Smart Comp: Jurnalnya Orang Pintar Komputer
Volume 15, Nomor 1, Januari 2026, hlm. 232-239
Terakreditasi Sinta, Peringkat 4, SK No. 105/E/KPT/2022
DOI : 10.30591/smartcomp.v13i1.9364

P-ISSN : 2089-676X
E-ISSN : 2549-0796

Evaluasi Terhadap Layanan LLM Dalam Pembuatan Soal Berbasis HOTS

Gusti Ghalib Ghazi[*1], Andhik Budi Cahyono[2]
[1)]Teknik Informatika, Fakultas Teknik Industri, Universitas Islam Indonesia
Email: [*1]gusti.ghalib@gmail.com, [2]andhikbudi@uii.ac.id

(Naskah masuk: *21 Juli 2025*, diterima untuk diterbitkan: *20 januari 2026*)

Abstrak: *Evaluasi ini memiliki tujuan untuk menguji kemampuan layanan LLM (Large Language Model) seperti ChatGPT, Gemini, dan Deepseek dalam menghasilkan soal bertipe HOTS (High Order Thinking Skills). Soal tipe HOTS ini memiliki peran dan dampak yang penting dalam meningkatkan kemampuan berpikir kritis. Kemampuan LLM dalam menghasilkan soal HOTS sesuai dengan prompt dan teks masukan perlu diteliti untuk mengetahui efektivitas dan kualitas dari layanan tersebut dilakukan dalam beberapa tahap termasuk penentuan LLM, pembuatan soal, dan penilaian soal. Hasil evaluasi yang telah dilakukan menunjukkan temuan yang signifikan. Secara keseluruhan, ketiga layanan LLM yang menjadi subjek penelitian, yaitu ChatGPT, Google Gemini, dan Deepseek AI, terbukti dapat menghasilkan soal yang sesuai dengan karakteristik tipe soal HOTS serta telah memenuhi syarat-syarat yang ditetapkan dalam kerangka Taksonomi Bloom. Namun ditemukan bahwa salah satu diantara layanan tersebut mampu menghasilkan soal namun dengan kualitas yang tidak sebaik lawannya. Temuan ini mengindikasikan bahwa LLM memiliki potensi besar untuk menjadi alat bantu yang efisien bagi para pendidik dalam mengembangkan instrumen penilaian yang berkualitas.*

Kata Kunci – *LLM; ChatGPT; pertanyaan; HOTS; taksonomi*

Evaluation of LLM Services in Creating HOTS-Based Questions

Abstract: *This evaluation aims to test the ability of Large Language Model (LLM) services such as ChatGPT and other services to generate HOTS (High Order Thinking Skills) questions. These HOTS-type questions play a significant role and impact in improving critical thinking skills. The LLM's ability to generate HOTS questions based on prompts and input text needs to be examined to determine the service's effectiveness and quality. This evaluation was conducted in several stages, including LLM determination, question creation, and question assessment. The evaluation results revealed significant findings. Overall, the three LLM services studied – ChatGPT, Google Gemini, and Deepseek AI – were proven to generate questions that align with the characteristics of HOTS questions and met the requirements set out in the Bloom's Taxonomy framework. However, one of these services was found to be capable of generating questions, but the quality was not as good as its competitors. These findings indicate that LLM has great potential to be an efficient tool for educators in developing high-quality assessment instruments.*

Keywords – *LLM; ChatGPT; questions; HOTS; Taxsonomy*

1. PENDAHULUAN

Penggunaan teknologi pengetahuan dan penyampaian pesanseperti smartphone dan komputer memengaruhi beberapa aspek kehidupan, salah satunya adalah pendidikan. Perubahan tersebut terjadi baik dalam perencanaan, pelaksanaan, evaluasi, dan monitoring dengan tujuan meningkatkan kualitas pendidikan agar bisa bersaing secara global. Salah satu upaya untuk meningkatkan kualitas pendidikan adalah dengan mengasah kemampuan berpikir kritis dan kreatif melalui HOTS (High Order Thinking Skills). Implementasi HOTS atau pada kurikulum belajar saat ini diharapkan mampu memperbaiki sistem pendidikan demi menciptakan generasi masa depan yang berkarakter dan menciptakan generasi yang unggul dan mampu bersaing di dunia internasional.

HOTS adalah suatu proses berpikir peserta didik dalam level kognitif yang lebih tinggi yang dikembangkan dari berbagai konsep dan metode kognitif dan taksonomi pembelajaran seperti

metode problem solving, taksonomi bloom, dan taksonomi pembelajaran, pengajaran, dan penilaian. Tujuan utama dari HOTS adalah bagaimana meningkatkan kemampuan berpikir peserta didik pada level yang lebih tinggi, terutama yang berkaitan dengan kemampuan untuk berpikir secara kritis dalam menerima berbagai jenis informasi, berpikir kreatif dalam memecahkan suatu masalah menggunakan pengetahuan yang dimiliki, serta membuat keputusan dalam situasi-situasi yang kompleks. HOTS (High Order Thinking Skills) dikemukakan oleh Susan M Brookhart, seorang penulis sekaligus Assosiate Professor dari Dusquance University. Dalam bukunya (Brookhart, 2010)didefinisikan model ini sebagai metode untuk transfer pengetahuan, berpikir kritis, dan memecahkan masalah. Tak sekedar model soal, HOTS juga termasuk didalamnya model pengajaran yang mencakup kemampuan berpikir, contoh, pengaplikasian pemikiran dan diadaptasikan dengan kebutuhan siswa yang berbeda-beda. (mukrodin & Mega Sasmita, 2021)

Salah satu bagian dari kemajuan teknologi saat ini adalah dikenalkannya kecerdasan buatan atau Artificial Intelligence (AI). AI mengacu kepada ilmu dan rekayasa untuk menciptakan sistem yang mampu melakukan tugas-tugas yang umumnya terkait dengan makhluk cerdas seperti pembelajaran, penilaian, dan pengambilan keputusan. Salah satu hasil dari pengembangan AI tersebut adalah dikenalkannya sebuah asisten virtual atau chatbot yang mampu memahami bahasa manusia dan menghasilkan informasi yang dapat dimengerti oleh manusia. Secara umum, tugas chatbot adalah melakukan tugas-tugas seperti seorang pelayan seperti menyapa, menjawab, dan mematuhi permintaan. Salah satu teknologi chatbot yang mulai digunakan oleh masyarakat sekarang adalah 'Chat Generative Pre-trained Transformer' atau ChatGPT. Dikenalkan pada November 2022, teknologi ini dikembangkan oleh OpenAI dan dilatih pada kumpulan data besar percakapan manusia yang memungkinkannya untuk melakukan tugas-tugas kompleks dan menghasilkan respons seperti manusia. (mukrodin & Mega Sasmita, 2021)

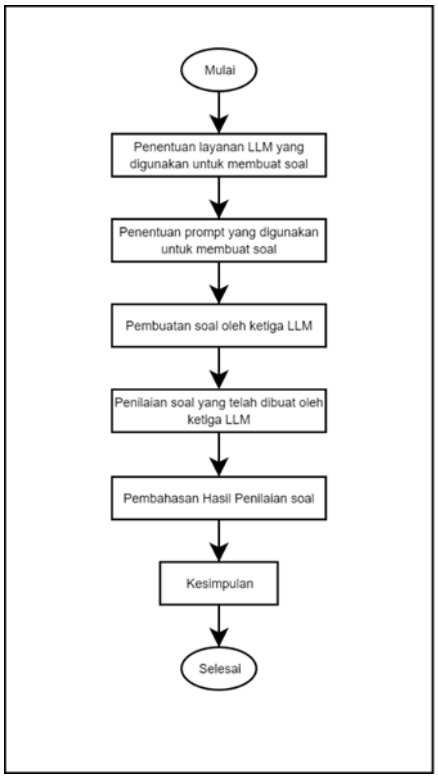
Dalam dunia pendidikan, penggunaan ChatGPT menjadi hal baru yang menjadikan proses belajar dan mengajar menjadi berbeda dengan sebelum digunakannya ChatGPT. Kemampuan dari teknologi ini membantu murid dan guru untuk mendapatkan informasi dan pengetahuan secara lebih cepat dan efisien dibandingkan sebelumnya. Penelitian-penelitian yang menggunakan teknologi ini juga mulai bermunculan seperti penelitian yang dilakukan oleh (Zhai, 2023). Pada penelitian tersebut didapatkan bahwa jurnal yang dibuat oleh ChatGPT bersifat lebih koheren, akurat, dan dapat diselesaikan hanya dalam waktu tiga jam saja.

Dengan meningkatnya popularitas dari layanan ChatGPT, ditambah dengan mulai banyaknya pengguna yang membutuhkan layanan yang serupa, muncullah layanan-layanan Large Language Model lain dari perusahaan-perusahaan teknologi lama maupun baru, seperti Google Gemini, Anthropic Claude, Microsoft Copilot, hingga Meta AI. Beberapa dari layanan baru tersebut bahkan berjalan pada bidang-bidang yang lebih spesifik seperti Blackbox AI dan Windsurf dalam bidang programming, Toolsaday dan PrepAI dalam bidang pendidikan, dan layanan-layanan lainnya.

Dari penjelasan yang telah dipaparkan, dilakukan sebuah penelitian yang ditujukan untuk mengevaluasi dan membandingkan layanan ChatGPT dengan layanan-layanan baru yaitu: Gemini dan Deepseek dalam konteks pembuatan soal yang sesuai dengan prinsip soal HOTS. Dengan melakukan perbandingan ini, akan diketahui sejauh mana teknologi ini dapat mengasilkan soal yang mampu mengukur keterampilan berpikir tingkat tinggi peserta didik. Evaluasi ini juga menjadi identifikasi perbedaan kemampuan antar layanan teknologi ini dan juga karakter dari hasil keluaran layanan-layanan ini dan memperkaya kajian tentang pemanfaatan layanan AI untuk pendidikan.

2. METODE PENELITIAN

Penelitian ini dilakukan dalam beberapa tahap. Tahap-tahap tersebut dapat dilihat pada gambar 1.



Gambar 1. Metode Penelitian

2.1. Penentuan Layanan LLM Untuk Membuat Soal

Dalam evaluasi ini, akan digunakan dua layanan LLM, selanjutnya disebut sebagai layanan, sebagai pembanding ChatGPT, layanan tersebut adalah Google Gemini dan Deepseek AI. Tiga layanan ini nantinya akan bertugas membuat soal dan saling menilai soal. Kemudian akan ada layanan keempat yang bertugas hanya menilai soal hasil ketiga layanan lainnya. Alasan pemilihan layanan tersebut adalah sebagai berikut.

1. Google Gemini dipilih karena layanan ini berasal dari perusahaan teknologi yang sudah lama ada di pasaran yaitu Google. Google menyatakan bahwa Gemini milik mereka lebih unggul 49% dibandingkan dengan ChatGPT.
2. Deepseek AI dipilih karena layanan ini meruapakan salah satu layanan dari Tiongkok yang dapat bersaing seimbang dengan ChatGPT. Menurut Deepseek sendiri, layanan mereka dapat berjalan seimbang dengan ChatGPT dengan menggunakan daya yang lebih kecil sehingga lebih efisien dibandingkan dengan ChatGPT.

Selain ketiga layanan tersebut akan digunakan layanan lain sebagai evaluator soal-soal yang dihasilkan oleh ketiga layanan lainnya, yaitu Copilot AI. Copilot AI dipilih layanan ini berasal dari perusahaan teknologi yang cukup lama berada di pasaran yaitu Microsoft. Diperkenalkan pada Januari 2024, layanan ini langsung mendapat sorotan dari publik mengingat Microsoft sudah memiliki reputasi yang baik dan kompeten di bidang teknologi dan informasi.

2.2. Penentuan Prompt yang Digunakan Untuk Membuat Soal

Prompt adalah teks masukan atau instruksi yang diberikan kepada ChatGPT untuk mendapatkan jawaban yang diinginkan. Dalam jurnalnya, (Mukhlis, 2024) mengemukakan bahwa dari hasil wawancara, diperoleh prompt yang digunakan untuk membuat soal sebagai berikut:

"Kamu akan berperan sebagai guru Bahasa Indonesia tingkat SMA/SMK yang akan melaksanakan assesment. Mulai sekarang kamu membuat soal objektif Bahasa Indonesia dengan dimensi kognitif C5, topiknya tentang cerita rakyat dan pastikan cuplikan cerita yang disajikan 2-3 paragraf"

Dalam prompt tersebut, ada istilah yang membuat prompt tersebut menghasilkan soal bertipe HOTS. Antara lain istilah "dimensi kognitif C5" dimana prompt tersebut akan menghasilkan soal yang menguji kemampuan siswa dalam membuat penilaian atau keputusan.

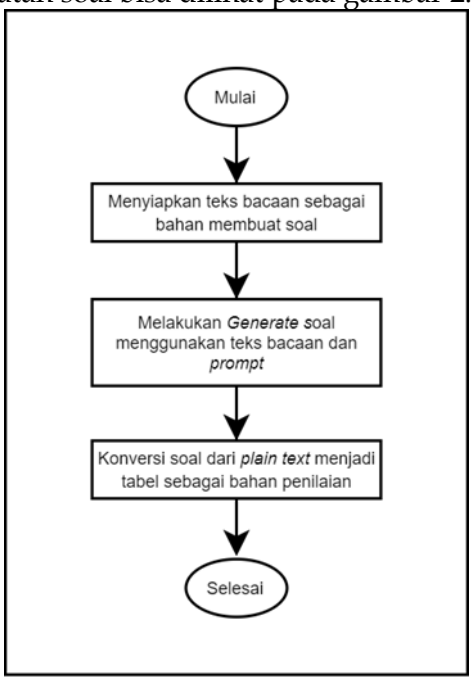
Dalam penelitian ini, prompt akan disederhanakan menjadi berikut:

"Buat 3 soal pilihan ganda objektif dengan dimensi kognitif C5, topiknya tentang dokumen yang telah diunggah"

Hasil prompt menunjukkan bahwa penyederhanaan tersebut menghasilkan output soal-jawaban yang relatif sama. Penyederhanaan soal menjadi penting karena bisa mengurangi penggunaan token. Hal ini penting dilakukan mengingat hasil penilitian ini akan digunakan untuk penelitian selanjutnya yaitu pengembangan aplikasi generate soal HOTS berbasis API dari layanan yang dievaluasi pada penelitian ini.

2.3. Pembuatan Soal oleh Ketiga LLM

Langkah-langkah pembuatan soal bisa dilihat pada gambar 2.



Gambar 2. Alur Pembuatan Soal

Dari Gambar 2, proses pertama adalah menyiapkan teks bacaan yang akan digunakan sebagai bahan utama dalam pembuatan soal. Teks bacaan yang digunakan adalah bacaan cerita rakyat yang biasa dibaca oleh siswa sekolah dasar. Teks bacaan didapat dari beberapa situs di internet. Ada sepuluh cerita rakyat yang memiliki variasi genre, panjang, dan judul yang akan digunakan dalam evaluasi ini.

Kemudian, masing-masing layanan diminta menghasilkan soal sesuai dengan prompt yang telah ditentukan sebelumnya.

Prompt yang nanti diujikan akan disertakan dengan sebuah dokumen teks yang berisi cerita rakyat yang telah dikumpulkan sebelumnya. Setiap layanan LLM akan diuji dengan sepuluh dokumen teks cerita rakyat dengan satu prompt instruksional, sehingga dalam evaluasi ini akan dilakukan 30 kali prompting dengan menggunakan ChatGPT, Gemini, dan Deepseek. Keluaran dari setiap proses prompting adalah 3 soal pilihan ganda dengan spesifikasi sesuai dengan prompt dan teks bacaan yang digunakan dalam format plain text.

Kemudian, proses selanjutnya adalah melakukan konversi dan pengelompokan soal yang telah dihasilkan kedalam tabel. Didalam tabel tersebut berisi soal-soal yang dihasilkan dari ketiga layanan terhadap satu dokumen teks bacaan. Tabel terdiri dari beberapa kolom, kolom pertama yaitu "LLM" berfungsi sebagai penanda untuk identifikasi layanan mana yang membuat soal tersebut, kolom sebelahnya yaitu "Pertanyaan" diikuti empat kolom disebelahnya lagi yaitu

"Pilihan a", "Pilihan b", Pilihan c", dan "Pilihan d" menampilkan opsi jawaban yang menyertai pertanyaan dan kolom terakhir yaitu "Jawaban" berisi jawaban dari pertanyaan. Setiap tabel memuat soal-soal yang dihasilkan terhadap satu dokumen teks bacaan sehingga dibuat sepuluh tabel yang nantinya digunakan dalam proses penilaian soal. Format tabel dapat dilihat pada Gambar 3.

LLM	pertanyaan	pilihan a	pilihan b	pilihan c	pilihan d	jawaban
LLM 1						
LLM 1						
LLM 1						
LLM 2						
LLM 2						
LLM 2						
LLM 3						
LLM 3						
LLM 3						

Gambar 3. Tabel Konversi Soal

2.4. Penilaian Soal

Setiap soal yang dihasilkan akan dinilai menggunakan mekanisme penilaian silang (cross evaluation), di mana setiap soal yang telah dibuat akan dinilai oleh model bahasa lainnya. Mekanisme ini menjadikan setiap model berperan ganda, baik sebagai penghasil soal maupun sebagai evaluator.

Secara rinci, seluruh soal yang telah dikonversi dalam sebuah tabel akan melalui proses evaluasi oleh Gemini, DeepSeek, dan ChatGPT. Prosedur berlaku hingga semua sepuluh tabel yang dibuat selesai dievaluasi. Penilaian final kemudian dilakukan oleh Copilot, sebuah model yang sengaja tidak diikutsertakan dalam tahap pembuatan soal untuk menjaga objektivitas. Proses evaluasi ini berpedoman pada kerangka acuan (prompt) dengan format seperti di bawah ini:

"Dokumen berikut berisi soal-soal yang dihasilkan oleh tiga llm terhadap perintah membuat soal pilihan ganda objektif dengan dimensi kognitif C5, berilah penilaian dari masing-masing llm berdasarkan keakuratan pertanyaan, pilihan, dan jawaban dengan dimensi kognitif C5 dalam taksonomi bloom, beri nilai antara 0 sampai 100"

Setiap nilai skor yang didapat kemudian disusun secara sistematis kedalam dua jenis tabel yang memudahkan dalam proses analisis dan perbandingan. Tabel pertama berisi nilai yang diterima oleh setiap layanan dari para penilai. Tabel ini memungkinkan analisis mendalam terhadap bagaimana kualitas soal dari satu model dipandang oleh model lainnya. Format tabel ini dapat dilihat pada gambar 4.

LLM 1 : ChatGPT				
	A	B	C	D
Penilai	ChatGPT	Gemini	Deepseek	Copilot
Dokumen 1				
Dokumen 2				
Dokumen 3				
Dokumen 4				
Dokumen 5				
Dokumen 6				
Dokumen 7				
Dokumen 8				
Dokumen 9				
Dokumen 10				
Nilai Total				
Rata-Rata				
Rata-Rata Total				

Gambar 4. Tabel Nilai Setiap LLM

Setelah semua model selesai dimasukkan dan dinilai ke dalam tabel pertama, langkah berikutnya adalah merekapitulasi seluruh hasil. Data ini kemudian diolah ke tabel kedua yang

menampilkan perbandingan performa akhir antar layanan. Tabel ini menggunakan nilai rata-rata dari setiap model sebagai metrik perbandingan utama untuk menentukan kualitas keseluruhan secara komparatif. Format tabel ini dapat dilihat pada gambar 5.

	LLM 1 : ChatGPT	LLM 2 : Gemini	LLM 3 : Deepseek
Nilai Total			
Rata-Rata			
Rata-Rata Total			
Peringkat			

Gambar 5. Tabel Nilai Total.

3. HASIL DAN PEMBAHASAN

Hasil penilaian yang telah diringkas dalam dua jenis tabel akan ditampilkan dan dilakukan analisis dalam bab ini. Analisis difokuskan pada perbandingan performa antar model dan pola-pola menarik yang muncul selama proses evaluasi. Analisis pertama dilakukan terhadap tabel jenis pertama pada Tabel 1, Tabel 2, dan Tabel 3, kemudian analisis tabel jenis kedua pada Tabel 4.

Tabel 1. Nilai ChatGPT

ChatGPT				
Penilai	**ChatGPT**	**Gemini**	**Deepseek**	**Copilot**
Nilai Total	854	945	883	787
Rata-Rata	85,4	94,2	88,3	78,7
Rata-Rata Total	86,7			

Dari Tabel 1 didapatkan bahwa ChatGPT secara keseluruhan dapat menghasilkan soal yang sesuai dengan permintaan yaitu dimensi kognitif C5 dalam Taksonomi Bloom dengan rata-rata total yaitu 86,7. Dalam tabel juga didapatkan bahwa ChatGPT sebagai penilai dari soal yang dihasilkan oleh ChatGPT tidak menilai soal dengan nilai yang sangat tinggi yang berarti ChatGPT tidak bersifat bias terhadap keluaran yang dihasilkan oleh layanan tersebut.

Analisis dari keempat LLM menyatakan bahwa ChatGPT ini menunjukkan pemahaman yang paling kuat dan akurat mengenai dimensi kognitif C5 dengan pertanyaan dan jawaban yang konsisten menantang evaluasi serta paling akurat dengan soal dan jawaban yang orisinal, mendalam, dan relevan. Mayoritas soal yang dihasilkan secara konsisten meminta peserta didik untuk membuat penilaian, memilih solusi, atau menilai tindakan berdasarkan kriteria etika dan tanggung jawab.

Namun, ChatGPT juga beberapa kali melakukan kesalahan. Salah satu analisis menyebutkan bahwa dari tiga pertanyaan yang ada, dua di antaranya secara akurat menargetkan level C5, namun satu pertanyaan sedikit meleset dan lebih condong ke level C4 (Analisis). Selain itu, ditemukan juga beberapa opsi jawaban yang melakukan pengulangan kalimat sehingga berpotensi membingungan murid yang mengerjakan soal.

Tabel 2. Nilai Gemini

Gemini				
Penilai	**ChatGPT**	**Gemini**	**Deepseek**	**Copilot**
Nilai Total	864	883	870	730
Rata-Rata	86,4	88,3	87	73
Rata-Rata Total	83,7			

Dari Tabel 2 didapatkan bahwa Gemini, sama seperti ChatGPT juga dapat menghasilkan soal yang sesuai yang terlihat dari rata-rata total yang diatas 80. Namun dalam tabel ini mulai terlihat sedikit kasus bias yang dimana Gemini sebagai penilai memberi penilaian yang lebih tinggi dibandingkan penilai lainnya.

Analisis keempat LLM terhadap Gemini ini juga kebanyakan positif, dinyatakan bahwa layanan ini dinilai sangat baik dan mampu secara konsisten serta brilian menghasilkan pertanyaan-pertanyaan yang tepat sasaran pada level C5. Setiap pertanyaan dirumuskan dengan menggunakan

kata kerja evaluatif yang kuat, seperti "evaluasi yang paling tepat", "bagaimana Anda menilai konsekuensi…", dan "pembenaran yang paling logis", yang secara eksplisit meminta adanya kritik atau justifikasi.

Namun, sama seperti ChatGPT, ada beberapa kesalahan yang juga dibuat oleh Gemini seperti beberapa soal yang dihasilkan cenderung mendekati level C4 (Menganalisis) karena lebih fokus pada pembedahan hubungan antar komponen daripada memberikan penilaian akhir sebagai kesimpulan. Ditambah, ditemukan soal yang bersifat repetitif, soal yang memiliki kesalahan ketik, dan soal yang memberikan pilihan jawaban yang tidak lengkap.

Tabel 3. Nilai Deepseek

Deepseek				
Penilai	**ChatGPT**	**Gemini**	**Deepseek**	**Copilot**
Nilai Total	724	355	833	655
Rata-Rata	72,4	35,5	83,3	65,5
Rata-Rata Total	64,2			

Dari Tabel 3 didapatkan bahwa walaupun deepseek mampu membuat soal sesuai dengan permintaan dengan rata-rata yang diatas 50, namun nilai tersebut lebih rendah dibandingkan dengan kedua LLM sebelumnya yang mampu menghasilkan nilai rata-rata diatas 80.

Analisis keempat LLM terhadap layanan ini berbeda dengan lawan-lawannya yang dimana keempat penilai sepakat manyatakan bahwa Deepseek ini belum berhasil memenuhi kriteria dimensi kognitif C5 secara konsisten. Beberapa dari mereka menyatakan bahwa mayoritas pertanyaan yang dihasilkan tidak mencapai tingkat evaluatif murni, melainkan berhenti pada level kognitif C1 hingga C4. Bahkan, dalam beberapa kasus, terjadi kegagalan di mana tidak ada satu pun pertanyaan dari sebuah teks bacaan yang berhasil mencapai dimensi C5.

Meskipun demikian, ada beberapa soal yang menunjukkan kemampuan dalam membuat soal level C5. Tetapi diantara soal-soal level C5 yang berhasil dibuat juga tidak memiliki kualitas yang sama dengan soal level C5 yang dihasilkan oleh lawan-lawannya. Analisis menunjukkan beberapa soal level C5 masih bersifat ambigu atau terlalu sederhana.

Tabel 4. Nilai Total

Penilai	**ChatGPT**	**Gemini**	**Deepseek**
Nilai Total	3469	3347	2567
Rata-Rata	86,7	83,7	64,2
Rata-Rata Total	78,2		
Peringkat	1	2	3

Pada tabel 4 ini semua nilai yang didapatkan dikumpulkan menjadi satu dan dihitung jumlah toal nilai dan rata-ratanya dan kemudian dari rata-rata tersebut diambil peringkat. Didapatkan bahwa ChatGPT memiliki kemampuan paling baik dibandingkan dengan lawannya dengan skor tertinggi, diikuti dengan Gemini dengan selisih yang sedikit. Kemudian Deepseek sebagai pemain baru memiliki skor yang paling rendah dan memiliki selisih yang besar..

4. KESIMPULAN

Secara keseluruhan, ketiga layanan LLM, yaitu ChatGPT, Google Gemini, dan Deepseek AI dapat menghasilkan soal yang sesuai dengan tipe soal HOTS dan memenuhi syarat Taksonomi Bloom. Kesimpulan setiap layanan LLM akan dijelaskan sebagai berikut.

1. ChatGPT, sebagai layanan LLM yang paling banyak digunakan dan sebagai LLM yang dibandingkan dalam evaluasi ini, dapat menghasilkan soal dengan skor paling baik. Menjadikan layanan ini sebagai rekomendasi dalam LLM dalam membuat soal bertipe HOTS.Deepseek AI dipilih karena layanan ini meruapakan salah satu layanan dari Tiongkok yang dapat bersaing seimbang dengan ChatGPT. Menurut Deepseek sendiri, layanan mereka dapat berjalan seimbang

dengan ChatGPT dengan menggunakan daya yang lebih kecil sehingga lebih efisien dibandingkan dengan ChatGPT.

2. Gemini, LLM buatan Google yang telah lama menguasai pasar layanan pencarian informasi, mampu bersaing dengan ChatGPT dengan skor yang tidak berselisih banyak. Layanan ini bisa dijadikan alternatif dalam membuat soal bertipe HOTS.
3. Deepseek, sebagai pemain baru dalam layanan LLM dan berasal dari negeri china, mampu melakukan tugas yang diberikan walaupun tidak sebaik kedua lawannya dan dengan kualitas soal yang lebih rendah dan terkadang tidak memenuhi syarat. Layanan ini dapat dijadikan alternatif lain selain kedua LLM besar sebelumnya namun harus diingat dengan kemampuan yang dimiliki maka keluaran yang dihasilkan tidak akan sebaik lawan-lawannya..

DAFTAR PUSTAKA

[1] Beddu, Sultan. (2019). Implementasi Pembelajaran Higher Order Thinking Skills (HOTS) Terhadap Hasil Belajar Peserta Didik.

[2] Farrokhnia, M. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research.

[3] Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEEInternet of Things Journal, 8(13), 10452–10473Zulkardi. (2002). Developing A Learning Environment on Realistic Mathematics Education for Indonesian Student Teachers. Published Dissertation. Enschede: University of Twente. https://doi.org/10.1109/JIOT.2021.3060508

[4] Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in Chatbots.Information, 13(1), 41. https://doi.org/10.3390/info13010041

[5] Susnjak, T. (2022). ChatGPT: The end of online exam integrity? ArXiv https://doi.org/10.48550/arXiv.2212.09292

[6] Zhai, Xiaoming. (2023). ChatGPT User Experience: Implications for Education.

[7] Setiawan, Adi., Luthfiyani, Ulfah Khairiyah. (2023). Penggunaan ChatGPT Untuk Pendidikan di Era Education 4.0: Usulan Inovasi Meningkatkan Keterampilan Menulis.

[8] Mukrodin, mukrodin, & Mega Sasmita, N. (2021). Artificial Inteligence Dalam Apilkasi Chatbot Sebagai Helpdesk Obyek Wisata Dengan Permodelan Natural Language Processing (Studi Kasus: Kabupaten Cilacap). Smart Comp :Jurnalnya Orang Pintar Komputer, 10(1). https://doi.org/10.30591/smartcomp.v10i1.2135

[9] M, Mukhlis. (2024). Persepsi Guru terhadap Pemanfaatan ChatGPT dalam Mengembangkan Soal Literasi Membaca: Studi Kasus pada Sekolah Menengah di Provinsi Riau.

[10] Brookhart, Susan M. (2010). How to Access Higher-order Thinking Skills in Your Classroom.