ISSN: 3089-9168 Volume-3, Edisi-1, Januari 2026

Deteksi Teks Akademik berbasis *Generative AI*: Studi Akurasi ZeroGPT dan Implikasinya terhadap Keadilan Evaluasi Pembelajaran

Fitrah Izul Falaq
Universitas Negeri Jakarta
fitrah.izul.falaq@unj.ac.id
Indonesia

***Abstrak* — Integrasi *generative ai* yang berkembang pesat dalam dunia pendidikan menimbulkan kekhawatiran terhadap integritas akademik, khususnya dalam evaluasi tugas tertulis. Penelitian ini bertujuan untuk menilai akurasi ZeroGPT sebagai alat deteksi teks yang dihasilkan kecerdasan buatan. Penelitian ini menggunakan pengujian dengan model evaluasi klasifikasi kinerja yang diinterpretasikan menggunakan analisa deskriptif kuantitatif dan kualitatif berdasarkan studi literatur. Objek penelitian berupa 80 teks akademik berbahasa Indonesia dengan topik materi evaluasi pembelajaran, secara rinci terdiri dari 40 teks hasil generasi AI dan 40 teks yang ditulis oleh mahasiswa program sarjana. Hasil pengujian menggunakan confussion matriks menunjukkan ZeroGPT mencapai akurasi keseluruhan sebesar 37,5%, dengan tingkat *recall* sebesar 55% untuk teks hasil AI serta tingkat kesalahan positif palsu *(false positive)* sebesar 80% pada teks yang ditulis manusia. Temuan ini menunjukkan keterbatasan reliabilitas ZeroGPT, sehingga menjadikannya tidak layak digunakan sebagai alat pendeteksi tunggal penilaian sumatif. Penelitian ini merekomendasikan agar pendidik: (1) menerapkan penelaahan manual dan pertimbangan profesional; (2) menggunakan ZeroGPT semata-mata sebagai alat pendukung; dan (3) merancang evaluasi yang menekankan proses, refleksi, dan pemahaman konseptual guna meminimalkan risiko otomatisasi.**

Kata kunci* — *artificial intelligence, ai detector, zero-gpt, integritas akademik, evaluasi pembelajaran

I. PENDAHULUAN

Perkembangan *generative ai* secara fundamental telah mentransformasikan praktik pendidikan tinggi, khususnya dalam ranah penulisan akademik. Model LLM seperti ChatGPT, Claude, dan sistem sejenis telah terbukti mampu menghasilkan teks akademik yang fasih, koheren, dan terstruktur dengan baik dalam hitungan detik. Teknologi ini memang dapat menawarkan manfaat pedagogis, seperti media pembelajaran yang dipersonalisasi serta bantuan pengembangan gagasan. Namun, disaat yang bersamaan juga menantang asumsi-asumsi yang telah lama mapan mengenai kepengarangan, orisinalitas, dan validitas penilaian pembelajaran [1].

Proses pembelajaran pada pendidikan tinggi menekankan tugas tertulis secara tradisional yang berfungsi sebagai instrumen utama untuk menilai pemahaman mahasiswa, kemampuan berpikir kritis, serta kapasitas mereka dalam mensintesis pengetahuan. Namun, munculnya AI generatif mempersulit hal tersebut. Peluang mahasiswa untuk menggunakan *generative AI* dalam mengerjakan tugas penulisan mengalami peningkatan. Mahasiswa berpotensi menyerahkan teks hasil generasi AI yang secara administratif memang memenuhi standar akademik, namun secara esensial tidak selalu mencerminkan keterlibatan kognitif. Situasi ini memperkuat kekhawatiran terkait integritas akademik, plagiarisme, dan keadilan dalam evaluasi pembelajaran [2].

Tantangan lain dalam pendidikan adalah kesulitan dalam membedakan teks yang ditulis oleh mahasiswa dengan teks yang dihasilkan oleh sistem AI. Berbeda dengan plagiarisme konvensional yang bergantung pada deteksi kesamaan teks, konten hasil AI umumnya bersifat baru dan tidak merujuk secara langsung pada sumber yang sudah ada. Akibatnya, perangkat pendeteksi plagiarisme tradisional seperti Turnitin memiliki keterbatasan dalam menangani bentuk pelanggaran akademik yang relatif baru ini [3]. Keterbatasan tersebut mendorong pengembangan alat deteksi teks berbasis AI yang dirancang untuk mengidentifikasi pola probabilistik dan linguistik yang diasosiasikan dengan teks hasil *generative AI*.

Salah satu alat pendeteksi teks *generative AI* yang populer di lingkungan akademik adalah ZeroGPT. Alat ini memiliki aksesibilitas, antarmuka yang ramah pengguna, dan dapat digunakan secara gratis. ZeroGPT pada laman resminya mengklaim mampu mendeteksi teks hasil AI dalam berbagai bahasa melalui analisis karakteristik seperti *perplexity*, *burstiness*, dan distribusi bahasa statistik. Bahkan, beberapa lembaga pendidikan menggunakan ZeroGPT sebagai tools resmi untuk melakukan deteksi teks oleh *generative AI*. Meskipun populer, masih terdapat pertanyaan mengenai sejauh mana ZeroGPT dapat bekerja secara andal ketika diterapkan pada konteks penulisan akademik yang autentik.

Sejumlah penelitian menunjukkan bahwa alat deteksi AI sering kali menunjukkan kinerja yang tidak konsisten pada berbagai genre, bahasa, dan tingkat kemahiran menulis. Tantangan ini khususnya tampak pada penulisan akademik formal oleh mahasiswa, yang umumnya menampilkan argumen terstruktur, nada yang konsisten, serta ekspresi personal yang minimal serta karakteristik yang dapat sangat menyerupai teks hasil AI. Tumpang tindih ini meningkatkan kemungkinan terjadinya kesalahan positif palsu *(false positive)*, yaitu ketika teks yang ditulis manusia secara keliru diklasifikasikan sebagai teks hasil AI [4] Dalam lingkungan pendidikan, kesalahan semacam ini sangat problematik karena dapat memicu tuduhan yang tidak berdasar dan merusak kepercayaan antara dosen dan mahasiswa.

Implikasi etis dari klasifikasi positif palsu dalam evaluasi pembelajaran tidak dapat diremehkan. Hasil penilaian memiliki konsekuensi akademik dan psikologis yang signifikan bagi mahasiswa, termasuk penalti nilai, sanksi disipliner, dan kerusakan reputas. Penggunaan alat deteksi AI tanpa validasi yang memadai menimbulkan kekhawatiran terkait transparansi, prosedur yang adil (due process), dan keadilan dalam penilaian pendidikan. Ketika keputusan algoritmik diperlakukan sebagai otoritatif tanpa disertai pertimbangan manusia, risiko ketidakadilan menjadi semakin besar [3].

Integrasi alat AI ke dalam pendidikan harus disertai dengan evaluasi empiris yang ketat dan kerangka kerja yang berlandaskan pedagogi [1]. Alih-alih mengadopsi teknologi deteksi sebagai instrumen yang bersifat menghukum, pendidik didorong untuk mempertimbangkan keselarasan alat tersebut dengan tujuan pembelajaran, validitas penilaian, dan pengembangan mahasiswa. Perspektif ini menunjukkan bahwa nilai alat deteksi AI tidak seharusnya dinilai semata-mata dari kemampuannya dalam menandai penggunaan AI, melainkan dari kontribusinya terhadap praktik penilaian yang adil dan bermakna.

Terlepas dari berbagai pertimbangan tersebut, bukti empiris mengenai kinerja ZeroGPT dalam konteks pendidikan tinggi masih terbatas. Klaim publik mengenai tingkat akurasinya sering kali didasarkan pada tolok ukur kepemilikan *(proprietary benchmarks)* yang tidak didokumentasikan secara transparan. Akibatnya, para pendidik kekurangan data independen yang andal untuk mendukung keputusan tentang apakah dan bagaimana ZeroGPT sebaiknya digunakan dalam evaluasi pembelajaran. Kesenjangan ini menegaskan perlunya studi empiris yang secara sistematis menilai kinerja klasifikasi ZeroGPT dengan menggunakan sampel tulisan mahasiswa yang nyata.

Isu krusial lainnya berkaitan dengan sifat *generative AI* yang terus berkembang dan semakin canggih. Perkembangan ini semakin mempersulit upaya deteksi, karena alat yang dilatih menggunakan model-model sebelumnya dapat mengalami kesulitan dalam melakukan generalisasi terhadap sistem yang lebih baru [2]. Oleh karena itu, setiap klaim mengenai tingkat akurasi deteksi yang tinggi harus ditafsirkan secara hati-hati, terutama ketika diterapkan pada konteks pendidikan yang berisiko tinggi.

Selain itu, ketergantungan pada alat deteksi otomatis berisiko mengalihkan fokus penilaian dari proses pembelajaran menuju praktik pengawasan dan kontrol. Penilaian pada proses pembelajaran seharusnya dirancang untuk mendorong pembelajaran, refleksi, dan keterlibatan kritis, bukan sekadar mengidentifikasi pelanggaran aturan [4]. Ketergantungan berlebihan pada pendeteksi AI dapat mendorong hubungan yang bersifat adversarial antara mahasiswa dan dosen, di mana tujuan utama menjadi menghindari deteksi alih-alih menunjukkan pemahaman.

Berdasarkan tantangan teoretis, etis, dan praktis tersebut, terdapat kebutuhan mendesak untuk mengevaluasi kinerja aktual alat deteksi teks AI dalam konteks evaluasi pembelajaran. Penelitian ini merespons kebutuhan tersebut dengan mengkaji secara empiris akurasi ZeroGPT dalam membedakan teks akademik hasil AI dan teks akademik yang ditulis oleh mahasiswa perguruan tinggi. Dengan menggunakan analisis berbasis matriks *confusion matrix*.

Secara khusus, penelitian ini menjawab dua pertanyaan utama: (1) sejauh mana ZeroGPT mampu mendeteksi teks akademik hasil AI secara akurat; dan (2) seberapa sering ZeroGPT salah mengklasifikasikan teks akademik yang ditulis manusia sebagai hasil AI, serta apa implikasi kesalahan tersebut terhadap keadilan dalam evaluasi pembelajaran.

Terdapat dua kontribusi utama dalam penelitian ini. Pertama, penelitian ini menyediakan bukti empiris mengenai keterbatasan kinerja ZeroGPT ketika diterapkan pada tulisan mahasiswa yang autentik. Kedua, temuan tersebut diposisikan dalam diskursus yang lebih luas mengenai integritas akademik, validitas penilaian, dan praktik pendidikan yang etis. Alih-alih memandang deteksi AI sebagai persoalan teknis semata, penelitian ini menekankan konsekuensi pedagogis dari pengambilan keputusan berbasis algoritma dalam pendidikan.

Harapannya, penelitian ini menjadi referensi dalam mendukung serta dapat menyajikan wawasan berbasis data mengenai kelayakan penggunaan ZeroGPT sebagai alat evaluasi pembelajaran, khususnya dalam kaitannya dengan keadilan dan integritas akademik.

II. METODE

Penelitian ini menggunakan desain penelitian deskriptif kuantitatif dengan pendekatan evaluasi kinerja klasifikasi untuk mengkaji akurasi ZeroGPT sebagai alat deteksi teks berbasis kecerdasan buatan dalam evaluasi pembelajaran. Penelitian deskriptif kuantitatif dianggap tepat ketika tujuan penelitian adalah mengukur dan menggambarkan kinerja suatu sistem melalui indikator numerik tanpa melakukan manipulasi variabel maupun penerapan perlakuan eksperimental [5]

Pendekatan evaluasi kinerja klasifikasi lazim digunakan dalam penelitian yang menilai sistem algoritmik, khususnya dalam bidang pembelajaran mesin dan temu kembali informasi. Pendekatan ini berfokus pada perbandingan antara prediksi yang dihasilkan sistem dengan label kebenaran dasar *(ground truth)* yang telah diverifikasi, guna menentukan tingkat akurasi klasifikasi serta jenis-jenis kesalahan [4].

Dalam penelitian ini, pendekatan tersebut diterapkan untuk mengevaluasi sejauh mana ZeroGPT mampu membedakan teks akademik hasil generasi AI dan teks akademik yang ditulis oleh manusia. Desain penelitian menekankan prinsip transparansi dan replikabilitas melalui pendefinisian yang jelas terhadap kategori klasifikasi, metrik kinerja, serta prosedur evaluasi.

Partisipan

Partisipan dalam penelitian ini terdiri atas 40 mahasiswa program sarjana yang terdaftar pada suatu mata kuliah di perguruan tinggi yang mensyaratkan penyelesaian tugas penulisan akademik. Pemilihan partisipan dilakukan menggunakan teknik *convenience sampling*, yaitu berdasarkan kemudahan akses dan relevansinya dengan konteks penelitian. Seluruh partisipan memiliki pengalaman sebelumnya dalam menyelesaikan tugas penulisan akademik sebagai bagian dari kegiatan perkuliahan. Untuk menjaga standar etika penelitian, identitas mahasiswa dianonimkan selama proses pengumpulan dan analisis data. Teks yang dikumpulkan dari partisipan digunakan semata-mata untuk kepentingan penelitian dan tidak dikaitkan dengan penilaian akademik maupun tindakan disipliner apa pun.

Data dan Bahan Penelitian

Dataset yang dianalisis dalam penelitian ini terdiri atas 80 teks akademik, yang dikelompokkan ke dalam dua kategori dengan jumlah yang sama, yaitu:

1. Teks hasil generasi AI (n = 40)

Teks-teks ini dihasilkan menggunakan alat kecerdasan buatan generatif yang mampu memproduksi tulisan bergaya akademik. *Prompt* yang diberikan kepada AI disesuaikan dengan topik yang sama seperti yang diberikan kepada mahasiswa, guna memastikan keselarasan tema dan ruang lingkup pembahasan.

2. Teks tulisan manusia (n = 40)

Teks-teks ini ditulis secara mandiri oleh mahasiswa program sarjana tanpa bantuan alat AI, berdasarkan tugas penulisan akademik yang berkaitan dengan mata kuliah yang diujikan.

Untuk memastikan keterbandingan dan meminimalkan bias sistematis dalam proses deteksi, seluruh teks distandarkan berdasarkan kriteria berikut:

1. Ditulis dalam bahasa Indonesia
2. Mengangkat topik akademik yang sebanding
3. Menggunakan genre teks yang sama (penulisan akademik ekspositori)
4. Memiliki jumlah kata yang relatif setara
5. Menggunakan gaya bahasa akademik formal
6. Panjang teks sekitar 400 kata
7. Teks AI dihasilkan menggunakan ChatGPT 5.0
8. Proses generasi teks AI dilakukan secara simultan

Instrumen utama yang digunakan dalam proses klasifikasi adalah ZeroGPT, yaitu alat deteksi teks berbasis AI berbasis daring yang mengklasifikasikan teks sebagai hasil AI atau tulisan manusia berdasarkan analisis pola statistik dan linguistik.

Prosedur Penelitian

Prosedur penelitian dilaksanakan secara sistematis dan berurutan untuk menjamin konsistensi penanganan data serta meminimalkan bias prosedural. Tahapan penelitian meliputi:

1. Pengumpulan teks tulisan manusia
   Teks akademik yang ditulis oleh mahasiswa dikumpulkan sebagai bagian dari ujian perkuliahan reguler. Teks-teks ini merepresentasikan produk penulisan autentik mahasiswa.
2. Generasi teks berbasis AI
   Dengan menggunakan topik akademik yang sama atau sangat serupa, teks-teks hasil AI dihasilkan untuk mencerminkan cakupan materi dan kompleksitas struktural yang sebanding dengan teks tulisan mahasiswa.
3. Input teks ke dalam ZeroGPT
   Seluruh 80 teks dimasukkan secara individual ke dalam sistem ZeroGPT untuk dianalisis. Keluaran sistem menunjukkan apakah masing-masing teks diklasifikasikan sebagai teks hasil AI atau tulisan manusia.

Kategorisasi Hasil Klasifikasi

Hasil keluaran ZeroGPT dicatat dan dikelompokkan ke dalam dua kategori klasifikasi, yaitu *"AI-generated"* dan *"Human-written"*.

TABEL 1. KATEGORI HASIL KLASISIFIKASI

Aspek	Deskripsi
Perbandingan dengan Ground Truth	Setiap hasil klasifikasi yang dihasilkan oleh ZeroGPT dibandingkan dengan sumber teks yang sebenarnya, yaitu teks yang dihasilkan oleh AI atau manusia. Perbandingan ini bertujuan untuk menentukan tingkat ketepatan klasifikasi serta mengidentifikasi jenis kesalahan klasifikasi yang terjadi.
Metode Analisis Data	Analisis data dilakukan menggunakan *confusion matrix*, yang merupakan metode evaluasi standar dalam menilai kinerja sistem klasifikasi.
Fungsi Confusion Matrix	*Confusion matrix* memungkinkan identifikasi rinci terhadap empat kemungkinan hasil klasifikasi, yaitu *true positives* (TP), *false positives* (FP), *true negatives* (TN), dan *false negatives* (FN).
Rujukan Metodologis	Pendekatan *confusion matrix* merujuk pada kerangka evaluasi klasifikasi. [6]

Berdasarkan confusion matrix tersebut, dihitung beberapa metrik kinerja utama, meliputi:

TABEL 2. MATRIK EVALUASI KINERJA ZERO GPT

Metrik Evaluasi	Deskripsi
Accuracy	Merepresentasikan proporsi keseluruhan klasifikasi yang dilakukan secara benar oleh sistem dibandingkan dengan total data yang dianalisis.
Recall (Sensitivity)	Mengukur kemampuan sistem dalam mendeteksi teks yang dihasilkan oleh AI secara tepat dari seluruh teks AI yang ada.
Specificity	Mengukur kemampuan sistem dalam mengidentifikasi teks yang ditulis oleh manusia secara benar dari seluruh teks manusia yang tersedia.
Error Rate	Menunjukkan proporsi keseluruhan kesalahan klasifikasi yang dilakukan oleh sistem terhadap total data yang diuji.

Penggunaan berbagai metrik kinerja ini dipandang penting karena akurasi semata dapat memberikan gambaran yang menyesatkan mengenai kinerja sistem, khususnya dalam konteks pendidikan di mana kesalahan positif palsu dapat menimbulkan implikasi etis dan pedagogis yang signifikan [4]. Seluruh hasil analisis ditafsirkan dengan mempertimbangkan implikasinya terhadap prinsip keadilan dan reliabilitas dalam evaluasi pembelajaran.

III. HASIL DAN DISKUSI

Evaluasi akurasi ZeroGPT dalam mengidentifikasi teks akademik yang dihasilkan oleh AI dan yang ditulis oleh manusia dievaluasi menggunakan *confusion matrix*. Adapun hasil perbandingan antara sumber teks asli *(ground truth)* dan prediksi yang dihasilkan oleh ZeroGPT dapat dilihat pada Tabel 1.

TABEL 1. HASIL KINERJA ZERO GPT

Jenis Teks Asli	Terdeteksi sebagai AI	Terdeteksi sebagai Manusia	Total
Teks hasil AI	22	18	40
Teks tulisan manusia	32	8	40
Total	54	26	80

Berdasarkan *confusion matrix* tersebut, diperoleh hasil klasifikasi sebagai berikut:

1. True Positives (TP): 22 teks hasil AI yang berhasil diidentifikasi dengan benar sebagai teks AI
2. False Negatives (FN): 18 teks hasil AI yang keliru diidentifikasi sebagai teks tulisan manusia
3. False Positives (FP): 32 teks tulisan manusia yang keliru diidentifikasi sebagai teks AI
4. True Negatives (TN): 8 teks tulisan manusia yang berhasil diidentifikasi dengan benar

Akurasi *(Accuracy)*

Akurasi merepresentasikan proporsi data yang diklasifikasikan dengan benar dibandingkan dengan jumlah keseluruhan data.

Accuracy = (TP + TN) / Total

Accuracy = (22 + 8) / 80 = 0,375 atau 37,5%

Hasil ini menunjukkan bahwa kurang dari setengah teks yang dianalisis berhasil diklasifikasikan dengan benar, sehingga mengindikasikan bahwa ZeroGPT memiliki reliabilitas keseluruhan yang rendah pada dataset ini.

***Recall (Sensitivity)* untuk Teks Hasil AI**

Recall mengukur kemampuan sistem dalam mendeteksi teks yang benar-benar dihasilkan oleh AI.

Recall = TP / (TP + FN)

Recall = 22 / 40 = 0,55 atau 55%

Temuan ini menunjukkan bahwa sebesar 45% teks hasil AI tidak berhasil terdeteksi dan lolos sebagai teks tulisan manusia. Keterbatasan ini secara signifikan melemahkan efektivitas ZeroGPT dalam menjaga integritas akademik.

Spesifisitas *(Specificity)* untuk Teks Tulisan Manusia

Spesifisitas mengukur kemampuan sistem dalam mengidentifikasi teks yang benar-benar ditulis oleh manusia.

Specificity = TN / (TN + FP)

Specificity = 8 / 40 = 0,20 atau 20%

Nilai spesifisitas yang sangat rendah ini menunjukkan bahwa 80% teks tulisan manusia secara keliru diklasifikasikan sebagai teks hasil AI, yang merepresentasikan tingkat kesalahan yang sangat tinggi.

***False Positive Rate* (FPR)**

False Positive Rate menunjukkan proporsi teks tulisan manusia yang salah diklasifikasikan sebagai teks AI.

FPR = FP / (FP + TN)

FPR = 32 / 40 = 0,80 atau 80%

Nilai ini menunjukkan bahwa sebagian besar kesalahan klasifikasi terjadi ketika karya tulis mahasiswa yang autentik secara keliru dilabeli sebagai hasil AI.

Error Rate

Error rate menunjukkan proporsi keseluruhan kesalahan klasifikasi yang terjadi.

Error Rate = (FP + FN) / Total

Error Rate = 50 / 80 = 62,5%

Tingginya tingkat kesalahan ini semakin menegaskan bahwa ZeroGPT tidak memiliki tingkat akurasi yang memadai untuk digunakan dalam konteks evaluasi pembelajaran berisiko tinggi *(high-stakes assessment)*.

Temuan penelitian ini menunjukkan bahwa ZeroGPT memiliki reliabilitas keseluruhan yang rendah ketika diterapkan pada teks akademik autentik yang dihasilkan dalam konteks pendidikan tinggi. Meskipun ZeroGPT berhasil mendeteksi sebagian teks hasil generasi AI, kinerjanya secara signifikan tereduksi oleh tingginya tingkat kesalahan *positif palsu (false positive)* serta tingkat kesalahan *negatif palsu (false negative)* yang cukup besar. Hasil ini memperkuat kekhawatiran yang telah diangkat dalam literatur terkini mengenai keterbatasan alat deteksi teks berbasis AI dalam lingkungan pendidikan [3].

Salah satu temuan paling signifikan adalah tingkat kesalahan positif palsu yang sangat tinggi pada teks yang ditulis oleh manusia. Dengan 80% teks karya mahasiswa yang secara keliru diklasifikasikan sebagai teks hasil AI, ZeroGPT menunjukkan bias yang kuat terhadap pelabelan tulisan akademik formal sebagai produk mesin. Temuan ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa sistem deteksi AI kerap bergantung pada isyarat linguistik seperti keteraturan kalimat, konsistensi leksikal, dan koherensi struktural yang juga lazim ditemukan pada tulisan akademik mahasiswa yang ditulis dengan baik [4].

Dari perspektif pedagogis, kesalahan positif palsu merupakan masalah yang sangat serius. Dalam evaluasi pembelajaran, pengklasifikasian keliru terhadap karya tulis manusia sebagai hasil AI dapat berujung pada sanksi akademik yang tidak adil serta merusak kepercayaan mahasiswa terhadap sistem penilaian. Berbeda dengan kesalahan negatif palsu yang memungkinkan konten hasil AI lolos dari deteksi, kesalahan positif palsu secara langsung merugikan mahasiswa yang telah mematuhi prinsip integritas akademik. Kondisi ini menimbulkan persoalan etis yang serius terkait keadilan, transparansi, dan pemenuhan prosedur yang adil *(due process)* dalam penilaian pendidikan [3].

Selain persoalan keadilan, rendahnya tingkat recall (55%) untuk teks hasil AI menunjukkan bahwa ZeroGPT juga gagal berfungsi sebagai mekanisme perlindungan yang efektif terhadap penyalahgunaan AI. Hampir setengah dari teks hasil AI tidak terdeteksi, sehingga ketergantungan pada ZeroGPT semata dapat menimbulkan rasa aman semu terkait integritas akademik. Keterbatasan ini semakin mengkhawatirkan mengingat perkembangan pesat model AI generatif yang menghasilkan keluaran semakin menyerupai tulisan manusia, baik dari segi gaya maupun struktur [2].

Kombinasi antara tingginya kesalahan positif palsu dan besarnya kesalahan negatif palsu mengindikasikan bahwa ZeroGPT tidak memiliki ketangguhan *(robustness)* sebagai sistem klasifikasi dalam konteks ini. Meskipun alat tersebut mungkin menunjukkan kinerja yang baik dalam skenario pengujian terkontrol, performanya menurun ketika diterapkan pada tulisan mahasiswa yang nyata. Ketidaksesuaian ini menunjukkan bahwa algoritma deteksi AI saat ini masih kesulitan melakukan generalisasi terhadap keragaman gaya penulisan akademik dan tingkat kemahiran penulis.

Temuan ini juga mempertanyakan penggunaan akurasi sebagai satu-satunya indikator kinerja. Walaupun akurasi merupakan metrik yang umum digunakan, metrik ini dapat

menutupi kesalahan klasifikasi yang krusial dalam konteks yang tidak seimbang atau berdampak tinggi. Dalam penilaian pendidikan, kesalahan positif palsu memiliki konsekuensi yang lebih berat dibandingkan kesalahan negatif palsu, namun perhitungan akurasi memberikan bobot yang sama pada keduanya [6]. Oleh karena itu, ketergantungan pada akurasi tanpa analisis kesalahan yang rinci berpotensi menghasilkan kesimpulan yang menyesatkan mengenai efektivitas suatu alat.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa ZeroGPT tidak seharusnya digunakan sebagai instrumen penentu dalam penilaian sumatif. Penggunaannya perlu dibatasi pada tahap penyaringan awal atau analisis eksploratif. Bahkan dalam konteks tersebut, keluaran ZeroGPT harus ditafsirkan secara hati-hati serta dilengkapi dengan pertimbangan manusia, pemahaman kontekstual, dan strategi penilaian alternatif.

Dari sudut pandang pendidikan yang lebih luas, temuan ini menegaskan perlunya peninjauan ulang desain penilaian di era AI generatif. Alih-alih hanya mengandalkan teknologi deteksi, pendidik perlu menekankan penilaian berbasis proses, penulisan reflektif, ujian lisan, dan tugas autentik yang relatif lebih resisten terhadap otomatisasi. Alat deteksi AI dapat berkontribusi pada peningkatan kesadaran dan dialog, namun tidak dapat menggantikan keahlian pedagogis dan tanggung jawab etis pendidik.

Prosedur penggunaan ZeroGPT sebagai Alat Pendeteksi Teks Generatif AI yang sesuai proses Pembelajaran

Penelitian menunjukkan bahwa format penilaian tradisional yang sangat bergantung pada esai tugas rumah semakin rentan terhadap penyalahgunaan AI [2]. Literatur yang berkembang mendukung pergeseran menuju penilaian autentik dan berorientasi proses, yang menekankan orisinalitas, relevansi kontekstual, dan refleksi mahasiswa [7]. Penilaian autentik mengurangi risiko pengumpulan tugas hasil AI dengan menuntut penerapan pengalaman personal, konteks spesifik disiplin, atau draf berulang yang sulit direplikasi secara meyakinkan oleh AI semata.

Studi-studi sebelumnya menunjukkan bahwa penugasan bertahap *(scaffolded assignments)*, proses penulisan multi-tahap, dan komentar reflektif secara signifikan meningkatkan keterlacakan proses belajar mahasiswa serta mengurangi peluang terjadinya pelanggaran akademik [8]. Pendekatan-pendekatan ini selaras dengan rekomendasi kerangka integritas akademik di berbagai institusi pendidikan tinggi yang menekankan penciptaan lingkungan penilaian di mana penyalahgunaan AI menjadi kurang memungkinkan [9].

Pentingnya Peran Pendidik dalam Interpretasi dan Pengambilan Keputusan terhadap Teks *Generative AI*

Tingginya tingkat kesalahan positif palsu yang ditemukan dalam penelitian ini mendukung temuan penelitian sebelumnya bahwa pendeteksi AI sering salah mengidentifikasi teks akademik yang ditulis manusia, terutama karya mahasiswa penutur non-native atau mahasiswa yang menggunakan bahasa akademik formal [10]. Oleh karena itu, pengawasan manusia tetap tidak tergantikan. Alat deteksi AI seharusnya hanya berfungsi sebagai mekanisme penandaan awal, bukan sebagai penentu akhir pelanggaran integritas akademik. Para pakar secara konsisten memperingatkan bahaya ketergantungan berlebihan pada alat algoritmik dalam konteks evaluatif, karena alat tersebut dapat memuat bias linguistik yang tidak disengaja, kurang transparan, serta menghasilkan klasifikasi yang tidak bermakna secara pedagogis [11], [12]. Pendidik perlu menerapkan pertimbangan profesional dalam menafsirkan hasil deteksi AI dan mengedepankan dialog dengan mahasiswa, alih-alih mengambil tindakan represif semata-mata berdasarkan keluaran otomatis.

Membangun Literasi AI bagi Mahasiswa dan Pendidik

Integritas akademik di era AI tidak dapat hanya mengandalkan deteksi, melainkan juga harus mencakup pendidikan literasi AI. Sejumlah penelitian menekankan pentingnya validasi independen terhadap alat deteksi AI, mengingat banyak di antaranya mengemukakan klaim akurasi berdasarkan dataset tertutup atau tolok ukur kepemilikan yang tidak diungkapkan [13]. Perguruan tinggi perlu membekali mahasiswa dengan pemahaman untuk menggunakan alat AI secara etis, bertanggung jawab, dan transparan, alih-alih sekadar melarang penggunaannya [14]. Pendidik juga perlu diperlengkapi dengan pengetahuan mengenai cara kerja AI, keterbatasannya, serta implikasi pedagogisnya. Inisiatif literasi AI terbukti mendorong pengambilan keputusan yang lebih berlandaskan pemahaman, mengurangi penyalahgunaan yang tidak disengaja, dan menumbuhkan budaya integritas [15]. Hal ini sejalan dengan rekomendasi global yang menekankan pemberdayaan dan kesadaran kritis dibandingkan pendekatan pengawasan semata.

Implikasi terhadap Evaluasi Pembelajaran

Berdasarkan temuan empiris penelitian ini diatas, beberapa implikasi dapat ditarik:

1. ZeroGPT tidak layak digunakan dalam penilaian sumatif, karena kesalahan klasifikasinya berpotensi menghasilkan keputusan akademik yang tidak adil. Alat deteksi AI tidak dapat dijadikan instrumen tunggal dalam penilaian berisiko tinggi akibat tingginya tingkat kesalahan positif palsu dan negatif palsu.
2. Alat ini memiliki risiko salah klasifikasi yang tinggi, khususnya terhadap tulisan akademik formal mahasiswa.
3. ZeroGPT paling jauh hanya dapat digunakan sebagai alat eksploratif atau penyaringan awal, dan itupun harus selalu dikombinasikan dengan pertimbangan dosen.

Oleh karena itu, evaluasi pembelajaran perlu terus menekankan:

1. Penilaian berbasis proses yang lebih menekankan pada penilaian ekploratif;
2. Menggunkaan ujian lisan atau wawancara reflektif;
3. Melakukan penilaian melalui praktikum terstruktur;
4. Penugasan yang terpersonalisasi dan berlandaskan konteks;
5. Tetap melakukan evaluasi secara manual untuk memastikan hasil yang telah dihasilkan oleh alat pendeteksi teks generative AI

Pendekatan-pendekatan tersebut mengurangi ketergantungan pada sistem deteksi otomatis dan mendukung penilaian yang lebih valid terhadap capaian belajar mahasiswa.

IV. KESIMPULAN

Berdasarkan temuan penelitian ini, dapat disimpulkan bahwa ZeroGPT tidak menunjukkan kinerja klasifikasi yang andal dalam membedakan teks hasil generasi AI dan teks akademik yang ditulis oleh manusia dalam konteks pendidikan tinggi. Tingkat akurasi keseluruhan sebesar 37,5%, yang disertai dengan tingkat kesalahan positif palsu *(false positive)* yang tinggi sebesar 80% pada teks tulisan manusia serta tingkat *recall* yang hanya mencapai 55% untuk teks hasil AI, mengindikasikan adanya keterbatasan yang signifikan dalam efektivitas alat tersebut.

Hasil penelitian menunjukkan bahwa ZeroGPT kerap salah mengklasifikasikan tulisan autentik mahasiswa sebagai teks hasil AI, yang berpotensi menimbulkan risiko serius terhadap prinsip keadilan dan integritas etis dalam evaluasi pembelajaran. Pada saat yang sama, alat ini juga gagal mendeteksi proporsi yang cukup besar dari teks yang benar-benar dihasilkan oleh AI, sehingga melemahkan fungsinya sebagai mekanisme penjaga integritas akademik. Dua keterbatasan ini secara bersamaan menunjukkan bahwa ZeroGPT tidak layak digunakan sebagai instrumen tunggal dalam penilaian sumatif, di mana kesalahan klasifikasi dapat menimbulkan konsekuensi akademik yang signifikan bagi mahasiswa.

Berdasarkan temuan tersebut, penelitian ini merekomendasikan agar pendidik: (1) Tetap menerapkan penelaahan manual serta pertimbangan profesional dalam mengevaluasi karya tulis mahasiswa.; (2) Menggunakan ZeroGPT semata-mata sebagai alat pendukung atau eksploratif, bukan sebagai mekanisme penilaian yang bersifat menentukan. (3) Merancang evaluasi pembelajaran yang menekankan proses, refleksi, dan pemahaman konseptual, seperti penulisan reflektif, ujian lisan, atau penugasan kontekstual yang relatif lebih resisten terhadap otomatisasi.

Lebih lanjut, temuan ini menegaskan pentingnya penerapan teknologi deteksi AI secara hati-hati dan bertanggung jawab dalam bidang pendidikan. Alat klasifikasi otomatis seharusnya berfungsi untuk mendukung pengambilan keputusan pedagogis, bukan menggantikannya, khususnya dalam konteks penilaian berisiko tinggi. Penelitian selanjutnya disarankan untuk melibatkan sampel yang lebih besar dan beragam, mencakup berbagai disiplin akademik, serta melakukan analisis komparatif terhadap berbagai alat deteksi teks berbasis AI. Dengan demikian, pemahaman yang lebih komprehensif mengenai reliabilitas, keterbatasan, dan implikasi etis teknologi deteksi AI dalam penilaian pendidikan dapat diperoleh.

REFERENSI

[1] E. Kasneci *et al.*, “ChatGPT for good? On opportunities and challenges of large language models for education,” *Learn. Individ. Differ.*, vol. 103, p. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.

[2] Y. Zhang and Q. Wang, “AI-generated academic writing and integrity in higher education,” *Computers & Education: Artificial Intelligence*, vol. 5, p. 100146, 2024, doi: 10.1016/j.caeai.2024.100146.

[3] M. Perkins, J. Roe, and D. Postma, “Detection of AI-generated text in higher education: Reliability and ethical considerations,” *Assess. Eval. High. Educ.*, vol. 48, no. 8, pp. 1239–1254, 2023, doi: 10.1080/02602938.2023.2182337.

[4] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[5] J. W. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*, 5th ed. Thousand Oaks: SAGE Publications, 2018.

[6] D. M. W. Powers, “Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[7] D. Vlachopoulos, “Authentic assessment in higher education: A systematic literature review,” *Assess. Educ.*, vol. 31, no. 1, pp. 1–21, 2024.

[8] P. Sotiriadou, D. Logan, A. Daly, and R. Guest, “The role of authentic assessment in fostering academic integrity,” *Assess. Eval. High. Educ.*, vol. 45, no. 3, pp. 416–431, 2020.

[9] T. Bretag, “A conceptual framework for academic integrity in higher education,” *Journal of Higher Education Research & Development*, vol. 38, no. 3, pp. 353–365, 2019.

[10] J. Roe and M. Perkins, “Generative AI in Self-Directed Learning: A Scoping Review,” *ArXiv*, 2024, doi: 10.48550/arXiv.2411.07677.

[11] L. Elsen and A. Rotaru, “Algorithmic detection of AI-generated text: Ethical risks and pedagogical implications,” *Comput. Educ.*, vol. 205, p. 104895, 2024.

[12] D. Kovacevic, “Use of ChatGPT in ESP Teaching Process,” *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–5, 2023, doi: 10.1109/INFOTEH57020.2023.10094133.

[13] W. Kim, “Analysis of the Educational Effects Regarding the Use of ChatGPT in Compulsory Basic Coding Subjects,” *The Korean Association of General Education*, 2023, doi: 10.46392/kjge.2023.17.5.113.

[14] W. Holmes, K. Porayska-Pomsta, and K. Holstein, “Ethics of AI in Education: Towards Responsible and Trustworthy AI in EdTech,” *British Journal of Educational Technology*, vol. 53, no. 6, pp. 1623–1639, 2022.

[15] E. Kessler and T. McLaughlin, “AI literacy as a foundation for academic integrity in the age of generative technologies,” *J. Acad. Ethics*, vol. 21, no. 4, pp. 789–803, 2023.