EVALUASI KOMPARATIF MANUSIA, CHATGPT, DAN GEMINI DALAM MENGANALISIS POLA KALIMAT PADA TEKS AKADEMIK

Ariesta Bagus Pramuwibowo[1*], Titik Dwi Ramthi Hakim [2]

Tadris Bahasa Indonesia, FTIK, UIN Sayyid Ali Rahmatullah Tulungagung[1,2]

ariesta.bagus@uinsatu.ac.id*, TitikDwiRamthi@uinsatu.ac.id,

Abstrak

Peningkatan minat siswa SMA dalam memanfaatkan aplikasi berbasis LLM (ChatGPT dan Gemini) sebagai alat belajar perlu dievaluasi. Hasil analisis kalimat dari ChatGPT dan Gemini yang sering dijadikan siswa sebagai karyanya perlu dikaji secara cermat. Oleh karena itu, evaluasi komparatif hasil analisis kalimat ChatGPT dan Gemini penting dilakukan untuk mengetahui keandalan dua aplikasi LLM tersebut. Penelitian ini menyajikan hasil evaluasi komparatif aplikasi ChatGPT dan Gemini dalam menganalisis pola kalimat yang didasarkan pada hasil analisis ahli bahasa Indonesia sebagai *gold standard*. Dalam penelitian ini, digunakan pendekatan kuantitatif komparatif. Data dalam penelitian ini yaitu 139 kalimat dalam teks akademik yang terdapat di buku siswa kelas XI Kemendikbudristekdikti yang sudah dikurasi. Dilakukan analisis uji kalsifikasi kesepakatan, uji performa antara ChatGPT dan Gemini. Tingkat kesepakatan ChatGPT tergolong hampir sempurna dengan angka koefisien 0,843, sedangkan kesepakatan Gemini tergolong sedang dengan angka koefisien 0,597. Hasil evaluasi ChatGPT menunjukkan bahwa nilai akurasi sebesar 89,93%, nilai *macro precision* sebesar 91,70%, nilai *macro recall* sebesar 87,97%, dan nilai *macro F1-score* sebesar 89,21%. Dalam ChatGPT, kesalahan klasifikasi terkonsentrasi pada pembedaan antara kalimat majemuk setara dan kalimat majemuk bertingkat. Selain itu, Hasil evaluasi Gemini menunjukkan bahwa nilai akurasi sebesar 73,38%, nilai *macro precision* sebesar 79,77%, nilai *macro recall* sebesar 72,42%, dan nilai *macro F1-score* sebesar 66,99%. Dalam Gemini, kesalahan klasifikasi terpusat pada klasifikasi kalimat majemuk setara. Dengan demikian, hasil analisis ahli bahasa Indonesia sebagai *gold standart* menunjukkan bahwa hasil evaluasi kesesuaian dan performa ChatGPT lebih tinggi dibandingkan dengan Gemini dalam hal menganalisis kalimat.

Kata Kunci: Komparasi, ChatGPT, Gemini, Kalimat

Abstract

The increasing interest of high school students in utilizing LLM-based applications (ChatGPT and Gemini) as learning tools needs to be evaluated. Sentence analysis results from ChatGPT and Gemini, which are frequently used by students in their work, require careful review. Therefore, a comparative evaluation of the sentence analysis results from ChatGPT and Gemini is important to determine the reliability of the two LLM applications. This study presents the results of a comparative evaluation of the ChatGPT and Gemini applications in analyzing sentence patterns, based on the analysis of Indonesian linguists as the “gold standard”. A comparative quantitative approach was used in this study. The data used were 139 sentences from academic texts found in the grade XI student textbooks curated by the Ministry of Education, Culture, Research, and Technology. Agreement classification tests and performance tests were conducted between ChatGPT and Gemini. The level of agreement for ChatGPT was classified as almost perfect with a coefficient of 0.843, while Gemini's agreement was classified as moderate with a coefficient of 0.597. The ChatGPT evaluation results showed an accuracy of 89.93%, a macro precision of 91.70%, a macro recall of 87.97%, and a macro F1-score of 89.21%. In ChatGPT, classification errors focused on distinguishing between compound-equivalent and compound-equivalent sentences. Furthermore, the Gemini evaluation results showed an accuracy of 73.38%, a macro precision of 79.77%, a macro recall of 72.42%, and a macro F1-score of 66.99%. In Gemini,

classification errors focused on the classification of compound-equivalent sentences. Thus, the analysis by Indonesian linguists, as the gold standard, indicates that ChatGPT's suitability and performance are superior to Gemini's in sentence analysis.

Keywords: Comparison, ChatGPT, Gemini, Sentences

PENDAHULUAN

Dalam dunia pendidikan, penggunaan aplikasi berbasis LLM dalam pembelajaran semakin meningkat. Aplikasi berbasis LLM, seperti ChatGPT dan Gemini, memudahkan guru untuk mempersiakan perencanaan pembelajaran, bahkan menjadi pendamping personal bagi siswa ketika belajar memahami materi pelajaran.(Damayanti & Zulkarnain, 2026). Tidak hanya itu, aplikasi tersebut mengubah cara siswa mengerjakan tugas. Dengan adanya aplikasi tersebut, mereka dimudahkan saat menganalisis pertanyaan sehingga memunculkan jawaban. Sayangnya, banyak siswa yang memilih untuk menjadikan jawaban yang muncul dari aplikasi tersebut sebagai pendapat mereka atau keputusan akhir (Fan et al., 2025). Sejumlah 89% siswa menggunakan ChatGPT dan Gemini untuk menyelesaikan tugas serta dijadikan sebagai referensi utama (Kurniahtunnisa et al., 2025). Pada pembelajaran Bahasa Indonesia, Gemini dan ChatGPT juga digunakan untuk menganalisis teks, mengoreksi unsur kebahasaan, dan mengembangkan kosakata saat menulis bagi siswa SMA (Wahyuni et al., 2024).

Di sisi lain, ada tantangan bagi keakuratan ChatGPT dan Gemini dalam menganalisis teks dan mengoreksi unsur kebahasaan dalam teks akademik. Hasil analisis yang dilakukan aplikasi berbasis LLM perlu dikaji lebih mendalam. Hal itu disebabkan karena keandalan hasil analisis yang dilakukan aplikasi berbasis LLM belum sesempurna hasil analisis yang dilakukan oleh manusia. Berdasarkan uji coba yang dilakukan, keakuratan aplikasi berbasis LLM dalam menganalisis unsur SPOK kalimat berbahasa Indonesia tergolong tinggi pada kalimat simpleks, tetapi rendah pada kalimat majemuk (Kiranawati et al., 2025). Dengan kata lain, aplikasi berbasis LLM belum akurat dalam menganalisis teks dibandingkan dengan kemampuan manusia dalam menganalisis teks. Aplikasi LLM belum bisa menggantikan sepenuhnya fungsi manusia sebagai evaluator utama dalam menganalisis teks (Bavaresco et al., 2025).

Pada pembelajaran bahasa Indonesia tingkat SMA, siswa diajarkan kemampuan menganalisis unsur kalimat. Siswa tersebut harus mampu menganalisis subjek, predikat, objek, pelengkap, dan keterangan (Dewi Marhamah Syadli & Dewi Anggraini, 2023; Nainggolan, 2025). Selain itu, siswa SMA harus mampu menganalisis jenis kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat. Pada kalimat majemuk setara, siswa harus mampu memahami lalu menelaah penciri kalimat majemuk setara, yaitu konjungsi koordinatif (Rosyada et al., 2025). Tidak hanya itu, siswa harus mampu memeriksa penciri konjungsi subrodinatif

pada kalimat majemuk bertingkat (Nur Gimas Tiaroh et al., 2026). Dengan memahami dan menguasai kedua penciri tersebut, siswa akan mudah memilah antara kalimat majemuk setara dan kalimat majemuk bertingkat.

Berdasarkan hasil wawancara dengan 127 siswa SMA se-Tulungagung, ditemukan bahwa 67% dari mereka lebih memilih untuk menggunakan aplikasi berbasis LLM, seperti ChatGPT dan Gemini. Mereka sangat puas dengan respons dari dua aplikasi tersebut (Kurniahtunnisa et al., 2025). Tingkat kepuasan siswa tersebut tidak serta merta menjadikan hasil respons ChatGPT dan Gemini yang berupa hasil analisis kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat lebih akurat dibandingkan dengan hasil analisis oleh manusia. Dengan demikian, diperlukan studi tentang distribusi klasifikasi pola kalimat-kalimat tersebut, tingkat kesesuaian hasil analisis, dan performa ChatGPT serta Gemini yang dibandingkan dengan hasil analisis manusia.

METODE

Penelitian ini berpendekatan kuantitatif dengan desain penelitian komparatif evaluatif. Studi ini bertujuan untuk membandingkan tingkat kesesuaian dan performa dari ChatGPT dan Gemini dalam menganalisis pola kalimat tunggal, majemuk setara, dan bertingkat berdasarkan hasil analisis manusia sebagai *gold standard*. Evaluasi dilaksanakan dengan pengukuran tingkat kesepakatan (*inter-rater agreement*) menggunakan koefisien Chones's Kappa. Selain itu, dilakukan pengukuran performa kalsifikasi menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *macro F1-score*.

Sumber data yang digunakan yaitu teks akademik yang terdapat dalam buku "Cerdas Cergas Berbahasa dan Bersastra Indonesia" siswa kelas XI Kemendikbudristekdikti. Dalam teks akademik berupa artikel yang berjudul "Status Kondisi Terumbu Karang di Teluk Ambon" terdapat 169 kalimat. Setelah dikurasi, data yang digunakan yaitu 139 kalimat yang termasuk kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat. kalimat tersebut dianalisis oleh ahli bahasa Indonesia, ChatGPT, serta Gemini. Untuk prompt, digunakan kalimat yang sama yaitu dimasukkan teori serta kalimat perintah agar aplikasi menganalisis kalimat tunggal, kalimat majemuk serta, dan kalimat majemuk bertingkat berdasarkan teori yang sudah tertulis..

Analisis yang digunakan ada empat tahap. Pertama, analisis klasifikasi kesepakatan yang dihitung menggunakan koefisien Cohen's Kappa (κ) serta Interpretasi nilai κ yang mengacu pada Landis dan Koch (J Richard Landis & Gary G Koch, 1977). Kedua, dianalisis kesalahan klasifikasi dengan Confusion Matrix. Ketiga, dihitung indikator performa berdasarkan akurasi, *macro precision*, *macro recall*, dan *macro F1-score*. Keempat, dilakukan analisis kesalahan klasifikasi yang memfokuskan pada

kekurangtepatan klasifikasi oleh ChatGPT dan Gemini yang didasarkan pada *gold standard*.

HASIL DAN PEMBAHASAN

Hasil penelitian ini terkait dengan evaluasi komparatif uji performa dan akurasi jenis kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat yang dihasilkan ChatGPT dan Gemini yang dibandingkan dengan hasil analisis ahli bahasa Indonesia sebagai *gold standard*.

Distribusi Klasifikasi Jenis Kalimat Tunggal, Kalimat Majemuk Setara, dan Kalimat Majemuk Bertingkat

Terdapat tiga hasil distribusi klasifikasi jenis kalimat yang didasarkan pada hasil analisis ahli bahasa Indonesia, ChatGPT, dan Gemini.

1. Distribusi Klasifikasi Jenis Kalimat oleh Ahli Bahasa Indonesia

Berdasarkan hasil analisis ahli bahasa Indonesia (*gold strandard*) yang mengacu pada teori kalimat tunggal, kalimat majemuk setara, kalimat majemuk bertingkat, diperoleh hasil yang tersaji dalam Tabel 1.

Tabel 1. Klasifikasi Pola Kalimat oleh Ahli Bahasa Indonesia

Nomor	Jenis Kalimat	Jumlah	Kalimat Ke-
1	Kalimat Tunggal	34	12, 18, 20, 25, 27, 28, 30, 31, 36, 37, 38, 39, 40, 42, 43, 44, 45, 54, 63, 64, 67, 70, 74, 78, 80, 81, 85, 114, 119, 126, 128, 129, 131, 136
2	Kalimat Majemuk Setara	46	1, 7, 8, 13, 15, 16, 19, 22, 23, 29, 32, 33, 34, 47, 48, 49, 55, 56, 58, 60, 71, 76, 77, 79, 82, 83, 84, 87, 88, 92, 93, 97, 99, 100, 110, 116, 117, 118, 120, 123, 124, 125, 133, 135, 138, 139
3	Kalimat Majemuk Bertingkat	59	2, 3, 4, 5, 6, 9, 10, 11, 14, 17, 21, 24, 26, 35, 41, 46, 50, 51, 52, 53, 57, 59, 61, 62, 65, 66, 68, 69, 72, 73, 75, 86, 89, 90, 91, 94, 95, 96, 98, 101, 102, 103, 104, 105, 106, 107, 108, 109, 111, 112, 113, 115, 121, 122, 127, 130, 132, 134, 137

Berdasarkan Tabel 1, teranalisis 139 kalimat oleh ahli bahasa Indonesia. Terdapat 34 kalimat tunggal, 46 kalimat majemuk setara, serta 59 kalimat majemuk bertingkat. Dengan demikian, persentase kalimat majemuk bertingkat tertinggi dibandingkan dengan kalimat tunggal dan kalimat majemuk setara.

2. Distribusi Klasifikasi Jenis Kalimat dengan ChatGPT

Berdasarkan hasil analisis dengan ChatGPT yang mengacu pada teori kalimat tunggal, kalimat majemuk setara, kalimat majemuk bertingkat, diperoleh hasil yang tersaji dalam Tabel 2.

Tabel 2. Klasifikasi Jenis Kalimat dengan ChatGPT

Nomor	Jenis Kalimat	Jumlah	Kalimat Ke-
1	Kalimat Tunggal	22	12, 18, 20, 27, 28, 30, 31, 36, 42, 43, 45, 54, 63, 64, 67, 70, 80, 81, 85, 114, 119, 126, 128, 129, 131, 136
2	Kalimat Majemuk Setara	41	1, 7, 8, 13, 15, 16, 19, 23, 29, 32, 33, 37, 38, 39, 40, 47, 48, 49, 55, 56, 58, 60, 71, 76, 77, 78, 79, 82, 83, 84, 87, 92, 93, 97, 99, 100, 105, 116, 117, 118, 120, 123, 124, 125, 133, 138, 139
3	Kalimat Majemuk Bertingkat	66	2, 3, 4, 5, 6, 9, 10, 11, 14, 17, 21, 22, 24, 25, 26, 34, 35, 41, 44, 46, 50, 51, 52, 53, 57, 59, 61, 62, 65, 66, 68, 69, 72, 73, 74, 75, 86, 88, 89, 90, 91, 94, 95, 96, 98, 101, 102, 103, 104, 106, 107, 108, 109, 110, 111, 112, 113, 115, 121, 122, 127, 130, 132, 134, 135, 137

Berdasarkan Tabel 2, teranalisis 139 kalimat oleh ChatGPT dengan prompt yang mengacu pada teori kalimat tunggal dan kalimat majemuk. Hasilnya yaitu ada 22 kalimat tunggal, 41 kalimat majemuk setara, serta 66 kalimat majemuk bertingkat. Dengan demikian, persentase kalimat majemuk bertingkat tertinggi dibandingkan dengan kalimat tunggal dan kalimat majemuk setara.

3. Distribusi Klasifikasi Jenis Kalimat dengan Gemini

Berdasarkan hasil analisis dengan Gemini yang mengacu pada teori kalimat tunggal, kalimat majemuk setara, kalimat majemuk bertingkat, diperoleh hasil yang tersaji dalam Tabel 3.

Tabel 3. Klasifikasi Jenis Kalimat dengan Gemini

Nomor	Jenis Kalimat	Jumlah	Kalimat Ke-
1	Kalimat Tunggal	59	1, 7, 8, 12, 15, 16, 18, 20, 23, 25, 27, 28, 30, 31, 33, 36, 37, 38, 39, 40, 42, 43, 45, 47, 48, 49, 54, 55, 56, 60, 64, 67, 70, 71, 76, 77, 78, 80, 81, 82, 83, 84, 85, 87, 92, 97, 99, 114, 116, 119, 120, 123, 126, 128, 129, 131, 136, 138, 139
2	Kalimat Majemuk Setara	12	13, 19, 29, 32, 58, 79, 93, 100, 118, 124, 125, 133
3	Kalimat Majemuk Bertingkat	68	2, 3, 4, 5, 6, 9, 10, 11, 14, 17, 21, 22, 24, 26, 34, 35, 41, 44, 46, 50, 51, 52, 53, 57, 59, 61, 62, 63, 65, 66, 68, 69, 72, 73, 74, 75, 86, 88, 89, 90, 91, 94, 95, 96, 98, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 115, 117, 121, 122, 127, 130, 132, 134, 135, 137

Berdasarkan Tabel 3, teranalisis 139 kalimat oleh Gemini dengan prompt yang mengacu pada teori kalimat tunggal dan kalimat majemuk. Hasilnya yaitu ada 59 kalimat tunggal, 12 kalimat majemuk setara, serta 68 kalimat majemuk bertingkat. Dengan demikian, persentase kalimat majemuk bertingkat tertinggi dibandingkan dengan kalimat tunggal dan kalimat majemuk setara.

Tingkat Kesesuaian Hasil Analisis Kalimat Tunggal, Kalimat Majemuk Setara, dan Kalimat Majemuk Bertingkat oleh Ahli Bahasa Indonesia (*Gold Standart*), ChatGPT, dan Gemini

Terdapat dua hasil tingkat kesesuaian dan interpretasi kesepakatan dari hasil analisis kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat oleh ahli bahasa Indonesia (*gold standart*), ChatGPT, dan Gemini. Analisis ini menggunakan koefisien Cohen's Kappa (κ).

Po adalah proporsi kesepakatan yang diamati (*observed agreement*) dengan rumus (Gambar 1), sedangkan Pe merupakan proporsi kesepakatan yang diperkirakan terjadi secara kebetulan (*expected agreement*) dengan rumus (Gambar 2). Setelah itu, dilakukan interpretasi nilai κ dengan mengacu pada kriteria dari Landis dan Koch yang ada di Tabel 4 (J Richard Landis & Gary G Koch, 1977).

$$P_o = \frac{\text{jumlah nilai diagonal}}{N}$$

Gambar 1. Rumus Po (*observed agreement*)

$$P_e = \sum_{i=1}^{k} \left(\frac{R_i}{N} \times \frac{C_i}{N} \right) = \frac{\sum (R_i \times C_i)}{N^2}$$

Gambar 2. Rumus Pe (*expected agreement*)

Tabel 4 Interpretasi Nilai κ Menurut Landis dan Koch

Nilai κ	Interpretasi
< 0,00	Kesepakatan sangat buruk
0,00—0,20	Kesepakatan lemah
0,21—0,40	Kesepakatan cukup
0,41—0,60	Kesepakatan sedang
0,61—0,80	Kesepakatan kuat
0,81—1,00	Kesepakatan hampir sempurna

1. **Tingkat Kesesuaian dan Interpretasi Kesepakatan dari Hasil Analisis Kalimat dari ChatGPT dan *Gold Standard***

Dari hasil analisis menggunakan SPSS, data pembandingan antara ChatGPT dan hasil analisis ahli bahasa Indonesia sebagai *gold standard* tertulis di Tabel 5.

Tabel 5. Manusia * ChatGPT Crosstabulation

			ChatGPT			Total
			Kalimat tunggal	Kalimat Majemuk Setara	Kalimat Majemuk Bertingkat	
Manusia	Kalimat tunggal	Count	26	5	3	34
		% within Manusia	76,5%	14,7%	8,8%	100,0%
		% within ChatGPT	100,0%	10,6%	4,5%	24,5%
	Kalimat Majemuk Setara	Count	0	41	5	46
		% within Manusia	0,0%	89,1%	10,9%	100,0%
		% within ChatGPT	0,0%	87,2%	7,6%	33,1%
	Kalimat Majemuk Bertingkat	Count	0	1	58	59
		% within Manusia	0,0%	1,7%	98,3%	100,0%
		% within ChatGPT	0,0%	2,1%	87,9%	42,4%
Total		Count	26	47	66	139
		% within Manusia	18,7%	33,8%	47,5%	100,0%
		% within ChatGPT	100,0%	100,0%	100,0%	100,0%

Berdasarkan tabel 5, proporsi kesepakatan yang diamati (*observed agreement*) hasil analisis ChatGPT dengan hasil analisis ahli bahasa Indonesia (*gold standard*) didasarkan pada pembagian antara jumlah kesepakatan dan jumlah seluruh data. Ditemukan bahwa jumlah kalimat yaitu 139, kalimat tunggal yaitu 26, kalimat majemuk setara 41, dan kalimat majemuk bertingkat yaitu 58. Berdasarkan hal tersebut, nilai Po adalah 0,8993.

Selain itu, berdasarkan tabel 5, proporsi kesepakatan yang diperkirakan terjadi secara kebetulan (*expected agreement*) hasil analisis ChatGPT dengan hasil analisis ahli bahasa Indonesia (*gold standard*) didasarkan pada total baris kolom dan total kolom per jenis kalimat yang tertulis di tabel 6.

Tabel 6 Data Total Baris dan Total Kolom per Jenis Kalimat
Hasil Analisis ChatGPT dan *Gold Standard*

Nomor	Jenis Kalimat	Total Baris	Total Kolom
1	Kalimat Tunggal	34	26
2	Kalimat Majemuk Setara	46	47

3	Kalimat Majemuk Bertingkat	59	66

Berdasarkan data dalam tabel 6, proporsi kesepakatan yang diperkirakan terjadi secara kebetulan (*expected agreement*) adalah 0,3592. Selain itu, hasil analisis koefisien Cohen's Kappa (κ) tertulis di tabel 7 yaitu 0,843.

Tabel 7 Symmetric Measures ChatGPT dan *Gold Standard*

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	,843	,040	13,818	,000
N of Valid Cases		139			

Berdasarkan hasil Po, Pe, dan Cohen's Kappa, dapat disimpulkan hasil klasifikasi kesepakatan hasil analisis kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat antara ChatGPT dengan hasil analisis ahli bahasa Indonesia (*gold standard*) hampir sempurna. Hal itu didasarkan pada hasil Cohen's Kappa (κ) sebesar 0,843.

2. Tingkat Kesesuaian dan Interpretasi Kesepakatan dari Hasil Analisis Kalimat dari Gemini dan *Gold Standard*

Dari hasil analisis menggunakan SPSS, data pembandingan antara Gemini dan hasil analisis ahli bahasa Indonesia sebagai *gold standard* tertulis di Tabel 8.

Tabel 8. Manusia * Gemini Crosstabulation

			Gemini			Total
			Kalimat tunggal	Kalimat Majemuk Setara	Kalimat Majemuk Bertingkat	
Manusia	Kalimat tunggal	Count	31	0	3	34
		% within Manusia	91,2%	0,0%	8,8%	100,0%
		% within Gemini	52,5%	0,0%	4,4%	24,5%
	Kalimat Majemuk Setara	Count	28	12	6	46
		% within Manusia	60,9%	26,1%	13,0%	100,0%
		% within Gemini	47,5%	100,0%	8,8%	33,1%
	Kalimat	Count	0	0	59	59

	Majemuk Bertingkat	% within Manusia	0,0%	0,0%	100,0%	100,0%
		% within Gemini	0,0%	0,0%	86,8%	42,4%
Total		Count	59	12	68	139
		% within Manusia	42,4%	8,6%	48,9%	100,0%
		% within Gemini	100,0%	100,0%	100,0%	100,0%

Berdasarkan tabel 8, proporsi kesepakatan yang diamati (*observed agreement*) hasil analisis Gemini dengan hasil analisis ahli bahasa Indonesia (*gold standard*) didasarkan pada pembagian antara jumlah kesepakatan dan jumlah seluruh data. Ditemukan bahwa jumlah kalimat yaitu 139, kalimat tunggal yaitu 31, kalimat majemuk setara 12, dan kalimat majemuk bertingkat yaitu 59. Berdasarkan hal tersebut, nilai Po adalah 0,7338.

Selain itu, berdasarkan tabel 8, proporsi kesepakatan yang diperkirakan terjadi secara kebetulan (*expected agreement*) hasil analisis ChatGPT dengan hasil analisis ahli bahasa Indonesia (*gold standard*) didasarkan pada total baris kolom dan total kolom per jenis kalimat yang tertulis di tabel 9.

Tabel 9 Data Total Baris dan Total Kolom per Jenis Kalimat
Hasil Analisis Gemini dan *Gold Standard*

Nomor	Jenis Kalimat	Total Baris	Total Kolom
1	Kalimat Tunggal	34	59
2	Kalimat Majemuk Setara	46	12
3	Kalimat Majemuk Bertingkat	59	68

Berdasarkan data dalam tabel 9, proporsi kesepakatan yang diperkirakan terjadi secara kebetulan (*expected agreement*) adalah 0,3400. Selain itu, hasil analisis koefisien Cohen's Kappa (κ) tertulis di tabel 10 yaitu 0,597.

Tabel 10 Symmetric Measures Gemini dan *Gold Standard*

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement Kappa	,597	,049	10,806	,000
N of Valid Cases	139			

Berdasarkan hasil Po, Pe, dan Cohen’s Kappa, dapat disimpulkan hasil klasifikasi kesepakatan hasil analisis kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat antara Gemini dengan hasil analisis

ahli bahasa Indonesia (*gold standard*) tergolong sedang. Hal itu didasarkan pada hasil Cohen's Kappa (κ) sebesar 0,597.

Performa ChatGPT serta Gemini dalam Menganalisis Kalimat Tunggal, Kalimat Majemuk Setara, dan Kalimat Majemuk Bertingkat

1. Performa ChatGPT dalam Menganalisis Kalimat Tunggal, Kalimat Majemuk Setara, dan Kalimat Majemuk Bertingkat

Berdasarkan tabel 5, nilai akurasi, *macro precision*, *macro recall*, dan *macro F1-score* performa ChatGPT dalam menganalisis jenis kalimat tunggal, kalimat majemuk setara, serta kalimat majemuk bertingkat tertulis dalam tabel 11.

Tabel 11. Nilai *Macro Precision*, *Macro Recall*, dan *Macro F1-score* Performa ChatGPT

Jenis Kalimat	Precision	Recall	F1-score
Kalimat Tunggal	100.00%	76.47%	86.67%
Kalimat Majemuk Setara	87.23%	89.13%	88.17%
Kalimat Majemuk Bertingkat	87.88%	98.31%	92.80%
Macro Average	91.70%	87.97%	89.21%
Akurasi	89,93%	-	-

Berdasarkan tabel 5 dan 11, ChatGPT berhasil untuk menggolongkan 125 dari 139 pola kalimat yang sesuai dengan anotasi hasil analisis kalimat yang dilakukan oleh ahli bahasa Indonesia. Nilai akurasinya sejumlah 89,93%. Selain itu, prediksi ketepatan hasil analisis yang dihasilkan ChatGPT tergolong tinggi dengan bukti nilai *macro precision* sejumlah 91,70%. Sementara itu, *macro recall* sejumlah 87,97% mengindikasikan bahwa hampir semua pola kalimat yang menjadi acuan oleh ahli bahasa Indonesia berhasil untuk dikenali secara tepat oleh ChatGPT. Kombinasi kedua metrik tersebut menghasilkan *macro F1-score* sejumlah 89,21%, yang menunjukkan keseimbangan yang baik antara ketepatan dan kelengkapan klasifikasi.

2. Performa Gemini dalam Menganalisis Kalimat Tunggal, Kalimat Majemuk Setara, dan Kalimat Majemuk Bertingkat

Berdasarkan tabel 8, nilai akurasi, *macro precision*, *macro recall*, dan *macro F1-score* performa ChatGPT dalam menganalisis jenis kalimat tunggal, kalimat majemuk setara, serta kalimat majemuk bertingkat tertulis dalam tabel 12.

Tabel 12. Nilai *Macro Precision*, *Macro Recall*, dan *Macro F1-score*

Performa Gemini

Jenis Kalimat	Precision	Recall	F1-score
Kalimat Tunggal	52,54%	91,18%	66,67%
Kalimat Majemuk Setara	100%	26,09%	41,38%
Kalimat Majemuk Bertingkat	86,76%	100,00%	92,91%
Macro Average	79,77%	72,42%	66,99%
Akurasi	73,38%	-	-

Berdasarkan tabel 8 dan 12, Gemini berhasil untuk menggolongkan 102 dari 139 pola kalimat yang sesuai dengan anotasi hasil analisis kalimat yang dilakukan oleh ahli bahasa Indonesia. Nilai akurasinya sejumlah 73,88%%. Selain itu, prediksi ketepatan hasil analisis yang dihasilkan gemini cukup tinggi dengan bukti nilai *macro precision* sejumlah 79,77%. Akan tetapi, *macro recall* sejumlah 72,42% mengindikasikan bahwa Gemini belum merata saat mengenali pola kalimat, terutama pada kategori kalimat majemuk setara yang *recall*-nya hanya 26,09%. Hal tersebut berpengaruh pada nilai *macro F1-score* sejumlah 66,99%, yang menunjukkan keseimbangan dan kelengkapan klasifikasi tergolong sedang.

Pembahasan Klasifikasi dan Performa ChatGPT dan Gemini dalam Menganalisis Kalimat dengan Analisis Ahli Bahasa Indonesia sebagai Gold Standard

Berdasarkan hasil penelitian, ChatGPT memiliki tingkat kesepakatan dan performa yang lebih tinggi dibandingkan dengan Gemini dalam menganalisis pola kalimat yang menghasilkan kalimat tunggal, kalimat majemuk setara, dan kalimat majemuk bertingkat. Pada kalimat tunggal, ChatGPT memperoleh *macro precision* sejumlah 100%, sedangkan Gemini hanya 52,54%. Hal ini menunjukkan bahwa hampir setengah klasifikasi dari Gemini yang kurang tepat dalam menganalisis pola kalimat tunggal. Hal itu disebabkan karena Gemini menghasilkan 59 prediksi kalimat tunggal, padahal hanya 31 prediksi yang tepat sebagai kalimat tunggal berdasarkan hasil analisis ahli bahasa Indonesia, sedangkan 28 diprediksi sebagai kalimat majemuk setara. Kekurangtepatan prediksi tersebut disebabkan karena ambiguitas struktural dan konjungsi dalam frasa yang menjadi pengecoh Gemini dalam menganalisis kalimat tunggal (Li et al., 2024).

Penanda konjungsi subordiatif serta penggolongan induk kalimat dan anak kalimat menjadi dasar analisis kalimat majemuk bertingkat (Owens et al., 2024). Dua penanda tersebut cenderung mudah dikenali. ChatGPT dan Gemini memilki performa yang baik saat menganalisis kalimat majemuk bertingkat. ChatGPT berhasil mengklasifikasikan 58 dari 59 kalimat majemuk bertingkat dengan benar sehingga memperoleh *macro recall* sejumlah 98,31%. Dengan kata lain, hanya terdapat 1 kalimat majemuk bertingkat yang kurang tepat dianalisis oleh ChatGPT. Sementara itu, Gemini berhasil

dalam mengklasifikasikan seluruh 59 kalimat majemuk bertingkat dengan benar sehingga memperoleh *recall* sejumlah 100,00%. Hal itu menunjukkan bahwa ChatGPT dan Gemini lebih mudah mengenali kalimat majemuk bertingkat dibandingkan dengan kalimat tunggal dan kalimat majemuk setara. Kemudahan tersebut disebabkan karena ChatGPT dan Gemini lebih mengenali penanda konjungsi subordinatif dibandingkan konjungsi koordinatif (Chen et al., 2021).

PENUTUP

KESIMPULAN

Dapat disimpulkan bahwa ChatGPT memiliki tingkat kesesuaian dan performa yang lebih baik dibandingkan Gemini dalam menganalisis pola kalimat pada teks akademik bahasa Indonesia. Hasil dari ChatGPT dan Gemini dibandingkan dengan hasil analisis ahli bahasa Indonesia sebagai *gold standard*. Berdasarkan uji Cohen's Kappa, ChatGPT memperoleh nilai κ sejumlah 0,843 yang termasuk kategori hampir sempurna, sedangkan Gemini memperoleh nilai κ sejumlah 0,597 yang termasuk kategori sedang. Dapat diartikan bahwa klasifikasi ChatGPT lebih konsisten dan lebih mendekati hasil analisis ahli bahasa Indonesia.

Evaluasi performa klasifikasi juga memperlihatkan keunggulan ChatGPT dengan akurasi sejumlah 89,93%, *macro precision* 91,70%, *macro recall* 87,97%, dan *macro F1-score* 89,21%. Sebaliknya, Gemini memperoleh akurasi 73,38%, *macro precision* 79,77%, *macro recall* 72,42%, dan *macro F1-score* 66,99%. Walaupun demikian, ChatGPT dan Gemini sangat baik dalam menganalisis atau mengenali kalimat majemuk bertingkat. Hal itu dibuktikan dengan *recall* ChatGPT mencapai 98,31% dan Gemini mencapai 100,00%. Dengan demikian, ChatGPT terbukti lebih andal sebagai alat bantu analisis pola kalimat bahasa Indonesia, sedangkan Gemini masih memerlukan peningkatan, terutama dalam membedakan kalimat tunggal dan kalimat majemuk setara yang memiliki karakteristik struktural serupa.

DAFTAR PUSTAKA

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2025). LLMs instead of Human Judges ? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, *2*, 238–255. https://doi.org/10.18653/v1/2025.acl-short.20

Chen, X., Alexopoulou, T., & Tsimpli, I. (2021). Automatic Extraction of Subordinate Clauses and Its Application in Second Language Acquisition Research. *Behavior Research Methods*, *53*(2), 803–817. https://doi.org/10.3758/s13428-020-01456-7

Damayanti, S., & Zulkarnain. (2026). Analisis Efektivitas Large Language Model (LLM) sebagai Asisten Praktikum pada Mata Kuliah Jaringan Komputer. *Klik - Jurnal Ilmu Komputer*, *7*(1), 30–38. https://doi.org/10.56869/klik.v7i1.778

Dewi Marhamah Syadli, & Dewi Anggraini. (2023). Kalimat Efektif dalam Teks Eksplanasi Siswa Kelas XI SMA Negeri 5 Sijunjung. *Jurnal Pendidikan, Bahasa Dan Budaya*, *2*(2), 01–11. https://doi.org/10.55606/jpbb.v2i2.1350

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance. *British Journal of Educational Technology*, *56*(2), 489–530. https://doi.org/10.1111/bjet.13544

J Richard Landis, & Gary G Koch. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/https://doi.org/10.2307/2529310

Kiranawati, B. I., Latifah, U., & Khusnah, N. K. (2025). Pemanfaatan Model AI dalam Analisis Sintaksis Bahasa Indonesia: Kajian Linguistik Komputasional. *JURNAL PROGRESIF*, *3*, 7–26. https://journal.univgresik.ac.id/index.php/progresif/article/view/443

Kurniahtunnisa, Manuel, M. Y., Aini, M., & Agustina, T. P. (2025). Persepsi dan Sikap Siswa terhadap Penggunaan Artificial Intelligence. *Scholaria: Jurnal Pendidikan Dan Kebudayaan*, *15*(1), 47–59. https://doi.org/10.24246/j.js.2025.v15.i1.p47-59

Li, M. Y., Liu, A., Wu, Z., & Smith, N. A. (2024). *A Taxonomy of Ambiguity Types for NLP*. https://doi.org/https://doi.org/10.48550/arXiv.2403.14072

Nainggolan, K. (2025). Sintakmatik dalam Analisis Struktur Kalimat Bahasa Indonesia. *JBSI: Jurnal Bahasa Dan Sastra Indonesia*, *5*(01), 215–222. https://doi.org/10.47709/jbsi.v5i01.6247

Nur Gimas Tiaroh, Lutfi Anifatun Wahidah, Yuwen Nurkhalifah, Atikah Nuha Hibatullah, Isnaini Muyassaroh, Asep Purwo Yudi Utomo, Rossi Galih Kesuma, & Zulfa Fahmy. (2026). Konjungsi Koordinatif dan Konjungsi Subordinatif dalam Teks Berita Kompas Terbitan September 2025. *Dinamika Pembelajaran : Jurnal Pendidikan Dan Bahasa*, *3*(1), 98–120. https://doi.org/10.62383/dilan.v3i1.2909

Owens, R. E., Pavelko, S. L., & Hahs-Vaughn, D. (2024). Growth of Complex Syntax: Coordinate and Subordinate Clause Use in Elementary School–Aged Children. *Language, Speech, and Hearing Services in Schools*, *55*(3), 714–723. https://doi.org/10.1044/2024_LSHSS-23-00102

Rosyada, A., Putri, A. N., Irawati, Y., Sari, R. R., Ghifari, I. Al, Utomo, A. P. Y., & Kesuma, R. G. (2025). Analisis Jenis Konjungsi Koordinatif pada Artikel Pendidikan dan Edukasi dalam Website kelasjuara.id Edisi Oktober 2024 sebagai Bahan Ajar Siswa SMA. *PUSTAKA: Jurnal Bahasa Dan Pendidikan*, *5*(3), 22–46. https://doi.org/https://doi.org/10.56910/pustaka.v5i3.3262

Suganda, A. (2023). Memilih AI yang Tepat untuk Guru: Perbandingan Fitur Gemini, ChatGPT, dan Claude A. *Jurnal Inovasi Teknologi Dan Edukasi Teknik*, *3*(11), 2023. https://doi.org/10.17977/um084.v3.i11.2023.2

Wahyuni, R. N., Suryani, I., & Warni, W. (2024). Pemanfaatan Gemini dalam Pembelajaran Bahasa Indonesia di SMA: Studi Literatur. *Diskursus: Jurnal Pendidikan Bahasa Indonesia*, *7*(3), 446. https://doi.org/10.30998/diskursus.v7i3.26571