KAMPUS AKADEMIK PUBLISING
Jurnal Ilmiah Penelitian Mahasiswa (JIPM)
Vol.4, No.6 Desember 2026
e-ISSN: 3025-5465, p-ISSN : 3025-7964, Hal 117-128
DOI: https://doi.org/10.61722/jipm.v4i6.3132

IMPLEMENTASI INDOBERT DENGAN AUGMENTASI DATA SINTETIS BERBASIS LARGE LANGUAGE MODEL UNTUK PENILAIAN ESAI OTOMATIS BAHASA INDONESIA

Nick Engelbert Mongkol[1], Ryan Putranda Kristianto[2]*
[1,2]Program Studi Ilmu Informatika, Fakultas Teknik, Universitas Katolik Darma Cendika, Surabaya
*Penulis Korespondensi: [1]nick.mongkol@student.ukdc.ac.id, [2]*ryan@ukdc.ac.id

Abstract. *Manual essay assessment is time-consuming and prone to inconsistency between raters, while the development of Automated Essay Scoring (AES) systems for Indonesian remains constrained by the limited availability of labeled datasets. This study implements an IndoBERT model with a regression approach to score student essay responses and applies Large Language Model (LLM)-based synthetic data augmentation to address class imbalance in the low-score categories. The primary dataset consisted of 556 essay responses from seventh-grade students across four Civic Education questions scored by a teacher, leaving 551 records after data cleaning. Augmentation using GPT-4o-mini produced 160 synthetic responses in the score 0 and score 1 categories, increasing the dataset to 711 records. The model was evaluated using 5-fold cross-validation with Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Spearman correlation across four testing conditions distinguishing seen and unseen questions. The results show that the augmented model achieved a QWK of 0.8942 and a Spearman correlation of 0.9063, up from 0.7884 and 0.8199 in the non-augmented model, although MAE slightly worsened from 0.3931 to 0.4113. However, when the same model was tested on 1,000 responses from 100 previously unseen questions, QWK dropped to 0.4832. These findings indicate that the availability of a question item in the training data is a more dominant determinant of performance than the source or volume of training data.*
Keywords: *Automated Essay Scoring; IndoBERT; Synthetic Data Augmentation; Quadratic Weighted Kappa; Large Language Model*

Abstrak. Penilaian esai secara manual membutuhkan waktu yang lama dan rentan terhadap inkonsistensi antarpenilai, sementara pengembangan sistem *Automated Essay Scoring* (AES) untuk Bahasa Indonesia masih terkendala keterbatasan dataset berlabel. Penelitian ini mengimplementasikan model IndoBERT dengan pendekatan regresi untuk menilai jawaban esai siswa dan menerapkan augmentasi data sintetis berbasis *Large Language Model* (LLM) guna mengatasi ketidakseimbangan kelas pada kategori skor rendah. Dataset primer terdiri atas 556 jawaban esai siswa kelas VII pada empat butir soal mata pelajaran Pendidikan Kewarganegaraan yang dinilai oleh guru, dan menyisakan 551 data setelah pembersihan. Augmentasi menggunakan GPT-4o-mini menghasilkan 160 jawaban sintetis pada kategori skor 0 dan 1 sehingga dataset meningkat menjadi 711 data. Model dievaluasi menggunakan *5-fold cross-validation* dengan metrik *Quadratic Weighted Kappa* (QWK), *Mean Absolute Error* (MAE), dan korelasi *Spearman* pada empat kondisi pengujian yang membedakan butir soal terlatih dan tidak terlatih. Hasil menunjukkan model hasil augmentasi memperoleh QWK 0,8942 dan *Spearman* 0,9063, meningkat dari 0,7884 dan 0,8199 pada model tanpa augmentasi, meskipun MAE sedikit memburuk dari 0,3931 menjadi 0,4113. Namun, ketika model yang sama diuji pada 1.000 jawaban dari 100 butir soal yang belum pernah dilatihkan, QWK turun menjadi 0,4832. Temuan ini menunjukkan bahwa ketersediaan butir soal pada data latih merupakan penentu performa yang lebih dominan dibandingkan sumber maupun jumlah data latih..
Kata kunci: Automated Essay Scoring; IndoBERT; Augmentasi Data Sintetis; Quadratic Weighted Kappa; Large Language Model

1. LATAR BELAKANG

Penilaian esai merupakan salah satu metode evaluasi yang digunakan untuk mengukur hasil pembelajaran yang bersifat kompleks (Putri et al., 2022; Kurniawati & Mega Pradnya, 2020). Namun, proses penilaian esai secara manual menghadapi sejumlah tantangan, di antaranya kebutuhan tenaga dan waktu yang besar untuk mengoreksi setiap

Naskah Masuk: 23 Agustus 2026; Revisi: 24 Agustus 2026; Diterima: 25 Agustus 2026; ; Terbit: 27 Agustus 2026

jawaban secara individual (Judijanto et al., 2025), inkonsistensi nilai antarpenilai yang berbeda (Putri et al., 2022), serta keterlambatan pemberian umpan balik yang berdampak langsung pada efektivitas pembelajaran siswa (Rahanra et al., 2025). Di Indonesia, permasalahan tersebut diperberat oleh rasio guru terhadap siswa yang dapat mencapai satu banding empat puluh atau lebih, sehingga guru sulit memberikan penilaian yang mendalam dan tepat waktu (Rahanra et al., 2025). Perbedaan standar antarpenilai juga menyebabkan kebingungan pada siswa dan menjadikan penilaian esai sebagai proses yang rumit (Zubaidi et al., 2025).

Automated Essay Scoring (AES) merupakan salah satu solusi yang memanfaatkan *Natural Language Processing* (NLP) untuk menilai esai secara otomatis. Sistem AES telah berkembang pesat dari pendekatan berbasis aturan dan fitur statistik menuju model *deep learning* berbasis *transformer* yang mampu memahami konteks semantik teks secara lebih mendalam (Misgna et al., 2025). Meskipun demikian, sebagian besar penelitian AES masih berfokus pada bahasa Inggris, sedangkan penelitian untuk Bahasa Indonesia masih sangat terbatas. Proporsi penelitian berbahasa Indonesia dalam studi *Large Language Model-Powered Automated Assessment* tercatat sangat kecil, dengan performa sistem pada konteks bahasa selain Inggris yang cenderung kurang stabil (Miller et al., 2025).

Penelitian AES untuk Bahasa Indonesia yang telah dilakukan umumnya terkendala oleh ukuran dataset. Studi komparatif terhadap sejumlah model *transformer* menggunakan 300 esai dari kuis mata kuliah Metodologi Penelitian menunjukkan bahwa varian indobert-base-p2 memberikan performa terbaik dengan QWK 0,6745, nilai yang masih berada di bawah ambang penerimaan umum pada penelitian AES (Prastyo et al., 2025). Penurunan performa IndoBERT yang tajam juga dilaporkan pada jawaban berkonteks panjang dengan dataset yang hanya berjumlah 100 esai (Amalia et al., 2025). Nilai QWK sebesar 0,931 memang pernah dicapai melalui penggabungan representasi IndoBERT dan IndoSBERT, namun dengan menggunakan dataset kompetisi hasil terjemahan alih-alih jawaban siswa Indonesia yang sebenarnya (Aliyah et al., 2025). Pendekatan kemiripan semantik berbasis IndoBERT juga telah diterapkan pada jawaban ujian perguruan tinggi, tetapi tanpa menggunakan QWK sebagai metrik sehingga hasilnya sulit dibandingkan lintas penelitian (Pradani & Suadaa, 2023).

Keterbatasan dataset tersebut tidak hanya menyangkut jumlah, tetapi juga distribusi. Pada data penilaian nyata, jawaban siswa cenderung terkonsentrasi pada skor menengah hingga tinggi, sedangkan jawaban berkualitas sangat rendah jarang muncul. Kondisi ini menyebabkan model kekurangan contoh untuk mempelajari pola kegagalan jawaban, sehingga berisiko memberikan skor yang terlalu tinggi pada jawaban yang sebenarnya tidak relevan. Augmentasi data sintetis menggunakan *Large Language Model* (LLM) merupakan salah satu pendekatan yang dapat digunakan untuk melengkapi kategori yang jarang muncul tersebut.

Berdasarkan uraian tersebut, penelitian ini mengimplementasikan model IndoBERT dengan pendekatan regresi untuk penilaian esai berbahasa Indonesia dan menerapkan augmentasi data sintetis berbasis GPT-4o-mini pada kategori skor rendah. Penelitian ini menjawab dua rumusan masalah, yaitu: (1) sejauh mana augmentasi data sintetis berbasis LLM memengaruhi kinerja model IndoBERT dalam penilaian esai berbahasa Indonesia; dan (2) sejauh mana kinerja tersebut bertahan ketika model diuji pada butir soal yang belum pernah dilatihkan. Sejalan dengan itu, penelitian ini bertujuan mengukur pengaruh augmentasi data sintetis terhadap kinerja model IndoBERT dan

mengevaluasi kemampuan generalisasi model pada butir soal terlatih maupun tidak terlatih.

2. KAJIAN TEORITIS

2.1 Automated Essay Scoring

Automated Essay Scoring (AES) adalah penerapan teknik komputasi dan NLP untuk mengevaluasi serta memberikan skor pada esai tertulis. AES pertama kali diperkenalkan melalui *Project Essay Grader* (PEG) pada tahun 1966 yang mengandalkan analisis gaya penulisan (Page, 1994). Secara umum, sistem AES menerima teks esai sebagai masukan, memproses fitur linguistik dan semantik di dalamnya, lalu menghasilkan skor numerik yang merepresentasikan kualitas esai tersebut. Model AES berbasis *deep learning* mampu mengenali pola kompleks dalam esai sehingga dapat menghasilkan prediksi skor yang akurat (Misgna et al., 2025). Penerapan AES pada bahasa selain Inggris juga telah dieksplorasi, seperti pada esai berbahasa Jepang dari penutur non-*native* (Li & Liu, 2024), serta melalui pendekatan *atomic evaluation* berbasis LLM untuk penilaian esai ujian berbahasa Indonesia (Winarta et al., 2024).

2.2 BERT dan IndoBERT

Transformer merupakan arsitektur *deep learning* yang menggunakan mekanisme *self-attention* sehingga hubungan antarkata dalam suatu kalimat dapat dipelajari secara langsung tanpa pemrosesan token secara berurutan (Vaswani et al., 2017). BERT (*Bidirectional Encoder Representations from Transformers*) memanfaatkan bagian *encoder* dari arsitektur tersebut dan melakukan *pre-training* secara dua arah, sehingga mampu memahami konteks kata dari sisi kiri dan kanan secara bersamaan serta dapat di-*fine-tune* untuk berbagai tugas NLP tanpa perubahan arsitektur yang besar (Devlin et al., 2019).

IndoBERT merupakan adaptasi BERT yang dilatih menggunakan korpus berbahasa Indonesia berskala besar, sehingga mampu merepresentasikan konteks bahasa Indonesia secara lebih akurat dibandingkan model *multilingual* umum (Koto et al., 2020). Model ini terbukti efektif pada berbagai tugas NLP Indonesia, termasuk analisis sentimen, dengan menghasilkan representasi teks yang lebih kaya secara semantik dibandingkan metode *machine learning* tradisional (Geni et al., 2023). Untuk tugas penilaian esai, IndoBERT dilengkapi dengan *regression head*, yaitu lapisan linier tunggal di atas representasi token khusus yang menghasilkan satu nilai kontinu. Berbeda dengan *classification head* yang menghasilkan distribusi probabilitas atas sejumlah kelas, *regression head* memperlakukan skor sebagai besaran kontinu sehingga jarak antartingkat skor tetap bermakna.

2.3 Augmentasi Data Sintetis Berbasis Large Language Model

Salah satu kendala utama pengembangan sistem penilaian otomatis Bahasa Indonesia adalah keterbatasan ketersediaan data berlabel. Pengumpulan jawaban siswa beserta anotasi skor dari guru memerlukan waktu dan tenaga yang besar, sehingga dataset yang tersedia umumnya berskala kecil dan terbatas pada satu domain. Augmentasi data sintetis menggunakan LLM merupakan salah satu pendekatan untuk mengatasi keterbatasan tersebut, yaitu dengan membangkitkan jawaban siswa buatan yang menyerupai karakteristik jawaban nyata. Perbandingan antara data berlabel manusia dan data hasil augmentasi LLM pada sepuluh tugas klasifikasi menunjukkan bahwa meskipun data berlabel manusia secara umum tetap unggul, augmentasi sintetis terbukti bermanfaat khususnya untuk meningkatkan performa pada kelas yang jarang muncul dalam tugas multikelas (Møller et al., 2024).

Kualitas data sintetis sangat bergantung pada cara pembangkitan dikondisikan. Pembangkitan tanpa acuan mengharuskan model menentukan sendiri kriteria kebenaran sekaligus gradasi antartingkat kualitas, sehingga standar tingkat skor berpotensi berbeda-beda antarbutir soal dan menghasilkan *noise* label. Pengondisian pada acuan penilaian, baik berupa kunci jawaban maupun rubrik poin, membuat gradasi antartingkat skor terkalibrasi secara lebih konsisten. Manfaat lain dari augmentasi juga telah dilaporkan, yaitu bahwa penambahan respons anomali seperti esai di luar topik dan jawaban tidak bermakna ke dalam data latih secara signifikan meningkatkan kemampuan model mendeteksi masukan anomali tanpa menurunkan performa penilaian secara keseluruhan (Doewes, 2026).

Di antara berbagai LLM yang tersedia, penelitian ini menggunakan GPT-4o-mini, yaitu varian ekonomis dari keluarga GPT-4o yang dirancang untuk tugas dengan kebutuhan biaya dan latensi rendah. Efektivitas model tersebut untuk augmentasi teks telah dibuktikan pada tugas deteksi stres, di mana data sintetis yang dibangkitkan melalui API GPT-4o-mini meningkatkan akurasi model BERT dari 73,94% pada *baseline* menjadi 84,74% pada dataset gabungan hasil augmentasi (Badran et al., 2025). Kemampuan LLM dalam konteks penilaian pendidikan berbahasa Indonesia juga telah diuji pada 646 lembar jawaban ujian siswa sekolah dasar dari enam sekolah di Indonesia, dengan temuan bahwa umpan balik yang dihasilkan tetap jelas dan benar secara faktual meskipun masukan yang diterima tidak sempurna (Aisyah et al., 2025). Kombinasi antara kemampuan mengikuti instruksi, dukungan terhadap Bahasa Indonesia, dan biaya yang rendah menjadikan model ini sesuai untuk pembangkitan data sintetis secara berulang pada penelitian berskala terbatas.

2.4 Metrik Evaluasi

Quadratic Weighted Kappa (QWK) merupakan metrik standar dalam penelitian AES untuk mengukur tingkat kesepakatan antara dua penilai, dengan memberikan penalti yang lebih besar terhadap perbedaan skor yang lebih jauh (Doewes et al., 2023). Formula QWK dinyatakan sebagai berikut.

$$\kappa = 1 - (\Sigma_{ij}\, w_{ij}\, O_{ij}) \,/\, (\Sigma_{ij}\, w_{ij}\, E_{ij})$$

dengan w_{ij} sebagai bobot kuadratik $(i - j)^2 / (N - 1)^2$, O_{ij} sebagai frekuensi observasi pada sel (i, j) matriks konfusi, E_{ij} sebagai frekuensi harapan berdasarkan distribusi marginal, dan N sebagai jumlah kategori skor. Nilai QWK diinterpretasikan mengikuti konvensi umum, yaitu 0,41–0,60 sebagai kesepakatan moderat, 0,61–0,80 sebagai kesepakatan substansial, dan 0,81–1,00 sebagai kesepakatan hampir sempurna.

Selain QWK, penelitian ini menggunakan *Mean Absolute Error* (MAE) untuk mengukur rata-rata selisih absolut antara skor prediksi model dan skor guru, serta korelasi peringkat *Spearman* untuk mengukur kesesuaian urutan kualitas antara prediksi model dan penilaian guru. Ketiga metrik tersebut bersifat saling melengkapi: QWK mengukur kesepakatan dengan penilai manusia, MAE mengukur besarnya kesalahan numerik, dan *Spearman* mengukur konsistensi urutan.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan data primer yang dianotasi oleh *expert*. Tahapan penelitian meliputi pengumpulan dataset, praproses data, augmentasi data sintetis, pengembangan model, serta pengujian model pada empat kondisi yang berbeda.

3.1 Dataset dan Praproses

Dataset primer diperoleh langsung dari guru mata pelajaran, berupa jawaban esai ujian mata pelajaran Pendidikan Kewarganegaraan untuk siswa kelas VII Sekolah Menengah Pertama. Dataset terdiri atas empat butir soal dengan kode Q01 hingga Q04, masing-masing berisi 139 jawaban siswa, sehingga total berjumlah 556 baris data. Butir Q01 dan Q03 memiliki rentang skor 0–3, sedangkan Q02 dan Q04 memiliki rentang skor 0–4. Contoh data primer disajikan pada Tabel 1.

Tabel 1. Contoh Data Primer Hasil Anotasi Guru

Kode Soal	Soal	Jawaban Siswa	Nilai Guru
Q01	Jelaskan apa yang dimaksud dengan "Negara" menurut pemahamanmu!	Negara adalah suatu wilayah atau tempat yang memiliki banyak masyarakat dan diakui oleh negara lain dan yang memiliki pemerintah.	3
Q03	Jelaskan apa yang dimaksud dengan keberagaman!	Keberagaman adalah perbedaan menurut agama, ras, asal-usul, suku, budaya, dll.	3
Q04	Sebutkan 2 dampak positif dan 2 dampak negatif dari keberagaman!	Dampak positif: masyarakat menjadi bisa menghormati agama, ras, dan asal-usul orang. Dampak negatif: terjadinya rasisme.	4

Praproses dilakukan melalui tiga tahap. Pertama, pembersihan data terhadap baris yang tidak terisi, di mana ditemukan lima baris bernilai "-" yang kemudian dihapus sehingga menyisakan 551 baris data valid. Kedua, konversi skor dari format *string* dengan koma desimal menjadi numerik *float*. Ketiga, normalisasi skor ke rentang 0–1 dengan membagi skor asli terhadap skor maksimum butir soal yang bersangkutan, agar butir soal dengan skala maksimum berbeda dapat dilatih dalam satu model.

3.2 Augmentasi Data Sintetis

Analisis awal terhadap dataset primer menunjukkan ketidakseimbangan distribusi yang signifikan, terutama pada kategori skor rendah. Pengujian awal mengonfirmasi dampaknya: pada butir Q04, jawaban yang sepenuhnya tidak relevan tetap diberi skor 2,0 oleh model, padahal seharusnya bernilai 0,0. Berdasarkan temuan tersebut, augmentasi diarahkan secara khusus pada rentang skor rendah.

Pembangkitan data sintetis dilakukan menggunakan GPT-4o-mini melalui API OpenAI. *Prompt* dikonstruksi dengan menyertakan tiga komponen sebagai konteks, yaitu identitas butir soal, teks pertanyaan secara utuh, dan kunci jawaban yang digunakan guru sebagai acuan penilaian. Penyertaan kunci jawaban merupakan keputusan rancangan yang penting: tanpa acuan tersebut, model harus menentukan sendiri kriteria kebenaran sekaligus gradasi antartingkat kualitas, sehingga standar tingkat skor berpotensi berbeda-beda antarbutir soal dan menghasilkan *noise* label.

Model tidak diminta membangkitkan jawaban secara bebas, melainkan diberi target skor beserta tipe kegagalan yang harus direpresentasikan. Untuk setiap butir soal diminta dua kategori keluaran, masing-masing berjumlah 20 jawaban, dengan spesifikasi sebagaimana disajikan pada Tabel 2.

Tabel 2. Spesifikasi Kategori dan Tipe Jawaban pada Prompt Augmentasi

Skor Target	Tipe Jawaban	Karakteristik yang Diminta	Jumlah per Butir Soal
0,0	*Gibberish* / asal ketik dan *out-of-domain*	Rangkaian karakter atau kata tanpa makna yang tidak membentuk kalimat dapat dinilai, serta kalimat yang gramatikal namun membahas topik yang sama sekali tidak berkaitan dengan pertanyaan	20
1,0	Menghindar / lupa / pasrah dan salah konsep / meleset jauh	Pengakuan tidak tahu atau lupa tanpa upaya menjawab substansi pertanyaan, serta jawaban yang berkaitan dengan topik namun keliru secara konseptual atau menyimpang jauh dari kunci jawaban	20

Pembedaan antara skor 0,0 dan skor 1,0 dirancang agar model tidak hanya belajar mengenali jawaban yang tidak valid, tetapi juga membedakan jawaban yang sepenuhnya tidak relevan dari jawaban yang masih memiliki keterkaitan dengan topik meskipun keliru. Pembedaan ini merupakan inti permasalahan yang ditemukan pada pengujian awal, ketika jawaban yang sepenuhnya tidak relevan pada butir Q04 tetap memperoleh skor 2,0. Dengan empat butir soal dan 40 jawaban per butir soal, proses ini menghasilkan 160 jawaban sintetis sehingga dataset meningkat dari 551 menjadi 711 baris.

Selain augmentasi untuk mengatasi ketidakseimbangan kelas, dilakukan pula pembangkitan dataset kedua yang digunakan khusus untuk menguji generalisasi model. Dataset ini terdiri atas 100 butir soal baru, yaitu 50 butir soal Ilmu Pengetahuan Alam dan 50 butir soal Pendidikan Pancasila, yang diambil dari buku pelajaran SMP/MTs Kelas VII. Struktur *prompt* yang digunakan sama dengan augmentasi sebelumnya, yaitu menyertakan identitas butir soal, teks pertanyaan, dan kunci jawaban sebagai konteks, serta menetapkan skor target beserta karakteristik jawaban yang harus direpresentasikan. Perbedaannya terletak pada rentang skor target: apabila augmentasi sebelumnya hanya menyasar skor 0,0 dan 1,0 untuk melengkapi kategori yang langka, pembangkitan dataset kedua menyasar seluruh rentang skor agar dapat berfungsi sebagai data latih sekaligus data uji yang utuh. Untuk setiap butir soal dibangkitkan 10 jawaban dengan skor yang terdistribusi merata, sehingga diperoleh 1.000 baris data.

3.3 Arsitektur dan Konfigurasi Pelatihan Model

Model yang digunakan adalah IndoBERT *base* (indobenchmark/indobert-base-p2) dengan tambahan *regression head*. Pemilihan varian tersebut didasarkan pada temuan bahwa indobert-base-p2 memberikan performa tertinggi dibandingkan varian base-p1 dan base-p3 pada tugas penilaian esai berbahasa Indonesia (Prastyo et al., 2025). Arsitektur model terdiri atas *encoder* dengan 12 lapisan *transformer*, 12 *attention head*, dan dimensi vektor tersembunyi sebesar 768; lapisan *pooling* yang mengambil representasi token [CLS]; serta *regression head* berupa lapisan *fully-connected* yang memetakan vektor 768 dimensi menjadi satu nilai skalar tanpa fungsi aktivasi.

Fungsi kerugian yang digunakan adalah *Mean Squared Error* (MSE), yang dipilih karena memberikan penalti kuadratik terhadap kesalahan besar. Konfigurasi pelatihan meliputi panjang maksimum urutan 512 token, *batch size* 8, *learning rate* 2×10^{-5} dengan

pengoptimal AdamW, *weight decay* 0,01, dan jumlah *epoch* maksimum 10 dengan *early stopping patience* sebesar 3 berdasarkan nilai QWK. Seluruh proses pelatihan dijalankan pada GPU dengan presisi FP16. Pada tahap inferensi, keluaran model dikembalikan ke skala asli melalui perkalian dengan skor maksimum butir soal, kemudian dibulatkan ke kelipatan 0,5 terdekat.

3.4 Rancangan Pengujian dan Metrik Evaluasi

Evaluasi dilakukan terhadap tiga model yang dilatih secara terpisah dan diuji pada empat kondisi. Kondisi B1 dan B2 merupakan dua pengujian terhadap model yang sama, yaitu model yang dilatih menggunakan gabungan data primer dan data augmentasi; perbedaannya terletak sepenuhnya pada data uji yang digunakan, bukan pada proses pelatihan. Rancangan keempat kondisi tersebut disajikan pada Tabel 3.

Tabel 3. Rancangan Model dan Kondisi Pengujian

Label	Model	Data Latih	Data Uji	Butir soal pernah dilatihkan?
A	Model primer	551 baris, 4 butir (primer)	5-fold CV pada data yang sama	Pernah
B1	Model augmentasi	711 baris, 4 butir (primer + augmentasi)	5-fold CV pada data yang sama	Pernah
B2	Model augmentasi (sama dengan B1)	711 baris, 4 butir (primer + augmentasi)	1.000 baris dari 100 butir soal baru	Belum pernah
C	Model sintetis	800 baris dari 100 butir soal baru	200 baris dari 100 butir soal baru	Pernah

Berdasarkan ketersediaan butir soal pada data latih, keempat kondisi terbagi menjadi dua kategori. Kondisi A, B1, dan C termasuk kategori *seen questions*, yaitu model menilai jawaban baru untuk butir soal yang telah dikenali selama pelatihan. Kondisi B2 termasuk kategori *unseen questions* atau *cross-prompt*, yaitu seluruh butir soal pada data uji belum pernah muncul dalam pelatihan. Pengelompokan ini penting karena kategori pertama diwakili oleh tiga model berbeda dengan sumber data latih yang berbeda pula, yaitu data primer tanpa augmentasi (A), data primer dengan augmentasi (B1), dan data sepenuhnya sintetis (C). Metrik yang digunakan adalah QWK sebagai metrik utama, serta MAE dan korelasi *Spearman* sebagai metrik pendukung.

4. HASIL DAN PEMBAHASAN

4.1 Karakteristik Dataset dan Hasil Augmentasi

Distribusi skor pada dataset primer menunjukkan ketidakseimbangan yang signifikan, di mana mayoritas jawaban terkonsentrasi pada skor tinggi maupun menengah. Pada Q01, skor 1,5 mendominasi dengan 52 sampel dan skor maksimal 3,0 sebanyak 51 sampel, sedangkan skor terendah yang tersedia hanya 1,0 dengan 4 sampel dan tidak ditemukan satu pun sampel berskor 0,0. Ketidakseimbangan paling ekstrem terlihat pada Q03 dan Q04: setelah pembersihan data, pada Q03 skor maksimal 3,0 mencakup 93 dari 138 sampel atau sekitar 67%, sedangkan pada Q04 skor maksimal 4,0

mencakup 120 dari 139 sampel atau sekitar 86%, dengan kategori skor rendah masing-masing hanya diwakili satu sampel. Perbandingan distribusi sebelum dan sesudah augmentasi disajikan pada Tabel 4.

Tabel 4. Perbandingan Distribusi Kategori Skor Sebelum dan Sesudah Augmentasi

Kategori Skor	Sebelum Augmentasi	Sesudah Augmentasi
Skor 0,0	3 (0,54%)	83 (11,67%)
Skor 1,0	15 (2,72%)	95 (13,36%)
Skor 1,5 – 4,0	533 (96,73%)	533 (74,96%)
Total	551	711

Contoh jawaban sintetis yang dihasilkan pada dataset pengujian generalisasi disajikan pada Tabel 5. Jawaban yang dibangkitkan menunjukkan variasi tipe kegagalan yang sesuai dengan tingkat skor yang ditargetkan, mulai dari jawaban yang sepenuhnya tidak berkaitan dengan pertanyaan hingga jawaban yang mendekati benar namun tidak lengkap.

Tabel 5. Contoh Jawaban Sintetis Hasil Augmentasi

Kode Soal	Mata Pelajaran	Soal	Jawaban Siswa (Sintetis)	Nilai
S001	IPA	Menurutmu, apakah yang dipelajari dalam ilmu Astronomi?	Astronomi itu tentang makanan enak.	0
S004	IPA	Apakah yang dimaksud dengan hipotesis pada proses melakukan percobaan?	Kalau tidak salah, hipotesis itu mungkin tentang hasil eksperimen, tapi saya lupa yang lainnya.	2
S054	PKN	Bagaimana proses penetapan Pancasila sebagai dasar negara dalam sidang PPKI?	Jadi, Pancasila ditetapkan sebagai dasar negara di sidang PPKI, tetapi saya bingung tentang tanggalnya, mungkin sekitar 18 Agustus 1945.	3

4.2 Hasil Evaluasi Model

Hasil evaluasi pada keempat kondisi pengujian disajikan pada Tabel 6. Nilai pada kondisi A dan B1 dilaporkan sebagai rerata dan simpangan baku antar-*fold* pada skema *5-fold cross-validation*, sedangkan kondisi B2 dan C dilaporkan sebagai nilai tunggal pada data uji terpisah.

Tabel 6. Hasil Evaluasi pada Empat Kondisi Pengujian

Label	Kategori	N uji	QWK	MAE	Spearman
A	Seen questions	551 (CV)	0,7884 ± 0,0267	0,3931 ± 0,0291	0,8199 ± 0,0249
B1	Seen questions	711 (CV)	0,8942 ± 0,0236	0,4113 ± 0,0656	0,9063 ± 0,0221

Label	Kategori	N uji	QWK	MAE	Spearman
B2	Unseen questions	1.000	0,4832	1,2125	0,7687
C	Seen questions	200	0,9342	0,3200	0,9143

Pada perbandingan antara model tanpa augmentasi (A) dan model hasil augmentasi (B1), terdapat peningkatan yang jelas pada metrik yang bersifat ordinal. Nilai QWK meningkat dari 0,7884 menjadi 0,8942, sehingga berpindah dari kategori kesepakatan substansial menjadi kesepakatan hampir sempurna. Korelasi *Spearman* juga meningkat dari 0,8199 menjadi 0,9063. Kedua metrik tersebut menunjukkan simpangan baku antar-*fold* yang lebih kecil pada model B1, yang mengindikasikan konsistensi yang lebih baik antar-*fold*.

Sebaliknya, pada metrik MAE, model A justru menunjukkan hasil yang sedikit lebih baik, yaitu 0,3931 dibandingkan 0,4113 pada model B1. Selisihnya relatif kecil, yaitu 0,0182, namun variabilitas MAE antar-*fold* pada model A jauh lebih kecil dibandingkan model B1 (simpangan baku 0,0291 berbanding 0,0656, atau sekitar 56% lebih rendah). Dengan demikian, keunggulan model hasil augmentasi tidak bersifat menyeluruh, melainkan terbatas pada aspek akurasi ordinal.

Temuan yang paling menonjol muncul pada perbandingan antarkategori kondisi. Ketiga kondisi berkategori *seen questions* menghasilkan nilai QWK yang mengelompok pada rentang 0,7884 hingga 0,9342, sedangkan satu-satunya kondisi berkategori *unseen questions* menghasilkan 0,4832. Selisih antara nilai terendah pada kategori pertama, yaitu kondisi A sebesar 0,7884, dengan kondisi B2 mencapai 0,3052, sementara selisih terhadap nilai tertinggi pada kondisi C mencapai 0,4510. Penurunan tersebut juga terlihat pada MAE, yang meningkat hampir tiga kali lipat menjadi 1,2125 pada kondisi B2.

PEMBAHASAN

Peningkatan QWK dari 0,7884 menjadi 0,8942 setelah augmentasi perlu dibaca bersama dengan memburuknya MAE pada model yang sama. Augmentasi yang dilakukan pada penelitian ini menambahkan 160 jawaban sintetis pada kategori skor 0 dan 1, yang sebagian besar berupa jawaban *gibberish*, jawaban di luar domain, dan kalimat menghindar. Jawaban dengan karakteristik demikian relatif mudah dibedakan oleh model, sehingga penambahannya melebarkan variansi skor pada data dan berkontribusi terhadap naiknya nilai QWK. Sebaliknya, MAE yang sedikit memburuk mengindikasikan bahwa ketepatan numerik pada rentang skor menengah tidak ikut membaik. Dengan demikian, kontribusi augmentasi pada penelitian ini lebih tepat dimaknai sebagai peningkatan kemampuan model mengenali jawaban yang tidak valid, sesuai dengan motivasi awal yang muncul dari temuan bahwa model sebelumnya memberi skor 2,0 pada jawaban yang sepenuhnya tidak relevan, dan bukan sebagai peningkatan akurasi penilaian secara menyeluruh. Interpretasi ini sejalan dengan temuan bahwa manfaat augmentasi sintetis terkonsentrasi pada kelas yang jarang muncul (Møller et al., 2024), serta bahwa penambahan respons anomali ke dalam data latih meningkatkan kemampuan model mendeteksi masukan anomali (Doewes, 2026).

Temuan kedua, yaitu kejatuhan performa pada kondisi *unseen questions*, memiliki implikasi yang lebih luas. Nilai penting dari pengelompokan pada Tabel 6 terletak pada keberagaman sumber data ketiga kondisi *seen questions*. Kondisi A menggunakan

jawaban siswa nyata, kondisi B1 menggunakan jawaban siswa nyata beserta tambahan sintetis, dan kondisi C menggunakan jawaban yang seluruhnya berasal dari pembangkitan LLM. Meskipun sumbernya berbeda secara fundamental, ketiganya tetap mengelompok pada rentang performa yang serupa. Sebaliknya, ketika model yang identik dengan B1 diuji pada butir soal yang belum pernah dilatihkan, nilai QWK turun hingga hampir setengahnya. Konvergensi tiga kondisi dari sumber yang berlainan tersebut menunjukkan bahwa faktor penentu utama performa pada penelitian ini bukan sumber maupun kualitas data latih, melainkan apakah butir soal pada data uji telah dilatihkan.

Temuan ini menempatkan hasil-hasil penelitian AES Bahasa Indonesia sebelumnya dalam perspektif yang lebih berhati-hati. Nilai QWK tinggi yang dilaporkan pada berbagai penelitian, termasuk 0,931 pada penelitian terdahulu (Aliyah et al., 2025) dan 0,8942 pada penelitian ini, umumnya diperoleh melalui skema *cross-validation* pada butir soal yang sama, sehingga mencerminkan kemampuan model menilai jawaban baru untuk soal yang telah dikalibrasi, bukan kemampuan menilai soal yang sama sekali baru. Perbedaan antara kedua kemampuan tersebut bersifat substansial, yaitu 0,8942 berbanding 0,4832 pada model yang identik. Konsekuensi praktisnya, sistem AES berbasis pendekatan ini lebih tepat diposisikan sebagai alat bantu penilaian untuk butir soal yang telah dilatihkan, misalnya pada ujian yang diselenggarakan berulang dengan bank soal tetap, dan belum dapat diterapkan secara langsung pada butir soal baru tanpa pelatihan ulang.

Perlu dicatat pula bahwa kondisi C memperoleh nilai tertinggi di antara seluruh kondisi, yaitu QWK 0,9342, meskipun seluruh data latih dan data ujinya berasal dari pembangkitan LLM. Hal ini tidak dapat ditafsirkan sebagai bukti bahwa data sintetis lebih baik daripada data nyata. Sebaliknya, hasil tersebut kemungkinan besar mencerminkan bahwa data sintetis memiliki *noise* yang lebih rendah dan gradasi antartingkat skor yang lebih teratur dibandingkan jawaban siswa yang sebenarnya, sehingga tugas penilaian menjadi lebih mudah. Karena itu, hasil pada data sintetis dilaporkan secara terpisah dan tidak digabungkan dengan hasil pada data nyata.

5. KESIMPULAN DAN SARAN

KESIMPULAN

Berdasarkan hasil implementasi dan evaluasi, dapat disimpulkan dua hal. Pertama, augmentasi data sintetis berbasis GPT-4o-mini pada kategori skor rendah meningkatkan kinerja ordinal model IndoBERT, yaitu QWK dari 0,7884 menjadi 0,8942 dan korelasi *Spearman* dari 0,8199 menjadi 0,9063, dengan simpangan baku antar-*fold* yang lebih kecil pada kedua metrik. Namun peningkatan tersebut tidak bersifat menyeluruh, karena MAE justru sedikit memburuk dari 0,3931 menjadi 0,4113. Manfaat augmentasi pada penelitian ini karenanya lebih tepat dimaknai sebagai peningkatan kemampuan model mengenali jawaban yang tidak valid, bukan peningkatan ketepatan numerik pada seluruh rentang skor.

Kedua, kinerja model tidak bertahan ketika diuji pada butir soal yang belum pernah dilatihkan. Model yang sama dengan kondisi B1 hanya memperoleh QWK 0,4832 pada 1.000 jawaban dari 100 butir soal baru, turun dari 0,8942 pada butir soal terlatih. Ketiga kondisi *seen questions* tetap mengelompok pada rentang 0,7884–0,9342 meskipun berasal dari tiga model dengan sumber data latih yang berbeda secara fundamental. Dengan demikian, penentu utama kinerja sistem pada penelitian ini adalah ketersediaan butir soal pada data latih, bukan sumber maupun jumlah data latih.

KETERBATASAN DAN SARAN

Penelitian ini memiliki beberapa keterbatasan. Pertama, label pada data sintetis ditentukan melalui *prompt* yang diberikan kepada LLM dan belum divalidasi ulang oleh penilai manusia, sehingga terdapat kemungkinan *noise* label pada 160 baris data augmentasi yang digunakan dalam pelatihan model B1. Kedua, evaluasi pada kondisi B1 dilakukan pada data gabungan antara jawaban nyata dan jawaban sintetis, sehingga nilai QWK yang diperoleh tidak sepenuhnya sebanding dengan nilai pada kondisi A yang hanya menggunakan data nyata. Ketiga, dataset primer hanya berasal dari satu mata pelajaran, satu jenjang kelas, dan satu orang guru penilai, sehingga hasilnya belum dapat digeneralisasi ke konteks penilaian yang lebih luas.

Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan untuk melakukan validasi manusia terhadap sampel data sintetis dan melaporkan tingkat kesesuaiannya dengan label yang dibangkitkan LLM, serta melaporkan hasil evaluasi model tanpa augmentasi dan model dengan augmentasi pada subset uji data nyata yang identik agar pengaruh augmentasi dapat diisolasi dari pengaruh pelebaran distribusi skor. Selain itu, disarankan pula untuk memperluas dataset ke berbagai mata pelajaran dan jenjang pendidikan, melibatkan lebih dari satu penilai agar tingkat kesepakatan antarpenilai manusia dapat dijadikan pembanding, serta mengeksplorasi pendekatan *cross-prompt* seperti penyertaan teks soal sebagai bagian dari masukan model guna meningkatkan kemampuan generalisasi pada butir soal baru.

DAFTAR REFERENSI

Aisyah, N., Al Kautsar, M. D., Hidayat, A., Chowdhury, R., & Koto, F. (2025). *Evaluating Vision–Language and Large Language Models for Automated Student Assessment in Indonesian Classrooms*. arXiv. https://doi.org/10.48550/arXiv.2506.04822

Aliyah, N. E., Sholikah, R. W., Firdausi, H., Ciptaningtyas, H. T., & Sabilla, I. A. (2025). Enhancing Automated Essay Scoring in Bahasa Indonesia with IndoBERT and IndoSBERT. *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, 1–7. https://doi.org/10.1109/SIML65326.2025.11080721

Amalia, A., Lydia, M. S., Muchtar, M. A., Manik, F. Y., & Gunawan, D. (2025). Mitigating Bias and Assessment Inconsistencies with BERT-Based Automated Short Answer Grading for the Indonesian Language. *IAENG International Journal of Computer Science*, *52*(3).

Badran, N., Le, J., Le, T., & Uchiya, T. (2025). Improving Stress Detection with Synthetic Datasets: GPT-4o-Mini and Transformer Model Evaluation. Dalam N. T. Nguyen et al. (Ed.), *Intelligent Information and Database Systems (ACIIDS 2025)*, *Lecture Notes in Computer Science*, vol. 15683. Springer. https://doi.org/10.1007/978-981-96-6008-7_9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Doewes, A. (2026). *Rethinking Automated Essay Scoring: Agreement, Fairness, and Feedback* [Disertasi doktoral, Eindhoven University of Technology].

Doewes, A., Kurdhi, N. A., & Saxena, A. (2023). Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. *Proceedings of the International Conference on Educational Data Mining*, 103–113. https://doi.org/10.5281/zenodo.8115784

Geni, L., Yulianti, E., & Sensuse, D. I. (2023). Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, *9*(3), 746–757. https://doi.org/10.26555/jiteki.v9i3.26490

Judijanto, L., Abdullah, G., Abdurahman, A., Lumbu, A., Tumober, R. T., Septikasari, D., Sogalrey, F. A. M., Mahliatussikah, H., & Subhaktiyasa, P. G. (2025). *Evaluasi Pembelajaran: Prinsip, Teknik, dan Aplikasi*. Sonpedia Publishing.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770.

Kurniawati, F. E., & Mega Pradnya, W. (2020). Implementasi Algoritma Winnowing pada Sistem Penilaian Otomatis Jawaban Esai pada Ujian Online Berbasis Web. *Jurnal Teknik Komputer AMIK BSI*, *6*(2). https://doi.org/10.31294/jtk.v4i2

Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-03209-9

Miller, I., Choe, Y., & Emirtekin, E. (2025). Large Language Model-Powered Automated Assessment: A Systematic Review. *Applied Sciences*, *15*(10), 5683. https://doi.org/10.3390/app15105683

Misgna, H., On, B. W., Lee, I., & Choi, G. S. (2025). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, *58*(2). https://doi.org/10.1007/s10462-024-11017-5

Møller, A. G., Pera, A., Dalsgaard, J., & Aiello, L. (2024). The Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 179–192. https://doi.org/10.18653/v1/2024.eacl-short.17

Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *The Journal of Experimental Education*, *62*(2), 127–142. https://doi.org/10.1080/00220973.1994.9943835

Pradani, K. A., & Suadaa, L. H. (2023). Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer untuk Penilaian Ujian Esai. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *10*(6), 1177–1184. https://doi.org/10.25126/jtiik.2023107338

Prastyo, P. H., Tungadi, E., & Zuhdi, S. (2025). Indonesian Automated Essay Scoring: A Comparative Study of Pretrained Transformer Models. *Information Technology Education Journal*, 120–130. https://doi.org/10.59562/intec.v4i2.8069

Putri, H., Susiani, D., Wandani, N. S., & Putri, F. A. (2022). Instrumen Penilaian Hasil Pembelajaran Kognitif pada Tes Uraian dan Tes Objektif. *Jurnal Papeda*, *4*(2).

Rahanra, N., Hossam, A., & Scott, J. (2025). Utilizing Artificial Intelligence (AI) for Automated Feedback on the English Essay Writing Skills of Indonesian University Students. *Journal International of Lingua and Technology*, *4*(3), 260–275. https://doi.org/10.55849/jiltech.v4i3.1121

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*. http://arxiv.org/abs/1706.03762

Winarta, I. M. W. P., Alfarozi, S. A. I., & Hidayah, I. (2024). Atomic Evaluation using Large Language Model for Automated Essay Exam Scoring. *2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, 21–27. https://doi.org/10.1109/COMNETSAT63286.2024.10862459

Zubaidi, A., Munip, A., Widodo, S. A., & Zerrouki, T. (2025). Enhancing Arabic writing skills using ChatGPT-based AI learning models: A tridimensional human-AI collaboration framework. *Indonesian Journal of Applied Linguistics*, *15*(1), 87–101. https://doi.org/10.17509/ijal.v15i1.75378