P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

EVALUASI KUANTITATIF RELIABILITAS DAN KECEPATAN INFERENSI PADA TIGA ARSITEKTUR LARGE LANGUAGE MODEL (ChatGPT, DeepSeek, Gemini) MENGGUNAKAN METRIK NLP BERBASIS PYTHON

Abubakar Basri[1]
[1]Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Djuanda, Bogor
E-mail: [1]abubakarbasri@unida.ac.id

ABSTRAK

Adopsi Large Language Model (LLM) yang masif dalam berbagai sektor menuntut evaluasi mendalam terhadap kapabilitas intrinsik sistem, terutama dalam merespons instruksi yang ambigu. Penelitian ini menyajikan analisis komparatif terstruktur terhadap tiga arsitektur LLM terkemuka, yaitu ChatGPT, DeepSeek, dan Gemini, dengan fokus mengukur reliabilitas faktual dan efisiensi komputasional (latensi inferensi) sebagai dampak langsung dari manipulasi ambiguitas prompt. Metodologi penelitian mengadopsi pendekatan Black Box Testing yang dikombinasikan dengan analisis kuantitatif, dilakukan secara sistematis terhadap 100 prompt yang telah diklasifikasikan ke dalam delapan kategori kompleksitas. Reliabilitas faktual diukur menggunakan metrik Cosine Similarity berbasis TF-IDF terhadap ground truth, sementara latensi direkam melalui pengukuran end-to-end berbasis stopwatch dan diproses menggunakan skrip Python pada lingkungan Google Colab. Hasil penelitian menunjukkan bahwa Gemini mencatatkan rata-rata akurasi tertinggi (66,91%), diikuti DeepSeek (64,89%) dan ChatGPT (62,10%), sementara DeepSeek unggul dalam kecepatan inferensi rata-rata (2,84 detik) berkat arsitektur Mixture of Experts. Uji korelasi Pearson pada ketiga model menghasilkan koefisien negatif (ChatGPT -0,200; Gemini -0,194; DeepSeek -0,030), yang membuktikan bahwa hipotesis speed–accuracy trade-off tradisional tidak terbukti pada ketiga arsitektur tersebut. Temuan ini diharapkan menjadi kompas empiris bagi pengguna dalam memilih model secara presisi serta literatur evaluatif bagi pengembang untuk merancang sistem AI yang lebih tangguh dan efisien.

Kata kunci: Large Language Model, halusinasi, ambiguitas prompt, cosine similarity, latensi inferensi

ABSTRACT

The massive adoption of Large Language Models (LLMs) across sectors demands an in-depth evaluation of the systems' intrinsic capabilities, particularly in responding to ambiguous instructions. This study presents a structured comparative analysis of three leading LLM architectures—ChatGPT, DeepSeek, and Gemini—focusing on factual reliability and computational efficiency (inference latency) as a direct consequence of prompt-ambiguity manipulation. The methodology adopts a Black-Box Testing approach combined with quantitative analysis, systematically applied to 100 prompts classified into eight complexity categories. Factual reliability was measured using a TF-IDF-based Cosine Similarity metric against ground truth, while latency was recorded through end-to-end stopwatch measurement and processed with Python scripts on Google Colab. Results show that Gemini achieved the highest average accuracy (66.91%), followed by DeepSeek (64.89%) and ChatGPT (62.10%), while DeepSeek led in average inference speed (2.84 seconds) owing to its Mixture-of-Experts architecture. Pearson correlation tests for all three models produced negative coefficients (ChatGPT −0.200; Gemini −0.194; DeepSeek −0.030), disproving the traditional speed–accuracy trade-off hypothesis. These findings are expected to serve as an empirical compass for users selecting models with precision and as evaluative literature for developers building more robust and efficient AI systems.

Keywords: Large Language Model, hallucination, prompt ambiguity, cosine similarity, inference latency

PENDAHULUAN

Kecerdasan buatan (Artificial Intelligence/AI) merupakan bidang ilmu komputer yang secara

fundamental berkomitmen pada pengembangan agen rasional yang dirancang untuk mengoptimalkan hasil yang diinginkan (Russell & Norvig, 2021). Evolusi teknologi informasi saat ini telah bertransisi dari era Deep Learning konvensional menuju era Generative AI, di mana Large Language Model (LLM) mutakhir seperti ChatGPT (OpenAI), DeepSeek AI, dan Gemini (Google) tidak lagi sekadar inovasi laboratorium, melainkan telah menjadi infrastruktur kritis bagi ekosistem ekonomi digital global (OpenAI, 2023).

Secara ekonomis, Bloomberg Intelligence (2023) memproyeksikan pasar Generative AI akan mencapai valuasi USD 1,3 triliun pada tahun 2032, sementara Goldman Sachs (2023) memperkirakan adopsi AI akan mendorong peningkatan PDB global sebesar 7% dalam sepuluh tahun ke depan. Namun demikian, realisasi potensi ekonomi tersebut sangat bergantung pada satu variabel kritis, yaitu reliabilitas. McKinsey & Company (2023) menekankan bahwa pada sektor berisiko tinggi (high-stakes domains) seperti layanan kesehatan, hukum, dan keuangan, toleransi terhadap kesalahan mendekati nol, sehingga tanpa akurasi, kecepatan dan efisiensi AI menjadi tidak relevan.

Sayangnya, LLM masih rentan terhadap fenomena halusinasi (hallucination), yaitu kondisi ketika model menghasilkan respons yang koheren secara linguistik namun secara faktual salah atau merupakan fabrikasi (Ji et al., 2023). Fenomena ini sejalan dengan teori Stochastic Parrots dari Bender et al. (2021), yang menyatakan bahwa LLM tidak memiliki pemahaman kognitif akan kebenaran, melainkan hanya mesin probabilistik yang memprediksi kata berdasarkan pola statistik. Ketika dihadapkan pada ambiguitas instruksi (prompt ambiguity), model cenderung melakukan "penebakan statistik" untuk mengisi kekosongan informasi, yang berujung pada produksi informasi palsu.

Permasalahan ini diperberat oleh variabilitas kinerja antararsitektur model. Pasar saat ini dibanjiri berbagai LLM yang masing-masing mengklaim sebagai yang terbaik, namun masih minim literatur yang menguji secara kuantitatif model mana yang paling tangguh dalam mempertahankan akurasi dan kecepatan ketika diberikan instruksi yang tidak ideal (White et al., 2023). Kesenjangan penelitian (research gap) ini mendasari perlunya studi eksperimental yang menggunakan pendekatan komputasional otomatis untuk membandingkan reliabilitas dan latensi antarmodel secara head-to-head.

Berdasarkan latar belakang tersebut, penelitian ini dirumuskan untuk menjawab empat pertanyaan utama: (1) bagaimana perbandingan akurasi luaran ChatGPT, Gemini, dan DeepSeek pada berbagai kategori prompt; (2) bagaimana perbedaan efisiensi waktu respons ketiga model tersebut; (3) bagaimana kecenderungan kerentanan masing-masing model terhadap penurunan akurasi pada instruksi kompleks; serta (4) apakah terdapat korelasi signifikan antara latensi respons dan skor akurasi faktual. Penelitian ini bertujuan memberikan landasan empiris bagi pengguna dan industri dalam memilih model generative AI yang paling akurat dan efisien, sekaligus memperkaya literatur Natural Language Processing (NLP) mengenai hubungan kausalitas antara ambiguitas input dan degradasi kinerja model.

Dibandingkan dengan penelitian terdahulu yang cenderung bersifat kualitatif atau hanya berfokus pada teknik prompting tanpa mengukur dampak teknisnya secara statistik (White et al., 2023), penelitian ini menawarkan kebaruan berupa kerangka kerja pengujian yang mengintegrasikan otomasi Python dengan uji statistik inferensial untuk mengukur dua dimensi performa sekaligus, yaitu reliabilitas faktual dan latensi inferensi, secara head-to-head pada tiga arsitektur LLM populer. Kontribusi ini juga melengkapi studi evaluasi hallucination pada ChatGPT yang telah dilakukan sebelumnya (Ji et al., 2023), dengan menghubungkan secara eksplisit faktor ambiguitas linguistik terhadap latensi komputasional menggunakan kalkulasi otomatis berbasis Python, sekaligus memberikan wawasan mengenai batasan kemampuan model AI generatif saat ini sebagai dasar pertimbangan etis bahwa output AI tidak boleh diterima tanpa verifikasi manusia, terutama pada sektor krusial seperti hukum dan kesehatan.

LANDASAN TEORI

Arsitektur Large Language Model

LLM modern dibangun di atas arsitektur Transformer yang memanfaatkan mekanisme

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

self-attention untuk memproses hubungan antarkata dalam suatu urutan teks secara paralel (Vaswani et al., 2017). Kapabilitas generatif LLM berskala besar seperti GPT semakin meningkat seiring bertambahnya jumlah parameter dan data pelatihan, sebagaimana ditunjukkan oleh Brown et al. (2020) melalui kemampuan few-shot learning. ChatGPT dibangun di atas arsitektur Dense Transformer yang mengaktifkan seluruh parameter pada setiap proses inferensi (OpenAI, 2023), Gemini merupakan model multimodal berkemampuan tinggi dari Google DeepMind (2023), sedangkan DeepSeek-V2 mengadopsi arsitektur Mixture-of-Experts (MoE) yang hanya mengaktifkan sebagian kecil parameter "pakar" yang relevan (sparse activation) sehingga secara teoretis lebih ekonomis dan efisien (DeepSeek-AI, 2024).

Halusinasi dan Ambiguitas Prompt

Sifat probabilistik dan stokastik LLM menyebabkan model rentan terhadap halusinasi ketika dihadapkan pada instruksi yang tidak jelas. Grice (1975) melalui Prinsip Kerjasama (Cooperative Principle) menegaskan bahwa komunikasi efektif menuntut kejelasan maksim kuantitas, kualitas, relevansi, dan cara; pelanggaran terhadap maksim tersebut pada sebuah prompt berpotensi menimbulkan multi-tafsir bagi model. Raghavan dan Behera (2023) mengklasifikasikan ambiguitas dalam NLP ke dalam tiga level, yaitu ambiguitas rendah (low), sedang (medium), dan tinggi (high), yang menjadi dasar rekayasa variabel independen pada penelitian ini. Liu dan Chilton (2022) turut menegaskan bahwa desain instruksi (prompt engineering) yang tepat merupakan determinan penting kualitas keluaran model bahasa besar.

Latensi Inferensi dan Speed–Accuracy Trade-off

Latensi inferensi didefinisikan sebagai durasi total yang dibutuhkan model sejak menerima instruksi hingga menghasilkan keluaran akhir (end-to-end latency). Miao et al. (2024) menjelaskan bahwa efisiensi decoding sangat dipengaruhi oleh arsitektur komputasi model, di mana teknik seperti speculative decoding dapat mereduksi latensi tanpa mengorbankan kualitas keluaran. Terdapat asumsi klasik mengenai speed–accuracy trade-off, yaitu bahwa waktu pemrosesan yang lebih lama memungkinkan model melakukan pencarian token dan inferensi konteks yang lebih komprehensif sehingga akurasi meningkat; asumsi ini yang diuji secara empiris melalui uji korelasi pada penelitian ini.

Cosine Similarity sebagai Metrik Evaluasi NLP

Untuk menjamin objektivitas pengukuran reliabilitas faktual, penelitian ini menggunakan metrik Cosine Similarity yang dikombinasikan dengan pembobotan TF-IDF (Term Frequency–Inverse Document Frequency) sebagaimana dijelaskan Salton (1989) dan Manning et al. (2008). Metode ini mengukur derajat kemiripan makna antara dua teks berdasarkan kedekatan arah vektornya, bukan sekadar pencocokan karakter secara persis (exact string match), sehingga mampu memproduksi skor reliabilitas yang lebih representatif terhadap kesesuaian semantik keluaran model dengan ground truth.

Penelitian Terdahulu dan Gap Penelitian

Bang et al. (2023, dalam Ji et al., 2023) telah melakukan evaluasi multitugas terhadap ChatGPT pada aspek reasoning, hallucination, dan interactivity, namun belum menghubungkan faktor ambiguitas linguistik dengan latensi komputasional secara otomatis. Mishra et al. (2024) serta Pratama dan Setiawan (2024) telah mengkaji efisiensi dan akurasi chatbot AI, tetapi masih terbatas pada satu arsitektur atau infrastruktur tertentu. Penelitian ini melengkapi kesenjangan tersebut dengan menyandingkan tiga arsitektur LLM populer secara head-to-head, menghubungkan variabel ambiguitas prompt dengan dua dimensi performa sekaligus (akurasi dan latensi) menggunakan kalkulasi Python otomatis berbasis Google Colab (Carneiro et al., 2018; Bisong, 2019), sehingga hasilnya lebih objektif dan bebas dari bias penilaian manual.

Karakteristik Teknis Ketiga Model yang Dievaluasi

ChatGPT dikembangkan oleh OpenAI melalui seri Generative Pre-trained Transformer (GPT). Sejak GPT-3 pada tahun 2020 yang terdiri atas 175 miliar parameter dengan kemampuan zero-shot learning, rilis ChatGPT pada November 2022 disusul GPT-4 pada Maret 2023 secara signifikan meningkatkan kapabilitas penalaran, kreativitas, dan akurasi

model (OpenAI, 2023). Arsitektur Dense Transformer yang digunakan ChatGPT mengaktifkan seluruh parameter pada setiap proses inferensi, sehingga cenderung menghasilkan keluaran yang stabil namun menuntut beban komputasi lebih besar pada instruksi yang kompleks.

DeepSeek-V2 dikembangkan dengan pendekatan arsitektur Mixture-of-Experts (MoE) yang secara ekonomis dan komputasional lebih efisien, karena hanya mengaktifkan sebagian kecil parameter "pakar" yang relevan untuk setiap token melalui mekanisme sparse activation (DeepSeek-AI, 2024). Karakteristik ini menjadikan DeepSeek unggul pada tugas restrukturisasi teks berbiaya komputasi rendah, tetapi berpotensi mengalami keterbatasan pada tugas yang menuntut sintesis pengetahuan lintas domain secara mendalam. Adapun Gemini, yang dikembangkan oleh Google DeepMind, merupakan model multimodal generasi terbaru yang dirancang dengan kapabilitas pemahaman lintas modalitas teks, gambar, dan kode secara terintegrasi (Google DeepMind, 2023), dan pada penelitian ini menunjukkan keunggulan relatif pada mayoritas kategori pengujian akurasi faktual maupun sintaksis.

Black-Box Testing dan Lingkungan Google Colaboratory

Pendekatan Black-Box Testing dipilih karena memungkinkan evaluasi kinerja model semata-mata berdasarkan hubungan antara input (prompt) dan output (respons), tanpa perlu mengetahui struktur internal parameter ketiga LLM yang bersifat proprietary. Seluruh eksekusi skrip Python untuk perhitungan metrik NLP dilakukan pada lingkungan Google Colaboratory, sebuah platform cloud-based programming yang terbukti efektif untuk mempercepat aplikasi deep learning tanpa memerlukan instalasi lokal (Carneiro et al., 2018; Bisong, 2019). Penggunaan Google Colab memungkinkan reproduksibilitas hasil analisis, karena seluruh proses tabulasi, perhitungan cosine similarity, dan visualisasi data dijalankan dalam satu ekosistem komputasi yang konsisten.

Hipotesis Penelitian

Berdasarkan kajian teori di atas, penelitian ini menguji hipotesis: (H1) terdapat perbedaan signifikan tingkat akurasi antara ChatGPT, DeepSeek, dan Gemini pada berbagai kategori prompt; (H2) terdapat perbedaan signifikan latensi respons antarketiga model; dan (H3) terdapat korelasi positif antara durasi latensi dan skor akurasi (speed–accuracy trade-off).

METODOLOGI

Penelitian ini menerapkan pendekatan mixed-methods yang menyinergikan data kualitatif dan kuantitatif, dengan fokus utama pada prosedur eksperimental berbasis Black-Box Testing (Sugiyono, 2019). Tahap awal dilakukan melalui survei semiterstruktur berbasis Google Form untuk memetakan profil demografis dan persepsi 44 responden terhadap reliabilitas dan ambiguitas instruksi AI generatif. Inti eksperimen dilakukan dengan memberikan 100 prompt yang terklasifikasi ke dalam delapan kategori kompleksitas—meliputi antara lain teks acak, leetspeak, teks kelebihan huruf, huruf hilang, geografi daerah pedesaan/terpencil, serta sejarah/histori—dan tiga gradien ambiguitas (rendah, sedang, tinggi) kepada ChatGPT, DeepSeek, dan Gemini melalui antarmuka web browser standar.

Data primer berupa teks keluaran dan waktu respons setiap model diinventarisasi ke dalam Microsoft Excel, kemudian diproses menggunakan skrip Python pada lingkungan Google Colaboratory (Carneiro et al., 2018) yang memanfaatkan pustaka NLP untuk perhitungan Cosine Similarity berbasis TF-IDF. Data sekunder berupa referensi faktual dari buku teks, jurnal terakreditasi, laporan resmi, serta technical paper pengembang model digunakan sebagai ground truth untuk memvalidasi keakuratan keluaran. Variabel latensi diukur sebagai end-to-end latency menggunakan instrumen stopwatch dan fungsi timestamp Python, dengan pengulangan pengujian untuk meminimalkan gangguan variabel jaringan.

Instrumen uji disusun sebanyak 100 prompt yang didistribusikan ke dalam delapan kategori kompleksitas, sebagaimana dirangkum pada Tabel 1, untuk menguji berbagai batasan kemampuan sintaksis dan kognitif ketiga arsitektur LLM secara proporsional.

Tabel 1. Distribusi Kategori Instrumen Prompt

Kategori Prompt	Jumlah
Teks acak	10

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Kategori Prompt	Jumlah
Kalimat leetspeak	10
Teks kelebihan huruf	10
Huruf hilang	10
Geografi pedesaan/terpencil	20
Sejarah/histori sub-urban	10
Kategori tambahan lain	30

Sumber: Output Google Colab

Analisis data dilakukan secara deskriptif untuk data survei, serta kuantitatif komparatif untuk data eksperimen. Tingkat halusinasi (hallucination rate) dihitung dengan membandingkan jumlah luaran tidak akurat terhadap total percobaan pada setiap gradien ambiguitas, sedangkan hubungan antara latensi dan akurasi diuji menggunakan koefisien korelasi Pearson. Sebelum pengujian utama dilakukan, peneliti melakukan kalibrasi perangkat (device calibration) untuk memastikan spesifikasi hardware tidak menimbulkan bias teknis pada pengukuran latensi antarmodel, sehingga seluruh prosedur pengujian bersifat adil (fair) dan dapat dipertanggungjawabkan secara ilmiah.

Sebelum data dianalisis lebih lanjut, dilakukan tahap eksplorasi data (exploratory data analysis) yang mencakup statistik deskriptif dan deteksi outlier untuk memastikan tidak ada nilai anomali yang mendistorsi interpretasi hasil. Selanjutnya, tahap prapemrosesan (preprocessing) dan pembersihan data (data cleaning) diterapkan pada teks keluaran model, meliputi normalisasi kapitalisasi, penghapusan tanda baca berlebih, dan tokenisasi, sebelum teks tersebut dikonversi menjadi representasi vektor TF-IDF untuk perhitungan cosine similarity. Seluruh tahapan ini dirancang untuk menjamin bahwa perbedaan skor akurasi yang teramati murni mencerminkan kapabilitas model, bukan artefak dari inkonsistensi format data.

Penelitian ini dibatasi pada versi publik/API standar ChatGPT (GPT-3.5 Turbo/GPT-4o-mini), DeepSeek (DeepSeek-V2/Chat), dan Gemini (Gemini 1.5 Flash/Pro) yang tersedia pada rentang waktu pengambilan data. Domain instrumen uji dibatasi pada pengetahuan umum berkebenaran mutlak (sejarah, geografi, sains dasar), sehingga tidak mencakup kemampuan kreatif, pemrograman, atau penalaran matematika kompleks. Batasan ini perlu dipertimbangkan dalam menggeneralisasi temuan penelitian ke domain aplikasi LLM lainnya.

HASIL DAN PEMBAHASAN

Karakteristik Demografi dan Persepsi Responden

Survei awal melibatkan 44 responden yang didominasi laki-laki (65,9%), berstatus mahasiswa aktif (84,1%), dan berada pada rentang usia 18–22 tahun (75,0%) yang merupakan kelompok pengadopsi awal (early adopters) teknologi AI generatif. Dari sisi preferensi arsitektur, ChatGPT menunjukkan hegemoni sebagai model yang paling banyak digunakan (63,6%), jauh mengungguli Gemini (20,5%) dan DeepSeek (2,3%), sehingga menimbulkan pertanyaan apakah dominasi popularitas ChatGPT sungguh berbanding lurus dengan superioritas performanya, atau justru arsitektur Mixture of Experts milik DeepSeek memiliki performa laten yang luput dari perhatian publik. Mayoritas responden juga mendelegasikan LLM untuk domain berisiko tinggi, yaitu riset akademik (27,3%) dan pemrograman (20,5%), yang menuntut akurasi faktual dan sintaksis ketat.

Terkait manifestasi kegagalan, 52,3% responden melaporkan pengalaman "salah memahami konteks" sebagai kendala paling dominan saat berinteraksi dengan AI generatif, diikuti oleh generalisasi berlebihan dan halusinasi ekstrinsik berupa fabrikasi fakta. Sejalan dengan itu, 63,6% responden menilai kemampuan mereka menyusun prompt baru berada pada taraf "cukup efektif", bukan "sangat efektif", yang menunjukkan bahwa celah ambiguitas (ambiguity gap) pada interaksi manusia-AI masih terbuka lebar. Temuan kualitatif ini memperkuat landasan empiris bagi pengujian kuantitatif pada sub-bab berikutnya, sekaligus menegaskan bahwa kendala ambiguitas instruksi merupakan realitas nyata yang dihadapi pengguna, bukan sekadar asumsi teoretis.

Tabel 2. Manifestasi Kegagalan Output AI Menurut Persepsi Pengguna

Manifestasi Kegagalan	Frekuensi	Persentase
Salah memahami konteks	23	52,3%
Generalisasi berlebihan	6	13,6%
Halusinasi ekstrinsik	5	11,4%

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Manifestasi Kegagalan	Frekuensi	Persentase
Kesalahan data numerik	4	9,1%
Abstain/tidak menjawab	6	13,6%

Sumber: Kuesioner Google Form

Dominasi kategori "salah memahami konteks" pada Tabel 2 merepresentasikan adanya degradasi pemahaman semantik yang linear dengan tingginya ambiguitas (entropy) pada ruang prompting, yang secara langsung relevan dengan variabel bebas ambiguitas prompt yang dimanipulasi pada tahap eksperimen kuantitatif berikutnya.

Kalibrasi Perangkat dan Standardisasi Lingkungan Pengujian

Sebelum pengujian utama dilaksanakan, peneliti melakukan kalibrasi perangkat dan standardisasi lingkungan observasi menggunakan perangkat mobile untuk memastikan bahwa durasi latensi yang tercatat murni berasal dari efisiensi komputasi cloud server masing-masing model, bukan akibat bottleneck spesifikasi perangkat lokal atau fluktuasi jaringan. Prosedur ini mencakup isolasi spesifikasi hardware, kalibrasi waktu ekstraksi berkas, pengujian time to first token pada variasi kecepatan jaringan, serta pembersihan sesi (session environment clearing) di antara setiap pengujian model untuk menghindari kontaminasi konteks percakapan sebelumnya. Melalui standardisasi ini, seluruh perbandingan latensi antar-ChatGPT, DeepSeek, dan Gemini dapat dianggap adil (fair) dan bebas dari bias teknis eksternal.

Analisis Akurasi Jawaban Model Berdasarkan Kategori Prompt

Evaluasi akurasi dilakukan dengan membandingkan 100 keluaran setiap model terhadap ground truth menggunakan Cosine Similarity berbasis TF-IDF. Secara keseluruhan, Gemini mencatatkan rata-rata akurasi tertinggi, disusul DeepSeek, kemudian ChatGPT, sebagaimana disajikan pada Tabel 3.

Tabel 3. Rata-rata Akurasi Keseluruhan per Model

Model	Rata-rata Akurasi (%)
ChatGPT	62,10
DeepSeek	64,89
Gemini	66,91

Sumber: Output Google Colab

Ditinjau per kategori, Gemini unggul pada lima dari delapan kategori, termasuk teks kelebihan huruf dan angka (94,49%), huruf hilang (76,76%), sejarah kurang populer (69,33%), geografi lokal (56,65%), dan terjemahan bebas (48,42%). DeepSeek unggul pada kategori pemrosesan teks yang terdistorsi secara visual, yakni teks kelebihan huruf (90,06%) dan teks terbalik (78,88%), namun mengalami penurunan signifikan pada kategori yang membutuhkan pemahaman konteks faktual seperti geografi lokal (49,97%) dan terjemahan bebas (43,66%). ChatGPT tidak menempati posisi tertinggi pada kategori manapun, tetapi tampil paling stabil dan konsisten di posisi tengah, misalnya pada sejarah (64,83%) dan geografi lokal (53,37%), tanpa fluktuasi ekstrem seperti yang dialami DeepSeek.

Tabel 4. Skor Tertinggi per Kategori Prompt

Kategori Prompt	Model Terbaik	Skor (%)
Kelebihan huruf & angka	Gemini	94,49
Huruf hilang	Gemini	76,76
Sejarah kurang populer	Gemini	69,33
Geografi lokal	Gemini	56,65
Terjemahan bebas	Gemini	48,42
Kelebihan huruf	DeepSeek	90,06
Teks terbalik	DeepSeek	78,88

Sumber: Output Google Colab

Analisis Waktu Respons Model

Evaluasi waktu respons menunjukkan bahwa DeepSeek secara konsisten mencatatkan durasi tercepat pada kategori modifikasi sintaksis, seperti teks kelebihan huruf (1,47 detik) dan terjemahan bebas (1,93 detik), berkat efisiensi arsitektur Mixture of Experts (MoE) yang hanya mengaktifkan sebagian parameter relevan (sparse activation). Sebaliknya, kategori yang menuntut penarikan memori faktual, seperti sejarah kurang populer dan geografi lokal, memberikan beban komputasi tertinggi dengan durasi terlama dicatatkan oleh Gemini (5,60 detik), DeepSeek (4,83 detik), dan ChatGPT (4,50 detik). Pola ini mencerminkan karakteristik arsitektur Dense Transformer yang memaksa model mengaktifkan parameter secara ekstensif untuk memverifikasi fakta dan memitigasi halusinasi.

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Tabel 5. Rata-rata Waktu Respons pada Kategori Tercepat dan Terberat (detik)

Kategori Prompt	ChatGPT	DeepSeek	Gemini
Kelebihan huruf (tercepat)	-	1,47	-
Terjemahan bebas	-	1,93	-
Sejarah/geografi (terberat)	4,50	4,83	5,60
Rata-rata keseluruhan	-	2,84	-

Sumber: Output Google Colab

Korelasi Waktu Respons dan Akurasi

Uji korelasi Pearson dilakukan untuk memvalidasi asumsi speed–accuracy trade-off, yaitu apakah waktu pemrosesan yang lebih lama berbanding lurus dengan akurasi yang lebih tinggi. Hasil pengujian pada Tabel 6 menunjukkan bahwa ketiga model justru mencatatkan koefisien korelasi negatif, yang berarti penambahan waktu pemrosesan tidak berbanding lurus dengan peningkatan akurasi keluaran.

Tabel 6. Koefisien Korelasi Pearson antara Latensi dan Akurasi

Model	Koefisien Korelasi (r)	Interpretasi
ChatGPT	-0,200	Negatif lemah
Gemini	-0,194	Negatif sangat lemah
DeepSeek	-0,030	Nyaris tidak berkorelasi

Sumber: Output Google Colab

Kecenderungan nilai korelasi negatif pada ketiga model mengindikasikan bahwa waktu decoding yang memanjang bukan mencerminkan proses penalaran yang lebih baik, melainkan menandakan model sedang mengalami hambatan komputasional dalam memecahkan prompt yang kompleks, seperti kebingungan menyusun konteks kalimat terdistorsi atau memvalidasi memori historis. Dengan demikian, tidak terdapat kompromi (trade-off) yang menguntungkan antara pengorbanan waktu pemrosesan dan perolehan tingkat akurasi pada evaluasi ketiga arsitektur generative AI ini, sehingga hipotesis H3 (korelasi positif) ditolak, sedangkan H1 dan H2 diterima.

Sintesis Hasil Pengujian

Secara umum, efektivitas generative AI tidak bersifat absolut, melainkan bergantung pada karakteristik dan tingkat kerumitan instruksi yang diberikan. Gemini dan DeepSeek unggul tajam pada instruksi berbasis modifikasi sintaksis dan restrukturisasi karakter visual, sementara ChatGPT menawarkan tingkat kestabilan akurasi yang lebih konsisten meskipun bukan yang tertinggi pada pengujian faktual. Dari sisi efisiensi komputasional, arsitektur Mixture of Experts pada DeepSeek terbukti unggul dalam kecepatan pemrosesan instruksi dasar, sementara arsitektur Dense Transformer menuntut beban komputasi lebih besar ketika memverifikasi fakta guna memitigasi halusinasi. Temuan paling krusial dalam penelitian ini adalah terpatahkannya hipotesis speed–accuracy trade-off tradisional: waktu "berpikir" yang lebih lama justru mengindikasikan hambatan komputasional dan kebingungan model (entropy), bukan peningkatan kualitas jawaban. Kinerja model AI generatif paling optimal dicapai ketika instruksi selaras dengan spesialisasi arsitektur dasarnya, di mana eksekusi dapat berjalan cepat tanpa mengorbankan akurasi.

Implikasi Teoritis dan Praktis

Secara teoretis, temuan ini memperkaya literatur NLP dengan bukti empiris kuantitatif mengenai hubungan kausalitas antara ambiguitas input dan degradasi kinerja model, sekaligus memberikan kontribusi terhadap teori kompleksitas komputasi melalui pembuktian bahwa fenomena speed–accuracy trade-off tidak berlaku universal pada arsitektur LLM kontemporer. Terpatahkannya hipotesis trade-off tradisional mengindikasikan bahwa waktu inferensi yang lebih lama pada instruksi ambigu lebih tepat dimaknai sebagai sinyal inefisiensi algoritma dalam menangani ketidakpastian (entropy), bukan sebagai proses deliberasi yang menghasilkan jawaban lebih baik.

Secara praktis, hasil penelitian ini menawarkan data benchmarking teknis yang dapat menjadi rujukan bagi pengembang dalam memilih arsitektur (tech stack) yang paling efisien untuk aplikasi real-time, sehingga dapat menyeimbangkan kecepatan layanan dengan biaya server (token cost). Bagi pengguna, khususnya pada domain berisiko tinggi seperti riset akademik, hukum, dan kesehatan, temuan ini menegaskan pentingnya verifikasi manusia (human in the loop) terhadap output AI, mengingat tidak satu pun

dari ketiga model yang diuji mampu mempertahankan akurasi tinggi secara konsisten pada seluruh kategori instruksi, khususnya kategori yang menuntut penarikan fakta spasial dan historis.

Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu menjadi perhatian pada studi lanjutan. Pertama, pengukuran latensi menggunakan instrumen stopwatch manual berpotensi menimbulkan variasi kecil akibat waktu reaksi pengamat, meskipun telah diminimalkan melalui pengulangan pengujian dan kalibrasi perangkat. Kedua, mengingat sifat LLM yang terus diperbarui secara berkelanjutan (continuous update), data penelitian ini merupakan snapshot pada rentang waktu tertentu sehingga perubahan performa model di luar rentang tersebut berada di luar cakupan penelitian. Ketiga, instrumen uji terbatas pada domain pengetahuan umum berkebenaran mutlak, sehingga generalisasi temuan pada domain kreatif, pemrograman, atau penalaran matematis kompleks memerlukan penelitian lanjutan dengan instrumen yang berbeda.

KESIMPULAN

Penelitian ini membuktikan bahwa reliabilitas faktual dan efisiensi komputasional LLM ChatGPT, DeepSeek, dan Gemini tidak bersifat absolut, melainkan sangat bergantung pada interaksi antara kompleksitas semantik instruksi (prompt ambiguity) dan arsitektur pemrosesan internal model. Gemini unggul dalam rata-rata akurasi semantik keseluruhan (66,91%), sedangkan DeepSeek mendominasi efisiensi kecepatan inferensi (rata-rata 2,84 detik) berkat arsitektur Mixture of Experts. Seluruh model mengalami degradasi performa dan peningkatan kerentanan halusinasi ketika dihadapkan pada instruksi yang menuntut penarikan memori faktual dan kontekstual mendalam, seperti sejarah dan geografi lokal. Temuan terpenting penelitian ini adalah terpatahkannya hipotesis speed–accuracy trade-off tradisional, di mana koefisien korelasi Pearson bernilai negatif pada ketiga model membuktikan bahwa durasi pemrosesan yang lebih lama tidak berkontribusi positif terhadap kualitas jawaban, melainkan mengindikasikan hambatan komputasional model. Dengan demikian, optimalisasi penggunaan generative AI menuntut keselarasan antara desain instruksi yang minim ambiguitas dan karakteristik arsitektur model, serta tetap memerlukan pengawasan manusia (human in the loop) guna memitigasi risiko kegagalan faktual pada domain-domain kritis seperti riset akademik, hukum, dan kesehatan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Djuanda, Bogor, atas dukungan fasilitas dan bimbingan akademik selama penelitian ini berlangsung, serta kepada seluruh responden yang telah berpartisipasi dalam survei pendahuluan penelitian ini.

DAFTAR PUSTAKA

Askell, A., Bai, Y., Chen, A., Dawn, B., Hernandez, D., & McCandlish, S. (2021). A general language assistant as a laboratory for alignment. Anthropic. https://arxiv.org/abs/2112.00861

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Bisong, E. (2019). Building machine learning and deep learning models on Google Cloud Platform: A comprehensive guide for beginners. Apress. https://doi.org/10.1007/978-1-4842-4470-8

Bloomberg Intelligence. (2023). Generative AI to become a $1.3 trillion market by 2032. Bloomberg Intelligence Report.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., & Dhariwal, P. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Carneiro, T., Da Nóbrega, R. V. M., Nepomuceno, T., Bian, G. B., De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications. IEEE Access, 6, 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767

Chowdhary, K. R. (2020). Natural language processing. In Fundamentals of artificial

intelligence. Springer. https://doi.org/10.1007/978-81-322-3972-7_19

DeepSeek-AI. (2024). DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model (Technical Report). https://arxiv.org/abs/2405.04434

Fromkin, V., Rodman, R., & Hyams, N. (2018). An introduction to language (11th ed.). Cengage Learning.

Goldman Sachs. (2023). The potentially large effects of artificial intelligence on economic growth. Goldman Sachs Economics Research.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Google DeepMind. (2023). Gemini: A family of highly capable multimodal models (Technical Report). https://arxiv.org/abs/2312.11805

Grice, H. P. (1975). Logic and conversation. In Speech acts (pp. 41–58). Brill.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed. draft). Stanford University.

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering with large language models. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3491102.3517732

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

McKinsey & Company. (2023). The economic potential of generative AI: The next productivity frontier. McKinsey Global Institute.

Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., & Zheng, Z. (2024). Speculative decoding: Exploiting speculative execution for accelerating seq2seq latency (arXiv preprint). https://arxiv.org/abs/2305.09781

Mishra, S., et al. (2024). Quantitative evaluation of LLMs: Efficiency, reliability, and accuracy. Journal of Artificial Intelligence Research, 78, 112–145.

OpenAI. (2023). GPT-4 technical report (arXiv:2303.08774).

Pratama, A., & Setiawan, B. (2024). Analisis performa kecepatan respons chatbot AI pada infrastruktur cloud dan lokal. Jurnal Ilmu Komputer dan Informatika, 5(1), 45–58.

Raghavan, R., & Behera, S. K. (2023). Ambiguity in natural language processing: A systematic review. Journal of Computer Science, 19(4), 512–528.

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193. https://doi.org/10.3390/info11040193

Russell, S. J., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Addison-Wesley.

Sugiyono. (2019). Metode penelitian kuantitatif, kualitatif, dan R&D. Alfabeta.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., & Gilbert, J. E. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT (arXiv preprint arXiv:2302.11382).