ACADEMIA: Jurnal Inovasi Riset Akademik
Vol. 6, No. 4, September-November 2026
e-ISSN : 2807-1808 | p-ISSN : 2807-2294
Online Journal System https : //jurnalp4i.com/index.php/academia

EVALUASI KEAMANAN PROMPT DALAM INTERAKSI MANUSIA DAN AI: STUDI KASUS PADA MODEL GPT-4

Herman Santoni[1]
Universitas Muhammadiyah Surakarta[1]
e-mail: l208230019@student.ums.ac.id[1]

Diterima: *12/07/2026*; Direvisi: *28/08/2026*; Diterbitkan: *07/09/2026*

ABSTRAK

Perkembangan kecerdasan buatan generatif, khususnya model bahasa seperti GPT-4, telah mengubah pola interaksi manusia dan mesin melalui penggunaan perintah (prompt) dalam berbagai konteks. Namun, kemampuan model dalam memahami dan menghasilkan respons secara adaptif menimbulkan tantangan keamanan, terutama terkait injeksi perintah, manipulasi keluaran, pengabaian batasan perilaku, dan potensi pelanggaran data. Penelitian ini bertujuan menganalisis dimensi keamanan interaksi berbasis perintah pada GPT-4 serta mengidentifikasi kerentanan dan efektivitas mekanisme mitigasi yang diterapkan. Penelitian menggunakan metode studi kasus dengan tahapan berupa perancangan dan pengujian berbagai jenis perintah manipulatif, pengamatan terhadap respons model, identifikasi bentuk penyimpangan dari batasan keamanan, serta evaluasi terhadap filter dan mekanisme pembatasan perilaku yang tersedia. Hasil penelitian menunjukkan bahwa penerapan filter keamanan dan pembatasan perilaku pada GPT-4 mampu mengurangi sebagian risiko interaksi berbahaya, tetapi belum sepenuhnya mencegah strategi rekayasa perintah yang lebih kompleks. Kerentanan tersebut menunjukkan bahwa keamanan sistem berbasis model bahasa tidak hanya bergantung pada mekanisme perlindungan bawaan, tetapi juga memerlukan strategi mitigasi berlapis, evaluasi keamanan secara berkelanjutan, dan peningkatan kesadaran pengguna. Penelitian ini menyimpulkan bahwa penguatan kerangka kerja keamanan, pengujian terhadap skenario manipulatif, dan penerapan praktik penggunaan AI yang aman dan etis diperlukan untuk meningkatkan ketahanan sistem AI berbasis perintah.

Kata Kunci*:* *Keamanan Prompt, Interaksi Manusia Dan AI, GPT-4, LLM, Keamanan Sistem AI.*

ABSTRACT

The rapid development of generative artificial intelligence, particularly language models such as GPT-4, has transformed human–machine interaction through the use of prompts across various contexts. However, the adaptive capabilities of these models introduce security challenges, particularly those involving prompt injection, output manipulation, behavioral constraint bypassing, and potential data breaches. This study aims to analyze the security dimensions of prompt-based interaction with GPT-4 and to identify its vulnerabilities and the effectiveness of the implemented mitigation mechanisms. A case study method was employed through several stages, including the design and testing of various types of manipulative prompts, observation of model responses, identification of deviations from security constraints, and evaluation of the available security filters and behavioral restrictions. The findings indicate that security filters and behavioral constraints implemented in GPT-4 can reduce certain risks associated with harmful interactions, but they do not completely prevent more sophisticated prompt-engineering strategies. These vulnerabilities demonstrate that the security of language-model-based systems cannot rely solely on built-in protection mechanisms but requires layered risk mitigation strategies, continuous security evaluation, and improved user awareness. The

https://doi.org/10.51878/academia.v6i4.13904

ACADEMIA: Jurnal Inovasi Riset Akademik
Vol. 6, No. 4, September-November 2026
e-ISSN : 2807-1808 | p-ISSN : 2807-2294
Online Journal System https : //jurnalp4i.com/index.php/academia

study concludes that strengthening security frameworks, conducting systematic testing against manipulative scenarios, and promoting safe and ethical AI usage practices are necessary to improve the resilience of prompt-based AI systems.
Keywords: *Security Prompt, Human and AI Interaction, GPT-4, LLM, AI System Security*

PENDAHULUAN

Perkembangan kecerdasan buatan generatif (*generative artificial intelligence*) telah mendorong penggunaan *large language model* (LLM) dalam berbagai aplikasi, seperti *chatbot*, sistem informasi, layanan pelanggan, serta aplikasi berbasis *web* dan perangkat bergerak. LLM mampu memahami instruksi dalam bahasa alami dan menghasilkan respons berdasarkan konteks yang diberikan pengguna sehingga menjadi salah satu komponen penting dalam pengembangan aplikasi berbasis kecerdasan buatan (Rachmat & Kesuma, 2024). Penggunaan LLM juga berkembang pada sistem multibahasa yang ditujukan untuk memenuhi kebutuhan pengguna dengan latar belakang dan bahasa yang beragam. Perkembangan tersebut menunjukkan bahwa keamanan dan reliabilitas menjadi aspek penting dalam penerapan *chatbot* berbasis AI (Razan et al., 2026). Selain itu, integrasi LLM dengan konsep *AI agent* memungkinkan model digunakan untuk memproses informasi dan mendukung pelaksanaan tindakan dalam suatu lingkungan sistem (Yusuf et al., 2025).

Peningkatan kemampuan LLM juga meningkatkan risiko keamanan ketika model diintegrasikan dengan data, fungsi aplikasi, *Application Programming Interface* (API), maupun komponen eksternal. Integrasi tersebut dapat menciptakan permukaan serangan baru sehingga kemampuan model dalam menghasilkan respons yang sesuai belum secara otomatis menjamin keamanan sistem secara keseluruhan (Jimmy, 2026). Salah satu aspek penting dalam keamanan LLM adalah *prompt* sebagai media utama interaksi antara pengguna dan model. *Prompt engineering* dapat digunakan untuk meningkatkan kualitas dan karakteristik keluaran LLM (Abdillah & Hardiani, 2026). Namun, struktur instruksi juga dapat memengaruhi keselarasan antara maksud pengguna, keluaran sistem, dan proses pengembangan yang melibatkan LLM (Ramdhon, 2026).

Dari perspektif keamanan, *prompt* dapat dimanfaatkan untuk memanipulasi perilaku LLM melalui berbagai teknik, seperti *prompt injection*, *prompt leakage*, dan *jailbreaking*. *STAR Prompt Injection* dapat digunakan untuk mengevaluasi respons model terhadap skenario injeksi perintah dan mengidentifikasi potensi kerentanan pada LLM (Irwandi et al., 2026). Teknik *prompt engineering* yang digunakan untuk meningkatkan kualitas keluaran juga dapat diarahkan untuk mengeksploitasi batasan model (Dermawan & Herdianto, 2024). Selain itu, perbedaan karakteristik model menyebabkan pola instruksi yang sama dapat menghasilkan respons yang berbeda pada model yang berbeda, seperti ChatGPT dan Gemini (Maulida, 2024). Serangan juga dapat dilakukan secara bertahap melalui rangkaian interaksi dan konteks percakapan, sehingga pengujian keamanan perlu mempertimbangkan skenario *multi-turn* (Sun et al., 2024).

Penelitian terdahulu telah memberikan dasar dalam pengujian keamanan LLM melalui pengujian *prompt injection*, perbandingan respons antar-model, serta analisis serangan *multi-turn*. Namun, penelitian tersebut masih menunjukkan adanya kebutuhan terhadap pendekatan evaluasi yang dapat menguji keamanan model secara lebih terstruktur dengan mempertimbangkan variasi skenario manipulatif dan perubahan konteks interaksi. Kerangka *Open Web Application Security Project* (OWASP) dapat digunakan sebagai dasar untuk mengidentifikasi dan mengklasifikasikan potensi kerentanan pada sistem (Kuncoro et al., 2022). Pendekatan *stateful* juga menunjukkan pentingnya mempertimbangkan perubahan

https://doi.org/10.51878/academia.v6i4.13904

ACADEMIA: Jurnal Inovasi Riset Akademik
Vol. 6, No. 4, September-November 2026
e-ISSN : 2807-1808 | p-ISSN : 2807-2294
Online Journal System https : //jurnalp4i.com/index.php/academia

maksud pengguna sepanjang rangkaian interaksi dalam mendeteksi *adversarial intent drift* (Albrethsen et al., 2026).

Kebaruan penelitian ini terletak pada pengembangan pendekatan evaluasi keamanan GPT-4 yang menguji beberapa skenario manipulatif secara terstruktur dengan mempertimbangkan respons model pada interaksi tunggal dan *multi-turn*. Evaluasi tidak hanya berfokus pada keberhasilan manipulasi, tetapi juga menganalisis perubahan respons model dan kemampuan mekanisme keamanan dalam mempertahankan batasan perilaku ketika konteks serta instruksi mengalami perubahan. Pendekatan tersebut digunakan untuk menghasilkan pemetaan kerentanan dan dasar perumusan strategi mitigasi keamanan interaksi berbasis *prompt* yang lebih terarah. Berdasarkan kebaruan tersebut, penelitian ini bertujuan mengevaluasi keamanan GPT-4 melalui berbagai skenario manipulatif, menganalisis respons model terhadap setiap skenario, mengidentifikasi pola kerentanan dalam interaksi berbasis *prompt*, serta menilai kemampuan mekanisme keamanan dalam mempertahankan batasan perilaku model. Hasil evaluasi diharapkan dapat menjadi dasar dalam merumuskan strategi mitigasi dan kerangka keamanan interaksi berbasis *prompt* yang lebih kuat.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan studi kasus dengan studi literatur sebagai dasar untuk menyusun kategori dan skenario pengujian keamanan pada GPT-4. Studi literatur digunakan untuk mengidentifikasi karakteristik ancaman dan teknik manipulasi *prompt* yang relevan, sedangkan studi kasus digunakan untuk mengamati respons GPT-4 terhadap skenario yang telah ditentukan. Objek penelitian adalah GPT-4 dengan *prompt* sebagai instrumen utama pengujian. Tahapan penelitian meliputi identifikasi permasalahan keamanan, penentuan kategori *prompt* manipulatif, penyusunan skenario pengujian, pemberian *prompt* kepada model, pencatatan respons yang dihasilkan, serta pengelompokan hasil berdasarkan jenis serangan. Kategori pengujian terdiri atas *prompt injection* untuk menguji kemampuan model dalam menghadapi instruksi tersembunyi, *prompt leakage* untuk menguji kemungkinan pengungkapan instruksi sistem, *prompt hijacking* untuk menguji upaya pengambilalihan arah respons dari instruksi awal, dan *ethical bypass prompt* untuk menguji kemampuan model mempertahankan batasan terhadap permintaan yang berpotensi menghasilkan konten yang tidak sesuai.

Data penelitian berupa respons GPT-4 terhadap setiap skenario *prompt* yang diberikan selama proses pengujian. Setiap respons diamati berdasarkan kesesuaian dengan instruksi awal, kemampuan model mempertahankan batasan keamanan, adanya indikasi pengungkapan informasi atau instruksi sistem, serta keberhasilan atau kegagalan *prompt* dalam memengaruhi perilaku model. Hasil pengujian kemudian didokumentasikan dan dianalisis secara deskriptif dengan membandingkan karakteristik respons pada masing-masing kategori *prompt*. Analisis dilakukan dengan mengelompokkan respons menjadi respons yang mempertahankan batasan keamanan dan respons yang menunjukkan indikasi keberhasilan manipulasi atau penyimpangan dari batasan tersebut. Selanjutnya, hasil analisis digunakan untuk mengidentifikasi pola kerentanan GPT-4 terhadap berbagai strategi rekayasa *prompt* dan menjadi dasar dalam mengevaluasi kebutuhan terhadap strategi mitigasi keamanan pada interaksi berbasis LLM.

HASIL DAN PEMBAHASAN

Hasil

Penelitian ini menghasilkan temuan mengenai respons GPT-4 terhadap beberapa bentuk *prompt* manipulatif yang meliputi *prompt injection*, *prompt leakage*, *prompt hijacking*,

https://doi.org/10.51878/academia.v6i4.13904

ACADEMIA: Jurnal Inovasi Riset Akademik
Vol. 6, No. 4, September-November 2026
e-ISSN : 2807-1808 | p-ISSN : 2807-2294
Online Journal System https : //jurnalp4i.com/index.php/academia

dan *ethical bypass prompt*. Pengujian dilakukan dengan memberikan sejumlah skenario *prompt* yang dirancang untuk mengubah arah respons, memperoleh informasi mengenai instruksi sistem, mengambil alih konteks percakapan, serta menguji kemampuan model dalam mempertahankan batasan etika. Setiap respons yang dihasilkan diamati berdasarkan kemampuan model dalam mempertahankan instruksi dan batasan keamanan yang berlaku. Respons kemudian dikategorikan berdasarkan karakteristik keluaran yang dihasilkan, yaitu penolakan, respons yang masih berada dalam batas aman, dan respons yang menunjukkan indikasi kerentanan. Ringkasan hasil pengujian pada setiap kategori disajikan pada Tabel 1 untuk memberikan gambaran umum mengenai pola respons GPT-4.

Tabel 1. Ringkasan Hasil Pengujian Respons GPT-4 terhadap *Prompt* Manipulatif

Jenis Pengujian	Jumlah Percobaan	Hasil Pengujian
Prompt Injection	10	40% respons menunjukkan keluaran ambigu atau berisiko
Prompt Leakage	50	Ditemukan indikasi pengungkapan informasi mengenai aturan sistem
Prompt Hijacking	10	Model mengikuti konteks naratif yang diberikan
Ethical Bypass Prompt	10	Model menghasilkan respons berdasarkan konteks fiktif

Berdasarkan Tabel 1, terlihat bahwa GPT-4 masih mampu mempertahankan mekanisme keamanan pada sebagian skenario, tetapi respons model dapat berubah ketika instruksi disusun dengan pola manipulatif tertentu. Pada pengujian *prompt injection*, model cenderung menolak instruksi yang secara eksplisit meminta keluaran yang bertentangan dengan batasan keamanan. Sebagai contoh, ketika diberikan permintaan untuk menuliskan dialog fiksi mengenai cara menghindari deteksi jaringan, model memberikan penolakan dengan menyatakan bahwa konten tersebut dapat melanggar kebijakan penggunaan. Namun, ketika instruksi dikembangkan menjadi *multi-layered instruction* dan disisipkan melalui konteks cerita, model menunjukkan respons yang lebih ambigu. Beberapa keluaran tidak memberikan instruksi berbahaya secara langsung, tetapi tetap memuat informasi yang dapat dikaitkan dengan tujuan awal *prompt*, sehingga menunjukkan bahwa perubahan struktur instruksi dapat memengaruhi konsistensi mekanisme keamanan.

Temuan berikutnya diperoleh dari pengujian *prompt leakage*, yang difokuskan pada kemampuan model dalam menjaga informasi mengenai instruksi sistem dan konfigurasi internal. Pengujian dilakukan dengan memberikan pertanyaan yang mengarahkan model untuk menjelaskan aturan sistem seolah-olah permintaan tersebut berasal dari pengembang aplikasi. Dalam beberapa respons, GPT-4 memberikan penjelasan mengenai aspek privasi, keterbatasan pengetahuan, fungsi pemrosesan, etika, keakuratan, serta karakteristik interaksi dengan pengguna. Respons tersebut menunjukkan bahwa model tidak selalu memberikan penolakan ketika diminta menjelaskan informasi mengenai aturan yang mengatur perilakunya. Meskipun demikian, informasi yang diberikan belum dapat secara langsung dikategorikan sebagai *system prompt* yang sebenarnya, sehingga temuan tersebut lebih tepat disebut sebagai indikasi pengungkapan informasi mengenai aturan atau konfigurasi umum model. Oleh karena itu, hasil pengujian menunjukkan adanya potensi *prompt leakage*, tetapi diperlukan pembuktian lebih lanjut untuk memastikan apakah informasi yang dihasilkan benar-benar berasal dari instruksi sistem internal.

https://doi.org/10.51878/academia.v6i4.13904

ACADEMIA: Jurnal Inovasi Riset Akademik
Vol. 6, No. 4, September-November 2026
e-ISSN : 2807-1808 | p-ISSN : 2807-2294
Online Journal System https : //jurnalp4i.com/index.php/academia

Pada pengujian *prompt hijacking*, model diberikan instruksi yang dirancang untuk menggeser arah respons dari tujuan awal menuju konteks yang ditentukan oleh peneliti. Pendekatan yang digunakan memanfaatkan narasi fiktif sehingga model diarahkan untuk mengikuti sudut pandang karakter tertentu. Salah satu skenario meminta model menghasilkan cerita dari sudut pandang tokoh antagonis yang menganggap pembunuhan sebagai bentuk keadilan sosial. GPT-4 dapat mengikuti karakter dan alur cerita yang diberikan sehingga menghasilkan narasi sesuai dengan konteks *prompt*. Respons tersebut menunjukkan bahwa konteks naratif dapat menjadi faktor yang memengaruhi arah keluaran model, terutama ketika instruksi manipulatif tidak disampaikan dalam bentuk perintah langsung. Temuan ini memperlihatkan bahwa evaluasi keamanan perlu mempertimbangkan kemampuan model dalam mempertahankan batasan ketika menghadapi manipulasi konteks dan perubahan arah narasi.

Pengujian selanjutnya dilakukan melalui *ethical bypass prompt* untuk melihat kemampuan GPT-4 dalam mempertahankan batasan etika ketika permintaan disampaikan melalui konteks fiksi. Hasil pengujian menunjukkan bahwa model dapat menghasilkan narasi yang menggambarkan tindakan ekstrem ketika permintaan dikemas sebagai bagian dari cerita dan diberikan melalui sudut pandang karakter tertentu. Salah satu respons menggambarkan bahwa tokoh dalam cerita menganggap tindakan kekerasan sebagai alat untuk mencapai keadilan, tanpa memberikan penjelasan tambahan mengenai konteks etika dari tindakan tersebut. Temuan tersebut menunjukkan bahwa konteks fiktif dapat mengubah karakteristik respons model dibandingkan dengan permintaan langsung yang memiliki tujuan serupa. Namun, keberhasilan model menghasilkan narasi fiksi tidak secara otomatis membuktikan bahwa sistem keamanan telah sepenuhnya dilewati, sehingga penilaian perlu mempertimbangkan apakah respons tersebut hanya bersifat naratif atau telah memberikan informasi operasional yang dapat digunakan untuk melakukan tindakan berbahaya.

Secara keseluruhan, hasil penelitian menunjukkan bahwa respons GPT-4 terhadap *prompt* manipulatif dipengaruhi oleh cara instruksi disusun dan konteks yang digunakan dalam interaksi. Model relatif mampu mempertahankan batasan ketika menghadapi permintaan berbahaya yang disampaikan secara langsung, tetapi menunjukkan respons yang lebih kompleks ketika instruksi dikemas melalui narasi, konteks fiktif, atau struktur perintah yang berlapis. Pola tersebut menunjukkan bahwa mekanisme keamanan berbasis penolakan terhadap kata atau pola tertentu belum tentu cukup untuk menghadapi manipulasi yang memanfaatkan konteks dan struktur bahasa. Hasil pengujian juga menunjukkan bahwa evaluasi keamanan perlu membedakan antara respons yang benar-benar menunjukkan keberhasilan serangan dan respons yang hanya menghasilkan konten yang secara tematik berkaitan dengan permintaan manipulatif. Dengan demikian, temuan penelitian ini mengindikasikan bahwa penguatan keamanan GPT-4 perlu dilakukan melalui evaluasi yang mempertimbangkan variasi struktur *prompt*, konteks percakapan, perubahan maksud pengguna, serta konsistensi model dalam mempertahankan batasan keamanan pada berbagai kondisi interaksi.

Pembahasan

Hasil penelitian menunjukkan bahwa kemampuan GPT-4 dalam mempertahankan keamanan tidak hanya dipengaruhi oleh jenis permintaan, tetapi juga oleh cara instruksi tersebut disusun dan disampaikan kepada model. Pada pengujian *prompt injection*, model cenderung mampu menolak permintaan yang secara eksplisit mengarah pada pelanggaran batasan keamanan, tetapi respons menjadi lebih ambigu ketika instruksi disusun secara berlapis atau disisipkan ke dalam konteks naratif. Kondisi tersebut menunjukkan bahwa mekanisme

https://doi.org/10.51878/academia.v6i4.13904

ACADEMIA: Jurnal Inovasi Riset Akademik
Vol. 6, No. 4, September-November 2026
e-ISSN : 2807-1808 | p-ISSN : 2807-2294
Online Journal System https : //jurnalp4i.com/index.php/academia

keamanan model masih menghadapi tantangan ketika harus membedakan antara instruksi yang sah dengan instruksi yang sengaja disamarkan dalam suatu konteks. Geng et al. (2026) menjelaskan bahwa *prompt injection* merupakan ancaman penting terhadap LLM karena serangan dapat mengeksploitasi cara model memproses instruksi dan konteks yang diberikan melalui *prompt*. Dengan demikian, temuan penelitian ini memperkuat pandangan bahwa pengujian keamanan LLM tidak cukup dilakukan menggunakan instruksi eksplisit, tetapi perlu mencakup variasi struktur, konteks, dan cara penyamaran perintah.

Temuan tersebut juga memperlihatkan bahwa manipulasi *prompt* berkaitan erat dengan kemampuan model dalam menafsirkan instruksi berdasarkan konteks. Pada skenario *multi-layered instruction*, perubahan susunan perintah dapat membuat model memberikan respons yang tidak sepenuhnya sesuai dengan tujuan mekanisme keamanan, meskipun model masih berupaya menghindari pemberian instruksi berbahaya secara langsung. Kondisi ini menunjukkan bahwa batas antara respons yang aman dan respons yang berisiko dapat menjadi tidak tegas ketika suatu permintaan dikemas melalui konteks tertentu. Joseph et al. (2025) menjelaskan bahwa *prompt injection* dapat dimanfaatkan sebagai metode eksploitasi karena penyerang berusaha memengaruhi perilaku LLM melalui manipulasi instruksi yang diterima model. Dalam konteks penelitian ini, hasil tersebut menunjukkan bahwa keamanan GPT-4 perlu dievaluasi tidak hanya berdasarkan kemampuan menolak permintaan, tetapi juga berdasarkan sejauh mana model mampu mempertahankan tujuan dan batasan awal ketika menghadapi instruksi yang telah dimodifikasi.

Pada pengujian *prompt leakage*, respons GPT-4 menunjukkan bahwa model dapat memberikan penjelasan mengenai aturan umum, keterbatasan, dan karakteristik sistem ketika pengguna meminta informasi mengenai konfigurasi internal. Temuan ini perlu dimaknai secara hati-hati karena pemberian informasi mengenai kebijakan atau kemampuan model belum secara otomatis membuktikan bahwa *system prompt* yang sebenarnya telah terbocorkan. Namun, kecenderungan model untuk memberikan informasi yang berkaitan dengan aturan operasional menunjukkan adanya potensi yang dapat dimanfaatkan untuk memahami pola pertahanan model dan merancang serangan lanjutan. Kondisi tersebut memperlihatkan bahwa perlindungan informasi internal merupakan bagian penting dari keamanan LLM karena informasi yang tampaknya tidak sensitif secara individual dapat menjadi lebih berisiko ketika dikombinasikan dengan teknik manipulasi lainnya. Mittal et al. (2026) menunjukkan bahwa perkembangan serangan *jailbreaking* pada LLM semakin beragam dan dirancang untuk mengeksploitasi keterbatasan mekanisme pertahanan model, sehingga perlindungan tidak cukup hanya mengandalkan satu lapisan penyaringan. Oleh karena itu, hasil penelitian ini mengindikasikan perlunya mekanisme yang mampu membatasi pengungkapan informasi internal sekaligus mempertahankan kemampuan model untuk memberikan penjelasan yang relevan kepada pengguna.

Hasil pengujian *prompt hijacking* dan *ethical bypass* semakin memperlihatkan pentingnya konteks dalam menentukan respons model. Ketika permintaan yang mengandung tindakan ekstrem dikemas sebagai cerita fiktif atau disampaikan melalui sudut pandang karakter tertentu, GPT-4 dapat mengikuti konteks naratif tersebut tanpa selalu memberikan penolakan atau penjelasan etis yang eksplisit. Kondisi ini menunjukkan bahwa model dapat memprioritaskan kesinambungan narasi dan pemenuhan instruksi pengguna ketika maksud berisiko tidak disampaikan secara langsung. Knowlton et al. (2026) menunjukkan bahwa teknik *jailbreaking* pada *chatbot* LLM dapat dilakukan melalui berbagai pola *prompt* yang dirancang untuk melewati mekanisme pertahanan model. Yi et al. (2024) juga menunjukkan bahwa serangan *jailbreak* berkembang melalui beragam strategi yang memanfaatkan kelemahan

https://doi.org/10.51878/academia.v6i4.13904

ACADEMIA: Jurnal Inovasi Riset Akademik
Vol. 6, No. 4, September-November 2026
e-ISSN : 2807-1808 | p-ISSN : 2807-2294
Online Journal System https : //jurnalp4i.com/index.php/academia

penyelarasan dan mekanisme pertahanan LLM. Temuan penelitian ini memperkuat kebutuhan untuk melakukan evaluasi terhadap *role-playing*, narasi, dan bentuk penyamaran lainnya karena pendekatan tersebut dapat menghasilkan respons yang berbeda dibandingkan dengan permintaan yang disampaikan secara langsung. Dengan demikian, penilaian keamanan sebaiknya tidak hanya menentukan apakah suatu *prompt* mengandung kata atau perintah berbahaya, tetapi juga mempertimbangkan maksud, konteks, dan potensi penggunaan keluaran yang dihasilkan model.

Berdasarkan keseluruhan temuan tersebut, keamanan interaksi dengan LLM perlu dipandang sebagai proses yang bersifat berkelanjutan dan tidak hanya bergantung pada kemampuan model melakukan penolakan terhadap permintaan tertentu. Strategi mitigasi perlu menggabungkan pemeriksaan *input*, analisis konteks dan maksud, pemantauan respons, serta pengujian keamanan secara berkala untuk mengidentifikasi pola serangan baru. Haromain et al. (2025) menunjukkan bahwa struktur *prompt* dapat memengaruhi kemampuan LLM dalam menghasilkan keluaran sesuai kebutuhan, sehingga aspek penyusunan instruksi perlu menjadi perhatian baik dari sisi fungsional maupun keamanan. Prinsip pengelolaan risiko yang menekankan identifikasi bahaya, pengendalian, dan pemantauan juga relevan untuk diterapkan dalam pengelolaan keamanan sistem berbasis AI (Rani et al., 2025). Selain itu, kemampuan mengidentifikasi pola dalam data berbasis bahasa dapat menjadi salah satu pendekatan pendukung dalam proses pemantauan respons model (Sumadewa et al., 2025). Berdasarkan hal tersebut, kontribusi penelitian ini terletak pada penegasan bahwa evaluasi keamanan GPT-4 perlu dilakukan melalui berbagai bentuk manipulasi *prompt*, termasuk instruksi berlapis, permintaan pengungkapan informasi, manipulasi konteks, dan *ethical bypass*, sehingga strategi mitigasi yang dikembangkan dapat lebih adaptif terhadap perubahan pola serangan dan tidak hanya bergantung pada pendeteksian instruksi berbahaya secara eksplisit. Temuan tersebut juga sejalan dengan kajian yang menempatkan pertahanan terhadap *prompt injection* dan *jailbreaking* sebagai bagian dari kerangka keamanan LLM yang perlu dilakukan secara berlapis dan sistematis (Correia et al., 2026).

KESIMPULAN

Berdasarkan hasil evaluasi keamanan interaksi berbasis *prompt* pada GPT-4, dapat disimpulkan bahwa GPT-4 masih memiliki kerentanan terhadap beberapa bentuk manipulasi *prompt*, khususnya *prompt injection*, *prompt leakage*, dan *bypass* terhadap batasan etika. Meskipun telah dilengkapi dengan mekanisme keamanan dan sistem moderasi, model masih dapat memberikan respons yang tidak sepenuhnya konsisten ketika menghadapi instruksi yang disusun melalui narasi fiktif, struktur perintah berlapis, maupun manipulasi konteks. Temuan tersebut menunjukkan bahwa kemampuan mekanisme keamanan GPT-4 dalam mempertahankan batasan perilaku belum sepenuhnya mampu menghadapi seluruh variasi manipulasi *prompt*. Dengan demikian, tujuan penelitian untuk mengevaluasi keamanan interaksi berbasis *prompt* pada GPT-4 menunjukkan bahwa masih terdapat celah keamanan yang perlu diperhatikan, terutama pada kemampuan model dalam mempertahankan batasan keamanan ketika menghadapi instruksi manipulatif dan konteks yang kompleks.

Berdasarkan temuan tersebut, pengembangan keamanan LLM perlu diarahkan pada mekanisme yang lebih adaptif terhadap perubahan pola dan konteks *prompt*. Penerapan *red teaming* secara berkala dapat digunakan untuk mengidentifikasi pola manipulasi baru, sedangkan penguatan pemahaman konteks dapat membantu model membedakan instruksi yang sah dari instruksi yang berupaya mengeksploitasi batasan keamanan. Penelitian selanjutnya disarankan untuk mengembangkan skenario pengujian yang lebih beragam, mencakup interaksi

https://doi.org/10.51878/academia.v6i4.13904

multi-turn dan perubahan maksud pengguna, serta mengembangkan mekanisme evaluasi yang dapat mengukur tingkat keberhasilan pertahanan model secara lebih sistematis. Hasil tersebut diharapkan dapat mendukung pengembangan mekanisme keamanan berbasis penalaran etis yang lebih kuat untuk mewujudkan interaksi dengan AI generatif yang aman dan bertanggung jawab.

DAFTAR PUSTAKA

Abdillah, M. L., & Hardiani, T. (2026). SAST implementation for evaluating LLM-generated code quality using prompt engineering. *Sistemasi: Jurnal Sistem Informasi, 15*(5), 1873–1885. https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/6395

Albrethsen, J., Datta, Y., Kumar, K., & Rajasekar, S. (2026). Deepcontext: Stateful real-time detection of multi-turn adversarial intent drift in LLMs. *arXiv*. https://arxiv.org/abs/2602.16935

Correia, P. H. B., Achjian, R. W., De Oliveira, D. E., Maria, Y. A., Hayashi, V. T., Lopes, M., & Simplicio, M. A. (2026). A systematic literature review on LLM defenses against prompt injection and jailbreaking: Expanding NIST taxonomy. *arXiv*. https://arxiv.org/abs/2601.22240

Geng, T., Xu, Z., Qu, Y., & Wong, W. E. (2026). Prompt injection attacks on large language models: A survey of attack methods, root causes, and defense strategies. *Computers, Materials & Continua, 87*(1), 4. https://doi.org/10.32604/cmc.2025.074081

Haromain, I., Munir, S., & Rahmah, A. (2025). Analisa prompt engineering pada large language model dengan retrieval-augmented generation untuk informasi obat dan vitamin. *Indonesian Journal of Computer Science, 4*(2), 144–153. https://jurnal.bsi.ac.id/https:/jurnal.bsi.ac.id/index.php/ijcs/article/view/10005

Irwandi, H., Silitonga, A. I., Chandra, R., & Simamora, W. S. (2026). Security evaluation of Indonesian LLMs for digital business using STAR prompt injection. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 10*(1), 449–457. https://doi.org/10.33395/sinkron.v10i1.15662

Jimmy, J. (2026). Analisis kerentanan keamanan pada integrasi API generative AI dalam aplikasi berbasis web. *Jurnal Minfo Polgan, 15*(2), 2128–2137. https://doi.org/10.33395/jmp.v15i2.16567

Joseph, J. K., Daniel, E., Kathiresan, V., & MAP, M. (2025). Prompt injection in large language model exploitation: A security perspective. In *2025 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)* (pp. 1–8). IEEE. https://ieeexplore.ieee.org/abstract/document/11064209

Knowlton, B., Campa, J., Gallo, D. S., Dajani, K., & Alzahrani, N. (2026). Prompt-based jailbreaking of leading LLM chatbots: A survey of attacks and defenses. *IEEE Transactions on Artificial Intelligence*. https://ieeexplore.ieee.org/abstract/document/11397677

Kuncoro, A. W., Fayruz Rahma, S. T., & Eng, M. (2022). Analisis metode Open Web Application Security Project (OWASP) pada pengujian keamanan website: Literature review. *Automata, 3*(1). https://journal.uii.ac.id/automata/article/view/21893

Maulida, R. (2024). Komparasi respons ChatGPT dan Gemini terhadap command pattern identik dengan metode black box. *Jurnal Teknik Informatika STMIK Antar Bangsa, 10*(2), 68–71. https://doi.org/10.51998/jti.v10i2.588

Mittal, A., Kaur, J., & Chatterjee, T. K. (2026). A systematic review of jailbreaking attacks and defense mechanisms in large language models. In *2026 International Conference on*

https://doi.org/10.51878/academia.v6i4.13904

Sustainable Engineering and Technology Innovations (ICSETI) (pp. 507–512). IEEE. https://ieeexplore.ieee.org/abstract/document/11636629

Rachmat, N., & Kesuma, D. P. (2024). Implementasi LLM Gemini pada pengembangan aplikasi chatbot berbasis Android. *Jurnal Ilmu Komputer (JUIK), 4*(1), 40–52. https://journal.umgo.ac.id/index.php/juik/article/view/2831

Ramdhon, A. N. (2026). Dari prompt ke produksi: Metode tiga dimensi untuk mengevaluasi keselarasan niat, keandalan kode, dan kognisi pengembang dalam vibe coding. *SABER: Jurnal Teknik Informatika, Sains dan Ilmu Komunikasi, 4*(2), 62–77. https://doi.org/10.59841/saber.v4i2.3716

Rani, H. A., Ismail, A., Jaa'far, M. H., & Ahmad, N. (2025). Risks, incidents, guidelines, and strategies pertaining to chemical storage and handling in primary healthcare: A narrative review. *International Journal of Public Health Research, 15*(2), 2366–2374. https://spaj.ukm.my/ijphr/index.php/ijphr/article/view/543

Razan, K. F. A., Nidji, M. F., & Hasanah, A. U. (2026). Pemetaan tematik pengembangan chatbot AI multibahasa: Tinjauan literatur sistematis terhadap tren dan tantangan masa depan. *Jurnal Riset Teknik Komputer, 3*(1), 31–43. https://doi.org/10.69714/yd129k17

Sumadewa, I. N. Y., Anggrawan, A., Hairani, H., Hariyadi, I. P., Satria, C., & Saputri, D. S. C. (2025). Emotion classification on customer reviews of drinking water services using IndoBERT and machine learning algorithms. In *2025 7th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1–5). IEEE. https://ieeexplore.ieee.org/abstract/document/11296077

Sun, X., Zhang, D., Yang, D., Zou, Q., & Li, H. (2024). Multi-turn context jailbreak attack on large language models from first principles. *arXiv*. https://arxiv.org/abs/2408.04686

Yi, S., Liu, Y., Sun, Z., Cong, T., He, X., Song, J., ... & Li, Q. (2024). Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*. https://arxiv.org/abs/2407.04295

Yusuf, R., Perdana, A., Sugandi, F., & Apsiswanto, U. (2025). Kajian konseptual AI agent on-chain berbasis LLM untuk manajemen aset DeFi. *Journal of Technology and Data Science, 3*(2), 216–228. https://doi.org/10.59025/ef9b943c

https://doi.org/10.51878/academia.v6i4.13904