

BAB V

PEMBAHASAN

A. Kualitas Soal Sumatif Bahasa Indonesia yang Dihasilkan *Artificial Intelligence* (ChatGPT dan DeepSeek) Berdasarkan Kriteria Kualitas Soal

Penelitian ini dilakukan untuk menganalisis kualitas soal sumatif Bahasa Indonesia yang dihasilkan oleh dua platform *Artificial Intelligence* (AI), yaitu ChatGPT dan DeepSeek. Berdasarkan hasil penelitian yang telah dipaparkan pada Bab IV, kedua model AI mampu menghasilkan soal sumatif sesuai dengan instruksi atau prompt yang diberikan oleh peneliti. Pemberian prompt yang sama kepada kedua AI bertujuan untuk memperoleh hasil yang objektif sehingga perbedaan kualitas soal yang dihasilkan dapat dibandingkan berdasarkan kemampuan masing-masing model dalam memahami instruksi.

Kualitas soal dalam penelitian ini dianalisis berdasarkan beberapa aspek, yaitu kesesuaian isi, konstruksi soal, penggunaan bahasa, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Aspek tersebut merupakan bagian penting dalam menentukan kualitas instrumen evaluasi pembelajaran. Menurut Arikunto (2019), instrumen evaluasi yang baik harus mampu mengukur kemampuan peserta didik sesuai dengan tujuan pembelajaran yang telah ditetapkan. Dengan demikian, sebuah soal tidak hanya dinilai berdasarkan bentuk dan jumlah butirnya, tetapi juga berdasarkan kemampuan soal dalam mengukur kompetensi peserta didik

secara tepat.

Hasil penelitian menunjukkan bahwa ChatGPT dan DeepSeek memiliki kemampuan dalam menghasilkan soal Bahasa Indonesia yang sesuai dengan materi pembelajaran kelas X. Kedua AI mampu memahami konteks materi, menyusun pertanyaan, memberikan alternatif jawaban, serta menentukan kunci jawaban berdasarkan perintah yang diberikan. Hal tersebut menunjukkan bahwa teknologi AI memiliki potensi sebagai alat bantu dalam proses penyusunan instrumen evaluasi pembelajaran.

Temuan penelitian ini memperkuat pendapat Rustamaji dan Hidayat (2023) yang menyatakan bahwa model bahasa besar atau Large Language Model (LLM) seperti ChatGPT memiliki kemampuan dalam memahami, mengolah, dan menghasilkan teks secara otomatis sehingga dapat dimanfaatkan dalam bidang pendidikan. Kemampuan tersebut memungkinkan AI digunakan dalam membantu guru membuat variasi soal, menyusun bahan evaluasi, serta mempercepat proses penyusunan instrumen pembelajaran.

Namun demikian, hasil penelitian juga menunjukkan bahwa penggunaan AI dalam penyusunan soal tetap membutuhkan proses validasi dari guru atau ahli evaluasi. Hal ini dikarenakan AI bekerja berdasarkan pola bahasa dan data yang tersedia, bukan berdasarkan pemahaman langsung terhadap kondisi peserta didik di kelas. Oleh karena itu, guru tetap memiliki peran penting dalam memastikan bahwa soal yang dihasilkan telah sesuai dengan capaian pembelajaran, tujuan pembelajaran, serta karakteristik peserta didik.

Hasil tersebut sejalan dengan penelitian Agustiana, Mayrita, dan Muchti mengenai analisis butir soal Bahasa Indonesia yang menyatakan bahwa kualitas soal perlu dianalisis secara mendalam melalui aspek validitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh agar instrumen evaluasi benar-benar mampu mengukur kemampuan siswa.

B. Perbedaan Kinerja ChatGPT dan DeepSeek dalam Menghasilkan Soal Sumatif Bahasa Indonesia

Berdasarkan hasil analisis data penelitian, terdapat perbedaan kecenderungan hasil antara soal yang dihasilkan oleh ChatGPT dan DeepSeek. Secara deskriptif, DeepSeek memperoleh nilai penilaian yang lebih tinggi dibandingkan ChatGPT. Hal tersebut menunjukkan bahwa peserta didik memberikan respons yang lebih positif terhadap soal yang dihasilkan oleh DeepSeek terutama pada aspek kesesuaian instruksi, kelengkapan komponen soal, dan ketepatan format soal.

Perbedaan tersebut menunjukkan bahwa meskipun ChatGPT dan DeepSeek sama-sama menggunakan teknologi AI berbasis bahasa, keduanya memiliki karakteristik pemrosesan informasi yang berbeda. ChatGPT memiliki kemampuan dalam menghasilkan bahasa yang lebih natural, komunikatif, dan mudah dipahami. Sementara itu, DeepSeek menunjukkan kemampuan yang lebih baik dalam mengikuti instruksi yang diberikan secara spesifik melalui prompt.

Hal ini dapat dikaitkan dengan teori penggunaan Artificial Intelligence dalam pendidikan yang menjelaskan bahwa setiap model AI memiliki kemampuan berbeda berdasarkan arsitektur dan cara pemrosesan

datanya. AI tidak hanya dinilai dari kemampuan menghasilkan jawaban, tetapi juga dari kemampuan memahami konteks, mengikuti instruksi, dan menghasilkan keluaran yang sesuai dengan kebutuhan pengguna.

Hasil penelitian ini memiliki kesamaan dengan penelitian sebelumnya mengenai perbandingan performa AI yang menyatakan bahwa tidak terdapat satu model AI yang selalu unggul dalam seluruh aspek. Keunggulan sebuah AI sangat dipengaruhi oleh jenis tugas yang diberikan. Model tertentu dapat lebih unggul dalam kemampuan bahasa, sedangkan model lain lebih unggul dalam mengikuti instruksi teknis atau menghasilkan struktur yang sesuai.

Dalam penelitian ini, DeepSeek menunjukkan keunggulan pada aspek kesesuaian dengan prompt penelitian. Hal tersebut menunjukkan bahwa DeepSeek mampu menerjemahkan instruksi yang diberikan menjadi bentuk soal yang lebih sesuai dengan kebutuhan penelitian. Meskipun demikian, ChatGPT tetap memiliki keunggulan dalam aspek penyajian bahasa dan keterbacaan soal.

C. Interpretasi Hasil Uji Statistik Perbandingan ChatGPT dan DeepSeek

Berdasarkan hasil pengujian statistik yang dilakukan, penelitian ini menggunakan beberapa tahap analisis yaitu uji normalitas, Mann-Whitney U Test, Chi-Square, dan Fisher's Exact Test. Pemilihan metode tersebut sesuai dengan pendekatan penelitian yang digunakan, yaitu kuantitatif komparatif untuk mengetahui perbedaan kinerja dua model AI dalam menghasilkan soal sumatif Bahasa Indonesia.

Hasil uji Mann-Whitney menunjukkan bahwa tidak terdapat

perbedaan signifikan antara hasil soal yang dihasilkan oleh ChatGPT dan DeepSeek. Hal tersebut menunjukkan bahwa secara statistik kedua AI memiliki kemampuan yang relatif sama dalam menghasilkan soal sumatif Bahasa Indonesia.

Tidak adanya perbedaan signifikan dapat dipahami karena kedua AI sama-sama termasuk teknologi kecerdasan buatan generatif yang memiliki kemampuan memahami bahasa dan menghasilkan teks berdasarkan pola data yang besar. Oleh sebab itu, meskipun terdapat perbedaan hasil secara deskriptif, perbedaan tersebut belum cukup menunjukkan bahwa salah satu AI secara mutlak lebih unggul dibandingkan yang lain.

Hasil ini mendukung pandangan bahwa AI dalam pendidikan bukanlah pengganti peran guru, melainkan sebagai teknologi pendukung yang membantu meningkatkan efektivitas pekerjaan guru. Guru tetap memiliki tanggung jawab dalam menilai kelayakan soal sebelum digunakan dalam kegiatan evaluasi.

Penelitian sebelumnya yang dilakukan oleh Alatas, Romadhon, dan Rachmayanti mengenai penggunaan ChatGPT dalam pembelajaran Bahasa Indonesia juga menunjukkan bahwa AI memberikan manfaat dalam membantu kegiatan akademik, tetapi tetap memiliki keterbatasan terutama dalam aspek akurasi informasi sehingga diperlukan pengawasan pengguna.

D. Hubungan Hasil Penelitian dengan Teori Evaluasi Pembelajaran

Dalam teori evaluasi pembelajaran, kualitas instrumen penilaian menjadi faktor penting dalam menentukan keberhasilan proses pengukuran hasil belajar. Instrumen yang baik harus memenuhi prinsip validitas,

reliabilitas, objektivitas, dan praktikalitas. Dalam penelitian ini, kualitas soal AI dikaji berdasarkan indikator tingkat kesukaran, daya pembeda, dan efektivitas pengecoh.

Menurut Haladyna (2004), tingkat kesukaran menunjukkan kemampuan soal dalam mengukur proporsi peserta didik yang mampu menjawab dengan benar. Sementara itu, daya pembeda menunjukkan kemampuan soal dalam membedakan peserta didik dengan kemampuan tinggi dan rendah. Efektivitas pengecoh juga menjadi indikator penting karena pilihan jawaban salah harus mampu menarik peserta didik yang belum memahami materi.

Berdasarkan hasil penelitian, AI mampu membantu proses awal penyusunan soal, tetapi belum sepenuhnya dapat menggantikan analisis evaluasi secara manual. Hal tersebut dikarenakan kualitas soal tidak hanya ditentukan oleh kemampuan menghasilkan teks, tetapi juga oleh kemampuan soal dalam mengukur kemampuan peserta didik secara nyata.

Dengan demikian, penggunaan AI dalam penyusunan soal perlu dilakukan melalui kombinasi antara kemampuan teknologi dan kompetensi profesional guru. AI berperan dalam mempercepat proses produksi soal, sedangkan guru berperan dalam melakukan validasi dan penyempurnaan instrumen.

E. Implikasi Penelitian terhadap Pemanfaatan *Artificial Intelligence* dalam Pendidikan

Hasil penelitian ini memberikan implikasi bahwa *Artificial Intelligence* memiliki peluang besar untuk diterapkan dalam bidang evaluasi

pendidikan, khususnya dalam membantu guru menyusun soal sumatif Bahasa Indonesia. Pemanfaatan AI dapat menghemat waktu, memberikan variasi soal, dan membantu guru menghasilkan instrumen evaluasi secara lebih efektif.

Berdasarkan hasil penelitian, DeepSeek menunjukkan kecenderungan lebih unggul berdasarkan penilaian peserta didik terhadap kualitas soal yang dihasilkan. Namun, hasil tersebut tidak berarti bahwa ChatGPT tidak dapat digunakan. Kedua AI memiliki kelebihan masing-masing sehingga penggunaannya dapat disesuaikan dengan kebutuhan guru.

ChatGPT dapat dimanfaatkan ketika guru membutuhkan penyusunan soal dengan bahasa yang lebih komunikatif dan fleksibel, sedangkan DeepSeek dapat digunakan ketika guru membutuhkan hasil yang lebih sesuai dengan instruksi tertentu.

Dengan demikian, hasil penelitian ini menegaskan bahwa *Artificial Intelligence* merupakan teknologi pendukung dalam evaluasi pembelajaran. Keberhasilan pemanfaatan AI tetap bergantung pada kemampuan guru dalam memberikan instruksi, melakukan evaluasi, dan mengambil keputusan terhadap hasil yang diberikan oleh AI.

Berdasarkan keseluruhan pembahasan dapat disimpulkan bahwa ChatGPT dan DeepSeek sama-sama memiliki kemampuan dalam menghasilkan soal sumatif Bahasa Indonesia. Perbedaan kinerja keduanya terlihat pada karakteristik hasil keluaran, di mana DeepSeek menunjukkan kecenderungan lebih baik dalam kesesuaian instruksi, sedangkan ChatGPT memiliki keunggulan pada aspek penyajian bahasa. Oleh karena itu, AI

dapat dijadikan sebagai alat bantu inovatif dalam penyusunan evaluasi pembelajaran, tetapi tetap memerlukan validasi manusia agar kualitas instrumen tetap terjamin.

F. Implikasi Teoritis dan Praktis Penelitian

1. Implikasi Teoretis

Penelitian ini memperluas kajian teori evaluasi pendidikan dengan memasukkan *Artificial Intelligence* sebagai variabel evaluatif. Selama ini evaluasi pendidikan lebih banyak dilakukan secara manual oleh guru atau ahli evaluasi pendidikan. Penelitian ini membuktikan bahwa AI dapat digunakan sebagai alat bantu evaluasi yang mampu mempercepat analisis soal, meningkatkan efisiensi kerja guru, membantu identifikasi kelemahan soal, memberikan rekomendasi revisi soal secara otomatis.

Selain itu, penelitian ini juga menunjukkan bahwa teori evaluasi pendidikan klasik masih relevan digunakan dalam era digital karena indikator kualitas soal seperti validitas, reliabilitas, daya pembeda, dan efektivitas pengecoh tetap menjadi dasar utama evaluasi.

2. Implikasi Praktis

Bagi guru Bahasa Indonesia, penelitian ini memberikan alternatif penggunaan AI dalam proses evaluasi pembelajaran. Guru dapat menggunakan DeepSeek untuk analisis teknis kualitas soal, sedangkan ChatGPT dapat digunakan untuk pengembangan materi dan penyusunan bahasa soal.

Bagi sekolah, penelitian ini dapat menjadi dasar pengembangan sistem evaluasi berbasis teknologi AI. Penggunaan AI dapat

meningkatkan efisiensi waktu dan kualitas analisis pembelajaran.

Bagi peneliti selanjutnya, penelitian ini membuka peluang kajian baru mengenai integrasi AI dalam evaluasi pendidikan berbasis kurikulum merdeka.

G. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan. Pertama, sampel penelitian hanya berasal dari satu kelas di SMAN 7 Kediri sehingga hasil penelitian belum dapat digeneralisasikan secara luas pada jenjang maupun sekolah lain. Kedua, penelitian ini hanya membandingkan dua model kecerdasan buatan, yaitu ChatGPT dan DeepSeek, tanpa melibatkan model AI lain seperti Gemini, Claude, maupun platform AI generatif lainnya sehingga ruang lingkup perbandingan masih terbatas. Ketiga, analisis penelitian hanya difokuskan pada soal pilihan ganda sehingga belum mencakup bentuk evaluasi lain seperti soal uraian, proyek, atau asesmen autentik. Selain itu, hasil analisis AI sangat dipengaruhi oleh kualitas prompt yang disusun peneliti sehingga perbedaan struktur prompt berpotensi memengaruhi output yang dihasilkan. Penelitian ini juga belum mengukur stabilitas dan konsistensi output AI dalam jangka panjang, mengingat model kecerdasan buatan terus mengalami pembaruan sistem dan perubahan performa secara berkala.

H. Rekomendasi dan Pengembangan Teori Baru

Berdasarkan hasil penelitian, dapat dirumuskan bahwa kecerdasan buatan (AI) dalam evaluasi pendidikan memiliki dua orientasi utama, yaitu AI komunikatif-pedagogis seperti ChatGPT yang unggul dalam

penyampaian bahasa dan penjelasan pembelajaran secara komunikatif, serta AI analitis-teknis seperti DeepSeek yang lebih kuat dalam analisis teknis evaluasi soal. Kedua orientasi tersebut dapat dikembangkan menjadi suatu konsep baru dalam evaluasi pendidikan digital yang disebut sebagai “Hybrid AI Evaluation Model”, yaitu model evaluasi pembelajaran yang mengintegrasikan kemampuan komunikatif dan kemampuan analitis AI secara bersamaan.

Model ini memungkinkan guru memperoleh analisis teknis yang lebih akurat, penjelasan pedagogis yang komunikatif, serta rekomendasi revisi soal yang lebih efektif dan sistematis. Dengan demikian, penelitian ini tidak hanya berfokus pada perbandingan dua model AI, tetapi juga memberikan kontribusi konseptual terhadap pengembangan evaluasi pendidikan berbasis kecerdasan buatan di era digital.

I. Perbandingan Temuan Penelitian dengan Penelitian Terdahulu

Pada sub-bab ini, temuan utama penelitian akan dibandingkan dengan hasil-hasil penelitian terdahulu yang relevan guna memperkuat posisi penelitian dalam khazanah keilmuan serta mengidentifikasi aspek konfirmasi maupun perbedaan yang muncul. Penelitian ini menemukan bahwa DeepSeek lebih unggul dalam analisis teknis pada aspek tingkat kesukaran dan daya pembeda, sementara ChatGPT lebih unggul dalam aspek kebahasaan pada efektivitas pengecoh. Temuan ini sejalan dengan penelitian yang dilakukan oleh Rahayu dan Sukenti (2024) yang menemukan bahwa analisis butir soal secara manual sering menghasilkan soal dengan tingkat kesukaran yang didominasi kategori "mudah". Dengan

demikian, penelitian ini mengonfirmasi bahwa pemanfaatan kecerdasan buatan, khususnya DeepSeek, dapat menjadi solusi alternatif untuk menghasilkan soal dengan tingkat kesukaran yang lebih bervariasi, termasuk soal-soal HOTS.

Selanjutnya, berkaitan dengan efektivitas pengecoh, penelitian ini menunjukkan bahwa ChatGPT unggul tipis dengan rata-rata skor 3,96 dibandingkan DeepSeek yang memperoleh 3,94, terutama dalam hal kemudahan fitur dan kesesuaian bahasa dengan gaya guru. Temuan ini berbeda dengan hasil penelitian Agustiana, Mayrita, dan Muchti (2018) yang menemukan bahwa sekitar 33% pengecoh dalam soal buatan guru tidak efektif. Perbedaan ini justru menunjukkan bahwa kecerdasan buatan, baik ChatGPT maupun DeepSeek, mampu meningkatkan efektivitas pengecoh secara signifikan dibandingkan dengan penyusunan soal secara manual oleh guru, sehingga AI dapat berperan sebagai alat bantu yang efektif dalam menyusun distraktor yang lebih berfungsi.

Adapun mengenai perbandingan kinerja antara ChatGPT dan DeepSeek, penelitian ini menemukan bahwa DeepSeek secara signifikan lebih unggul dibandingkan ChatGPT berdasarkan uji Mann-Whitney dengan nilai signifikansi $p < 0,001$. Temuan ini menarik karena berbeda dengan hasil penelitian Chow dkk. (2025) yang justru menemukan bahwa ChatGPT-4 lebih unggul dari DeepSeek-R1 dalam menjawab soal ujian bidang onkologi radiasi dengan akurasi 89% berbanding 84%. Perbedaan hasil ini mengindikasikan bahwa keunggulan suatu model kecerdasan buatan sangat bergantung pada domain tugas yang diujikan. Dalam konteks

evaluasi soal Bahasa Indonesia yang menekankan aspek kebahasaan dan analisis teks, DeepSeek terbukti lebih unggul, sementara dalam konteks ujian medis yang menuntut akurasi faktual dan pengetahuan klinis, ChatGPT lebih unggul. Hal ini memperkuat argumentasi bahwa performa AI bersifat domain-spesifik dan tidak ada satu model pun yang mutlak unggul dalam segala hal.

Sementara itu, terkait dengan persepsi pengguna terhadap kecerdasan buatan, penelitian ini menemukan bahwa ChatGPT dinilai lebih komunikatif dan mudah digunakan karena bahasanya yang natural dan terasa "mirip guru". Temuan ini mengonfirmasi hasil penelitian Anjarsari dan Dewi (2023) serta Alatas, Romadhon, dan Rachmayanti (2024) yang menemukan bahwa mahasiswa lebih memilih ChatGPT karena dianggap lebih nyaman, mudah digunakan, dan mampu memberikan jawaban yang cepat serta jelas. Dengan demikian, keunggulan utama ChatGPT memang terletak pada aspek komunikatif-pedagogis, bukan pada kedalaman analisis teknis. Hal ini menjadi pembeda yang jelas antara ChatGPT yang lebih bersifat ramah dan komunikatif layaknya seorang guru, dengan DeepSeek yang lebih bersifat teliti dan mendalam layaknya seorang analis. Secara keseluruhan, perbandingan dengan penelitian terdahulu menunjukkan bahwa penelitian ini tidak hanya mengonfirmasi temuan-temuan sebelumnya terkait kemudahan dan popularitas ChatGPT, tetapi juga memberikan kontribusi baru dengan membuktikan bahwa DeepSeek lebih unggul dalam tugas analisis teknis evaluasi soal Bahasa Indonesia, sehingga pemilihan model AI yang tepat harus disesuaikan dengan tujuan dan

konteks penggunaannya dalam dunia pendidikan.

BAB IV

PENUTUP

A. Kesimpulan

Berdasarkan hasil penelitian mengenai analisis komparatif kualitas soal sumatif Bahasa Indonesia yang dihasilkan oleh *Artificial Intelligence* (AI), yaitu ChatGPT dan DeepSeek di SMAN 7 Kediri, dapat disimpulkan bahwa kedua platform AI tersebut memiliki kemampuan dalam membantu proses penyusunan soal sumatif Bahasa Indonesia. Pemberian prompt yang sama kepada kedua AI menunjukkan bahwa ChatGPT dan DeepSeek mampu menghasilkan soal sesuai dengan instruksi yang diberikan, baik dari aspek kesesuaian materi, penyusunan pertanyaan, penggunaan bahasa, maupun kelengkapan komponen soal.

Hasil validasi instrumen menggunakan uji Aiken's V menunjukkan bahwa instrumen penelitian memiliki tingkat validitas yang tinggi sehingga layak digunakan untuk mengukur kualitas soal yang dihasilkan oleh kedua AI. Hal tersebut menunjukkan bahwa indikator penilaian yang digunakan telah mampu menggambarkan aspek-aspek penting dalam membandingkan kinerja ChatGPT dan DeepSeek.

Berdasarkan hasil analisis deskriptif, terdapat perbedaan kecenderungan hasil penilaian siswa terhadap kedua AI. DeepSeek memperoleh nilai rata-rata lebih tinggi dibandingkan ChatGPT. Hal tersebut menunjukkan bahwa soal yang dihasilkan DeepSeek

memperoleh respons lebih positif dari siswa terutama pada aspek ketepatan, kelengkapan, dan kesesuaian dengan instruksi yang diberikan.

Hasil uji statistik menggunakan Mann-Whitney U Test menunjukkan bahwa tidak terdapat perbedaan yang signifikan secara statistik antara kinerja ChatGPT dan DeepSeek dalam menghasilkan soal sumatif Bahasa Indonesia. Hal ini menunjukkan bahwa meskipun DeepSeek memperoleh skor rata-rata lebih tinggi, perbedaan tersebut belum cukup menunjukkan bahwa salah satu AI memiliki keunggulan mutlak dibandingkan AI lainnya.

Selanjutnya, hasil uji hubungan kategori menggunakan Fisher's Exact Test menunjukkan adanya hubungan yang signifikan antara jenis AI dengan kategori penilaian siswa. Hal tersebut menunjukkan bahwa jenis *Artificial Intelligence* yang digunakan memiliki keterkaitan terhadap kecenderungan kualitas hasil soal yang dihasilkan. DeepSeek lebih banyak memperoleh kategori penilaian tinggi sehingga menunjukkan performa yang lebih baik berdasarkan persepsi siswa.

Dengan demikian, penelitian ini menunjukkan bahwa *Artificial Intelligence* dapat dimanfaatkan sebagai alat bantu dalam penyusunan soal sumatif Bahasa Indonesia. ChatGPT dan DeepSeek sama-sama memiliki potensi dalam mendukung proses evaluasi pembelajaran, tetapi penggunaannya tetap membutuhkan peran guru untuk melakukan validasi, penyempurnaan, dan penyesuaian terhadap kebutuhan pembelajaran.

B. Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat diberikan sebagai berikut.

1. Bagi Guru Bahasa Indonesia

Guru Bahasa Indonesia dapat memanfaatkan *Artificial Intelligence* seperti ChatGPT dan DeepSeek sebagai alat bantu dalam menyusun variasi soal sumatif. Namun, hasil keluaran AI tidak dapat langsung digunakan tanpa melalui proses pemeriksaan dan validasi. Guru tetap perlu melakukan analisis terhadap kesesuaian materi, tingkat kognitif, bahasa, serta karakteristik peserta didik agar soal yang digunakan benar-benar berkualitas.

2. Bagi Sekolah

Sekolah diharapkan dapat memberikan ruang bagi pemanfaatan teknologi *Artificial Intelligence* secara bijak dalam proses pembelajaran dan evaluasi. Penggunaan AI dapat menjadi inovasi untuk meningkatkan efektivitas kerja guru, khususnya dalam penyusunan perangkat pembelajaran, dengan tetap memperhatikan prinsip profesionalisme dan etika akademik.

3. Bagi Pengembang *Artificial Intelligence*

Pengembang AI diharapkan dapat terus meningkatkan kemampuan model dalam memahami konteks pendidikan, khususnya dalam menghasilkan instrumen evaluasi yang sesuai dengan standar pembelajaran. Pengembangan AI yang lebih adaptif terhadap kurikulum dan kebutuhan pendidikan akan meningkatkan

kualitas penggunaannya di lingkungan sekolah.

4. Bagi Peneliti Selanjutnya

Penelitian selanjutnya disarankan untuk memperluas objek penelitian dengan melibatkan lebih banyak sekolah, jenjang pendidikan, atau membandingkan lebih banyak platform *Artificial Intelligence* lainnya. Selain itu, penelitian berikutnya dapat mengkaji lebih mendalam aspek kualitas butir soal seperti tingkat kesukaran, daya pembeda, validitas empiris, serta efektivitas pengecoh agar memperoleh gambaran yang lebih komprehensif mengenai kemampuan AI dalam bidang evaluasi pembelajaran.

5. Bagi Pengguna Teknologi Pendidikan

Penggunaan *Artificial Intelligence* dalam pendidikan hendaknya dipahami sebagai teknologi pendukung, bukan sebagai pengganti peran pendidik. Keputusan akhir dalam penggunaan soal tetap harus berada pada guru atau ahli evaluasi karena kualitas pembelajaran tetap membutuhkan pertimbangan pedagogis dan pemahaman terhadap karakteristik peserta didik.