BAB I

PENDAHULUAN

A. Latar Belakang

Evaluasi hasil belajar merupakan komponen esensial dalam menentukan keberhasilan proses pembelajaran di sekolah. Dalam mata pelajaran Bahasa Indonesia, penilaian berfungsi untuk mengukur kemahiran siswa dan menentukan sejauh mana pemahaman siswa terhadap materi telah tercapai. Untuk memastikan keakuratan pengukuran, instrumen tes, khususnya soal-soal ujian sumatif, harus memiliki kualitas yang baik dan memenuhi kriteria psikometri yang valid.

Kualitas soal dievaluasi berdasarkan tiga indikator utama, yaitu tingkat kesukaran, daya pembeda (daya beda), dan fungsi pengecoh (distraktor). Analisis butir soal merupakan langkah wajib untuk menilai apakah soal telah memenuhi standar kualitas yang diharapkan. Penelitian terdahulu dengan pendekatan deskriptif kuantitatif menunjukkan bahwa analisis butir soal sangat perlu dilakukan, karena masih ditemukan sejumlah besar soal dengan kategori kualitas yang kurang baik, misalnya soal dengan tingkat kesukaran yang dominan mudah atau daya beda yang jelek. Selain itu, efektivitas evaluasi pembelajaran Bahasa Indonesia secara umum juga telah menjadi fokus penelitian kuantitatif, yang menekankan pentingnya instrumen penilaian yang efektif.

Pesatnya perkembangan teknologi, terutama di bidang Kecerdasan

Buatan (Artificial Intelligence/AI), telah membawa inovasi signifikan ke dalam dunia pendidikan. Mahasiswa dan pelajar sebagai digital native menunjukkan respons yang cepat terhadap integrasi AI dalam kegiatan akademik sehari-hari, karena AI terbukti mampu meningkatkan produktivitas dan efektivitas dalam belajar.

Dua model AI generatif yang paling banyak menarik perhatian dan adopsi adalah ChatGPT dan DeepSeek. Dalam konteks akademik, kedua alat ini berpotensi untuk digunakan dalam tugas-tugas evaluasi dan analisis konten, termasuk potensi untuk mempercepat proses analisis butir soal yang secara konvensional membutuhkan waktu dan upaya manual yang substansial. Kemampuan AI untuk memproses dan mengevaluasi data teks yang besar menjadikannya alternatif yang menarik untuk diuji efektivitas dan akurasinya dalam tugas evaluasi soal.

Penelitian dengan metode kuantitatif telah dilakukan untuk membandingkan penggunaan dan preferensi siswa terhadap AI seperti ChatGPT dan DeepSeek dalam menunjang pembelajaran mandiri. Hasil studi tersebut secara kuantitatif menunjukkan bahwa ChatGPT lebih dipilih dan dianggap lebih nyaman oleh sebagian besar responden dibandingkan DeepSeek, karena dianggap lebih mudah digunakan dan mampu memberikan jawaban yang cepat dan jelas. Temuan ini mengindikasikan adanya perbedaan kinerja dan pengalaman pengguna yang signifikan antara kedua model AI tersebut.

Namun, fokus perbandingan kinerja kedua AI tersebut dalam

penelitian terdahulu masih berada pada ranah efektivitas dan efisiensi belajar siswa secara umum. Belum terdapat penelitian kuantitatif yang secara spesifik menguji dan membandingkan kinerja objektif ChatGPT dan DeepSeek dalam konteks yang lebih sempit dan teknis, yaitu evaluasi kualitas butir soal mata pelajaran Bahasa Indonesia pada tingkat sekolah menengah. Padahal, akurasi AI dalam menganalisis kriteria soal (tingkat kesukaran, daya beda dan pengecoh / distraktor) sangat krusial untuk menentukan kualitas pendidikan.

Berdasarkan paparan di atas, terlihat adanya kebutuhan mendesak untuk mengintegrasikan teknologi terkini dalam upaya peningkatan kualitas soal, sekaligus adanya kekosongan data kuantitatif mengenai komparasi kinerja AI generatif dalam tugas evaluasi soal. Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan melakukan evaluasi soal sumatif menggunakan metode kuantitatif untuk menguji dan membandingkan secara empiris model AI mana (ChatGPT atau DeepSeek) yang menunjukkan kinerja lebih unggul dan konsisten dalam membuat butir soal sumatif Bahasa Indonesia untuk siswa SMAN 7 Kediri. Hasil penelitian ini diharapkan dapat memberikan landasan ilmiah yang kuat bagi guru dan pemangku kebijakan di SMAN 7 Kediri dalam mengadopsi dan memanfaatkan AI secara bijaksana sebagai alat bantu validasi dan perancangan instrumen evaluasi pembelajaran.

B. Rumusan Masalah

1. Bagaimana kualitas soal sumatif mata pelajaran Bahasa Indonesia di AI ChatGPT dan DeepSeek berdasarkan kriterian tingkat kesukaran,

daya pembeda dan pengecoh?

2. Apakah terdapat perbedaan signifikan dalam hasil analisis kualitas butir soal sumatif Bahasa Indonesia di SMAN 7 Kediri antara yang dilakukan oleh ChatGPT dan DeepSeek?

C. Tujuan Penelitian

1. Untuk mendeskripsikan kualitas butir soal sumatif mata pelajaran Bahasa Indonesia SMAN 7 Kediri di tinjau dari tingkat kesukaran, daya pembeda, dan pengecoh yang dianalisis menggunakan ChatGPT dan DeepSeek.
2. Untuk mengetahui dan menguji perbedaan signifikan hasil evaluasi kualitas butir soal sumatif Bahasa Indonesia SMAN 7 Kediri antara yang dihasilkan oleh ChatGPT dan DeepSeek.

D. Manfaat Penelitian

Penelitian ini diharapkan memberikan dua manfaat yaitu:

1. Manfaat Teoritis
    a. Kontribusi pada ilmu evaluasi pendidikan menambahkan khazanah keilmuan dalam bidang evaluasi pendidikan, khususnya mengenai metode analisis butir soal. Penelitian ini akan memberikan data empiris kuantitatif mengenai sejauh mana model-model Large Language Model (LLM) seperti ChatGPT dan DeepSeek dapat mereplikasi atau menggantikan proses analisis butir soal konvensional.
    b. Pengambilan kajian Artificial Intelligence (AI) dalam pendidikan menjadi referensi akademik yang spesifik mengenai

komparasi kinerja dua model AI generatif yang berbeda (CharGPT & DeepSeek) dalam tugas teknis pembuatan soal. Hasil penelitian ini dapat digunakan sebagai dasar bagi penelitian lanjutan tentang optimalisasi pemanfaatan AI untuk jaminan kualitas instrumen penilaian.

2. Manfaat Praktis
    a. Bagi guru mata pelajaran bahasa Indonesia. Memberikan informasi yang konkret mengenai efisiensi dan akurasi relatif antara AI ChatGPT dan AI DeepSeek. Guru dapat menentukan salah satu AI tersebut sebagai alat bantu efektif dan cepat untuk memvalidasi kualitas soal yang telah mereka buat sebelum diujikan. Membantu guru dalam melakukan revisi soal-soal sumatif secara lebih terstruktur dan berdasar data, sehingga instrumen evaluasi menjadi lebih valid dan reliabel.
    b. Bagi sekolah. Menyediakan rekomendasi berbasis data mengenai integrasi teknologi AI yang paling tepat (ChatGPT & DeepSeek) dalam sistem penjaminan mutu sekolah, khususnya pada proses evaluasi dan analisis soal. Meningkatkan efisien waktu dan sumber daya sekolah yang dialokasikan untuk kegiatan analisis butir soal, memungkinkan fokus yang lebih besar pada kegiatan peningkatan kualitas pembelajaran.
    c. Bagi peneliti selanjutnya. Hasil penelitian komperatif (perbandingan) ini dapat digunakan sebagai landasan awal (penelitian terdahulu) untuk mengembangkan model-model soal

otomatis berbasis AI yang lebih canggih dan spesifik, baik untuk mata pelajaran Bahasa Indonesia maupun mata pelajaran yang lainnya. Memberikan peta jalan tentang batasan dan keunggulan masing-masing AI dalam konteks analisis data kuantitatif pendidikan.

E. Batasan Penelitian

Untuk menghindari perluasan ruang lingkup penelitian dan agar penelitian dapat dilaksanakan secara terarah, sistematis, dan mendalam, maka penelitian ini dibatasi pada siswa kelas X 8 SMAN 7 Kediri. Adapun aspek-aspek batasan penelitian:

1. Penelitian ini hanya membandingkan kinerja dua platform kecerdasan buatan, yaitu ChatGPT dan DeepSeek, dalam melakukan analisis dan evaluasi terhadap soal sumatif Bahasa Indonesia kelas X 8 SMAN 7 Kediri.
2. Objek penelitian ini berupa soal sumatif Bahasa Indonesia yang telah digunakan sebagai instrumen penelitian hasil belajar peserta didik kelas X 8 SMAN 7 Kediri.
3. Evaluasi kualitas soal dilakukan berdasarkan kriteria penyusunan promt peneliti yang meliputi kesesuaian dengan capaian pembelajaran, tujuan pembelajaran, tingkat kognitif, konstruksi soal, dan pengguna bahasa.
4. Perbandingan kinerja ChatGPT dan DeepSeek dibatasi pada kualitas hasil evaluasi yang dihasilkan meliputi ketepatan identifikasi tingkat kesukaran, pengecoh dan daya pembeda.

5. Penelitian ini tidak membahas aspek teknis yang berkaitan dengan pengembangan sistem, dan algoritma.
6. Temuan penelitian dibatasi pada konteks evaluasi soal sumatif Bahasa Indonesia kelas X 8 SMAN 7 Kediri sehingga hasil penelitian tidak dimaksudkan untuk digeneralisasikan secara luas pada mata pelajaran, jenjang pendidikan, atau satuan pendidikan lainnya.

Dengan batasan tersebut, peneliti ini diharapkan mampu memberikan gambaran yang lebih fokus mengenai perbandingan kinerja ChatGPT dan DeepSeek dalam menghasilkan soal sumatif Bahasa Indonesia pada materi yang sudah di jelaskan guru Bahasa Indonesia kepada peserta didik kelas X-8 SMAN 7 Kediri.

F. Penelitian Terdahulu

1. Penelitian yang dilakukan oleh Andini Nisa Dewi dan Risza Tri Anjarsari yang berjudul Perkembangan dan Penggunaan AI Di kalangan Mahasiswa Untuk Menunjang Pembelajaran. Pesatnya perkembangan teknologi kecerdasan buatan (AI), terutama kemunculan AI generatif seperti ChatGPT, telah secara fundamental mengubah lanskap pendidikan tinggi. Artikel yang disusun oleh Risza Tri Anjarsari dan Andini Nisa Dewi ini membahas bagaimana mahasiswa tidak hanya menyaksikan perkembangan ini, tetapi juga secara aktif menggunakannya untuk menunjang proses pembelajaran mereka. Mahasiswa memandang AI sebagai alat bantu (tools) yang sangat efektif untuk meningkatkan efisiensi dan

produktivitas akademik. Dalam praktik sehari-hari, AI dimanfaatkan untuk berbagai fungsi, mulai dari mencari dan mengumpulkan referensi tambahan, meringkas materi teks yang panjang, hingga membantu menyusun kerangka awal bagi tugas-tugas penulisan akademik seperti esai atau makalah. Penggunaan ini didorong oleh kemampuan AI yang cepat dalam memproses informasi dan menghasilkan draft yang koheren. Meskipun membawa dampak positif yang signifikan seperti membantu mahasiswa memahami konsep sulit dan mempercepat start penulisan penggunaan AI ini juga menimbulkan kekhawatiran serius di lingkungan akademik. Risiko utama yang disoroti adalah potensi peningkatan plagiarisme serta ancaman terhadap kemampuan berpikir kritis mahasiswa jika mereka terlalu bergantung pada AI untuk menyelesaikan tugas tanpa usaha kognitif yang memadai. Menghadapi kenyataan ini, para peneliti menyimpulkan bahwa respons institusi akademik sebaiknya tidak berupa pelarangan total. Sebaliknya, dosen dan universitas didorong untuk mengintegrasikan AI secara etis ke dalam kurikulum. Tantangan utamanya adalah mengembangkan pedoman yang jelas agar mahasiswa menggunakan AI sebagai pendorong atau penguat (enhancement tool) kemampuan berpikir mereka, dan bukan sebagai substitusi mutlak terhadap proses berpikir kritis. Secara keseluruhan, AI kini telah menjadi realitas yang tak terhindarkan, menuntut adaptasi kebijakan akademik untuk menjamin penggunaan teknologi yang bertanggung jawab dan

suportif dalam pendidikan tinggi**.**

2. Penelitian yang dilakukan oleh Mega Agustiana, Hastari Mayrita, dan Andina Muchti yang berjudul Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Bahasa Indonesia Kelas XI. Bertujuan menganalisis secara mendalam kualitas butir soal Ujian Akhir Semester (UAS) mata pelajaran Bahasa Indonesia untuk kelas XI. Analisis ini, yang melibatkan penggunaan program statistik seperti SPSS dan Anates, mengungkapkan temuan beragam mengenai tingkat kesukaran, daya pembeda, validitas, dan reliabilitas soal. Secara umum, butir soal pilihan ganda memiliki kecenderungan tingkat kesukaran yang tinggi, di mana mayoritas soal diklasifikasikan sebagai sukar (25 butir) dan bahkan sangat sukar (12 butir). Hanya sebagian kecil yang masuk kategori mudah atau sangat mudah. Berdasarkan daya pembeda, mayoritas soal pilihan ganda dinilai memiliki daya pembeda yang cukup baik (32 butir), meskipun ada 16 butir yang diklasifikasikan kurang dan 2 butir yang memiliki daya pembeda negatif. Terkait dengan validitas, soal pilihan ganda yang digunakan sebagian besar telah memenuhi kriteria validitas (31 butir), namun terdapat 19 butir yang dianggap tidak valid. Di sisi lain, seluruh butir soal uraian (5 butir) menunjukkan hasil yang positif karena semuanya diklasifikasikan valid dan memiliki daya pembeda yang cukup hingga baik. Aspek terakhir, yaitu reliabilitas, menunjukkan hasil yang kontras. Butir soal pilihan ganda memiliki koefisien reliabilitas yang sangat tinggi

(0.898), yang menandakan konsistensi pengukuran yang baik. Namun, hasil yang berbeda ditunjukkan oleh soal uraian yang justru diklasifikasikan tidak reliabel dengan koefisien yang sangat rendah (0.093). Secara keseluruhan, penelitian ini menyimpulkan bahwa meskipun soal pilihan ganda memiliki reliabilitas yang tinggi, masih diperlukan perbaikan signifikan, terutama pada butir-butir yang tidak valid, terlalu sukar, atau memiliki daya pembeda yang negatif, demi memastikan alat ukur hasil belajar yang lebih berkualitas

3. Penelitian yang dilakukan oleh Imam Syafi'i dan Lukman Hakim yang berjudul Efektivitas Evaluasi Hasil Belajar Bahasa Indonesia Melalui Aplikasi Google From. Artikel ini membahas efektivitas penggunaan Google form sebagai media evaluasi hasil belajar dalam pembelajaran Bahasa Indonesia. Penulis menunjukkan bahwa Google Form memberikan banyak kemudahan bagi guru, seperti pembuatan soal yang cepat, penyebaran yang praktis, serta koreksi otomatis yang membatu efisiensi waktu. Melalui penelitian yang dilakukan, Google form terbukti mampu meningkatkan ketepatan penilaian, kecepatan memperoleh hasil, dan transparansi skor bagi siswa. Siswa juga merespons positif karena antarmuka Google form mudah digunakan, dapat diakses melalui berbagai perangkat, dan memberikan pengalaman evaluasi yang lebih modern. Penelitian ini menyimpulkan bahwa Google form efektif digunakan sebagai alat evaluasi pembelajaran Bahasa Indonesia karena meningkatkan efisiensi, akurasi penilaian, serta motivasi belajar siswa selama

proses evaluasi.

4. Penelitian yang dilakukan oleh Mochamad Arifin Alatas, Sahrul Romadhon, Irma Rachmayanti yang berjudul Penggunaan ChatGPT dalam Pembelajaran Bahasa: Perspektif Mahasiswa Tadris Bahasa Indonesia IAIN Madura. Artikel ini mengeksplorasi bagaimana mahasiswa program studi Tadris Bahasa Indonesia di IAIN Madura memandang penggunaan ChatGPT dalam pembelajaran bahasa. Penelitian ini menggunakan metode campuran Kuantitatif (angket) kualitatif (wawancara, observasi, dokumentasi). Hasil menunjukkan bahwa sebagian besar mahasiswa merasa terbantu oleh ChatGPT dalam mengerjakan tugas, mencari referensi, memahami materi, dan mengembangkan keterampilan menulis. Namun, mereka juga menyadari tantangan seperti keakuratan informasi. ChatGPT sebagai aplikasi kecerdasan buatan (AI) menawarkan berbagai manfaat dalam proses pembelajaran, termasuk membantu untuk menyelesaikan tugas, memahami materi dan meningkatkan keterampilan bahasa. Hasil penelitian ini menunjukkan bahwa ChatGPT memiliki dampak positif signifikan, membantu mahasiswa dalam mengerjakan tugas dan mencari referensi materi. Mayoritas mahasiswa merasa terbantu dan menilai positif penggunaan ChatGPT, meskipun ada tantangan keakuratan informasi. Sebagian besar responden merasa bahwa ChatGPT cukup atau sangat mempengaruhi hasil belajar dan membantu memahami materi, serta efektif dalam pengembangan keterampilan menulis.

5. Penelitian yang dilakukan oleh Eric Jonthan, Anak Agung Ugrasena, Axel Theo Winata ursia yang berjudul Comparative Analysis Based On Taks Oriented Evaluation On ChatGPT-4 And Gemini (2,5 Flash). Penelitian ini menyajikan analisis perbandingan berbasis tugas antara dua model bahasa besar (LLM) terkemuka. ChatGPT - 4 dari OpenAI dan Gemini 2.5 Flash dari Google. Tujuannya adalah mengevaluasi kinerja kedua AI ini dalam menyelesaikan serangkaian tugas praktis dan terarah (task-oriented) yang mereplikasi skenario penggunaan dunia nyata. Evaluasi difokuskan pada berbagai kategori tugas, yang sering digunakan oleh pengguna, termasuk kemampuan penalaran (reasoning), pembuatan kode (coding), dan pemahaman multimodal (walaupun Gemini 2.5 Flash adalah model yang ringan, kinerja dasarnya tetap diuji). Secara umum, temuan penelitian menyoroti adanya kekuatan yang saling melengkapi antara kedua model:
    a. ChatGPT-4 cenderung menunjukkan keunggulan yang lebih konsisten dalam tugas-tugas yang membutuhkan kedalaman konteks, penalaran kompleks, dan menghasilkan respons dengan nuansa kreativitas yang lebih baik. Reputasi ChatGPT dalam memberikan hasil yang hampir menyerupai bahasa manusia tetap menjadi keunggulannya.
    b. Gemini 2.5 Flash, sebagai model yang dioptimalkan untuk kecepatan dan efisiensi, menonjol dalam tugas-tugas yang memerlukan akses informasi terkini (berkat integrasinya dengan

Google Search) dan pemrosesan yang cepat. Meskipun Gemini 2.5 Flash adalah versi "ringan," model ini tetap menunjukkan kapabilitas yang baik, terutama dalam pemecahan masalah yang memerlukan integrasi data atau dukungan ekosistem Google Workspace.

Para peneliti menyimpulkan bahwa tidak ada pemenang mutlak di antara keduanya. Pilihan model yang lebih unggul sangat bergantung pada kebutuhan spesifik pengguna dan jenis tugas yang dihadapi. Untuk tugas yang menuntut keakuratan tinggi, kreativitas, dan penjelasan teknis yang mendalam, ChatGPT-4 sering menjadi pilihan utama. Untuk tugas yang membutuhkan kecepatan, efisiensi biaya, integrasi dengan layanan Google, dan informasi real-time, Gemini 2.5 Flash menawarkan solusi yang sangat kompeten. Kesimpulannya, persaingan antara ChatGPT-4 dan Gemini 2.5 Flash memberikan manfaat bagi pengguna dengan menawarkan model-model yang semakin terspesialisasi satu unggul dalam kedalaman, satu unggul dalam kecepatan dan integrasi ekosistem.

6. Penelitian yang dilakukan oleh Luluk Puji Rahayu, dan Desi Sukenti Rangkuman artikel yang berjudul Kualitas Soal Bahasa Indonesia Kelas XI SMAN 2 Bangko Pusako: Analisis Butir Soal. Penelitian ini bertujuan mengetahui kualitas butir soal ujian semester ganjil mata pelajaran Bahasa Indonesia kelas XI SMAN 2 Bangko Pusako (tahun ajaran 2023/2024) ditinjau dari tingkat kesukaran, daya beda, kualitas pengecoh. Jenis penelitian ini deskriptif kuantitatif. Data

berupa naskah soal, kunci jawaban, dan hasil jawaban siswa yang dianalisis menggunakan anates versi 4. Hasil utama mayoritas soal tergolong mudah, distribusi kualitas daya beda berkisar dari jelek sampai baik dan pengecoh sebagian besar efektif. Berdasarkan tingkat kesukaran soal terdapat 17 (57%) soal kategori mudah. Berdasarkan daya beda terdapat 6 (20%) soal kategori jelek, 14 (47%) soal berkategori sedang, dan 10 (33%) soal kategori baik. Berdasarkan kualitas pengecoh terdapat 67% efektif dan 33% tidak efektif.

7. Penelitian yang dilakukan oleh Ronald Chow, Ajay Zheng, Chenxi Gao, Milo Vermeulen, Francis Yu, Irini Yacoub, Arpit M.Chabra, J.Isabelle Choi, Haibo Lin, Gilmer Valdes, dan Charles B Simone yang berjudul ChatGPT vs. DeepSeek: Assessing AI Performance On Radiation Oncology exam Questions (Evaluasi Kinerja AI pada Soal Ujian Onkologi Radiasi). Penelitian ini merupakan studi pertama yang mengevaluasi akurasi model bahasa besar (LLM) baru bernama DeepSeek-R1 dalam menjawab pertanyaan medis, khususnya dalam bidang Onkologi Radiasi, dan membandingkannya dengan berbagai model ChatGPT (termasuk model terbaru ol). Para peneliti menguji DeepSeek R-1 dan beberapa model ChatGpt dengan 600 soal ujian pilihan ganda dari ujian in-service Onkologi Radiasi nasional. Soal-soal ini mencakup pengetahuan tentang anatomi, perencanaan perawatan, epidemiologi kanker, dan uji klinis penting (Landmark trials). Hasil utama

menunjukkan bahwa:

a. ChatGPT ol adalah yang paling akurat, menjawab 89.0% pertanyaan dengan benar. Akurasinya tidak berbeda signifikan di seluruh kategori pertanyaan, bahkan mencapai tingkat tertinggi (93.5%) pada pertanyaan tentang studi penting (landmark studies)
b. DeepSeek-R1 menjawab 84.0% pertanyaan dengan benar, menjadikannya secara signifikan kurang akurat dibandingkan ChatGPT ol (p=0.011). menariknya, DeepSeek-R1 paling tidak akurat pada kategori pertanyaan tentang studi penting (hanya 74.2%)

Meskipun kalah dari model terbaru, DeepSeek-R1 secara signifikan lebih akurat dari pada model ChatGPT sebelumnya, seperti ChatGPT 4o (72.2%) dan ChatGPT 3.5 (53.8%). Aspek penting lain yang dianalisis adalah efisiensi waktu dan biaya penggunaan model. Waktu: DeepSeek-R1 membutuhkan waktu rata-rata 59 detik per pertanyaan, sementara ChatGPT ol jauh lebih cepat, hanya membutuhkan 10 detik per pertanyaan. Total waktu proses DeepSeek-R1 (9 jam 47 menit) hampir enam kali lebih lama daripada ChatGPT ol (1 jam 38 menit). Para peneliti menyimpulkan bahwa DeepSeek-R1 adalah model yang kurang akurat dan lebih lambat dalam merespons dibandingkan dengan ChatGPT ol. Namun, keunggulannya terletak pada biayanya yang jauh lebih rendah. Kondisi ini menciptakan skenario di mana tidak ada model yang benar-benar dominan (lebih murah

sekaligus lebih efektif). Oleh karena itu, diperlukan analisis dan pertimbangan yang cermat mengenai kinerja dan biaya saat memutuskan implementasi salah satu model untuk memastikan keselarasan optimal dengan tujuan peningkatan akurasi dan efisiensi yang dituju.

G. Definisi Operasional

Variabel Penelitian	Devinisi Operasional	Indikator Pengukuran dan Sub Indikator
Soal sumatif bahasa Indonesia di SMAN 7 Kediri	Soal pilihan ganda yang digunakan sebagai instrumen penilaian untuk mengukur pencapaian kompetensi beberapa siswa di SMAN 7 Kediri pada mata pelajaran Bahasa Indonesia	Beberapa soal ujian sumatif Bahasa Indonesia (termasuk kunci jawaban AI dan jawaban siswa) yang digunakan sebagai data masukan (input) dalam analisis AI
Kualitas butir soal	Tingkat kelayakan suatu butir soal berdasarkan kriteria psikometri dasar yang diukur secara kuantitatif. Kualitas ini merupakan hasil keluaran (output) yang dilakukan oleh ChatGPT dan DeepSeek	Tingkat Kesukaran (TK) mengkategorikan soal menjadi Mudah, Sedang, Sukar. Daya Pembeda (DP) mengkategorikan soal menjadi cukup, baik dan sangat baik. Efektivitas Pengecoh (EP) mengkategorikan pengecoh (distraktor) menjadi efektif dan tidak efektif.
ChatGPT (X1)	Model kecerdasan buatan yang digunakan untuk menghasilkan soal sumatif Bahasa Indonesia berdasarkan prompt yang sama dengan DeepSeek.	Butir soal yang dihasilkan ChatGPT. Meliputi jumlah soal, struktur soal, variasi level kognitif, kelengkapan komponen soal.
DeepSeek (X2)	Model kecerdasan buatan alternatif yang digunakan untuk menghasilkan soal sumatif Bahasa Indonesia. Dengan perintah yang sama dengan ChatGPT.	Butir soal yang dihasilkan DeepSeek. Meliputi jumlah soal, struktur soal, variasi level kognitif, kelengkapan komponen soal.
Validitas Isi (Y1)	Kesesuaian hasil soal AI dengan kompetensi, indikator, dan materi pembelajaran Bahasa Indonesia di SMAN 7 Kediri	Kesesuaian materi juga indikator, dan kecakupan materi Bahasa Indonesia di SMAN 7 Kediri. Kecocokan dengan CP/KD, kesesuaian dengan topik, keakuratan konsep, kesesuaian soal dengan tujuan pembelajaran, tidak keluar dari materi cakupan sesuai level.
Validitas konstruksi (Y2)	Ketepatan struktur, bentuk, dan susunan soal yang dihasilkan AI	Pokok soal (stem) kejelasan stem, tidak ambigu, tidak

	sehingga sesuai deangan prompt yang di unggah pada kedua AI (ChatGPT dan DeepSeek)	menimbulkan penafsiran ganda. Pilihan jawaban hanya satu kunci benar panjang opsi seimbang. Kaidah teknis, format sesuai standar, bebas kesalahan logika.
Validitas bahasa (Y3)	Ketepatan penggunaan bahasa Indonesia dalam soal yang dihasilkan AI, baik dari segi struktur, kelancaran, maupun kebakuan	Kebakuan bahasa menggunakan EYD/PUEB pilihan kata tepat. Kejelasan kalimat, kalimat tidak berbelit, tidak ambigu. Keterbacaan mudah dipahami, tidak ada kesalahan ejaan.
Tingkat kesukaran (Y4)	Seberapa mudah atau sulit soal yang dijawab berdasarkan jumlah siswa yang menjawab benar saat uji coba	Proporsi siswa menjawab benar, soal mudah ($p > 0,70$), sedang ($0,30$-$0,70$), sulit ($p < 0,30$)
Daya pembeda (Y5)	Kemampuan soal membedakan siswa berkemampuan tinggi dan rendah dari hasil uji coba.	Indeks diskriminasi baik ($D > 0,40$), cukup ($0,20 – 0,39$), buruk ($D < 0,20$).
Efektivitas Pengecoh (Y6)	Seberapa efektif opsi salah (pengecoh) menarik siswa yang kurang memahami materi.	Persentase pengecoh dipilih siswa. Efektif jika dipilih ≥5% siswa, tidak efektif jika tidak dipilih
Level Kognitif Bloom (Y7)	Kategori kemampuan berpikir yang diukur soal AI berdasarkan taksonomi Bloom revisi (Anderson & Krathwohl,2001)	C1 – Mengingat (Fakta, definisi, informasi dasar), C2 – Memahami (Menjelaskan, meringkas, mereformulasi), C3 – Menerapkan (Menggunakan konsep ke situasi baru), C4 – Menganalisis (Membedakan, mengidentifikasi ide utama, menemukan hubungan)
Kinerja AI dalam Pembuatan Soal (Z1)	1. ChatGPT (OpenAI): Dioperasionalkan sebagai model AI berbasis arsitektur Transformer yang dikembangkan oleh OpenAI. Dalam variabel ini, kinerja ChatGPT dinilai dari kemampuannya memproses prompt (instruksi) butir soal Bahasa Indonesia dengan fokus pada aspek kelogisan bahasa, ketepatan identifikasi level kognitif sesuai taksonomi Bloom, serta akurasi perhitungan indeks tingkat kesukaran dan daya pembeda berdasarkan data empiris yang diberikan. DeepSeek (DeepSeek-AI):	Keakuratan (Sesuai materi & indikator), Keterbacaan (Bahasa jelas, tidak ambigu), Struktur soal (Kaidah soal dipenuhi), dan Kualitas butir (Indeks kesukaran, pembeda, pengecoh). ChatGPT dan DeepSeek dalam penelitian ini berperan sebagai representasi teknologi AI yang memiliki basis pengetahuan luas dan kemampuan naratif yang terstruktur. Keduanyan berperan sebagai pembanding kritis untuk melihat apakah perbedaan arsitektur model berpengaruh signifikan terhadap kualitas evaluasi

	Dioperasionalkan sebagai model yang dikembangkan oleh DeepSeek-AI, Dalam variabel ini, kinerja DeepSeek dinilai dari kemampuannya memproses prompt (instruksi) butir soal Bahasa Indonesia dengan fokus pada aspek kelogisan bahasa, ketepatan identifikasi level kognitif sesuai taksonomi Bloom, serta akurasi perhitungan indeks tingkat kesukaran dan daya pembeda berdasarkan data empiris yang diberikan.	soal sumatif di jenjang SMA. Kedua model AI tersebut akan diberikan prompt yang identik untuk dianalisis kinerjanya berdasarkan empat indikator utama, yaitu: keakuratan materi, keterbacaan bahasa, Kesesuaian struktur konstruksi soal, dan ketepatan statistik kualitas butir (indeks kesukaran, daya pembeda, dan pengecoh).
Perbandingan Kinerja ChatGPT & DeepSeek (Z2)	Perbedaan hasil kualitas soal kedua model berdasarkan validitas, kualitas butir, dan level kognitif.	Selisih nilai validitas (Selisih skor isi, konstruksi, bahasa), Selisih kualitas butir (Perbandingan tingkat kesukaran, daya pembeda, pengecoh), dan Dominasi level kognitif (Level Bloom yang paling banyak muncul).

Tabel 1.1 Devinisi Oprasional 1

Kinerja dalam penelitian ini adalah perbandingan hasil butir soal yang di hasilkan oleh ChatGPT dan DeepSeek dengan indikator validitas dan kualitas soal.