



Akurasi *Artificial Intelligence* Berbasis NLP dalam Mendeteksi Kesalahan Sintaksis Bahasa Indonesia pada Makalah Akademik

Bustomi¹

¹Fakultas Tarbiyah dan Keguruan, UIN Sunan Gunung Djati Bandung
bustomi@uinsgd.ac.id

Moh. Azhar²

²Jawaharlal Nehru University, India
azharjnu1@gmail.com

Ahsan Arrizqi³

³Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati Bandung
1257050047@student.uinsgd.ac.id

Akhmad Darul Qutni⁴

⁴Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati Bandung
1257050035@student.uinsgd.ac.id

Azmi Fadilah⁵

⁵Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati Bandung
1257050027@student.uinsgd.ac.id

Dheriel Elbazariel⁶

⁶Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati Bandung
1257050098@student.uinsgd.ac.id

Abstrak: Kesalahan sintaksis dalam makalah akademik mahasiswa merupakan permasalahan yang berulang dan berdampak signifikan terhadap kualitas karya tulis ilmiah. Kehadiran teknologi *Artificial Intelligence* (AI) berbasis *Natural Language Processing* (NLP) membuka peluang untuk mendeteksi dan mengoreksi kesalahan tersebut secara otomatis. Penelitian ini bertujuan untuk menganalisis secara sistematis tingkat akurasi sistem AI berbasis NLP dalam mendeteksi kesalahan sintaksis pada makalah akademik berbahasa Indonesia berdasarkan kajian literatur yang komprehensif. Metode yang digunakan dalam penelitian ini ialah analisis isi dan evaluasi komparatif terhadap metrik kinerja berbagai model AI. Hasil penelitian menunjukkan bahwa profil kesalahan sintaksis mahasiswa didominasi oleh tataran kalimat (67,02%), khususnya kalimat tanpa subjek dan predikat serta ketidaklogisan struktur. Model Large Language Model (LLM), seperti GPT-4, mencapai F1-Score sebesar 89,79% dalam deteksi koreksi, sementara model Grammatical Error Correction (GEC) khusus, yaitu SAN-GEC dengan fine-tuning, meraih F1-Score sebesar 71,94% dan BLEU-Score sebesar 78,13%. Secara diagnostik, AI paling andal dalam mendeteksi kesalahan morfologi, dengan tingkat

akurasi sebesar 84,4%, tetapi menunjukkan kinerja paling rendah dalam penalaran numerik, yaitu 45,0%. Penelitian ini menyimpulkan bahwa AI berbasis NLP berpotensi besar sebagai asisten penyuntingan akademik, tetapi masih memerlukan pengembangan lebih lanjut, khususnya untuk menangani kompleksitas semantik dan penalaran kontekstual dalam bahasa Indonesia.

Kata Kunci: akurasi deteksi, bahasa Indonesia, kesalahan sintaksis, makalah akademik, *Natural Language Processing*

Abstract: Syntactic errors in students' academic papers are a recurring issue that significantly impacts the quality of scholarly writing. The emergence of Artificial Intelligence (AI) technology based on Natural Language Processing (NLP) offers the potential to automatically detect and correct such errors. This study aims to systematically analyze the accuracy of NLP-based AI systems in detecting syntactic errors in Indonesian-language academic papers, drawing upon a comprehensive literature review. The methodology employed includes content analysis and a comparative evaluation of the performance metrics of various AI models. The findings indicate that the profile of students' syntactic errors is dominated by sentence-level issues (67.02%), specifically sentences lacking subjects or predicates, as well as illogical structural constructions. Large Language Models (LLMs), such as GPT-4, achieved an F1 Score of 89.79% for error detection and correction. On the contrary, a specialized Grammatical Error Correction (GEC) model – SAN-GEC – after fine-tuning attained an F1-Score of 71.94% and a BLEU-Score of 78.13%. Diagnostically, AI proved most reliable in detecting morphological errors, with an accuracy of 84.4%, yet demonstrated the lowest performance in numerical reasoning at 45.0%. The study concludes that while NLP-based AI holds significant potential as an academic editing assistant, further development is required – particularly to address the complexities of semantics and contextual reasoning in Indonesian.

Keywords: detection accuracy, bahasa Indonesia, syntactic errors, academic papers, *Natural Language Processing*

PENDAHULUAN

Bahasa dalam konteks akademik berfungsi sebagai sarana utama untuk menyampaikan gagasan ilmiah secara logis dan sistematis. Salah satu aspek fundamental yang menentukan kualitas sebuah tulisan ilmiah adalah penguasaan sintaksis, yakni kaidah yang mengatur penyusunan kata menjadi frasa, klausa, dan kalimat yang gramatikal, karena tata bahasa pada dasarnya merupakan kumpulan aturan yang memformalkan hubungan antarkata dalam suatu bahasa (Casal & Lee, 2019; Musyafa et al., 2022). Seiring berkembangnya teknologi komputasi bahasa, *Artificial Intelligence* (AI) berbasis *Natural Language Processing* (NLP) kini hadir sebagai metode komputasional yang memungkinkan mesin memahami dan

mengoreksi struktur bahasa manusia secara otomatis. Perkembangan ini tercermin dari hadirnya berbagai sumber daya NLP untuk bahasa Indonesia, seperti *benchmark* IndoNLG untuk tugas pembangkitan bahasa (Cahyawijaya et al., 2021) dan model pralatih IndoBERT yang dikembangkan melalui *benchmark* IndoLEM untuk berbagai tugas pemahaman bahasa Indonesia, termasuk tataran morfosintaksis (Koto et al., 2020).

Secara ideal, setiap makalah akademik yang dihasilkan mahasiswa seharusnya memenuhi standar kalimat efektif dan kaidah sintaksis Bahasa Indonesia yang baku, mengingat penggunaan kalimat efektif merupakan salah satu indikator utama keberhasilan sebuah artikel ilmiah dalam menyampaikan gagasan secara jelas, padat, dan logis (Diani et al., 2025; Nababan et al., 2025). Penguasaan gramatikal mahasiswa secara konseptual juga seharusnya meningkat setelah menempuh mata kuliah kebahasaan seperti sintaksis, morfologi, dan Keterampilan Menulis Akademik (Arifin, 2008). Namun, kenyataan di lapangan menunjukkan adanya kesenjangan yang nyata. Kesalahan sintaksis paling umum justru terjadi pada aspek kesatuan gagasan, koherensi, dan penekanan kalimat (Nababan et al., 2025), sementara lemahnya pemahaman aturan sintaksis dan kurangnya pembiasaan berbahasa baku turut menyebabkan kesalahan yang berulang dalam tulisan ilmiah mahasiswa (Arifin, 2008). Permasalahan ini diperparah oleh keterbatasan alat *spell-checker* konvensional, yang umumnya hanya mendeteksi kesalahan pada level ejaan kata dan belum mampu menjangkau kesalahan struktural pada level kalimat. Sebuah tantangan juga diakui dalam riset *Grammatical Error Correction* (GEC) yang mengindikasikan bahwa bahasa Indonesia masih tergolong sebagai *low-resource language* dengan keterbatasan data latih dibandingkan dengan bahasa-bahasa besar seperti Inggris dan Mandarin (Musyafa et al., 2022).

Kondisi tersebut menegaskan urgensi untuk menguji sejauh mana teknologi AI berbasis NLP dapat memberikan solusi yang lebih akurat dan andal dalam mendeteksi kesalahan sintaksis pada teks akademik berbahasa Indonesia. Pengujian akurasi menjadi krusial karena performa model AI sangat bergantung pada ketersediaan data latih dan karakteristik bahasa yang diproses, sebagaimana

ditunjukkan oleh upaya pengembangan model GEC berbasis Transformer yang melaporkan hasil *F1-score* dan *BLEU-score* yang kompetitif untuk bahasa Indonesia, namun tetap menyatakan bahwa kesalahan semantik dan sintaksis yang kompleks masih menjadi tantangan tersendiri bagi model tersebut.

Penelitian-penelitian terdahulu telah mengkaji isu ini dari berbagai sudut pandang. Dari ranah linguistik deskriptif, penelitian tentang kesalahan sintaksis pada artikel ilmiah mahasiswa di 12 perguruan tinggi (Nababan et al., 2025) serta kesalahan gramatikal pada teks akademik mahasiswa Universitas Nias (Arifin & Yusnita, 2022) memetakan jenis dan pola kesalahan secara manual melalui analisis isi kualitatif. Dari ranah pengembangan teknologi, penelitian seperti IndoNLG (Mahendra et al., 2021) dan IndoLEM/IndoBERT (Koto et al., 2020) membangun *benchmark* dan model bahasa untuk bahasa Indonesia secara umum, sementara penelitian khusus pada tugas GEC mengusulkan model berbasis Transformer dengan mekanisme *copy* (Musyafa et al., 2022), serta kerangka kerja untuk konstruksi korpus evaluasi GEC yang juga menguji potensi LLM seperti GPT-3.5-Turbo dan GPT-4 (N. Lin et al., 2025). Meskipun demikian, penelitian-penelitian tersebut umumnya berfokus pada pengembangan model atau pemetaan kesalahan secara deskriptif, dan belum secara spesifik mengukur tingkat akurasi sistem AI/NLP dalam mendeteksi kesalahan sintaksis dalam konteks makalah akademik bahasa Indonesia.

Berdasarkan kesenjangan tersebut, kebaruan penelitian ini terletak pada fokusnya yang spesifik, yaitu mengkaji dan menganalisis secara sistematis tingkat akurasi AI berbasis NLP dalam mengidentifikasi kesalahan sintaksis pada makalah akademik bahasa Indonesia berdasarkan temuan dari berbagai literatur terdahulu. Hal ini berbeda dengan penelitian terdahulu yang lebih berfokus pada pemetaan jenis kesalahan secara manual atau pada pengembangan model GEC tanpa kajian akurasi pada ranah sintaksis makalah akademik secara khusus.

METODE PENELITIAN

Penelitian ini menggunakan metode studi literatur dengan pendekatan deskriptif kualitatif untuk mengkaji kemampuan teknologi kecerdasan buatan

dalam mendeteksi kesalahan sintaksis pada karya tulis akademik mahasiswa. Data penelitian bersumber dari literatur ilmiah yang relevan, tugas mahasiswa Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati Bandung, serta dokumentasi tolok ukur pemrosesan bahasa alami bahasa Indonesia, seperti IndoLEM dan IndoNLG. Data yang dianalisis meliputi karakteristik kesalahan sintaksis mahasiswa yang telah diidentifikasi secara manual, hasil evaluasi model bahasa, serta temuan penelitian terdahulu mengenai kinerja berbagai teknologi AI, termasuk large language models (LLM), seperti GPT-3.5 dan GPT-4, serta model berbasis arsitektur Transformer, seperti IndoBERT dan SAN-GEC. Analisis data dilakukan melalui analisis isi dan perbandingan temuan penelitian untuk mengidentifikasi kemampuan, keterbatasan, serta konsistensi model AI dalam mendeteksi berbagai jenis kesalahan sintaksis, seperti ketidaklengkapan unsur subjek dan predikat, penggunaan konjungsi yang tidak tepat, serta kesalahan struktur frasa. Dengan mengikuti pendapat Lee & Bang (2025), apabila tersedia, metrik evaluasi kuantitatif berupa *precision*, *recall*, dan *F1-score* digunakan sebagai informasi pendukung untuk membandingkan kinerja model berdasarkan hasil evaluasi yang dilaporkan dalam sumber rujukan. Selanjutnya, temuan tersebut diinterpretasikan secara deskriptif dengan mempertimbangkan kompleksitas struktur kalimat dalam bahasa Indonesia serta karakteristik bahasa akademik. Melalui prosedur ini, penelitian diharapkan dapat memberikan gambaran yang sistematis mengenai potensi dan keterbatasan teknologi AI dalam mendukung identifikasi kesalahan sintaksis pada karya tulis akademik mahasiswa.

HASIL DAN PEMBAHASAN

Natural Language Processing (NLP) adalah cabang kecerdasan buatan yang memungkinkan komputer memahami, menginterpretasikan, dan menghasilkan teks dalam bahasa manusia (Chowdhary, 2020; Fanni et al., 2023). Dalam konteks deteksi kesalahan tata bahasa, NLP bekerja melalui serangkaian tahapan pemrosesan yang saling bergantung: dimulai dari tokenisasi (pemecahan teks menjadi unit kata atau subkata), dilanjutkan dengan analisis morfologis

(identifikasi bentuk dan imbuhan kata), analisis sintaksis (penguraian struktur kalimat menjadi pohon sintaksis), hingga analisis semantik (penafsiran makna). Setiap lapisan analisis ini berperan dalam mengidentifikasi kelas kesalahan yang berbeda, sehingga kapabilitas sebuah sistem sangat ditentukan oleh seberapa dalam ia mampu memproses teks hingga lapisan tersebut (Qasim et al., 2022).

Dalam tugas *Grammatical Error Correction* (GEC), terdapat dua paradigma utama yang mendominasi penelitian saat ini. Paradigma pertama adalah pendekatan berbasis aturan yang mengaktifkan sistem untuk mencocokkan teks masukan dengan seperangkat aturan gramatikal yang telah dikodekan secara eksplisit oleh pakar bahasa (Bryant et al., 2023). Paradigma ini memiliki presisi tinggi untuk jenis kesalahan yang terpola, namun tidak mampu menangani variasi konteks yang kompleks. Paradigma kedua adalah pendekatan berbasis model bahasa (*language model-based*), yang mencakup model Transformer seperti BERT (Qasim et al., 2022) dan GPT (Mao et al., 2025). Model-model ini tidak mengikuti aturan eksplisit, melainkan mempelajari pola distribusi bahasa dari miliaran data teks sehingga mampu menangkap nuansa kontekstual yang jauh lebih kaya. Khusus untuk arsitektur Transformer, mekanisme *self-attention* memungkinkan setiap token dalam kalimat untuk secara langsung memperhatikan dan mengintegrasikan informasi dari token lain di seluruh rangkaian kalimat, sehingga menjadikannya sangat efektif untuk tugas-tugas yang memerlukan pemahaman konteks jarak jauh (*long-range dependency*).

Dalam ekosistem bahasa Indonesia, tantangan bagi sistem NLP menjadi berlipat ganda karena sifat aglutinatif bahasa Indonesia yang kaya akan afiksasi. Sebuah kata dasar seperti 'tulis' dapat berubah menjadi puluhan bentuk derivatif seperti 'menulis', 'ditulis', 'tertulis', 'penulisan', 'kepenulisan', dan sebagainya, masing-masing dengan makna gramatikal yang berbeda. Kondisi ini menciptakan ruang morfologis yang sangat besar yang harus dipelajari oleh model AI, dan menjadikan bahasa Indonesia sebagai salah satu *low-resource language* yang memerlukan investasi khusus dalam pembangunan data dan model (Cahyawijaya et al., 2021; Koto et al., 2020). Pemahaman atas konteks teknis dan linguistik inilah yang menjadi prasyarat untuk menginterpretasikan temuan-temuan berikut secara

tepat.

Data hasil penelitian ini dibagi menjadi tiga kategori utama untuk memberikan gambaran yang komprehensif: (1) Profil kesalahan sintaksis pada makalah akademik, (2) Performa model AI (LLM dan Transformer) dalam mendeteksi dan mengoreksi kesalahan, serta (3) Akurasi AI berdasarkan fenomena linguistik tertentu.

Kesalahan Sintaksis dalam Makalah Mahasiswa

Sebelum mengevaluasi kapabilitas AI, penting untuk terlebih dahulu memetakan jenis kesalahan sintaksis yang paling sering muncul pada makalah mahasiswa yang terdeteksi oleh sistem AI, yakni GPT-4. Berdasarkan hasil evaluasi terhadap karya ilmiah yang ditulis oleh mahasiswa jurusan Teknik Informatika di UIN Sunan Gunung Djati Bandung pada Juni 2026, profil kesalahan sintaksis mahasiswa dapat dilihat pada Tabel 1 berikut.

Tabel 1. Profil Kesalahan Sintaksis pada Makalah Akademik Mahasiswa

Kategori Kesalahan	Persentase/Frekuensi	Rincian Kesalahan
Tataran kalimat	67,02%	Kalimat tidak berpredikat, kalimat tanpa subjek, kalimat buntung (tanpa S-P), kalimat tidak logis, serta redundansi konjungsi
Tataran frasa	32,98%	Penggunaan preposisi tidak tepat, susunan kata tidak benar, serta penggunaan bentuk jamak ganda.

Data pada Tabel 1 menunjukkan bahwa kesalahan sintaksis pada tataran kalimat mendominasi dengan persentase 67,02%, mencakup kalimat tidak berpredikat, kalimat buntung tanpa unsur subjek-predikat (S-P), ketidaklogisan argumen, serta redundansi konjungsi. Sementara itu, kesalahan pada tataran frasa mencapai 32,98% yang meliputi penggunaan preposisi yang tidak tepat, susunan kata yang salah, serta pembentukan jamak ganda yang tidak baku. Hal ini sejalan dengan temuan (Nababan et al., 2025) yang menegaskan bahwa aspek kesatuan gagasan dan koherensi antarkalimat merupakan titik lemah terbesar dalam karya tulis ilmiah mahasiswa Indonesia. Kondisi ini mencerminkan bahwa mahasiswa belum sepenuhnya menginternalisasi kaidah kalimat efektif sebagai standar

penulisan akademik yang formal.

Arifin & Yusnita (2022) mengemukakan bahwa lemahnya pemahaman aturan sintaksis sering bersumber dari minimnya pembiasaan berbahasa Indonesia baku dalam kehidupan sehari-hari mahasiswa, yang berdampak pada kesulitan menerapkan kaidah gramatikal secara konsisten dalam teks akademik. Profil kesalahan ini menjadi acuan diagnostik yang penting untuk menilai sejauh mana sistem AI mampu mengidentifikasi dan mengoreksi setiap kategori kesalahan tersebut.

Kesalahan dalam kalimat yang dibuat oleh mahasiswa sering disebabkan oleh kebiasaan mereka menggunakan bahasa yang tidak resmi dalam komunikasi sehari-hari. Anugrah et al. (2026) dan Almira Apsarini Ramadhani et al. (2026) menjelaskan bahwa kebiasaan tersebut dapat memengaruhi kemampuan mahasiswa dalam menerapkan kaidah bahasa Indonesia yang baku, terutama saat mereka menulis teks karya ilmiah. Dalam komunikasi informal, penggunaan kata-kata tidak baku, struktur kalimat yang kurang tepat, penghilangan unsur kalimat, serta penggunaan ungkapan yang tidak sesuai dengan konteks formal cenderung dianggap wajar. Namun, ketika kebiasaan tersebut terbawa ke dalam penulisan karya ilmiah, kalimat yang dihasilkan berpotensi mengandung kesalahan dari aspek tata bahasa, pilihan kata, keefektifan kalimat, dan ketepatan makna. Kesalahan-kesalahan tersebut dapat dilihat pada contoh temuan berikut.

Tabel 2. Kesalahan Sintaksis Mahasiswa dalam Menulis Karya Ilmiah

Kalimat Salah	Kalimat Benar	Label Error	Keterangan
Di dalam penelitian ini mengusulkan arsitektur jaringan saraf konvolusional baru.	Penelitian ini mengusulkan arsitektur jaringan saraf konvolusional baru.	ERR_NO_SUBJ	Penggunaan kata depan “di dalam” menyebabkan subjek hilang karena membentuk keterangan.
Berdasarkan hasil pengujian menunjukkan bahwa throughput jaringan meningkat sebesar 20%.	Hasil pengujian menunjukkan bahwa throughput jaringan meningkat sebesar 20%. Berdasarkan hasil pengujian, throughput jaringan meningkat sebesar 20%.	ERR_NO_SUBJ	Kata “berdasarkan” menyebabkan subjek dalam kalimat itu hilang.
Sistem berbasis <i>cloud computing</i> yang	Sistem berbasis <i>cloud computing</i> mampu	ERR_NO_PRED	Penggunaan kata “yang”

mampu mengintegrasikan data dari berbagai sensor IoT secara <i>real-time</i> .	mengintegrasikan data dari berbagai sensor IoT secara <i>real-time</i> .		menyebabkan kalimat ini tidak memiliki predikat karena hanya berupa frasa nomina atau klausa yang belum lengkap.
Pengembangan aplikasi berbasis Android yang menggunakan framework Flutter untuk mempercepat proses.	Pengembangan aplikasi berbasis Android menggunakan framework Flutter untuk mempercepat proses pengembangan.	ERR_NO_PRED	Penggunaan kata “yang” dalam ungkapan ini mengakibatkan predikatnya hilang.
Meskipun performa server menurun, tetapi sistem tidak mengalami crash.	Meskipun performa server menurun, sistem tidak mengalami crash.	ERR_DBL_CONJ	Kalimat ini menjadi rusak strukturnya karena menggunakan dua konjungsi secara bersamaan, yaitu “maupun” dan “tetapi”.
Karena algoritma enkripsi sangat kompleks, maka waktu eksekusi menjadi lebih lama.	Karena algoritma enkripsi sangat kompleks, waktu eksekusi menjadi lebih lama.	ERR_DBL_CONJ	Kalimat ini juga rusak strukturnya karena menggunakan dua konjungsi secara bersamaan, yaitu “karena” dan “maka”.
Administrator jaringan dapat mengakses ke server utama melalui protokol SSH.	Administrator jaringan dapat mengakses server utama melalui protokol SSH.	ERR_TRAN_PREP	Kata kerja “mengakses” tidak membutuhkan preposisi “ke” karena merupakan kata kerja transitif.

Berdasarkan evaluasi GPT-4 sebagaimana paparan tabel di atas, kesalahan ketidaklengkapan subjek (ERR_NO_SUBJ) ditemukan pada kalimat yang diawali dengan ungkapan *di dalam* dan *berdasarkan*, sehingga unsur yang seharusnya berfungsi sebagai subjek tidak hadir secara jelas dalam struktur kalimat. Sementara itu, kesalahan ketidaklengkapan predikat (ERR_NO_PRED) terjadi ketika digunakan kata *yang* menyebabkan konstruksi kalimat hanya berupa frasa atau klausa, sehingga belum membentuk kalimat lengkap. Kesalahan penggunaan konjungsi ganda (ERR_DBL_CONJ) tampak pada penggunaan pasangan konjungsi *meskipun–tetapi* dan *karena–maka* secara bersamaan dalam konstruksi yang tidak memerlukan keduanya secara bersamaan.

Selain itu, ditemukan kesalahan penggunaan preposisi (ERR_TRAN_PREP), yaitu penambahan kata *ke* setelah verba transitif *mengakses*, yang seharusnya langsung diikuti objek tanpa preposisi. Temuan tersebut juga dikeluhkan oleh Umrah et al., (2026) dalam penelitiannya bahwa mahasiswa masih mengalami kesulitan dalam memahami hubungan antarkomponen sintaksis serta menerapkan kaidah pembentukan kalimat bahasa Indonesia dengan tepat. Berkaitan dengan kesalahan-kesalahan sintaksis tersebut, kesalahan-kesalahan mahasiswa berkaitan dengan kebiasaan menggunakan struktur bahasa yang informal, kurangnya penguasaan tata bahasa, serta kecenderungan menyusun kalimat berdasarkan pola yang lazim digunakan dalam komunikasi sehari-hari.

Akurasi AI dalam Mendeteksi dan Memvalidasi Koreksi (Sub-task GEC)

Pengujian terhadap model-model LLM (Large Language Model) dalam konteks GEC (General Error Correction) dilakukan melalui dua *subtask* dengan tingkat kompleksitas yang berbeda. *Subtask* 1 menguji kemampuan model dalam memvalidasi apakah koreksi tata bahasa yang diajukan sudah benar, sementara *Subtask* 2 menguji kemampuan model dalam memverifikasi apakah keseluruhan kalimat hasil koreksi benar-benar bebas dari semua jenis kesalahan. Perbedaan antara dua *subtask* ini mencerminkan perbedaan antara kemampuan deteksi lokal (per-koreksi) dan pemahaman holistik pada tingkat kalimat. Tabel 3 menyajikan hasil evaluasi komparatif tersebut.

Tabel 3. Performa Model LLM dalam Deteksi dan Validasi Koreksi GEC

Model AI	Precision	Recall	F1-Score	Keterangan Performa
GPT-4	0,9074	0,8886	0,8979	Memiliki tingkat presisi tertinggi dalam deteksi.
GPT-3.5-Turbo	0,8636	0,9922	0,9234	Sangat sensitif (<i>Recall</i> tinggi) dalam mengenali koreksi.
InternLM2-20b	0,8609	0,9858	0,9191	Kompetitif dengan GPT dalam tugas deteksi awal.
GPT-3.5 (Subtask 2)	0,3600	0,1395	0,2011	Akurasi menurun drastis saat memproses hasil akhir.

Data pada Tabel 3 mengungkapkan temuan menarik mengenai kapabilitas LLM dalam konteks GEC. GPT-4 menunjukkan presisi tertinggi sebesar 0,9074, yang berarti dari seluruh deteksi yang dilakukannya, 90,74% di antaranya

merupakan deteksi yang tepat. Namun, GPT-3.5-Turbo justru mencapai F1-Score tertinggi sebesar 0,9234, didorong oleh nilai *recall* yang sangat tinggi (0,9922), yang mengindikasikan kemampuannya untuk mengenali hampir seluruh koreksi yang valid. Performa kompetitif juga ditunjukkan oleh model sumber terbuka InternLM2-20b dengan F1-Score 0,9191, yang membuktikan bahwa model non-proprietary pun mampu mendekati kapabilitas model komersial terkemuka (L. Lin et al., 2025).

Temuan yang paling krusial adalah penurunan performa yang sangat drastis pada Subtask 2, di mana GPT-3.5 hanya mampu mencapai F1-Score sebesar 0,2011. Penurunan ini mengindikasikan bahwa meskipun LLM cukup andal dalam mendeteksi apakah sebuah koreksi sudah benar secara lokal (Shardlow et al., 2021), model tersebut mengalami kesulitan signifikan dalam memverifikasi bahwa keseluruhan kalimat hasil koreksi bebas dari semua jenis kesalahan. Fenomena ini sejalan dengan temuan penelitian tentang keterbatasan LLM dalam tugas penalaran multi-langkah ketika model harus mempertimbangkan semua aspek gramatikal sekaligus dalam satu kalimat yang utuh. Beban kognitif komputasionalnya meningkat secara eksponensial, sehingga mengakibatkan penurunan kinerja yang signifikan.

Berbeda dari LLM generalis, tugas GEC menunjukkan profil performa yang lebih stabil dan terkalibrasi. Model-model ini umumnya mengadopsi arsitektur *sequence-to-sequence* berbasis Transformer, di mana kalimat yang mengandung kesalahan diperlakukan sebagai kalimat sumber dan kalimat yang sudah terkoreksi sebagai kalimat target, serupa dengan cara kerja sistem penerjemahan mesin. Penelitian Musyafa et al. (2022) mengungkapkan bahwa beberapa perangkat AI memiliki tingkat akurasi yang berbeda dalam mengevaluasi model-model GEC dalam konteks bahasa Indonesia.

Tabel 4. Performa Model AI Berbasis Transformer untuk Koreksi Gramatikal (GEC)

Model / Framework	F1-Score	BLEU Score	Keunggulan Utama
SAN-GEC + Fine-Tuning	71,94	78,13	Model terbaik mampu menangani kata-kata asing (UNK) melalui mekanisme copy.
SAN-GEC + BPEChar 1k	69,87	77,21	Sangat efektif dalam menangani kesalahan ejaan.
CNN-GEC	43,53	56,76	Proses pelatihan dan decoding lebih cepat dibandingkan dengan RNN.

Berdasarkan Tabel 4, model SAN-GEC dengan *fine-tuning* merupakan model terbaik dengan F1-Score 71,94 dan BLEU-Score 78,13. Keunggulan model ini terletak pada penerapan mekanisme copy yang memungkinkan model untuk secara langsung menyalin kata-kata di luar kosakata (*out-of-vocabulary*/UNK), termasuk kata asing, nama ilmiah, dan istilah teknis yang kerap muncul dalam teks akademik, tanpa harus menggantinya dengan kata lain yang tidak tepat konteksnya (Musyafa et al., 2022). Kemampuan ini sangat relevan untuk menangani teks akademik berbahasa Indonesia yang sering mengandung terminologi internasional.

Model SAN-GEC dengan BPEChar 1k menunjukkan keunggulan dalam menangani kesalahan ejaan berkat representasi bagian katanya, namun tetap sedikit di bawah model *yang telah di-fine-tune* pada metrik secara keseluruhan. Sementara CNN-GEC, meski memiliki skor terendah, menawarkan efisiensi komputasi yang signifikan dibandingkan dengan model berbasis RNN. Perbandingan ini mengindikasikan bahwa terdapat trade-off antara akurasi dan efisiensi komputasi yang perlu dipertimbangkan dalam implementasi praktis sistem GEC untuk bahasa Indonesia. Temuan ini menegaskan bahwa arsitektur khusus dengan strategi penanganan data yang terbatas memberikan hasil yang lebih stabil dibandingkan dengan model umum yang tidak dirancang khusus untuk karakteristik morfologis bahasa Indonesia.

Akurasi AI Berdasarkan Fenomena Linguistik (Analisis Diagnostik)

Untuk memperoleh pemahaman yang lebih rinci dan jelas, analisis diarahkan pada performa AI berdasarkan fenomena linguistik tertentu. Pendekatan diagnostik ini penting karena kesalahan sintaksis bukanlah entitas yang homogen.

Setiap jenis kesalahan memiliki tingkat kompleksitas komputasional yang berbeda, tergantung pada seberapa banyak konteks yang diperlukan untuk mengidentifikasinya. Tabel 4 merangkum temuan diagnostik hasil karya ilmiah mahasiswa yang dideteksi menggunakan GPT-4 dan GPT-3.5-Turbo yang sangat presisi dan sensitif dalam mendeteksi kesalahan sintaksis.

Tabel 5. Akurasi AI Berdasarkan Fenomena Linguistik

Fenomena Linguistik	Akurasi AI (%)	Keterangan
Morfologi (Imbuhan)	84,4%	paling mudah dideteksi oleh AI
Konjungsi Koordinatif	80,7%	mudah sekali dideteksi oleh AI
Struktur Sintaksis	79,0%	mudah dideteksi oleh AI
Logika & Kuantifikasi	75,7%	Kesulitan AI dalam mendeteksi fenomena linguistik ini sedang
Perbandingan (Komparatif)	59,5%	Sulit dideteksi oleh AI
Penalaran Numerik	45,0%	sangat sulit dideteksi oleh AI (titik lemah AI dalam mendeteksi aspek fenomena linguistik)

Data pada Tabel 5 menampilkan gradasi akurasi yang sangat signifikan, mencerminkan hierarki kompleksitas linguistik dari perspektif komputasional. AI paling andal dalam mendeteksi kesalahan morfologi (imbuhan) dengan akurasi 84,4%. Capaian ini dapat dijelaskan secara teknis: kesalahan morfologis memiliki pola yang sangat reguler dan dapat diidentifikasi berdasarkan konteks lokal (hanya memerlukan pengamatan satu hingga dua kata), sehingga model yang telah dilatih pada korpus besar bahasa Indonesia dapat dengan mudah mengenalinya melalui pencocokan pola statistik (Koto et al., 2020). Konjungsi koordinatif juga terdeteksi dengan baik (80,7%) karena kesalahannya umumnya bersifat leksikal, yaitu mendeteksi penggunaan 'sedangkan' yang seharusnya 'sehingga'; misalnya, hal ini hanya memerlukan pemahaman konteks pada klausa yang berdekatan.

Akurasi mulai menurun secara signifikan ketika AI dihadapkan pada fenomena yang memerlukan penalaran tingkat yang lebih tinggi. Kesalahan struktur sintaksis berada pada akurasi 79,0% dan logika serta kuantifikasi pada 75,7%, yang mengindikasikan bahwa AI mulai kesulitan mengevaluasi apakah relasi logis antarunsur kalimat sudah koheren. Penurunan paling drastis terjadi pada perbandingan komparatif (59,5%) dan penalaran numerik (45,0%). Akurasi penalaran numerik yang hanya 45,0%, bahkan lebih rendah daripada peluang acak

pada tugas biner, mengindikasikan bahwa AI tidak hanya gagal mendeteksi kesalahan jenis ini, tetapi juga secara aktif membuat keputusan yang keliru. Hal ini mengindikasikan bahwa model belum memiliki representasi internal yang memadai untuk memahami relasi kuantitatif dalam teks berbahasa Indonesia (Cahyawijaya et al., 2021).

Temuan ini berkorelasi kuat dengan batasan teknis dalam arsitektur model saat ini. Mekanisme *self-attention* pada Transformer, meski sangat efektif untuk menangkap relasi linguistik, belum dibekali kemampuan penalaran simbolik yang diperlukan untuk memahami ekspresi numerik dan perbandingan kuantitatif secara akurat. Sebagaimana ditunjukkan dalam evaluasi model berbasis *pre-training* seperti BERT (Qasim et al., 2022), tugas yang memerlukan pemahaman relasi numerik dan komparatif secara konsisten menjadi titik lemah model bahasa, dan hal ini tercermin secara nyata dalam konteks bahasa Indonesia.

Peluang dan Tantangan AI-NLP dalam Konteks Akademik Bahasa Indonesia

Temuan-temuan yang disintesis dalam penelitian ini secara kolektif mengonstruksi gambaran yang nuansatif mengenai posisi teknologi AI berbasis NLP dalam ekosistem penulisan akademik bahasa Indonesia. Di satu sisi, capaian F1-Score LLM di atas 0,89 pada *Subtask 1* GEC merupakan sinyal yang sangat menjanjikan bahwa teknologi ini telah melampaui ambang batas untuk menjadi alat bantu yang praktis dan andal. Di sisi lain, penurunan drastis pada *Subtask 2* dan kelemahan akurasi dalam penalaran numerik (45,0%) mengingatkan bahwa sistem ini masih jauh dari sempurna.

Salah satu faktor struktural yang menjelaskan keterbatasan AI dalam bahasa Indonesia adalah posisinya sebagai bahasa bersumber daya rendah dalam ekosistem NLP global. Dibandingkan dengan bahasa Inggris yang memiliki ribuan dataset *benchmark* dan model yang telah diuji secara ekstensif, sumber daya untuk bahasa Indonesia masih sangat terbatas (Musyafa et al., 2022). Upaya pembangunan *benchmark* seperti IndoLEM (Koto et al., 2020) dan IndoNLG (Cahyawijaya et al., 2021) merupakan langkah penting, namun masih belum mencukupi kebutuhan pengembangan model GEC yang robust untuk mencakup

seluruh fenomena linguistik dalam bahasa Indonesia.

Tantangan lain yang tidak kalah penting adalah kompleksitas bahasa Indonesia yang inheren. Sistem afiksasi yang kaya (prefiks, sufiks, infiks, konfiks, dan kombinasinya) menciptakan ruang morfologis yang sangat besar (A. Amalia et al., 2021), sementara fleksibilitas urutan kata dalam kalimat bahasa Indonesia, yang berbeda dari struktur SVO yang lebih kaku dalam bahasa Inggris, menambah dimensi kompleksitas yang harus dimodelkan oleh sistem AI. Karakteristik inilah yang membuat transfer *learning* dari model berbahasa Inggris tidak serta-merta menghasilkan performa optimal pada bahasa Indonesia dan menggarisbawahi pentingnya pengembangan model yang dilatih secara khusus pada korpus bahasa Indonesia berkualitas tinggi.

Dari perspektif praktis, penelitian ini merekomendasikan pendekatan *hybrid* yang memaksimalkan kelebihan masing-masing komponen: AI berbasis NLP dapat difungsikan sebagai penyaring otomatis (*automatic pre-screening*) untuk kesalahan morfologi, ejaan, dan konjungsi, sementara penyunting manusia yang terampil tetap harus menangani aspek-aspek yang memerlukan pemahaman mendalam terhadap logika argumen, konsistensi terminologi, dan konteks disiplin ilmu. Pendekatan berlapis seperti ini tidak hanya lebih efektif secara kualitas, tetapi juga lebih efisien dalam hal waktu dibandingkan dengan penyuntingan manual secara penuh.

PENUTUP

Penelitian ini berhasil menganalisis secara sistematis akurasi AI berbasis NLP dalam mendeteksi kesalahan sintaksis pada makalah akademik berbahasa Indonesia. Tiga kesimpulan utama dapat ditarik dari kajian ini. Pertama, profil kesalahan sintaksis pada makalah akademik mahasiswa didominasi oleh kesalahan tataran kalimat (67,02%), terutama kalimat tidak berpredikat dan kalimat buntung, yang merepresentasikan target deteksi paling kritis bagi sistem AI. Kedua, model LLM generasi terkini seperti GPT-4 menunjukkan akurasi yang tinggi dalam tugas validasi koreksi (F1-Score 0,8979), sementara model GEC khusus SAN-GEC terbukti lebih stabil dan andal dalam tugas koreksi otomatis langsung dengan F1-

Score 71,94. Ketiga, analisis diagnostik mengungkap ketidakmerataan performa yang signifikan: AI sangat kuat pada aspek morfologi (84,4%), namun sangat lemah pada penalaran numerik (45,0%), yang mengindikasikan bahwa AI belum mampu menggantikan peran penyunting manusia secara menyeluruh.

Berdasarkan temuan tersebut, strategi paling optimal dalam pemanfaatan AI untuk penyuntingan akademik adalah pendekatan berlapis (*layered editing*): memanfaatkan AI sebagai penyaring otomatis untuk kesalahan mekanis dan morfologis, dilanjutkan dengan LLM untuk penyempurnaan struktur sintaksis, dan diakhiri dengan validasi manusia sebagai otoritas akhir. Peneliti selanjutnya disarankan untuk mengembangkan korpus GEC Bahasa Indonesia yang lebih besar dan representatif, serta mengeksplorasi teknik *fine-tuning* berbasis instruksi (*instruction fine-tuning*) pada model LLM dengan data teks akademik Indonesia agar performa pada aspek penalaran kontekstual dapat ditingkatkan.

DAFTAR RUJUKAN

- Amalia, A., Sitompul, O. S., Mantoro, T., & Nababan, E. B. (2021). Morpheme embedding for Bahasa Indonesia using modified byte pair encoding. *IEEE Access*, 9, 155699–155710. <https://doi.org/10.1109/ACCESS.2021.3128439>
- Anugrah, G., Cahyani, G. I., Murni, B. C., Pratama, M. H., & Zuve, F. O. (2026). Slang language: Its impact on standard Indonesian language proficiency among Indonesian language and literature education students at Universitas Negeri Padang. *Jurnal Skripta*, 12(1), 55–65. <https://doi.org/10.31316/skripta.v12i1.8998>
- Arifin, Z. (2008). *Sintaksis*. Grasindo. https://books.google.com/books?hl=id&lr=&id=RTMoOGc5urAC&oi=fnd&pg=PA1&dq=Sintaksis+adalah&ots=CWo1hOmNgR&sig=akAa3tuBnt-X_V-dny3_UEN_OJE
- Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3), 643–701. https://doi.org/10.1162/coli_a_00478
- Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8875–8898. <https://doi.org/10.18653/v1/2021.emnlp-main.699>
- Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, 44, 51–62. <https://doi.org/10.1016/j.jslw.2019.03.005>
- Chowdhary, K. R. (2020). Natural language processing. In K. R. Chowdhary (Ed.), *Fundamentals of artificial intelligence* (pp. 603–649). Springer. https://doi.org/10.1007/978-81-322-3972-7_19
- Diani, D. R., Samudra, D. F., & Sabrina, Y. V. (2025). Analisis kesalahan bidang ejaan dan morfologi pada poster di kampus UIN Raden Mas Said Surakarta. *Narasi: Jurnal*

- Kajian Bahasa, Sastra Indonesia, dan Pengajarannya*, 3(2), 177–187.
<https://doi.org/10.30762/narasi.v3i2.5546>
- Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural language processing. In M. E. Klontzas, S. C. Fanni, & E. Neri (Eds.), *Introduction to artificial intelligence* (pp. 87–99). Springer. https://doi.org/10.1007/978-3-031-25928-9_5
- 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.
<https://doi.org/10.18653/v1/2020.coling-main.66>
- Lee, S.-M., & Bang, J. (2025). Transforming language education through AI: Artificial intelligence digital textbook (AIDT). In H. Reinders, J.-K. Park, & J. S. Lee (Eds.), *Innovation in language learning and teaching: The case of Korea* (pp. 49–72). Springer. https://doi.org/10.1007/978-3-031-83561-2_3
- Lin, L., Zhang, S., Fu, S., & Liu, Y. (2025). FD-LLM: Large language model for fault diagnosis of complex equipment. *Advanced Engineering Informatics*, 65, 103208.
<https://doi.org/10.1016/j.aei.2025.103208>
- Lin, N., Zeng, M., Huang, W., Jiang, S., Xiao, L., & Yang, A. (2025). A simple yet effective corpus construction framework for Indonesian grammatical error correction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(1), 1–15. <https://doi.org/10.1145/3704264>
- Mahendra, R., Aji, A. F., Louvan, S., Rahman, F., & Vania, C. (2021). IndoNLI: A natural language inference dataset for Indonesian. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10511–10527.
<https://doi.org/10.18653/v1/2021.emnlp-main.821>
- Mao, Y., Ge, Y., Fan, Y., Xu, W., Mi, Y., Hu, Z., & Gao, Y. (2025). A survey on LoRA of large language models. *Frontiers of Computer Science*, 19(7), 197605.
<https://doi.org/10.1007/s11704-024-40663-9>
- Musyafa, A., Gao, Y., Solyman, A., Wu, C., & Khan, S. (2022). Automatic correction of Indonesian grammatical errors based on transformer. *Applied Sciences*, 12(20), 10380. <https://doi.org/10.3390/app122010380>
- Nababan, E. B., Rianto, Kusuma, D., Rizkina, T., & Sastromihardjo, A. (2025). Analisis kesalahan sintaksis pada artikel jurnal ilmiah mahasiswa [Analysis of syntax errors in student scientific journal articles]. *Polyglot: Jurnal Ilmiah*, 21(1), 51–68.
<https://doi.org/10.19166/pji.v21i1.9127>
- Qasim, R., Bangyal, W. H., Alqarni, M. A., & Almazroi, A. A. (2022). A fine-tuned BERT-based transfer learning approach for text classification. *Journal of Healthcare Engineering*, 2022, 1–17. <https://doi.org/10.1155/2022/3498123>
- Ramadhani, A. A., Ramadhan, D. A., Artanti, F. D., & Parhusip, K. A. (2026). Analisis bentuk dan faktor kesalahan berbahasa Indonesia pada konten Instagram remaja usia 17-20 tahun. *Jurnal Bintang Pendidikan Indonesia*, 4(2), 74–83.
<https://doi.org/10.55606/jubpi.v4i2.4341>
- Shardlow, M., Evans, R., Paetzold, G. H., & Zampieri, M. (2021). SemEval-2021 Task 1: Lexical complexity prediction. In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 1–16). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.semeval-1.1>
- Umrah, U., Nurhikma, N., Nensilanti, N., & Saleh, M. (2026). Analisis kesalahan penulisan tata bahasa baku siswa dalam pembelajaran bahasa Indonesia. *DEIKTIS: Jurnal Pendidikan, Bahasa dan Sastra*, 6(2), 1762–1772.
<https://doi.org/10.53769/deiktis.v6i2.3431>