



Benchmarking Performa dan Efisiensi Respon Model AI ChatGPT dan DeepSeek Menggunakan Metode Eksperimen Komparatif

Performance and Response Efficiency Benchmarking of ChatGPT and DeepSeek AI Models Using a Comparative Experimental Method

Siti Intan Nurfadilah*, Al Imran, Sandi Firmansyah, Muhammad Alamsyah, Farizi Ilham

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: ^{1,*}oaku4544@gmail.com, ²imranalimran160703@gmail.com, ³sandifirmansyah17507@gmail.com,

⁴malamsyah149@gmail.com, ⁵dosen02954@unpam.ac.id

Email Penulis Korespondensi: oaku4544@gmail.com

Received: 04/10/2026, Revised: 17/10/2026, Accepted: 23/10/2026, Published: 23/10/2026

Abstrak—Pesatnya perkembangan *Large Language Model* (LLM) melahirkan berbagai platform kecerdasan buatan, seperti ChatGPT dan DeepSeek, yang diandalkan untuk efisiensi pemrosesan data. Namun, perbedaan arsitektur antar platform memicu permasalahan terkait variasi efisiensi respon dan keandalan informasi saat menangani permintaan pengguna. Penelitian ini bertujuan untuk melakukan *benchmarking* guna mengukur performa dan efisiensi respon dari model AI ChatGPT dan DeepSeek secara obyektif. Solusi yang diterapkan menggunakan metode eksperimen komparatif melalui analisis terhadap 10.000 sampel interaksi dari dataset sintesis di Kaggle. Tiga metrik evaluasi utama yang diukur meliputi kecepatan respon dalam detik, tingkat akurasi respon dalam persentase, serta kepuasan pengguna (*User Rating* skala 1–5). Hasil analisis *benchmarking* menunjukkan bahwa DeepSeek memiliki hasil yang lebih baik dibandingkan ChatGPT pada ketiga indikator yang dianalisis. DeepSeek mampu memproses respon dengan rata-rata kecepatan 1,24 detik, jauh lebih cepat daripada ChatGPT yang memerlukan 3,44 detik. Pada metrik akurasi, DeepSeek meraih tingkat ketepatan 89,97%, sementara ChatGPT berada pada angka 80,26%. Hasil tersebut sejalan dengan kepuasan pengguna, di mana DeepSeek memperoleh *User Rating* sebesar 4,80 dari 5,00 dan ChatGPT sebesar 4,00. Penelitian ini memberikan dua kontribusi. Pertama, acuan empiris mengenai performa komparatif kedua model AI pada tiga indikator sekaligus dalam dataset yang sama, guna mendukung pemilihan sistem pemrosesan bahasa yang optimal. Kedua, perhitungan perbandingan yang dilengkapi rumus dan Langkah hitung sehingga hasilnya dapat ditelusuri dan diperiksa ulang. Hasil berlaku pada data yang dianalisis.

Kata Kunci: Benchmarking; Performa AI; Efisiensi Respon; ChatGPT; DeepSeek; Eksperimen Komparatif

Abstract—The rapid development of Large Language Models (LLMs) has led to the emergence of various artificial intelligence platforms, such as ChatGPT and DeepSeek, which are relied upon for data processing efficiency. However, architectural differences between platforms raise issues regarding variations in response efficiency and information reliability when handling user queries. This study aims to conduct a benchmarking analysis to objectively measure the performance and response efficiency of ChatGPT and DeepSeek. The implemented solution employs a comparative experimental method through the analysis of 10,000 interaction records from a synthetic dataset available on Kaggle. Three primary evaluation metrics were measured: response speed (in seconds), response accuracy (in percentage), and user satisfaction (User Rating on a 1–5 scale). The benchmarking analysis results indicate that DeepSeek performs better than ChatGPT across the three evaluated indicators. DeepSeek processes responses with an average speed of 1.24 seconds, significantly faster than ChatGPT, which requires 3.44 seconds. In terms of accuracy, DeepSeek achieves an accuracy rate of 89.97%, whereas ChatGPT reaches 80.26%. This is consistent with user satisfaction, where DeepSeek earns a User Rating of 4.80 out of 5.00 compared to ChatGPT's 4.00. This study makes two contributions. First, it provides empirical benchmarks of the comparative performance of both AI models on three indicators within a single dataset, to support the selection of an optimal language processing system. Second, it provides comparison calculations with formulas and calculation steps, so that the results can be traced and re-checked. The results apply to the analyzed data.

Keywords: Benchmarking; AI Performance; Response Efficiency; ChatGPT; DeepSeek; Comparative Experiment

1. PENDAHULUAN

Model bahasa besar (*Large Language Model*, LLM) kini dipakai di banyak bidang, mulai dari menjawab pertanyaan, menulis dan merangkum teks, sampai membantu pemrograman dan layanan kesehatan [1]. ChatGPT dari OpenAI, yang berbasis keluarga model GPT [2], menjadi salah satu yang paling dikenal. Belakangan DeepSeek muncul sebagai pesaing lewat model V3 dan R1 yang menekankan efisiensi dan kemampuan penalaran [3], [4]. Kehadirannya ramai dibicarakan karena dianggap lebih hemat sumber daya komputasi [5].

Bagi pengguna, dua hal yang paling terasa adalah seberapa cepat jawaban muncul dan seberapa tepat isinya. Jawaban yang lambat menghambat pekerjaan yang melibatkan banyak permintaan, misalnya pengolahan dokumen atau layanan pelanggan otomatis. Jawaban yang cepat tetapi keliru memaksa pengguna memeriksa ulang, dan model bahasa memang diketahui bisa menghasilkan jawaban yang meyakinkan padahal salah (halusinasi) [6]. Penilaian pengguna terhadap jawaban juga ikut menentukan apakah sebuah model terus dipakai. Meski begitu, pilihan antara ChatGPT dan DeepSeek sering didasarkan pada kebiasaan atau kesan pribadi, bukan pada data yang terukur.

Membandingkan model bahasa secara terukur termasuk dalam kajian evaluasi LLM, yang umumnya menjawab tiga pertanyaan: apa yang dievaluasi, di mana (data atau tolok ukur yang dipakai), dan bagaimana mengevaluasinya [7]. Penelitian ini menggunakan pendekatan *benchmarking* komparatif dengan membandingkan nilai indikator performa ChatGPT dan DeepSeek yang tercatat dalam satu dataset sintesis. Perbandingan dilakukan berdasarkan kecepatan



respons, akurasi respons, dan penilaian pengguna sebagaimana tercantum dalam dataset. ChatGPT bersifat tertutup, sedangkan DeepSeek merilis bobot modelnya ke publik [4]. Rilis bobot pun belum berarti sebuah model terbuka sepenuhnya [8]. Karena itu, hasil perbandingan sebaiknya dibaca sebagai hasil pada data dan versi model tertentu. Penggunaan dataset yang sama memungkinkan ketiga indikator dibandingkan dalam satu kerangka analisis. Namun, interpretasi hasil tetap perlu mempertimbangkan karakteristik data, termasuk perbedaan topik kueri antarplatform dan sifat sintesis dataset. Oleh karena itu, temuan penelitian ini dibatasi pada data yang dianalisis dan tidak langsung menggambarkan performa aktual kedua layanan dalam seluruh kondisi penggunaan.

Sejumlah penelitian sudah membandingkan ChatGPT dan DeepSeek. Meyer dkk. [9] menguji tiga chatbot pada 200 soal pilihan ganda kedokteran berbahasa Jerman. Akurasi DeepSeek 96%, Gemini 94%, dan ChatGPT 92,5%, tetapi selisihnya tidak signifikan secara statistik. Etaiwi dan Alhijawi [10] membandingkan keduanya pada lima tugas NLP (analisis sentimen, klasifikasi topik, perangkuman, terjemahan, dan *textual entailment*) dan melaporkan DeepSeek lebih stabil pada klasifikasi dan penalaran logis, sedangkan ChatGPT lebih baik pada tugas yang menuntut pemahaman bernuansa. Jiang dkk. [11] menguji DeepSeek, ChatGPT, dan Claude pada soal komputasi ilmiah dan *scientific machine learning*, sementara Singh dkk. [12] dan Rahman dkk. [13] membandingkan arsitektur serta fitur model-model tersebut secara lebih umum. Kajian lain menyoroti DeepSeek di bidang kesehatan dan riset ilmiah [14]–[16], pada soal oftalmologi [17], [18], dan dari sisi keamanan [19]. Selain itu, performa ChatGPT diketahui berbeda antar versi [20], [22] dan dipengaruhi karakteristik soal, misalnya panjang pertanyaan [21].

Dari penelitian-penelitian tersebut ada tiga hal yang terlihat. Pertama, sebagian besar menguji soal pada bidang tertentu, terutama kedokteran, sehingga hasilnya sulit digeneralisasi. Kedua, ukuran yang dipakai umumnya akurasi, sedangkan kecepatan respon dan penilaian pengguna jarang dilaporkan bersamaan pada sampel yang sama. Ketiga, hasil antar studi tidak seragam: DeepSeek unggul pada sebagian tugas, ChatGPT pada tugas lain, dan pada [9] selisih akurasi tidak signifikan. Berdasarkan studi yang ditinjau penulis, masih ada ruang untuk perbandingan yang memuat kecepatan, akurasi, dan kepuasan pengguna sekaligus pada satu kumpulan sampel, dengan perhitungan yang bisa ditelusuri ulang.

Tiga indikator yang dipakai dalam penelitian ini dipilih dari kebutuhan pengguna yang paling langsung. Kecepatan respon menggambarkan efisiensi, yaitu berapa lama pengguna menunggu sampai jawaban selesai. Akurasi respons menggambarkan nilai ketepatan jawaban sebagaimana tercatat dalam dataset dan merupakan salah satu indikator yang digunakan dalam evaluasi LLM [7], [9]. Kepuasan pengguna menggambarkan penilaian pengguna terhadap jawaban, yang dalam penelitian ini berupa rating pada skala 1–5. Ketiganya dibandingkan memakai rata-rata dan selisih relatif, sehingga perbedaan antar platform dapat dibaca dalam satuan yang sama dan perhitungannya dapat ditelusuri ulang.

Tujuan penelitian ini ada tiga, yaitu (1) mengukur dan membandingkan rata-rata kecepatan respon ChatGPT dan DeepSeek, (2) membandingkan rata-rata akurasi responnya, dan (3) membandingkan rata-rata kepuasan pengguna (*user rating*), semuanya pada 10.000 baris data interaksi dari dataset sintesis di Kaggle. Kontribusi penelitian ini ada dua. Pertama, penelitian ini menyajikan perbandingan ChatGPT dan DeepSeek pada tiga indikator sekaligus dalam satu dataset yang sama. Kedua, setiap angka perbandingan dilengkapi rumus, langkah hitung, simpangan baku, dan uji statistik, serta sifat sintesis dataset dan batas kesimpulannya dinyatakan secara terbuka, sehingga pembaca dapat menelusuri dan memeriksa ulang hasilnya.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif komparatif untuk membandingkan indikator performa ChatGPT dan DeepSeek berdasarkan data yang tersedia dalam satu dataset sintesis. Penelitian dilakukan sebagai analisis terhadap dataset yang sudah tersedia, bukan dengan menguji kedua layanan secara langsung. Bagian ini menjelaskan objek penelitian, data, tahapan penelitian, serta cara pengukuran dan perhitungan.

2.1 Objek Penelitian

Objek penelitian adalah dua platform LLM, yaitu ChatGPT (OpenAI) dan DeepSeek (DeepSeek-AI). Keduanya dipilih karena sama-sama dipakai luas untuk menjawab pertanyaan berbasis teks, tetapi dikembangkan oleh organisasi yang berbeda dan memiliki pendekatan rilis yang berbeda [4], [8]. Pada dataset, ChatGPT tercatat dengan versi model GPT-4-turbo dan DeepSeek dengan versi DeepSeek-Chat 1.5 (kolom AI_Model_Version). Versi model penting dicatat karena hasil dapat berubah ketika model diperbarui [9]. Sesuai kata kunci pada abstrak, objek lain yang diamati adalah tiga indikator performa, yaitu kecepatan respon, akurasi respon, dan kepuasan pengguna, yang menjadi dasar perbandingan.

2.2 Data Penelitian

Data berasal dari dataset “DeepSeek vs ChatGPT: AI Platform Comparison” karya Aakif Khan yang tersedia di Kaggle [23]. Pada halaman dataset dan berkas README, pembuatnya menyatakan bahwa data ini dibangun secara sintesis (memakai Python dan SQL) dan bukan berasal dari log pengguna sebenarnya. Dengan demikian, nilai kecepatan, rating, dan akurasi pada dataset adalah nilai simulasi, bukan hasil pengukuran langsung terhadap layanan ChatGPT atau DeepSeek. Dataset ini dipakai karena menyediakan ketiga indikator dalam satu tabel dengan jumlah baris yang besar. Keterbatasan akibat sifat sintesisnya dibahas pada Subbab 3.10.



Berkas utama `deepseek_vs_chatgpt.csv` berisi 10.000 baris dan 28 kolom, dengan rentang tanggal 1 Juli 2023 sampai 1 Februari 2025. Setiap baris mewakili satu interaksi dan antara lain memuat kolom `AI_Platform`, `AI_Model_Version`, `Query_Type`, `Topic_Category`, `Response_Speed_sec`, `User_Rating`, dan `Response_Accuracy`. Tiga kolom terakhir dipakai sebagai indikator. Komposisi data per platform ditunjukkan pada Tabel 1, dan cuplikan beberapa baris data ditunjukkan pada Tabel 2.

Tabel 1. Komposisi data per platform

Keterangan	ChatGPT	DeepSeek
Jumlah baris	5.076	4.924
Versi model (<code>AI_Model_Version</code>)	GPT-4-turbo	DeepSeek-Chat 1.5
<code>Query_Type</code> = General	5.076	1.246
<code>Query_Type</code> = Technical	0	3.678
<code>Topic_Category</code>	Best Practices, Content Creation, Education, Professional Writing	Algorithm Explanation, Code Optimization, Debugging, Implementation Help
<code>Response_Accuracy</code> kosong	182 baris (3,59%)	197 baris (4,00%)
Baris akurasi yang dipakai	4.894	4.727

Tabel 2. Cuplikan data (tiga baris pertama tiap platform, kolom yang dipakai)

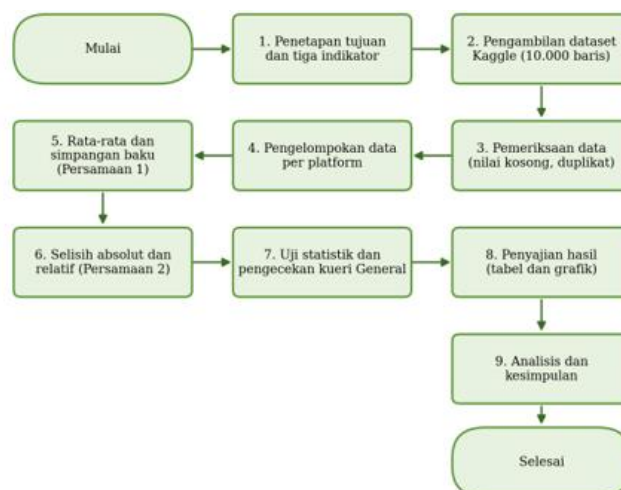
Date	AI_Platform	Query_Type	Response_Speed_sec	User_Rating	Response_Accuracy
2024-09-21	ChatGPT	General	3,30	4	0,7842
2024-09-21	ChatGPT	General	3,28	4	0,8194
2024-09-21	ChatGPT	General	3,07	3	0,8090
2024-05-16	DeepSeek	Technical	1,48	5	0,9366
2024-05-16	DeepSeek	Technical	1,80	5	0,8992
2024-05-16	DeepSeek	General	1,66	5	0,9623

Dua hal pada Tabel 1 perlu diperhatikan. Pertama, seluruh kueri ChatGPT bertipe General, sedangkan kueri DeepSeek bertipe General (1.246 baris) dan Technical (3.678 baris). Kedua, kategori topik kedua platform tidak beririsan sama sekali. Artinya, kedua platform tidak diuji dengan kumpulan kueri yang sama, sehingga selisih yang ditemukan dapat bercampur dengan perbedaan jenis kueri. Untuk memeriksa pengaruh jenis kueri, perbandingan diulang hanya pada kueri bertipe General (Subbab 3.7).

Pemeriksaan data menunjukkan tidak ada baris yang duplikat secara utuh dan tidak ada nilai kosong pada kolom kecepatan respon maupun rating. Nilai kosong hanya ada pada kolom `Response_Accuracy`, sebanyak 379 baris (3,79%); deskripsi dataset menyebut nilai kosong ini sengaja dimasukkan untuk latihan pembersihan data. Baris yang kosong dikeluarkan hanya pada perhitungan akurasi dan tidak diisi nilai pengganti, sehingga rata-rata kecepatan respon dan rating memakai seluruh 10.000 baris.

2.3 Tahapan Penelitian

Tahapan penelitian digambarkan pada Gambar 1 dan dijelaskan satu per satu di bawah ini. Seluruh perhitungan dikerjakan dengan Python (pustaka `pandas` dan `SciPy`).



Gambar 1. Diagram alur tahapan penelitian



Tahap 1, penetapan tujuan dan indikator. Peneliti menetapkan tujuan perbandingan dan tiga indikator yang diukur, yaitu kecepatan respon, akurasi respon, dan kepuasan pengguna. Ketiganya dipilih karena menggambarkan dua kebutuhan utama pengguna (cepat dan tepat) ditambah penilaian pengguna sendiri. Tahap 2, pengambilan dataset. Berkas CSV dataset diunduh dari Kaggle [23]. Berkas ini berisi 10.000 baris dan 28 kolom seperti dijelaskan pada Subbab 2.2. Tahap 3, pemeriksaan data. Berkas diperiksa untuk menemukan nilai kosong dan baris duplikat. Hasilnya, tidak ada baris duplikat dan terdapat 379 nilai kosong pada kolom Response_Accuracy yang dikeluarkan hanya dari perhitungan akurasi. Tahap 4, pengelompokan data. Data dibagi menurut kolom AI_Platform menjadi kelompok ChatGPT (5.076 baris) dan kelompok DeepSeek (4.924 baris), supaya kedua kelompok dihitung dengan cara yang sama. Tahap 5, perhitungan rata-rata dan simpangan baku. Untuk setiap platform dan setiap indikator dihitung rata-rata aritmetika memakai Persamaan (1) serta simpangan baku sampel. Tahap 6, perhitungan selisih. Rata-rata kedua platform dibandingkan melalui selisih absolut dan selisih relatif pada Persamaan (2). Tahap 7, uji statistik dan pengecekan kueri General. Selisih rata-rata diuji dengan uji t Welch dan uji Mann–Whitney U, lalu perbandingan diulang hanya pada kueri bertipe General untuk memeriksa pengaruh jenis kueri. Tahap 8, penyajian hasil. Hasil disajikan dalam tabel (Tabel 4 sampai Tabel 8) dan grafik batang (Gambar 2). Tahap 9, analisis dan kesimpulan. Hasil dibahas per indikator, dibandingkan dengan penelitian terdahulu, lalu dirangkum dalam kesimpulan yang dibatasi pada data yang dianalisis.

2.4 Indikator, Metode Pengukuran, dan Perhitungan

Tiga indikator yang digunakan beserta satuannya dirangkum pada Tabel 3. Ketiganya diambil langsung dari kolom dataset tanpa pengukuran ulang.

Tabel 3. Indikator dan metrik evaluasi performa AI

Indikator	Metrik pengukuran	Satuan	Deskripsi operasional
Kecepatan respon	Average of Response_Speed_sec	Detik	Waktu respon yang tercatat pada kolom Response_Speed_sec. Semakin kecil semakin baik.
Kepuasan pengguna	Average of User_Rating	Skala 1–5	Rating bilangan bulat 1–5 pada kolom User_Rating. Semakin besar semakin baik.
Akurasi respon	Average of Response_Accuracy ($\times 100$)	Persen (%)	Skor ketepatan jawaban pada rentang 0–1 (kolom Response_Accuracy) yang dikali 100. Semakin besar semakin baik.

Kolom Response_Speed_sec menyatakan waktu respon dalam detik. Kolom User_Rating berisi rating 1 (sangat tidak puas) sampai 5 (sangat puas); pada data ini hanya muncul nilai 3 sampai 5. Kolom Response_Accuracy berisi skor ketepatan jawaban 0–1 menurut README dataset, yang dikali 100 agar dinyatakan dalam persen. README tidak menjelaskan bagaimana skor akurasi dan rating dibangkitkan atau siapa yang menilainya, sehingga penelitian ini memperlakukannya sebagai nilai yang sudah diberikan oleh dataset.

Rata-rata setiap indikator dihitung dengan Persamaan (1):

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{k,i} \quad (1)$$

dengan k adalah platform (ChatGPT atau DeepSeek), n_k jumlah sampel pada platform tersebut, dan $x_{k,i}$ nilai indikator pada sampel ke- i .

Simpangan baku dihitung sebagai simpangan baku sampel, yaitu akar dari jumlah kuadrat selisih tiap nilai terhadap rata-rata dibagi $(n - 1)$ pada masing-masing platform. Perbandingan memakai dua ukuran. Selisih absolut (Δ) adalah rata-rata DeepSeek dikurangi rata-rata ChatGPT, dalam satuan indikator masing-masing. Selisih relatif (R) pada Persamaan (2) menunjukkan besar selisih itu dibandingkan nilai ChatGPT.

$$R = \frac{X_{\text{DeepSeek}} - X_{\text{ChatGPT}}}{X_{\text{ChatGPT}}} \times 100\% \quad (2)$$

Di sini X_{DeepSeek} dan X_{ChatGPT} adalah rata-rata kedua platform dari Persamaan (1). R bernilai negatif pada kecepatan respon berarti waktu DeepSeek lebih singkat. R bernilai positif pada rating dan akurasi berarti nilai DeepSeek lebih tinggi. Untuk akurasi, selisih absolut ditulis dalam poin persentase supaya tidak tertukar dengan selisih relatif.

Selisih rata-rata juga diuji dengan uji t Welch, yang tidak mengandaikan ragam kedua kelompok sama, dan dengan uji Mann–Whitney U sebagai pembanding karena rating berupa skala ordinal. Hasil uji dilaporkan sebagai statistik t beserta derajat bebasnya, nilai p , selisih rata-rata dengan selang kepercayaan 95%, dan ukuran efek Cohen's d (selisih rata-rata dibagi simpangan baku gabungan). Nilai p di bawah 0,001 ditulis $p < 0,001$.

3. HASIL DAN PEMBAHASAN

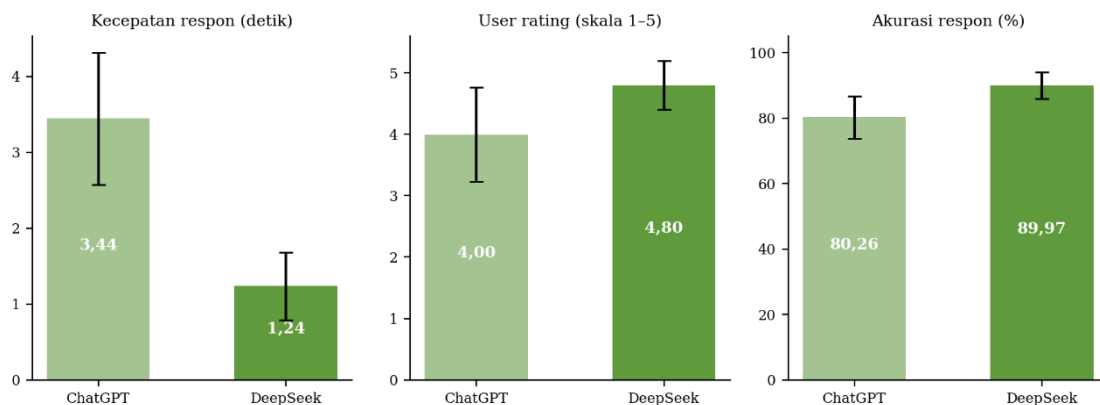
3.1 Statistik Deskriptif

Tabel 4 merangkum jumlah data, rata-rata, simpangan baku, median, dan rentang nilai ketiga indikator pada tiap platform. Nilai rata-rata yang sama ditampilkan sebagai grafik batang pada Gambar 2.



**Tabel 4.** Statistik deskriptif ketiga indikator per platform

Indikator	Platform	n	Rata-rata	Simpangan baku	Median	Min–Maks
Kecepatan respon (detik)	ChatGPT	5.076	3,44	0,87	3,43	1,82–5,19
	DeepSeek	4.924	1,24	0,45	1,24	0,33–2,18
Kepuasan pengguna (1–5)	ChatGPT	5.076	4,00	0,77	4,00	3–5
	DeepSeek	4.924	4,80	0,40	5,00	4–5
Akurasi respon (%)	ChatGPT	4.894	80,26	6,46	80,38	65,42–94,67
	DeepSeek	4.727	89,97	4,06	89,97	80,55–99,72

**Gambar 2.** Rata-rata (± simpangan baku) ketiga indikator pada ChatGPT dan DeepSeek

Ketiga indikator memiliki satuan yang berbeda, jadi grafik pada Gambar 2 hanya dipakai untuk membandingkan kedua platform pada indikator yang sama, bukan untuk membandingkan tinggi batang antar indikator.

3.2 Penelusuran Perhitungan Perbandingan

Agar angka perbandingan dapat diperiksa ulang, Tabel 5 menunjukkan langkah hitung dari rata-rata tiap platform (ditulis empat desimal) memakai Persamaan (2). Selisih pada teks dan tabel lain dihitung dari rata-rata empat desimal ini, lalu dibulatkan, sehingga dapat berbeda 0,01 dari selisih rata-rata yang sudah dibulatkan pada Tabel 4.

Tabel 5. Langkah perhitungan selisih absolut dan selisih relatif

Indikator	ChatGPT	DeepSeek	Selisih absolut (Δ)	Selisih relatif (R)	Hasil R
Kecepatan respon	3,4433	1,2365	$1,2365 - 3,4433 = -2,2068$ detik	$(-2,2068 \div 3,4433) \times 100\%$	-64,09%
Kepuasan pengguna	3,9988	4,8028	$4,8028 - 3,9988 = +0,8040$	$(0,8040 \div 3,9988) \times 100\%$	+20,11%
Akurasi respon	80,2574	89,9684	$89,9684 - 80,2574 = +9,7110$ poin	$(9,7110 \div 80,2574) \times 100\%$	+12,10%

Pada kecepatan respon, ChatGPT memerlukan $3,4433 \div 1,2365 \approx 2,78$ kali waktu DeepSeek. Sebagai gambaran, bila 10.000 kueri diproses berurutan, total waktu respon ChatGPT adalah $10.000 \times 3,4433 = 34.433$ detik (sekitar 9 jam 34 menit), sedangkan DeepSeek $10.000 \times 1,2365 = 12.365$ detik (sekitar 3 jam 26 menit). Selisihnya 22.068 detik atau sekitar 6 jam 8 menit. Hitungan ini hanya ilustrasi dari nilai rata-rata, tidak memperhitungkan pemrosesan paralel maupun variasi antar kueri.

3.3 Uji Statistik

Tabel 6 menunjukkan hasil uji t Welch dan uji Mann–Whitney U untuk ketiga indikator.

Tabel 6. Hasil uji statistik perbandingan DeepSeek terhadap ChatGPT

Indikator	t Welch (derajat bebas)	p	Cohen's d	Selisih Δ [selang kepercayaan 95%]	Mann–Whitney p
Kecepatan respon	-160,26 (7.635,6)	< 0,001	-3,18	-2,21 [-2,23; -2,18] detik	< 0,001
Kepuasan pengguna	65,95 (7.662,4)	< 0,001	1,31	+0,80 [0,78; 0,83]	< 0,001
Akurasi respon	88,54 (8.278,4)	< 0,001	1,79	+9,71 [9,50; 9,93] poin	< 0,001



Ketiga selisih signifikan secara statistik ($p < 0,001$) dengan ukuran efek yang besar menurut patokan umum ($|d| > 0,8$). Perlu diingat bahwa dengan ribuan baris, selisih yang kecil pun mudah menjadi signifikan, dan data yang dianalisis bersifat sintesis. Karena itu, hasil uji ini menunjukkan bahwa selisih pada dataset ini bukan akibat variasi acak di dalam dataset, bukan bahwa selisih yang sama pasti terjadi pada layanan sebenarnya.

Nilai p yang diperoleh menunjukkan bahwa perbedaan indikator pada dataset cukup kuat secara statistik untuk menolak hipotesis tidak adanya perbedaan rata-rata pada uji yang digunakan. Namun, interpretasi hasil tidak cukup hanya berdasarkan nilai p . Ukuran efek Cohen's d juga perlu diperhatikan karena memberikan gambaran mengenai besarnya perbedaan relatif terhadap variasi data. Pada ketiga indikator, nilai absolut Cohen's d berada di atas 1, sehingga perbedaan dalam dataset tergolong besar menurut patokan umum. Meskipun demikian, ukuran efek tersebut tetap menggambarkan data sintesis yang dianalisis, bukan ukuran keunggulan kedua layanan dalam penggunaan nyata. Selain itu, hasil uji statistik tidak menjelaskan penyebab perbedaan. Perbedaan topik kueri, karakteristik data, dan mekanisme pembentukan skor dapat ikut memengaruhi pola yang terlihat. Oleh sebab itu, hasil statistik perlu dibaca bersama statistik deskriptif dan keterbatasan dataset.

3.4 Analisis Kecepatan Respon

Rata-rata waktu respon DeepSeek adalah 1,24 detik, sedangkan ChatGPT 3,44 detik. Selisihnya 2,21 detik, atau DeepSeek 64,09% lebih singkat (Tabel 4 dan Tabel 5). Sebarannya hampir tidak tumpang tindih: waktu respon DeepSeek berkisar 0,33–2,18 detik, sedangkan ChatGPT 1,82–5,19 detik. Hanya 6,9% respon ChatGPT yang waktunya berada di dalam rentang DeepSeek ($\leq 2,18$ detik), dan 11,9% respon DeepSeek berada di dalam rentang ChatGPT ($\geq 1,82$ detik). Simpangan baku ChatGPT (0,87 detik) juga lebih besar daripada DeepSeek (0,45 detik).

Pada DeepSeek, waktu respon hampir sama untuk kueri General (1,25 detik, $n = 1.246$) dan Technical (1,23 detik, $n = 3.678$), sehingga jenis kueri pada data ini tidak tampak memengaruhi kecepatannya. Perbandingan serupa tidak bisa dilakukan pada ChatGPT karena seluruh kuerinya bertipe General. Laporan teknis DeepSeek menyebut pemakaian arsitektur *Mixture-of-Experts* [3], yang hanya mengaktifkan sebagian parameter untuk tiap masukan [13]. Hal itu mungkin berkaitan dengan hasil yang diperoleh, tetapi penelitian ini tidak menguji penyebabnya, apalagi data waktu responnya bersifat sintesis.

3.5 Analisis Kepuasan Pengguna

Rata-rata *user rating* DeepSeek adalah 4,80 dan ChatGPT 4,00 pada skala 1–5. Selisihnya +0,80 poin, atau DeepSeek 20,11% lebih tinggi. Sebaran rating menunjukkan perbedaannya: pada DeepSeek, 80,3% rating bernilai 5 dan 19,7% bernilai 4, sedangkan pada ChatGPT 29,5% bernilai 5, 40,9% bernilai 4, dan 29,6% bernilai 3. Tidak ada rating 1 atau 2 pada kedua platform, sehingga rating pada data ini terpusat di bagian atas skala.

Data hanya berisi angka rating tanpa alasan penilaian, sehingga penyebab perbedaannya, misalnya gaya bahasa, panjang jawaban, atau kejelasan, tidak bisa disimpulkan dari penelitian ini. Pada tiap platform, rating hampir tidak berkorelasi dengan kecepatan respon maupun akurasi pada baris yang sama (korelasi Spearman $|p| \leq 0,04$), jadi selisih rating antar platform tidak dapat dijelaskan oleh selisih kecepatan atau akurasi dalam data ini. Preferensi pengguna juga bisa berubah menurut jenis tugas; penelitian lain melaporkan ChatGPT lebih baik pada tugas yang menuntut pemahaman bernuansa [10]. Hasil rating ini sebaiknya tidak dibaca sebagai bukti bahwa DeepSeek selalu lebih disukai.

3.6 Analisis Akurasi Respon

Rata-rata akurasi DeepSeek adalah 89,97% dan ChatGPT 80,26%. Selisihnya 9,71 poin persentase, atau 12,10% lebih tinggi secara relatif. Kedua angka ini perlu dibedakan: 9,71 poin adalah beda angka persen, sedangkan 12,10% adalah beda dibandingkan nilai ChatGPT. Sebaran akurasi ChatGPT juga lebih lebar (simpangan baku 6,46 poin, rentang 65,42–94,67%) dibanding DeepSeek (4,06 poin, rentang 80,55–99,72%). Nilai terendah DeepSeek (80,55%) hampir sama dengan rata-rata ChatGPT.

Arah hasil ini sejalan dengan penelitian lain yang menempatkan DeepSeek setara atau sedikit lebih baik, misalnya pada soal kedokteran [9]. Pada [9] selisih akurasinya tidak signifikan secara statistik, sedangkan di sini selisihnya signifikan dan berukuran besar. Perbedaan ini wajar, karena [9] memakai 200 soal sungguhan, sedangkan penelitian ini memakai ribuan baris data sintesis. Akurasi juga bergantung pada jenis kueri: hasil ChatGPT dan DeepSeek dapat berbalik menurut jenis tugas [10], berubah menurut versi model [20], [22], dan dipengaruhi karakteristik soal seperti panjang pertanyaan [21]. Terakhir, jawaban yang salah tetapi terdengar yakin tetap terjadi pada model bahasa [6], sehingga pada akurasi 89,97% pun masih ada jawaban yang perlu diperiksa pengguna.

3.7 Pengecekan pada Kueri Bertipe General

Karena seluruh kueri ChatGPT bertipe General, perbandingan diulang dengan hanya memakai baris DeepSeek bertipe General (1.246 baris, 1.190 di antaranya memiliki nilai akurasi). Hasilnya pada Tabel 7 hampir sama dengan perbandingan penuh: DeepSeek tetap lebih cepat, lebih tinggi ratingnya, dan lebih akurat, dengan seluruh nilai $p < 0,001$ pada uji t Welch. Jadi selisih utama tidak disebabkan oleh adanya kueri Technical pada DeepSeek. Meski begitu, kategori topik kedua platform tetap berbeda (Tabel 1), dan pengecekan ini tidak dapat menghilangkan pengaruh perbedaan topik.

**Tabel 7.** Perbandingan hanya pada kueri bertipe General

Indikator	ChatGPT	DeepSeek (General)	Selisih absolut (Δ)	Selisih relatif (R)	Cohen's d
Kecepatan respon (detik)	3,44	1,25	-2,20	-63,76%	-2,73
Kepuasan pengguna (1-5)	4,00	4,80	+0,80	+20,02%	1,13
Akurasi respon (%)	80,26	89,94	+9,69 poin	+12,07%	1,60

3.8 Konsistensi Menurut Tahun

Untuk melihat apakah selisih bertahan sepanjang periode data, rata-rata dihitung ulang per tahun pada Tabel 8. Data 2023 mencakup Juli sampai Desember, data 2024 satu tahun penuh, dan data 2025 hanya sampai 1 Februari sehingga jumlahnya paling sedikit.

Tabel 8. Rata-rata ketiga indikator menurut tahun (n = jumlah baris; akurasi memakai baris yang tidak kosong)

Tahun	Platform	n	Kecepatan respon (detik)	User rating (1-5)	Akurasi (%)
2023	ChatGPT	1.605	3,37	4,01	81,22
	DeepSeek	1.268	1,29	4,80	89,91
2024	ChatGPT	3.226	3,49	3,99	79,77
	DeepSeek	3.381	1,20	4,80	90,07
2025	ChatGPT	245	3,30	3,97	80,31
	DeepSeek	275	1,41	4,83	89,02

Pada ketiga tahun, DeepSeek tetap lebih cepat (1,20–1,41 detik dibanding 3,30–3,49 detik), lebih tinggi ratingnya (4,80–4,83 dibanding 3,97–4,01), dan lebih akurat (89,02–90,07% dibanding 79,77–81,22%). Rata-rata kedua platform bergeser tipis antar tahun, tetapi arah dan besar selisihnya relatif sama, sehingga hasil keseluruhan tidak ditentukan oleh satu periode saja. Pada tahun 2025 jumlah barisnya kecil (245 dan 275), sehingga angkanya kurang stabil. Secara ringkas, pada seluruh pemeriksaan (perbandingan penuh, kueri General saja, dan per tahun), DeepSeek tercatat lebih unggul pada ketiga indikator pada data yang dianalisis.

3.9 Perbandingan dengan Penelitian Terdahulu

Dari segi arah, temuan di sini sejalan dengan [9], tempat DeepSeek mencatat akurasi tertinggi di antara tiga chatbot (walau tidak signifikan), dan dengan [17], yang melaporkan DeepSeek-R1 unggul atas model OpenAI pada penalaran oftalmologi dwibahasa. Di sisi lain, [10] menemukan keunggulan DeepSeek hanya pada klasifikasi dan penalaran logis, sedangkan ChatGPT lebih baik pada tugas bernuansa. Perbedaan ini wajar karena jenis soal, bahasa, versi model, dan cara penilaian tiap studi tidak sama. Yang membedakan penelitian ini adalah penggunaan tiga indikator sekaligus pada satu dataset, lengkap dengan simpangan baku dan uji statistik. Meski begitu, karena datanya sintetis dan domain kuerinya tidak seragam dengan studi-studi tersebut, angka antar studi tidak dapat dibandingkan langsung.

Perbandingan tersebut memperlihatkan bahwa penilaian performa model bahasa perlu mempertimbangkan konteks evaluasinya. Hasil akurasi yang tinggi pada suatu dataset belum tentu menunjukkan hasil serupa pada bidang atau jenis pertanyaan lain. Demikian pula, kecepatan respons dan rating pengguna tidak selalu cukup untuk menjelaskan kualitas jawaban secara menyeluruh. Penelitian ini melengkapi perbandingan dari sisi penyajian tiga indikator secara bersamaan, tetapi belum dapat memastikan apakah pola yang ditemukan akan bertahan ketika kedua platform menerima pertanyaan identik dan dinilai dengan prosedur yang sama. Dengan demikian, kontribusi utamanya terletak pada perbandingan terukur dalam dataset yang dianalisis serta penyajian langkah perhitungan yang dapat diperiksa ulang, bukan pada penetapan bahwa satu platform selalu lebih unggul dalam semua situasi.

3.10 Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan. Pertama dan terpenting, dataset dibangkitkan secara sintetis [23], sehingga hasilnya menggambarkan data tersebut dan bukan performa nyata ChatGPT atau DeepSeek. Kedua, kedua platform tidak diuji dengan kueri yang sama: kueri ChatGPT seluruhnya bertipe General, kategori topiknya berbeda dari DeepSeek, dan pengecekan pada Subbab 3.7 tidak menghilangkan perbedaan topik. Ketiga, cara skor akurasi dan rating dibangkitkan tidak dijelaskan oleh dataset; rating hanya bernilai 3–5 untuk ChatGPT dan 4–5 untuk DeepSeek, dan rating hampir tidak berkorelasi dengan kecepatan maupun akurasi, sehingga hubungan antar indikator pada data nyata belum dapat dinilai. Keempat, 379 nilai akurasi yang kosong (3,79%) dikeluarkan dari perhitungan akurasi. Kelima, nama versi model mengikuti label pada dataset dan tidak diverifikasi terhadap layanan aslinya. Keenam, signifikansi statistik pada ribuan baris mudah tercapai, sehingga ukuran efek dan konteks data perlu dibaca bersama nilai p. Penelitian berikutnya dapat menguji kedua layanan secara langsung dengan kueri yang sama dan penilai yang jelas.

4. KESIMPULAN

Penelitian ini bertujuan membandingkan performa ChatGPT dan DeepSeek berdasarkan tiga indikator, yaitu kecepatan respons, akurasi respons, dan kepuasan pengguna, menggunakan 10.000 baris data interaksi dari dataset sintetis yang tersedia di Kaggle. Berdasarkan hasil analisis, DeepSeek menunjukkan rata-rata yang lebih baik daripada ChatGPT pada





ketiga indikator dalam dataset yang diteliti. Pada indikator kecepatan respons, DeepSeek memiliki rata-rata waktu 1,24 detik, sedangkan ChatGPT 3,44 detik, dengan selisih 2,21 detik atau sekitar 64,09% lebih singkat. Pada indikator akurasi respons, DeepSeek memperoleh rata-rata 89,97%, sementara ChatGPT 80,26%, dengan selisih 9,71 poin persentase. Sementara itu, pada indikator kepuasan pengguna, DeepSeek memperoleh rata-rata *user rating* 4,80 dari 5,00, sedangkan ChatGPT memperoleh 4,00, dengan selisih 0,80 poin. Hasil pengujian statistik menunjukkan bahwa perbedaan ketiga indikator antara kedua platform signifikan secara statistik pada data yang dianalisis, dengan nilai $p < 0,001$. Pola perbedaan juga tetap terlihat pada analisis tambahan yang membatasi data pada kueri bertipe *General* serta pada perbandingan berdasarkan tahun 2023–2025. Dengan demikian, ketiga tujuan penelitian telah terjawab melalui perbandingan rata-rata, perhitungan selisih, dan pengujian statistik. Penelitian ini juga memberikan perbandingan tiga indikator dalam satu dataset serta menyajikan rumus dan langkah perhitungan agar hasilnya dapat ditelusuri dan diperiksa kembali. Meskipun demikian, hasil penelitian perlu ditafsirkan sesuai dengan karakteristik data yang digunakan. Dataset bersifat sintesis dan bukan berasal dari pengujian langsung terhadap layanan ChatGPT dan DeepSeek. Selain itu, jenis dan topik kueri kedua platform tidak sepenuhnya sama, sehingga perbedaan yang ditemukan tidak dapat dikaitkan hanya dengan platform yang dibandingkan. Mekanisme pembentukan nilai akurasi dan *user rating* dalam dataset juga tidak dijelaskan secara terperinci. Oleh karena itu, hasil penelitian ini menunjukkan pola perbedaan pada dataset yang dianalisis, bukan bukti bahwa DeepSeek selalu lebih unggul daripada ChatGPT dalam penggunaan nyata atau pada seluruh jenis tugas. Penelitian selanjutnya disarankan melakukan pengujian langsung terhadap kedua platform menggunakan kumpulan kueri yang sama, kondisi pengujian yang terkontrol, serta kriteria penilaian yang konsisten. Pengujian juga dapat dilakukan pada beberapa jenis tugas dan versi model untuk mengetahui apakah perbedaan performa tetap terlihat pada kondisi yang berbeda. Dengan prosedur tersebut, perbandingan diharapkan dapat memberikan gambaran yang lebih kuat mengenai kecepatan respons, ketepatan jawaban, dan penilaian pengguna pada masing-masing platform.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: S. I. N.;

Metodologi: A. I., S. F., dan F. I.;

Perangkat Lunak: M. A.;

Validasi: S. I. N.;

Analisis Formal: A. I.;

Investigasi: A. I.;

Sumber Daya: S. F.;

Kurasi Data: M. A.;

Penulisan Draf Awal: S. I. N.;

Penulisan, Peninjauan, dan Penyuntingan: S. I. N., dan A. I.;

Visualisasi: S. I. N.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang dianalisis tersedia secara publik di Kaggle (dataset “DeepSeek vs ChatGPT: AI Platform Comparison”, Aakif Khan) [23]. Hasil perhitungan dapat direproduksi dari berkas `deepseek_vs_chatgpt.csv` dengan langkah pada Subbab 2.4.

5.3 Pendanaan

Penelitian ini tidak menerima dukungan pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.





REFERENCES

- [1] A. J. Thirunavukarasu, D. S. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. Ting, "Large language models in medicine," *Nat. Med.*, vol. 29, no. 8, pp. 1930–1940, 2023, doi: 10.1038/s41591-023-02448-8.
- [2] OpenAI, "GPT-4 technical report," arXiv:2303.08774, 2023, doi: 10.48550/arXiv.2303.08774.
- [3] DeepSeek-AI, "DeepSeek-V3 technical report," arXiv:2412.19437, 2024, doi: 10.48550/arXiv.2412.19437.
- [4] DeepSeek-AI, "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, vol. 645, pp. 633–638, 2025, doi: 10.1038/s41586-025-09422-z.
- [5] M. Poo, "Reflections on DeepSeek's breakthrough," *Natl. Sci. Rev.*, vol. 12, no. 3, art. no. nwaf044, 2025, doi: 10.1093/nsr/nwaf044.
- [6] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.
- [7] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, 2024, doi: 10.1145/3641289.
- [8] A. Liesenfeld and M. Dingemans, "Rethinking open source generative AI: Open-washing and the EU AI Act," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil, 2024, pp. 1774–1784, doi: 10.1145/3630106.3659005.
- [9] A. Meyer, Y. Karay, A. U. Steinbicker, T. Streichert, and R. Overbeek, "Performance of DeepSeek-R1, ChatGPT (GPT-o3-mini), and Gemini 2.0 Flash on German medical multiple-choice questions: Comparative evaluation," *JMIR Form. Res.*, vol. 9, art. no. e77357, 2025, doi: 10.2196/77357.
- [10] W. Etaiwi and B. Alhijawi, "Comparative evaluation of ChatGPT and DeepSeek across key NLP tasks: Strengths, weaknesses, and domain-specific performance," arXiv:2506.18501, 2025, doi: 10.48550/arXiv.2506.18501.
- [11] Q. Jiang, Z. Gao, and G. E. Karniadakis, "DeepSeek vs. ChatGPT vs. Claude: A comparative study for scientific computing and scientific machine learning tasks," arXiv:2502.17764, 2025, doi: 10.48550/arXiv.2502.17764.
- [12] S. Singh, S. Bansal, A. El Saddik, and M. Saini, "From ChatGPT to DeepSeek AI: A comprehensive analysis of evolution, deviation, and future implications in AI-language models," arXiv:2504.03219, 2025, doi: 10.48550/arXiv.2504.03219.
- [13] A. Rahman *et al.*, "Comparative analysis based on DeepSeek, ChatGPT, and Google Gemini: Features, techniques, performance, future prospects," arXiv:2503.04783, 2025, doi: 10.48550/arXiv.2503.04783.
- [14] Y. Peng *et al.*, "From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare," *J. Biomed. Inform.*, vol. 163, art. no. 104791, 2025, doi: 10.1016/j.jbi.2025.104791.
- [15] A. Temsah *et al.*, "DeepSeek in healthcare: Revealing opportunities and steering challenges of a new open-source artificial intelligence frontier," *Cureus*, vol. 17, no. 2, art. no. e79221, 2025, doi: 10.7759/cureus.79221.
- [16] M. E. Kayaalp *et al.*, "DeepSeek versus ChatGPT: Multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation—redefining innovation in research and practice," *Knee Surg. Sports Traumatol. Arthrosc.*, vol. 33, no. 5, pp. 1553–1556, 2025, doi: 10.1002/ksa.12628.
- [17] P. Xu *et al.*, "DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning," *Adv. Ophthalmol. Pract. Res.*, vol. 5, no. 3, pp. 189–195, 2025, doi: 10.1016/j.aopr.2025.05.001.
- [18] D. Mikhail *et al.*, "Performance of DeepSeek-R1 in ophthalmology: An evaluation of clinical decision-making and cost-effectiveness," *Br. J. Ophthalmol.*, vol. 109, no. 9, pp. 976–981, 2025, doi: 10.1136/bjo-2025-327360.
- [19] A. Arrieta, M. Ugarte, P. Valle, J. A. Parejo, and S. Segura, "o3-mini vs DeepSeek-R1: Which one is safer?," arXiv:2501.18438, 2025, doi: 10.48550/arXiv.2501.18438.
- [20] M. Liu *et al.*, "Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis," *J. Med. Internet Res.*, vol. 26, art. no. e60807, 2024, doi: 10.2196/60807.
- [21] M. Alfertshofer *et al.*, "Analyzing question characteristics influencing ChatGPT's performance in 3000 USMLE-style questions," *Med. Sci. Educ.*, vol. 35, no. 1, pp. 257–267, 2025, doi: 10.1007/s40670-024-02176-9.
- [22] A. Meyer, J. Riese, and T. Streichert, "Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: Observational study," *JMIR Med. Educ.*, vol. 10, art. no. e50965, 2024, doi: 10.2196/50965.
- [23] A. Khan, "DeepSeek vs ChatGPT: AI Platform Comparison," Kaggle, dataset versi 8. [Online]. Available: <https://www.kaggle.com/datasets/khanaakif/deepseek-vs-chatgpt-ai-platform-comparison>. Accessed: Oct. 16, 2026.
- [24] W.-L. Chiang *et al.*, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," *ICML*, vol. 235, pp. 8359–8388, 2024. <https://proceedings.mlr.press/v235/chiang24b.html>
- [25] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *NeurIPS*, vol. 36, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract.html