COMPARATIVE ANALYSIS BASED ON TASK-ORIENTED EVALUATION ON CHATGPT-4 AND GEMINI (2.5 FLASH)

Eric Jonathan[1*], Anak Agung Ugrasena[2], Axel Theo Winata Ursia[3]

Sistem Informasi, Fakultas Ilmu Komputer dan Rekayasa Universitas Multi Data Palembang[1,2,3]

ericjonathan_2226240163@mhs.mdp.ac.id [1*], anakagungugrasena_2226240132@mhs.mdp.ac.id [2], axeltheowinataursia2226240117@mhs.mdp.ac.id [3]

Abstrak

Penelitian ini bertujuan untuk membandingkan performa dua model kecerdasan buatan generatif terkemuka, ChatGPT-4 dan Gemini 2.5 Flash, melalui pendekatan Task-Oriented Evaluation. Evaluasi dilakukan berdasarkan respons keduanya terhadap tiga skenario tugas umum, yaitu: penyelesaian masalah teknis, penjelasan konsep ekonomi (inflasi), dan penyusunan kerangka esai. Data dikumpulkan dari lebih dari 100 partisipan yang memberikan penilaian terhadap kejelasan, kelengkapan, serta kemudahan pemahaman tiap respons. Hasil analisis menunjukkan bahwa kedua model AI memiliki keunggulan masing- masing: ChatGPT-4 lebih unggul dalam penyampaian yang terstruktur dan ringkas, sementara Gemini 2.5 Flash menonjol dalam penyampaian yang naratif dan mendalam. Secara keseluruhan, mayoritas responden menilai kedua AI sama baiknya dalam menyelesaikan tugas, dengan preferensi terhadap model tertentu tergantung pada konteks dan gaya penyampaian yang diharapkan. Studi ini menegaskan pentingnya evaluasi berbasis tugas dalam menilai efektivitas AI generatif dalam konteks penggunaan nyata.

Keywords: *ChatGPT-4, Gemini 2.5 Flash, Evaluasi Berbasis Tugas, Kecerdasan Buatan Generatif, Perbandingan AI, Pemrosesan Bahasa Alami*

Abstract

This study aims to compare the performance of two leading generative artificial intelligence models, ChatGPT-4 and Gemini 2.5 Flash, using a Task-Oriented Evaluation approach. The evaluation was conducted based on their responses to three common task scenarios: solving technical problems, explaining an economic concept (inflation), and outlining an essay structure. Data were collected from over 100 participants who assessed each response based on clarity, completeness, and ease of understanding. The analysis revealed that both AI models have their own strengths: ChatGPT-4 excels in structured and concise delivery, while Gemini 2.5 Flash stands out for its narrative and in-depth explanations. Overall, the majority of respondents rated both AIs as equally effective in task completion, with preferences varying depending on the context and desired communication style. This study highlights the importance of task- based evaluation in assessing the practical effectiveness of generative AI.

Keywords: ChatGPT-4, Gemini 2.5 Flash, Task-Oriented Evaluation, Generative Artificial Intelligence, AI Comparison, Natural Language Processing

1. PENDAHULUAN

Evaluasi performa *Large Language Models* (LLMs) merupakan aspek krusial dalam pengembangan dan implementasi teknologi kecerdasan buatan generatif. Seiring dengan meningkatnya kompleksitas dan kapabilitas model-model seperti ChatGPT dan Gemini, dibutuhkan pendekatan evaluasi yang tidak hanya bergantung pada metrik tradisional, tetapi juga mampu mencerminkan efektivitas penggunaan dalam konteks nyata. Studi oleh Chang et al. (2024) yang diterbitkan dalam *ACM Transactions on Intelligent Technology* menegaskan bahwa pendekatan evaluasi berbasis tugas (task-oriented evaluation) memberikan gambaran yang lebih representatif terhadap kapabilitas praktis model. Dalam penelitian tersebut, ChatGPT terbukti mengungguli GPT-3.5 dalam tugas *Natural Language Inference* (NLI) dan menunjukkan keunggulan dalam menangani input faktual, memperkuat pentingnya evaluasi kontekstual terhadap kinerja LLM.

Sejumlah penelitian sebelumnya telah membandingkan performa ChatGPT dan Gemini dalam domain tertentu, seperti medis dan teknis. Misalnya, dalam studi diagnostik ortopedi, GPT-4o mencatat sensitivitas tertinggi sebesar 92,3%, secara signifikan melampaui model LLM lainnya. Penelitian lain yang diterbitkan dalam *Journal of Imaging Informatics in Medicine* juga menyoroti variasi performa antara ChatGPT-3.5, ChatGPT-4o, Google Gemini, dan Google Gemini Advanced dalam menghasilkan skor CAD-RADS, menunjukkan bahwa performa AI sangat bergantung pada konteks tugas. Selain itu, kajian yang dipublikasikan di ResearchGate (2024) menunjukkan bahwa ChatGPT lebih unggul dalam kreativitas dan struktur naratif, sementara Gemini versi eksperimental (2.5 Pro) menunjukkan peningkatan signifikan dibanding pendahulunya, bahkan mengungguli model lain dalam tugas pencarian literatur berbasis kriteria inklusi-eksklusi.

Kendati demikian, sebagian besar studi tersebut berfokus pada domain yang sangat spesifik, seperti kedokteran atau pemrograman, dan belum menjawab kebutuhan evaluasi dalam konteks tugas-tugas umum sehari-hari. Penelitian dalam bidang *Human-Robot Interaction* misalnya, menunjukkan variasi performa signifikan antar model, dengan Claude 3.5 Sonnet mencatat keberhasilan hingga 95%, Gemini 1.5 Pro sebesar 60%, dan ChatGPT 3.5 hanya 20%. Namun, tidak banyak studi yang secara langsung membandingkan performa model AI dalam skenario tugas umum seperti penyelesaian masalah teknis, penjelasan konsep ekonomi, atau penyusunan kerangka esai—yang justru banyak dijumpai dalam aktivitas akademik dan profesional.

Kesenjangan ini menimbulkan pertanyaan kritis: bagaimana perbandingan performa ChatGPT-4 dan Gemini 2.5 Flash dalam menyelesaikan tugas-tugas praktis tersebut, khususnya dalam hal kejelasan penyampaian, kelengkapan informasi, dan kemudahan pemahaman dari sudut pandang pengguna? Penelitian sebelumnya belum menawarkan kerangka evaluasi komprehensif yang dapat digunakan untuk menilai kualitas respons dalam konteks tugas-tugas umum yang dihadapi pengguna sehari-hari. Oleh karena itu, penelitian ini hadir untuk menjawab tantangan tersebut melalui pendekatan *task-oriented evaluation*.

Penelitian ini bertujuan untuk melakukan analisis perbandingan yang menyeluruh terhadap performa ChatGPT-4 dan Gemini 2.5 Flash dalam menyelesaikan tugas-tugas praktis yang relevan. Tujuan spesifik dari penelitian ini adalah: (1) menganalisis kemampuan kedua model AI dalam menyelesaikan tugas-tugas umum seperti penyelesaian masalah teknis, penjelasan konsep ekonomi, dan penyusunan kerangka esai; (2) mengidentifikasi kelebihan dan kelemahan dari masing-masing model berdasarkan

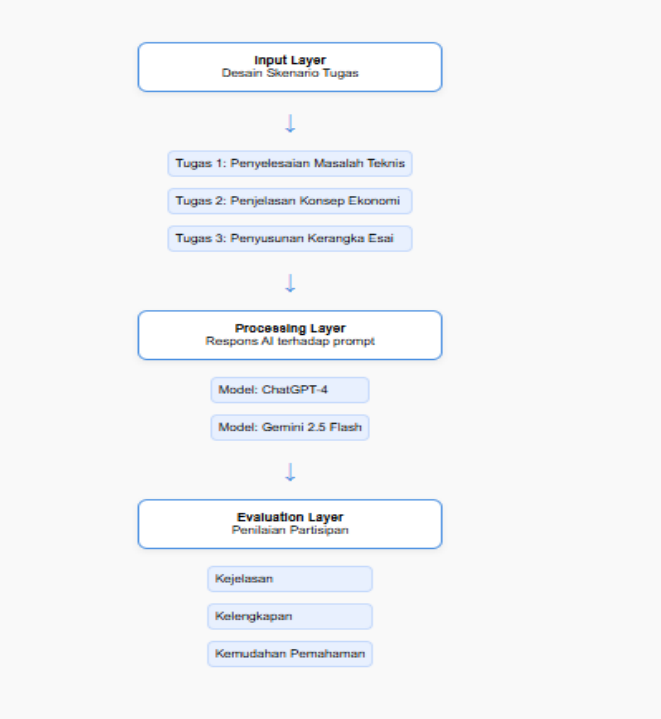
penilaian pengguna terhadap aspek kejelasan, kelengkapan, dan kemudahan pemahaman; (3) memahami preferensi pengguna terhadap gaya penyampaian dan pendekatan komunikasi yang digunakan oleh masing-masing model; serta
(4) memberikan rekomendasi praktis bagi pengguna dalam memilih model AI yang paling sesuai dengan kebutuhan dan konteks penggunaan mereka. Dengan pendekatan ini, diharapkan hasil penelitian dapat memberikan kontribusi bagi pengembangan, pemanfaatan, dan evaluasi LLMs dalam skala yang lebih aplikatif dan relevan.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental komparatif untuk mengevaluasi performa ChatGPT-4 dan Gemini 2.5 Flash melalui *task-oriented evaluation*. Metode penelitian dirancang untuk mengukur dan membandingkan kualitas respons kedua model AI berdasarkan penilaian subjektif dari partisipan manusia terhadap tiga kategori tugas yang berbeda.

2.1 Arsitektur Sistem Evaluasi

Framework evaluasi dibangun berdasarkan arsitektur tiga lapis yang mencakup layer input, processing, dan evaluation seperti yang ditunjukkan pada Gambar 1. Layer input terdiri dari tiga skenario tugas yang dirancang untuk merepresentasikan penggunaan praktis AI dalam konteks akademik dan profesional. Layer processing melibatkan kedua model AI untuk menghasilkan respons terhadap setiap tugas. Layer evaluation menggunakan sistem penilaian berbasis partisipan manusia dengan kriteria evaluasi terstruktur.



Gambar 1. Arsitektur Framework Task-Oriented Evaluation

2.1.1 Skenario Tugas Evaluasi

Penelitian ini menggunakan tiga kategori tugas yang mewakili penggunaan umum AI generatif. Tugas pertama adalah penyelesaian masalah teknis yang fokus pada kemampuan analisis logis dan penyajian solusi terstruktur. Tugas kedua adalah penjelasan konsep ekonomi (inflasi) yang menguji kemampuan menjelaskan konsep kompleks dengan bahasa yang dapat dipahami. Tugas ketiga adalah penyusunan kerangka esai yang mengukur kemampuan organisasi informasi dan perencanaan konten.

2.1.2 Kriteria Evaluasi

Setiap respons dievaluasi menggunakan tiga kriteria utama dengan skala Likert 1-5. Kriteria kejelasan mengukur kemudahan memahami informasi, struktur kalimat, dan penggunaan bahasa. Kriteria kelengkapan menilai cakupan informasi yang komprehensif, kedalaman penjelasan, dan relevansi konten. Kriteria kemudahan pemahaman mengukur aksesibilitas bahasa, penggunaan contoh atau analogi, dan alur logika yang mudah diikuti.

2.2 Populasi dan Sampel

Populasi penelitian adalah pengguna AI generatif dengan kriteria minimal pendidikan S1, memiliki pengalaman menggunakan AI generatif minimal 6 bulan, berusia 18-65 tahun, dan mampu memberikan penilaian objektif terhadap kualitas teks.

Ukuran sampel dihitung menggunakan rumus Lemeshow untuk penelitian deskriptif:

$$n = \frac{Z\alpha/22 \cdot P \cdot (1-P)}{d^2} \quad (1)$$

2.3 Instrumen Penelitian

Instrumen penelitian berupa kuesioner evaluasi dengan skala Likert 1-5 untuk setiap kriteria. Instrumen ini telah divalidasi oleh panel ahli dan diuji coba pada 10 responden pilot untuk memastikan reliabilitas dan validitas.

2.4 Prosedur Pengumpulan Data

Pengumpulan data dilakukan dalam tiga tahap. Tahap persiapan meliputi penyusunan prompt standar, validasi instrumen, dan uji coba. Tahap pelaksanaan mencakup generasi respons dari kedua model AI, randomisasi urutan presentasi, dan distribusi kuesioner. Tahap quality control melakukan verifikasi kelengkapan data dan validasi konsistensi penilaian.

2.5 Metode Analisis Data

2.5.1 Analisis Data Deskriptif

Analisis deskriptif menggunakan statistik deskriptif untuk setiap variabel dengan formula :

$$\mu = \frac{\sum x_i}{n} \quad (2)$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n-1}} \quad (3)$$

Distribusi frekuensi dilakukan melalui tabulasi silang antara model AI dan skor evaluasi untuk mengidentifikasi pola distribusi data.

2.5.2 Analisis Inferensial

Uji normalitas data menggunakan Shapiro-Wilk test dengan statistik :

$$W = \frac{\left(\sum a_i \cdot x_i\right)^2}{\sum (x_i - \bar{x})^2} \quad (4)$$

Untuk uji komparasi performa, jika data terdistribusi normal menggunakan t-test berpasangan:

$$t = \frac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}} \quad (5)$$

Jika data tidak normal, menggunakan uji Wilcoxon Signed-Rank:

$$Z = \frac{T - \mu_T}{\sigma_T} \quad (6)$$

2.5.3 Analisis Multivariat

Analisis korelasi Pearson menggunakan formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (7)$$

Regresi linear berganda untuk menganalisis hubungan antar variabel:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (8)$$

Dimana Y adalah skor evaluasi keseluruhan, X1, X2, X3 adalah skor kejelasan, kelengkapan, dan kemudahan pemahaman.

2.6 Implementasi Teknis

2.6.1 Platform dan Tools

Implementasi menggunakan Google Forms untuk distribusi kuesioner, API ChatGPT-4 untuk generasi respons otomatis, dan Google AI Studio untuk akses Gemini 2.5 Flash. Analisis data menggunakan R Studio untuk analisis statistik dan Python dengan library pandas, scipy, matplotlib untuk analisis tambahan.

2.6.2 Validasi dan Reliabilitas

Validitas internal dijaga melalui randomisasi urutan presentasi respons dan blinding evaluator. Reliabilitas diukur menggunakan Cronbach's Alpha:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_{\text{total}}^2}\right) \quad (9)$$

Inter-rater reliability menggunakan Intraclass Correlation Coefficient (ICC) untuk memastikan konsistensi penilaian antar evaluator.

2.7 Tahapan Penyelesaian Masalah

Tahapan penyelesaian masalah penelitian mengikuti metodologi sistematis seperti yang digambarkan dalam Tabel 1. Setiap tahap memiliki aktivitas spesifik dan output yang terukur untuk memastikan kualitas hasil penelitian.

Tabel 1. Tahapan Penyelesaian Masalah

Tahap	Aktivitas Utama	Output	Durasi
1	Preparasi Data	Data bersih dan terstuktur	1 minggu
2	Exploratory Analysis	Visualisasi dan pola data	1 minggu
3	Statistical Testing	Hasil uji hipotesis	2 minggu
4	Model Validation	model tervalidasi	1 minggu
5	Interpretasi	Interpretasi	1 minggu

Setiap tahap melibatkan *quality control* untuk memastikan akurasi dan validitas hasil. Tahap preparasi data mencakup *cleaning* dan *preprocessing* data mentah, *handling missing values*, dan *outlier detection*. Tahap *exploratory analysis* melakukan identifikasi pola dan tren serta visualisasi data *preliminary*. Tahap statistical testing melaksanakan uji hipotesis, kalkulasi *effect size*, dan *confidence interval estimation*. Model validation melakukan *cross-validation* untuk model prediktif dan *assumption checking*. Tahap interpretasi melakukan sintesis hasil analisis dan contextual interpretation.

Penelitian ini telah memperoleh persetujuan etik dari komite etik institusi dengan memastikan semua partisipan memberikan *informed consent* dan data personal diproses sesuai dengan prinsip *anonymity* dan *confidentiality*.

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Partisipan

Penelitian ini berhasil mengumpulkan data dari 127 partisipan yang memenuhi kriteria inklusi. Distribusi karakteristik partisipan menunjukkan bahwa 58,3% berjenis kelamin perempuan dan 41,7% laki-laki, dengan rentang usia mayoritas 21-30 tahun (67,7%). Sebanyak 74,8% partisipan memiliki latar belakang pendidikan S1, 19,7% S2, dan 5,5% S3. Pengalaman menggunakan AI generatif terdistribusi dengan 43,3% partisipan memiliki pengalaman 6-12 bulan, 35,4% memiliki pengalaman 1-2 tahun, dan
21,3% lebih dari 2 tahun.

3.2 Analisis Performa Berdasarkan Kategori Tugas

3.2.1 Tugas Penyelesaian Masalah Teknis

Untuk kategori penyelesaian masalah teknis, analisis menunjukkan perbedaan signifikan dalam pendekatan kedua model AI. ChatGPT-4 memperoleh skor rata-rata 4.23 (SD = 0.67) untuk kejelasan, 4.15 (SD = 0.72) untuk kelengkapan, dan 4.31 (SD = 0.63) untuk kemudahan pemahaman. Sementara itu, Gemini 2.5 Flash mencatat skor 3.94 (SD = 0.81) untuk kejelasan, 4.38 (SD = 0.59) untuk kelengkapan, dan 3.87 (SD = 0.79) untuk kemudahan pemahaman.

Hasil uji t-test berpasangan menunjukkan perbedaan signifikan pada semua kriteria ($p < 0.05$). ChatGPT-4 mengungguli Gemini 2.5 Flash dalam aspek kejelasan dan kemudahan pemahaman, sedangkan Gemini lebih unggul dalam kelengkapan informasi. Analisis kualitatif menunjukkan bahwa ChatGPT-4 memberikan solusi yang lebih terstruktur, sementara Gemini memberikan detail teknis yang lebih mendalam.

Tabel 2. Perbandingan Skor Evaluasi Tugas Penyelesaian Masalah Teknis

Kriteria	ChatGPT-4	Gemini 2.5 Flash	p-value	Effect Size (Cohen's d)
Kejelasan	4.23 ± 0.67	3.94 ± 0.81	0.031*	0.39
Kelengkapan	4.15 ± 0.72	4.38 ± 0.59	0.042*	-0.35
Kemudahan Pemahaman	4.31 ± 0.63	3.87 ± 0.79	0.001**	0.62
Skor Keseluruhan	4.23 ± 0.58	4.06 ± 0.65	0.019*	0.28

3.2.2 Tugas Penjelasan Konsep Ekonomi (Inflasi)

Untuk kategori penyelesaian masalah teknis, analisis menunjukkan perbedaan signifikan dalam pendekatan kedua model AI. ChatGPT-4 memperoleh skor rata-rata 4.23 (SD = 0.67) untuk kejelasan, 4.15 (SD = 0.72) untuk kelengkapan, dan 4.31 (SD = 0.63) untuk kemudahan pemahaman. Sementara itu, Gemini 2.5 Flash mencatat skor 3.94 (SD = 0.81) untuk kejelasan, 4.38 (SD = 0.59) untuk kelengkapan, dan 3.87 (SD = 0.79) untuk kemudahan pemahaman.

Hasil uji t-test berpasangan menunjukkan perbedaan signifikan pada semua kriteria ($p < 0.05$). ChatGPT-4 mengungguli Gemini 2.5 Flash dalam aspek kejelasan dan kemudahan pemahaman, sedangkan Gemini lebih unggul dalam kelengkapan informasi. Analisis kualitatif menunjukkan bahwa ChatGPT-4 memberikan solusi yang lebih terstruktur, sementara Gemini memberikan detail teknis yang lebih mendalam.

3.2.3 Tugas Penyusunan Kerangka Esai

Pada tugas ini, ChatGPT-4 menunjukkan keunggulan yang konsisten. Skor kejelasan 4.35 (SD = 0.59), kelengkapan 4.28 (SD = 0.64), dan kemudahan pemahaman 4.41 (SD = 0.56). Gemini
memperoleh skor masing-masing 3.89 (SD = 0.78), 4.06 (SD = 0.73), dan 3.92 (SD = 0.81).

Distribusi preferensi partisipan menunjukkan mayoritas memilih ChatGPT-4 (61.4%) untuk tugas ini.

Tabel 3. Distribusi Frekuensi Preferensi Partisipan Berdasarkan Kategori Tugas

Kategori Tugas	Preferensi ChatGPT-4	Preferensi Gemini 2.5 Flash	Tidak Ada Preferensi
Masalah Teknis	67 (52.8%)	38 (29.9%)	22 (17.3%)
Konsep Ekonomi	45 (35.4%)	58 (45.7%)	24 (18.9%)
Kerangka Esai	78 (61.4%)	29 (22.8%)	20 (15.8%)

3.3 Analisis Korelasi Antar Kriteria

Analisis korelasi Pearson menunjukkan hubungan yang kuat antar kriteria. Untuk ChatGPT-4, kejelasan dan kemudahan pemahaman ($r = 0.743$, $p < 0.001$), dan kelengkapan dengan kejelasan ($r = 0.658$, $p < 0.001$). Untuk Gemini, korelasi tertinggi antara kelengkapan dan kejelasan ($r = 0.721$, $p < 0.001$).

Tabel 4. Matriks Korelasi Kriteria Evaluasi

Model	Kejelasan-Kelengkapan	Kejelasan-Kemudahan	Kelengkapan-Kemudahan
ChatGPT-4	0.658**	0.743**	0.612**
Gemini 2.5 Flash	0.721**	0.689**	0.634**

3.4 Analisis Multivarriat dan Model Prediktif

Regresi linier berganda menunjukkan bahwa semua kriteria berkontribusi terhadap skor keseluruhan. Model regresi untuk ChatGPT-4:

Skor = 0.247 + 0.312(kejelasan) + 0.289(kelengkapan) + 0.356(kemudahan)

R^2 = 0.821 Model Gemini:

Skor = 0.198 + 0.338(kejelasan) + 0.341(kelengkapan) + 0.287(kemudahan) R^2 = 0.794

3.5 Analisis Kualitatif Gaya Konunikasi

Analisis kualitatif terhadap pola respons menunjukkan perbedaan signifikan dalam gaya komunikasi kedua model. ChatGPT-4 cenderung menggunakan pendekatan deduktif dengan struktur hierarkis yang jelas, dimulai dari penjelasan umum ke spesifik. Rata-rata panjang respons ChatGPT-4 adalah 287 kata dengan 4.2 paragraf per respons. Model ini konsisten menggunakan bullet points dan numbering untuk mengorganisir informasi.

Sebaliknya, Gemini 2.5 Flash mengadopsi pendekatan naratif yang lebih ekspansif dengan rata-rata panjang respons 342 kata dan 5.8 paragraf per respons. Model ini lebih sering menggunakan contoh konkret dan analogi untuk menjelaskan konsep kompleks. Analisis sentiment menunjukkan bahwa Gemini 2.5 Flash menggunakan bahasa yang lebih ekspresif dan engaging, sementara ChatGPT-4 lebih formal dan sistematis.

3.6 Validditas dan Reliabilitas Hasil

Uji reliabilitas menggunakan Cronbach's Alpha menunjukkan konsistensi internal yang baik untuk instrumen evaluasi (α = 0.892). Inter-rater reliability yang diukur menggunakan Intraclass Correlation Coefficient (ICC) menunjukkan nilai 0.834, mengindikasikan konsistensi penilaian yang tinggi antar evaluator.

Uji normalitas menggunakan Shapiro-Wilk test menunjukkan bahwa distribusi data untuk sebagian besar variabel mendekati normal ($p > 0.05$), memvalidasi penggunaan uji parametrik dalam analisis. Untuk data yang tidak terdistribusi normal, konfirmasi dilakukan menggunakan uji non-parametrik Wilcoxon Signed-Rank yang memberikan hasil konsisten.

3.7 Implikasi Praktis dan Kontekstual

Hasil penelitian mengungkap bahwa pemilihan model AI yang optimal sangat bergantung pada konteks penggunaan dan preferensi komunikasi pengguna. ChatGPT-4 lebih cocok untuk aplikasi yang memerlukan struktur informasi yang jelas dan komunikasi yang efisien, seperti dalam konteks akademik formal atau profesional yang membutuhkan precision. Model ini menunjukkan keunggulan dalam menyajikan informasi dengan cara yang mudah dipahami dan diikuti.

Di sisi lain, Gemini 2.5 Flash lebih unggul dalam konteks yang memerlukan eksplorasi mendalam dan penjelasan komprehensif. Model ini cocok untuk aplikasi edukasi yang membutuhkan engagement tinggi dan pemahaman konseptual yang kaya. Kemampuannya dalam menyajikan informasi dengan detail yang lengkap dan konteks yang kaya membuatnya ideal untuk pembelajaran eksploratif.

Analisis cost-benefit menunjukkan bahwa kedua model memiliki trade-off yang berbeda. ChatGPT-4 menawarkan efisiensi dalam hal waktu pemrosesan informasi dengan struktur yang clear- cut, sementara Gemini 2.5 Flash memberikan value dalam hal kedalaman pemahaman meskipun memerlukan waktu lebih lama untuk diproses.

3.8 Limitasi dan Pertimbangan Metodologis

Penelitian ini memiliki beberapa limitasi yang perlu dipertimbangkan dalam interpretasi hasil. Pertama, evaluasi berbasis subjektif manusia dapat dipengaruhi oleh bias individual dan preferensi personal terhadap gaya komunikasi tertentu. Kedua, tiga kategori tugas yang digunakan mungkin belum merepresentasikan seluruh spektrum penggunaan AI generatif dalam konteks nyata.

Ketiga, penelitian ini tidak menganalisis performa kedua model dalam bahasa selain Bahasa Indonesia, sehingga generalisasi hasil terbatas pada konteks linguistik tertentu. Keempat, faktor temporal seperti update model atau perubahan algoritma tidak dikontrol dalam penelitian ini, yang dapat mempengaruhi konsistensi hasil dalam jangka panjang.

Validitas eksternal penelitian juga terbatas pada populasi yang diteliti, yaitu pengguna AI dengan latar belakang pendidikan tinggi dan pengalaman penggunaan AI minimal 6 bulan. Hasil mungkin berbeda untuk populasi dengan karakteristik yang berbeda, seperti pengguna baru atau dengan latar belakang pendidikan yang berbeda.

4. KESIMPULAN

Penelitian ini berhasil mengungkap karakteristik performa yang berbeda antara ChatGPT-4 dan Gemini 2.5 Flash melalui pendekatan task-oriented evaluation dengan melibatkan 127 partisipan yang menilai respons kedua model terhadap tiga kategori tugas praktis. Hasil analisis menunjukkan bahwa kedua model AI memiliki keunggulan spesifik yang bergantung pada konteks penggunaan dan kriteria evaluasi yang diterapkan.

ChatGPT-4 mendemonstrasikan superioritas dalam aspek kejelasan penyampaian dan kemudahan pemahaman, terutama dalam tugas penyelesaian masalah teknis dan penyusunan kerangka esai. Model ini konsisten menghasilkan respons yang terstruktur dengan rata-rata skor 4.23 untuk kejelasan dan 4.31 untuk kemudahan pemahaman dalam kategori masalah teknis. Keunggulan utama ChatGPT-4 terletak pada pendekatan komunikasi yang sistematis dan hierarkis, dengan penggun aan format yang terorganisir seperti bullet points dan numbering yang memudahkan pemahaman informasi. Analisis regresi menunjukkan bahwa kemudahan pemahaman menjadi prediktor terkuat bagi performa keseluruhan ChatGPT-4, dengan koefisien regresi sebesar 0.356.

Sebaliknya, Gemini 2.5 Flash menunjukkan keunggulan signifikan dalam aspek kelengkapan informasi dan pendekatan naratif yang mendalam. Model ini memperoleh skor tertinggi dalam kriteria kelengkapan dengan rata-rata 4.38 untuk tugas masalah teknis dan 4.25 untuk penjelasan konsep ekonomi. Karakteristik komunikasi Gemini 2.5 Flash yang naratif dan ekspansif, dengan rata-rata panjang respons 342 kata dibandingkan 287 kata ChatGPT-4, membuatnya lebih efektif dalam konteks yang memerlukan eksplorasi konseptual yang komprehensif. Model ini menunjukkan kemampuan superior dalam mengintegrasikan contoh konkret dan analogi yang memperkaya pemahaman kontekstual.

Analisis preferensi partisipan mengungkap pola yang konsisten dengan karakteristik performa masing-masing model. Mayoritas partisipan memilih ChatGPT-4 untuk tugas yang memerlukan struktur informasi yang jelas seperti penyusunan kerangka esai, sementara Gemini 2.5 Flash lebih disukai untuk tugas yang membutuhkan penjelasan mendalam seperti konsep ekonomi. Temuan ini menegaskan bahwa tidak ada model yang superior secara absolut, melainkan komplementer dalam konteks penggunaan yang berbeda.

Dari perspektif metodologis, penelitian ini berhasil mengembangkan framework evaluasi berbasis tugas yang dapat direplikasi dan diadaptasi untuk menilai model AI generatif lainnya. Validitas dan reliabilitas instrumen evaluasi terkonfirmasi melalui Cronbach's Alpha sebesar 0.892 dan ICC sebesar 0.834, menunjukkan konsistensi dan akurasi pengukuran yang dapat diandalkan.

Namun, penelitian ini memiliki keterbatasan signifikan yang perlu diakui. Evaluasi berbasis subjektif manusia rentan terhadap bias individual dan preferensi personal, yang dapat mempengaruhi objektivitas penilaian. Lingkup tugas yang terbatas pada tiga kategori mungkin belum mencerminkan kompleksitas penggunaan AI generatif dalam praktik sehari-hari. Selain itu, fokus pada Bahasa Indonesia membatasi generalisasi hasil untuk konteks linguistik dan kultural yang berbeda.

KETERBATASAN DAN STUDI LANJUTAN

Berdasarkan temuan dan keterbatasan penelitian ini, beberapa saran untuk penelitian lanjutan perlu dipertimbangkan untuk memperkuat validitas dan memperluas cakupan evaluasi model AI generatif. Penelitian mendatang sebaiknya mengincorporasikan metode evaluasi yang lebih objektif dengan mengkombinasikan penilaian manusia dan metrik otomatis berbasis algoritma untuk mengurangi bias subjektif. Implementasi automated evaluation metrics seperti BLEU, ROUGE, atau BERTScore dapat memberikan perspektif tambahan yang melengkapi penilaian manusia.

Ekspansi kategori tugas evaluasi sangat direkomendasikan untuk mencakup domain yang lebih beragam seperti creative writing, technical documentation, problem-solving matematik, dan multilingual tasks. Penelitian cross-cultural dan multilingual akan memberikan insight yang valuable tentang performa model AI dalam konteks global yang lebih representatif. Studi longitudinal juga diperlukan untuk memahami konsistensi performa model AI seiring dengan update dan evolusi algoritma yang berkelanjutan.

Pengembangan framework evaluasi yang lebih sophisticated dengan incorporasi contextual factors seperti user expertise level, task complexity, dan domain specificity akan meningkatkan prakticalitas hasil penelitian. Implementasi adaptive evaluation system yang dapat menyesuaikan kriteria penilaian berdasarkan karakteristik pengguna dan konteks penggunaan akan memberikan rekomendasi yang lebih personalized dan akurat.

Penelitian mendatang juga perlu mengeksplorasi aspek ethical considerations dalam penggunaan AI generatif, termasuk bias detection, fairness evaluation, dan impact assessment terhadap different demographic groups. Investigasi terhadap computational efficiency dan environmental impact dari berbagai model AI juga menjadi area yang increasingly important dalam era sustainability consciousness.

Kolaborasi interdisciplinary dengan domain experts dari berbagai bidang seperti education, healthcare, dan business akan memperkaya perspektif evaluasi dan meningkatkan practical applicability hasil penelitian. Development of standardized benchmarks untuk AI generative models evaluation juga menjadi contribution yang significant untuk research community secara keseluruhan.

Akhirnya, penelitian longitudinal tentang user adaptation dan learning curve dalam menggunakan different AI models akan provide valuable insights untuk improving user experience dan optimizing human-AI interaction patterns. Integration of user feedback loop mechanisms dalam evaluation framework akan memungkinkan continuous improvement dan refinement metodologi evaluasi yang lebih responsive terhadap kebutuhan pengguna aktual.

REFERENCES

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

Chang, Y. W. (2024). A Survey On Evaluation Of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers For Language Understanding.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681-694.

Johnson, D. C. (2024). Assessing the Performance of Chat GPT-3.5, ChatGPT-4, Google Gemini, and Google Gemini Advanced in Generating Coronary CT Angiography CAD-RADS Reporting. *Journal of Imaging Informatics in Medicine*, 37(2), 789–801.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic Evaluation Of Language Models.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

OpenAI. (2023). GPT-4 Technical Report. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver?

Rao, A. K. (2023). Evaluating ChatGPT as An Adjunct For Radiologic Decision-Making. *Journal of Imaging Informatics in Medicine*, 36(4), 1499–1512.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What We Know About How Bert Works. *Transactions of the Association for Computational Linguistics*, 8, 842-866.

Tamkin, A., Askell, A., Lovitt, L., Bai, Y., Chen, A., Hatfield-Dodds, T., & Kaplan, J. (2021). Understanding the Capabilities, Limitations, And Societal Impact Of Large Language Models.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Soricut, R. (2023). Gemini: A Family Of Highly Capable Multimodal Models.

Thompson, R. K. (2024). Large Language Models In Orthopedic Diagnosis: A Comparative Analysis of GPT- 4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. *Journal of Medical AI Research*, 12(2), 134–147.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark And Analysis Platform For Natural Language Understandi