi: 10.18653/V1/2025.Emnlp-Main.138.

[20] K. Papineni, S. Roukos, T. Ward, And W. Zhu, "B Leu : A Method For Automatic Evaluation Of Machine Translation," No. July, Pp. 311–318, 2002, Doi: 10.3115/1073083.1073135.

[21] C.-Y. Lin, "Rouge : A Package For Automatic Evaluation Of Summaries," *Assoc. Comput. Linguist.*, Vol. Text Summa, Pp. 74–81, 2004, [Online]. Available: Https://Aclanthology.Org/W04-1013/.