



Analisis Kualitatif Kerentanan AI Agent terhadap Serangan *Prompt Injection* dan Strategi Mitigasinya

Ardian Fachreza

Teknik Informatika, Universitas Wahid Hasyim Semarang, Indonesia

*Penulis Korespondensi: ardian.fachreza@unwahas.ac.id

Abstract. The development of Large Language Model (LLM)-based AI agents in enterprise environments is driving increased digital automation while expanding modern cybersecurity risks. One of the most discussed threats in AI security developments is prompt injection, a natural language-based instruction manipulation technique that can influence AI agent behavior and trigger corporate data leaks. This study aims to analyze AI agents' vulnerability to prompt injection attacks, identify their impact on corporate data security, and examine security mitigation strategies based on global industry standards. The study used a descriptive qualitative method with a threat modeling approach and contemporary cyber case studies. The results show that AI agents are highly vulnerable to direct and indirect prompt injection, especially in systems connected to tool-calling, cloud storage, corporate email, and internal databases. Analysis based on the OWASP Top 10 for LLM Applications and MITRE ATLAS frameworks shows that prompt injection is closely related to sensitive information disclosure, excessive agency, and data extraction in enterprise operational environments. Mitigation strategies such as context isolation, input validation, role-based access control, adversarial testing, and strengthening AI governance are used to reduce the risk of exploitation of modern AI agents. The research implications point to the need to develop AI security standards that are more adaptive to natural language-based threats in enterprise digital ecosystems.

Keywords: AI Agent; Company Data; Cyber; Large Language Models; Prompt Injection.

Abstrak. Perkembangan AI Agent berbasis Large Language Model (LLM) dalam lingkungan perusahaan mendorong peningkatan otomatisasi digital sekaligus memperluas risiko keamanan siber modern. Salah satu ancaman yang banyak dibahas dalam perkembangan keamanan AI adalah *prompt injection*, yaitu teknik manipulasi instruksi berbasis bahasa alami yang dapat memengaruhi perilaku AI Agent dan memicu kebocoran data perusahaan. Penelitian ini bertujuan untuk menganalisis kerentanan AI Agent terhadap serangan *prompt injection*, mengidentifikasi dampaknya terhadap keamanan data perusahaan, serta mengkaji strategi mitigasi keamanan berbasis standar industri global. Penelitian menggunakan metode kualitatif deskriptif dengan pendekatan *threat modeling* dan studi kasus siber kontemporer. Hasil penelitian menunjukkan bahwa AI Agent memiliki kerentanan tinggi terhadap *direct* dan *indirect prompt injection*, terutama pada sistem yang terhubung dengan *tool-calling*, *cloud storage*, email perusahaan, dan database internal. Analisis berdasarkan kerangka OWASP Top 10 for LLM Applications dan MITRE ATLAS memperlihatkan bahwa *prompt injection* berkaitan erat dengan *sensitive information disclosure*, *excessive agency*, dan *data extraction* dalam lingkungan operasional perusahaan. Strategi mitigasi seperti *context isolation*, *input validation*, *role-based access control*, *adversarial testing*, serta penguatan *AI governance* digunakan untuk mengurangi risiko eksploitasi terhadap AI Agent modern. Implikasi penelitian menunjukkan perlunya pengembangan standar keamanan AI yang lebih adaptif terhadap ancaman berbasis bahasa alami dalam ekosistem digital perusahaan.

Kata Kunci: AI Agent; Data Perusahaan; Large Language Model; Prompt Injection; Siber.

1. LATAR BELAKANG

Perkembangan teknologi kecerdasan buatan berbasis Large Language Model (LLM) mendorong perubahan besar dalam pola interaksi manusia dengan sistem digital. Integrasi model generatif ke dalam AI Agent melahirkan sistem yang mampu menjalankan tugas secara otonom, seperti membaca surel, mengakses basis data, melakukan pencarian web, hingga menjalankan perintah berbasis *Application Programming Interface* (API) (Chaitanya & Rolla, 2024). Pada tahun 2024, berbagai perusahaan teknologi global mulai mengembangkan *agentic AI* yang terhubung langsung dengan layanan produktivitas digital dan sistem *cloud* perusahaan.

OpenAI memperkenalkan pengembangan *AI Agent* berbasis *tool usage*, sedangkan Microsoft mengintegrasikan Copilot ke dalam ekosistem kerja perusahaan melalui layanan Microsoft 365. Adopsi tersebut menjadi bukti bahwa *AI Agent* mulai ditempatkan sebagai bagian dari infrastruktur operasional digital dalam berbagai sektor industri (Chen et al., 2024).

Peningkatan penggunaan *AI Agent* di lingkungan bisnis memunculkan perhatian baru terhadap aspek keamanan siber. Sistem berbasis LLM memproses seluruh teks sebagai konteks percakapan, sehingga batas antara instruksi sistem, data pengguna, dan informasi eksternal menjadi semakin sulit dipisahkan. Karakteristik tersebut berbeda dari sistem perangkat lunak konvensional yang memisahkan input dan command secara lebih tegas (Rane, Tawde, Choudhary, & Rane, 2023). Pada tahun 2025, OWASP menempatkan *Prompt Injection* sebagai ancaman utama dalam daftar OWASP Top 10 for LLM Applications karena tingginya potensi manipulasi perilaku model generatif. Perhatian terhadap kerentanan tersebut meningkat seiring penggunaan *AI Agent* yang semakin luas pada sistem perusahaan dan layanan digital publik.

Prompt injection berkembang sebagai bentuk serangan yang memanfaatkan kelemahan pemrosesan bahasa alami pada model AI generatif. Penyerang dapat menyisipkan instruksi tersembunyi melalui dokumen, halaman web, email, maupun file yang dibaca *AI Agent* selama proses interaksi. Instruksi tersebut kemudian diproses sebagai bagian dari konteks internal model sehingga memengaruhi respons maupun tindakan sistem (Rababah et al., 2024). Dalam beberapa kasus, *prompt injection* digunakan untuk memanipulasi *agent* agar mengabaikan *system prompt*, membocorkan data sensitif, atau menjalankan perintah yang tidak sesuai dengan kebijakan keamanan perusahaan. Pola serangan seperti ini mulai banyak dibahas dalam laporan keamanan siber global sejak meningkatnya penggunaan *AI Agent* berbasis *tool-calling* pada tahun 2023 dan 2024 (Sarker, 2024).

Risiko keamanan menjadi semakin kompleks ketika *AI Agent* memiliki akses terhadap perangkat eksternal dan sumber data perusahaan. Integrasi AI dengan *cloud storage*, dokumen internal, sistem komunikasi digital, dan dashboard perusahaan memperluas permukaan serangan siber terhadap organisasi. Pada tahun 2024, beberapa penelitian keamanan AI menunjukkan bahwa *agent* berbasis LLM dapat diarahkan untuk melakukan *data exfiltration* melalui manipulasi prompt tersembunyi yang terdapat pada dokumen atau situs web eksternal. Kebocoran data yang berpotensi terjadi meliputi kredensial sistem, histori komunikasi internal, data pelanggan, hingga informasi strategis perusahaan. Situasi tersebut mulai menjadi perhatian dalam pengembangan kebijakan keamanan AI di berbagai perusahaan teknologi global.

Perkembangan *agentic AI* juga meningkatkan kompleksitas ancaman keamanan karena model mulai beroperasi secara semi-otonom dalam menjalankan tugas digital (Mathew, 2024). Beberapa *AI Agent* modern telah dirancang untuk melakukan *browsing* internet, membaca file, mengirim email, membuat ringkasan laporan, hingga menjalankan *workflow* otomatis tanpa intervensi manusia secara langsung. Pada tahun 2025, berbagai laporan keamanan menunjukkan bahwa *indirect prompt injection* memiliki tingkat keberhasilan yang tinggi pada *AI Agent* yang terhubung dengan layanan pihak ketiga. Penyerang memanfaatkan sumber eksternal yang terlihat aman untuk menyisipkan instruksi manipulatif agar dibaca oleh model selama proses eksekusi tugas. Kondisi tersebut berkaitan dengan meningkatnya ketergantungan perusahaan terhadap sistem AI yang memiliki akses lintas platform digital.

Di sisi lain, pengembangan strategi mitigasi terhadap *prompt injection* masih menghadapi berbagai tantangan teknis. Penggunaan *filtering* sederhana dinilai belum mampu menghadapi pola serangan yang bersifat adaptif dan kontekstual. Beberapa pengembang mulai menerapkan pendekatan seperti *context isolation*, *instruction hierarchy*, *role separation*, *adversarial testing*, serta *information flow control* untuk membatasi pengaruh *prompt* berbahaya terhadap perilaku *AI Agent*. Pada tahun 2025, MITRE melalui kerangka MITRE ATLAS mulai memperluas pembahasan mengenai teknik serangan terhadap sistem AI, termasuk manipulasi model generatif dan eksploitasi *prompt* berbasis bahasa alami. Pengembangan standar keamanan tersebut berlangsung seiring meningkatnya integrasi AI generatif dalam ekosistem bisnis modern.

Beberapa penelitian terdahulu telah membahas kerentanan *Large Language Model* (LLM) dan *AI Agent* terhadap serangan *prompt injection* dalam berbagai skenario keamanan digital. Penelitian Agarwal et al. (2024) menunjukkan bahwa kombinasi manipulasi percakapan dan efek *sycophancy* model mampu meningkatkan keberhasilan serangan hingga lebih dari 80 persen pada beberapa skenario pengujian. Sementara itu, Debenedetti et al. (2024) mengembangkan lingkungan evaluasi untuk mengukur ketahanan *AI Agent* terhadap *prompt injection* dan menemukan bahwa sebagian besar agent masih gagal mempertahankan keamanan ketika menerima instruksi manipulatif dari sumber eksternal.

Kedua penelitian tersebut masih berorientasi pada pengujian teknis keamanan model generatif, sedangkan pembahasan mengenai dampak *prompt injection* terhadap kebocoran data perusahaan, analisis ancaman berbasis OWASP Top 10 for LLM Applications dan MITRE ATLAS, serta strategi mitigasi keamanan yang dijabarkan secara kualitatif dalam lingkungan operasional perusahaan masih memerlukan pembahasan lebih komprehensif. Berdasarkan kondisi tersebut, penelitian ini bertujuan menganalisis kerentanan *AI Agent* terhadap serangan

prompt injection, mengidentifikasi potensi risiko kebocoran data perusahaan, serta mengkaji strategi mitigasi keamanan berbasis standar industri global dalam ekosistem digital modern.

2. METODE PENELITIAN

Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan metode kualitatif deskriptif dengan pendekatan threat modeling dan studi kasus siber kontemporer. Pendekatan tersebut digunakan untuk menganalisis pola kerentanan *AI Agent* terhadap serangan *prompt injection*, mekanisme eksploitasi yang digunakan penyerang, serta potensi dampaknya terhadap kebocoran data perusahaan dalam lingkungan digital modern. Analisis dilakukan melalui penelaahan berbagai kasus keamanan AI terkini, laporan ancaman siber global, serta kerangka keamanan industri seperti OWASP Top 10 for LLM Applications dan MITRE ATLAS guna memetakan pola ancaman dan strategi mitigasi keamanan *AI Agent*.

Sumber Data Penelitian

Data penelitian berasal dari data sekunder yang diperoleh melalui berbagai sumber akademik dan laporan keamanan siber internasional. Literatur yang digunakan diprioritaskan pada publikasi tahun 2021–2025 agar pembahasan selaras dengan perkembangan terbaru *AI security* dan *agentic AI*. Sumber data penelitian meliputi: Jurnal internasional terindeks Scopus dan IEEE. Prosiding konferensi keamanan siber dan AI. Laporan resmi OWASP dan MITRE. Dokumentasi teknis keamanan *Large Language Model* (LLM). Laporan industri dari Gartner dan perusahaan keamanan siber global.

Teknik Pengumpulan Data

Pengumpulan data dilakukan melalui studi dokumentasi dan studi literatur. Tahapan pengumpulan data dilakukan sebagai berikut: Mengidentifikasi publikasi yang membahas *AI Agent* dan *prompt injection*. Menyeleksi literatur berdasarkan relevansi tema dan tahun publikasi. Mengelompokkan data sesuai kategori ancaman keamanan AI. Mengumpulkan dokumentasi mengenai strategi mitigasi *prompt injection*. Mencatat pola kerentanan yang berkaitan dengan kebocoran data perusahaan

Teknik Analisis Data

Analisis data dilakukan secara kualitatif deskriptif dengan beberapa tahapan berikut:

Tabel 1. Teknik Analisis Data.

Tahapan Analisis	Uraian
Reduksi data	Menyaring literatur dan laporan yang relevan dengan fokus penelitian
Kategorisasi data	Mengelompokkan ancaman berdasarkan jenis serangan prompt injection
Interpretasi data	Menafsirkan hubungan antara kerentanan <i>AI Agent</i> dan risiko <i>data leakage</i>
Analisis kerangka keamanan	Memetakan ancaman berdasarkan OWASP Top 10 for LLM Applications dan MITRE ATLAS
Penarikan kesimpulan	Menyusun pola ancaman dan strategi mitigasi keamanan <i>AI Agent</i>

Kerangka Analisis Penelitian

Penelitian ini menggunakan dua kerangka utama dalam proses analisis, yaitu: OWASP Top 10 for LLM Applications yang digunakan untuk mengidentifikasi kategori kerentanan pada sistem berbasis LLM, khususnya *Prompt Injection*, *Sensitive Information Disclosure*, *Insecure Output Handline*, dan *Excessive Agency*. MITRE ATLAS digunakan untuk menganalisis pola serangan adversarial terhadap sistem AI, meliputi *reconnaissance* terhadap model AI, manipulasi input berbasis *prompt*, *data extraction*, dan *exploitation* terhadap *AI Agent* berbasis *tool access*.

Validitas Data

Validitas data dilakukan melalui teknik triangulasi sumber dengan membandingkan temuan dari jurnal ilmiah, laporan organisasi keamanan siber, dan dokumentasi teknis *AI security*. Proses tersebut diarahkan untuk menjaga konsistensi data serta memperkuat interpretasi terhadap pola kerentanan *AI Agent* dalam lingkungan operasional perusahaan.

3. HASIL DAN PEMBAHASAN

Hasil Penelitian

Kerentanan AI Agent terhadap Serangan Prompt Injection dalam Lingkungan Operasional Perusahaan

Implementasi *AI Agent* dalam lingkungan perusahaan mengalami peningkatan pesat sejak berkembangnya teknologi *Large Language Model* (LLM) pada tahun 2023–2025. *AI Agent* digunakan untuk membantu pengelolaan dokumen, analisis data, layanan pelanggan, pengolahan email, hingga otomatisasi *workflow digital* berbasis *cloud*. Integrasi tersebut memperluas kemampuan sistem AI karena agent dapat terhubung langsung dengan database internal, layanan penyimpanan *cloud*, serta aplikasi perusahaan melalui API dan *tool-calling system*. Kondisi tersebut juga memperbesar permukaan serangan siber karena *AI Agent* memproses data dan instruksi dari berbagai sumber eksternal secara simultan.

Laporan keamanan dari OWASP menempatkan *prompt injection* sebagai ancaman utama dalam OWASP Top 10 for LLM Applications 2025. Ancaman tersebut muncul ketika penyerang menyisipkan instruksi manipulatif agar *AI Agent* menjalankan tindakan yang berbeda dari kebijakan sistem yang telah ditetapkan. Dalam lingkungan perusahaan modern, *prompt injection* berkembang menjadi ancaman serius karena *AI Agent* memiliki kemampuan membaca dokumen internal, melakukan *browsing* otomatis, mengakses email perusahaan, hingga menjalankan perintah berbasis sistem eksternal.

Tabel 2. Bentuk Kerentanan *AI Agent* terhadap Serangan *Prompt Injection*.

Kerangka Analisis	Bentuk Ancaman	Dampak terhadap Perusahaan
OWASP LLM01: <i>Prompt Injection</i>	Manipulasi instruksi melalui input langsung atau sumber eksternal	Perubahan perilaku <i>AI Agent</i> dan pelanggaran kebijakan sistem
OWASP LLM02: <i>Sensitive Information Disclosure</i>	Kebocoran data akibat respons AI yang tidak terkontrol	Bocornya dokumen internal, kredensial, dan data pelanggan
OWASP LLM08: <i>Excessive Agency</i>	<i>AI Agent</i> memiliki akses <i>tool</i> dan API berlebihan	Eksekusi tindakan tanpa validasi keamanan
MITRE ATLAS – <i>Prompt Manipulation</i>	Eksplorasi konteks percakapan berbasis bahasa alami	Manipulasi respons dan pengambilan keputusan AI
MITRE ATLAS – <i>Data Extraction</i>	Pengambilan data sensitif melalui interaksi <i>prompt</i>	Kebocoran informasi strategis perusahaan
MITRE ATLAS – <i>Reconnaissance</i>	Pengamatan pola respons model AI	Penyusunan <i>prompt</i> manipulatif yang lebih efektif

Berdasarkan analisis pada Tabel 1, *prompt injection* memiliki hubungan langsung dengan risiko kebocoran data perusahaan. *AI Agent* bekerja dengan memproses seluruh teks sebagai bagian dari konteks interaksi, sehingga model sering kali gagal membedakan instruksi sah dan instruksi manipulatif. Pola tersebut menciptakan celah keamanan ketika *AI Agent* membaca halaman web, dokumen, spreadsheet, maupun email yang telah disisipi *prompt* tersembunyi oleh penyerang. Dalam temuan Zhan et al. (2024), *indirect prompt injection* menunjukkan tingkat keberhasilan yang cukup tinggi pada sistem *agentic AI* yang memiliki kemampuan *browsing* dan *tool access* otomatis.

Serangan *indirect prompt injection* umumnya dilakukan melalui sumber eksternal yang terlihat aman bagi sistem. Penyerang dapat menanamkan instruksi tersembunyi pada metadata file, halaman website, atau dokumen digital yang kemudian diproses oleh *AI Agent*. Ketika *AI Agent* membaca dokumen tersebut, model dapat diarahkan untuk mengabaikan *system prompt*, membuka akses terhadap informasi internal perusahaan, atau mengirimkan data sensitif kepada pihak eksternal.

Karakteristik ancaman tersebut berbeda dari serangan siber konvensional karena eksploitasi dilakukan melalui manipulasi bahasa alami dan isi percakapan. Dalam kerangka MITRE ATLAS, *prompt injection* dikategorikan sebagai bentuk *adversarial AI attack* yang memanfaatkan kelemahan pemrosesan bahasa alami pada model AI generatif. Teknik tersebut

sering dikombinasikan dengan *reconnaissance* untuk memahami pola respons *AI Agent* sebelum penyerang menyusun *prompt* manipulatif yang lebih efektif. Pendekatan tersebut menunjukkan bahwa ancaman terhadap *AI Agent* berkembang melalui kombinasi eksploitasi linguistik, manipulasi konteks, dan penyalahgunaan akses digital dalam sistem perusahaan modern.

Dampak *prompt injection* terhadap perusahaan berkaitan erat dengan kebocoran data internal organisasi. *AI Agent* yang memiliki akses terhadap *cloud storage*, email perusahaan, dashboard administrasi, maupun database pelanggan berpotensi membocorkan informasi sensitif ketika model dipengaruhi instruksi manipulatif. Yue et al. (2024) menunjukkan bahwa *AI Agent* dapat diarahkan untuk menampilkan isi dokumen rahasia, histori komunikasi internal, hingga kredensial sistem melalui pola interaksi berbasis *prompt*. Risiko tersebut menjadi perhatian besar pada sektor keuangan, kesehatan, dan layanan digital yang memiliki volume data sensitif tinggi.

Beberapa bentuk risiko kebocoran data akibat *prompt injection* dalam lingkungan operasional perusahaan meliputi: Kebocoran dokumen internal perusahaan. Pengungkapan kredensial dan token akses sistem. Penyebaran data pelanggan dan histori transaksi. Manipulasi *workflow* otomatis berbasis *AI Agent*. Pengiriman informasi sensitif melalui email otomatis. Pengambilan data strategis perusahaan oleh pihak eksternal

Gartner pada tahun 2026 menunjukkan bahwa peningkatan penggunaan *agentic AI* dalam perusahaan mendorong perubahan pola ancaman keamanan siber berbasis AI generatif. Perusahaan mulai mengintegrasikan *AI Agent* ke dalam berbagai aktivitas operasional seperti pengelolaan dokumen, layanan pelanggan, analisis data, pengolahan email, hingga otomatisasi *workflow* berbasis *cloud*. Integrasi tersebut memperluas kemampuan sistem AI karena agent dapat terhubung langsung dengan database internal, platform komunikasi digital, dan layanan pihak ketiga secara simultan. Di sisi lain, peningkatan konektivitas tersebut memperbesar peluang eksploitasi ketika *AI Agent* menerima instruksi manipulatif dari sumber eksternal yang tidak terverifikasi.

Kondisi tersebut berkaitan dengan meningkatnya ketergantungan organisasi terhadap sistem otomatisasi digital yang memiliki akses lintas platform dan data internal perusahaan. Perubahan pola ancaman keamanan tersebut mendorong organisasi global mulai memperkuat *AI governance* dan pengawasan keamanan terhadap implementasi *AI Agent* dalam lingkungan perusahaan modern. Berbagai perusahaan teknologi mulai menerapkan pembatasan akses *tool-calling*, pemisahan data sensitif, pengawasan aktivitas AI secara *real-time*, serta mekanisme validasi instruksi berbasis *multilayer security*. Pengembangan kebijakan keamanan AI juga

mulai diarahkan pada penguatan kontrol akses, audit sistem AI, dan pembentukan standar operasional penggunaan *agentic AI* dalam lingkungan kerja digital. *Prompt injection* dipandang sebagai ancaman yang sulit dieliminasi sepenuhnya karena berkaitan dengan karakteristik dasar *Large Language Model* yang memproses bahasa alami sebagai bagian utama dari sistem pengambilan keputusan AI.

Pembahasan Penelitian

Strategi Mitigasi Keamanan AI Agent Berbasis OWASP Top 10 for LLM Applications dan MITRE ATLAS

Peningkatan ancaman *prompt injection* terhadap *AI Agent* mendorong perusahaan dan organisasi teknologi mulai mengembangkan berbagai strategi mitigasi keamanan berbasis standar industri global. Penguatan keamanan AI modern berfokus pada pengendalian konteks percakapan, validasi instruksi, pembatasan akses *tool-calling*, serta pengawasan alur data yang diproses oleh *Large Language Model* (LLM). Pendekatan tersebut berkembang seiring meningkatnya penggunaan *agentic AI* dalam sistem operasional perusahaan sejak tahun 2023.

Laporan keamanan OWASP menempatkan mitigasi *prompt injection* sebagai prioritas utama dalam pengembangan keamanan sistem berbasis LLM. Strategi mitigasi diarahkan untuk membatasi kemampuan model dalam memproses instruksi manipulatif yang berasal dari pengguna maupun sumber eksternal. Sementara itu, kerangka MITRE ATLAS digunakan untuk memetakan pola serangan adversarial terhadap AI serta membantu organisasi memahami tahapan eksploitasi yang digunakan penyerang dalam lingkungan *agentic AI*.

Tabel 3. Strategi Mitigasi Keamanan AI Agent terhadap *Prompt Injection*.

Strategi Mitigasi	Bentuk Implementasi	Tujuan Keamanan
<i>Context Isolation</i>	Memisahkan instruksi sistem dan input pengguna	Mengurangi manipulasi konteks percakapan
<i>Input Validation</i>	<i>Filtering prompt</i> dan deteksi instruksi berbahaya	Mencegah <i>prompt</i> manipulatif masuk ke sistem
<i>Role-Based Access Control</i>	Pembatasan akses <i>AI Agent</i> terhadap <i>tool</i> dan data	Mengurangi risiko <i>excessive agency</i>
<i>Human-in-the-Loop</i>	Validasi manusia pada tindakan sensitif <i>AI Agent</i>	Mengontrol eksekusi otomatis sistem
<i>Output Monitoring</i>	Pemeriksaan respons AI sebelum diproses sistem lain	Mencegah kebocoran data dan <i>output</i> berbahaya
<i>Adversarial Testing</i>	Simulasi serangan <i>prompt injection</i> secara berkala	Mengidentifikasi kelemahan keamanan AI
<i>Data Segmentation</i>	Pemisahan data sensitif perusahaan	Membatasi akses AI terhadap informasi kritis

Berdasarkan Tabel 2, *context isolation* menjadi salah satu strategi mitigasi yang banyak digunakan dalam pengembangan *AI Agent* modern. Pendekatan tersebut dilakukan dengan memisahkan *system prompt*, instruksi pengguna, dan data eksternal agar model tidak memproses seluruh input sebagai konteks yang memiliki prioritas setara. Dalam beberapa

implementasi *enterprise AI*, mekanisme tersebut dikombinasikan dengan *instruction hierarchy* untuk menjaga agar *AI Agent* tetap mengikuti kebijakan internal sistem meskipun menerima prompt manipulatif dari sumber eksternal.

Strategi *input validation* juga digunakan untuk mendeteksi pola *prompt injection* sebelum diproses oleh model AI. Sistem melakukan *filtering* terhadap instruksi yang mengandung pola manipulatif, *command* tersembunyi, maupun upaya eksploitasi berbasis bahasa alami. Pada beberapa *platform AI enterprise*, *filtering* dilakukan menggunakan kombinasi *rule-based detection* dan *machine learning classification* untuk mengenali *prompt* berbahaya secara lebih adaptif. Pendekatan tersebut berkembang karena pola serangan *prompt injection* terus mengalami perubahan dan penyesuaian terhadap mekanisme pertahanan model AI.

Dalam kerangka OWASP Top 10 for LLM Applications, *excessive agency* menjadi salah satu ancaman utama terhadap *AI Agent* modern. Kondisi tersebut muncul ketika *AI Agent* memiliki akses berlebihan terhadap *tool*, *API*, *cloud storage*, maupun database perusahaan tanpa mekanisme pembatasan yang memadai. Oleh sebab itu, penerapan *role-based access control* dan *principle of least privilege* mulai digunakan untuk membatasi ruang akses *AI Agent* hanya pada fungsi yang benar-benar dibutuhkan dalam operasional sistem.

Pendekatan *human-in-the-loop* juga mulai diterapkan pada berbagai sistem AI perusahaan, terutama untuk aktivitas yang berkaitan dengan data sensitif dan pengambilan keputusan operasional. Dalam mekanisme tersebut, *AI Agent* tetap dapat menjalankan proses otomatisasi, namun tindakan tertentu memerlukan persetujuan manusia sebelum dieksekusi. Strategi tersebut digunakan untuk mengurangi risiko kebocoran data, pengiriman informasi sensitif, maupun manipulasi *workflow* akibat *prompt injection*.

Berdasarkan kerangka MITRE ATLAS, *adversarial testing* menjadi bagian penting dalam penguatan keamanan *AI Agent*. Organisasi melakukan simulasi serangan *prompt injection* untuk mengidentifikasi kelemahan sistem sebelum dimanfaatkan oleh pihak eksternal. Pengujian dilakukan melalui berbagai skenario, seperti *indirect prompt injection*, *data extraction*, *context manipulation*, dan *privilege escalation* berbasis AI. Hasil pengujian tersebut kemudian digunakan untuk memperbarui kebijakan keamanan dan memperkuat mekanisme validasi instruksi pada *AI Agent*. Beberapa langkah mitigasi yang mulai diterapkan perusahaan dalam pengelolaan *AI Agent* meliputi: Membatasi akses AI terhadap data sensitif perusahaan. Memisahkan lingkungan operasional AI dan database utama. Menggunakan *sandbox environment* untuk proses *tool-calling*. Menerapkan monitoring terhadap aktivitas *AI*

Agent secara *real-time*. Melakukan audit keamanan AI secara berkala. Mengembangkan kebijakan AI governance dalam organisasi

Laporan Gartner menunjukkan bahwa pengembangan keamanan *agentic AI* akan menjadi prioritas utama perusahaan global dalam beberapa tahun mendatang. Peningkatan penggunaan AI generatif dalam sistem operasional perusahaan berkaitan dengan kebutuhan penguatan *AI governance*, pengawasan akses data, dan pengembangan standar keamanan baru terhadap ancaman berbasis *prompt injection*. Situasi tersebut memperlihatkan bahwa mitigasi keamanan *AI Agent* berkembang sebagai bagian dari transformasi keamanan siber modern dalam ekosistem digital berbasis kecerdasan buatan.

4. KESIMPULAN

Perkembangan *AI Agent* dalam lingkungan operasional perusahaan meningkatkan efisiensi sistem digital sekaligus memperluas risiko keamanan siber berbasis kecerdasan buatan. Serangan *prompt injection* berkembang sebagai ancaman utama terhadap sistem berbasis *Large Language Model* (LLM) karena mampu memanipulasi perilaku *AI Agent* melalui instruksi berbasis bahasa alami. Analisis berdasarkan kerangka OWASP Top 10 for LLM Applications dan MITRE ATLAS menunjukkan bahwa *prompt injection* memiliki hubungan erat dengan risiko kebocoran data perusahaan, terutama pada *AI Agent* yang terhubung dengan *tool-calling system*, *cloud storage*, email perusahaan, dan database internal organisasi. Strategi mitigasi seperti *context isolation*, *input validation*, *role-based access control*, *adversarial testing*, serta penguatan *AI governance* mulai digunakan untuk mengurangi potensi eksploitasi terhadap sistem AI modern dalam lingkungan digital perusahaan.

Perusahaan dan pengembang AI perlu memperkuat mekanisme keamanan *AI Agent* melalui pembatasan akses sistem, validasi instruksi berbasis *multilayer security*, serta pengawasan aktivitas AI secara berkala dalam lingkungan operasional digital. Pengembangan kebijakan *AI governance* juga perlu diarahkan pada pengendalian penggunaan *tool-calling* dan perlindungan data sensitif perusahaan agar risiko kebocoran informasi dapat diminimalkan. Penelitian selanjutnya dapat mengembangkan analisis berbasis studi kasus empiris pada implementasi *AI Agent* di sektor keuangan, kesehatan, maupun layanan publik untuk memperoleh gambaran yang lebih spesifik mengenai pola eksploitasi *prompt injection* dan efektivitas strategi mitigasi keamanan AI generatif.

DAFTAR REFERENSI

- Agarwal, D., Fabbri, A. R., Risher, B., Laban, P., Joty, S., & Wu, C.-S. (2024). Prompt leakage effect and defense strategies for multi-turn LLM interactions. *arXiv preprint*, 1(1), 1–10.
- Chaitanya, K., & Rolla, K. J. (2024). The evolution and impact of large language models in artificial intelligence. In *Algorithms in Advanced Artificial Intelligence* (pp. 410–417). CRC Press.
- Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., ... Wang, X. (2024). When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4), 42.
- Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems*, 37, 82895–82920.
- Gartner. (2026). *Gartner predicts 25% of all enterprise GenAI applications will experience at least five minor security incidents per year by 2028*. <https://www.gartner.com/en/newsroom/press-releases/2026-04-09-gartner-predicts-25-percent-of-all-enterprise-gen-ai-applications-will-experience-at-least-five-minor-security-incidents-per-year-by-2028>
- Mathew, E. (2024). Enhancing security in large language models: A comprehensive review of prompt injection attacks and defenses. *Authorea Preprints*.
- OWASP. (2024). *OWASP Top 10 for LLM applications 2025*.
- Rababah, B., Wu, S. T., Kwiatkowski, M., Leung, C. K., & Akcora, C. G. (2024). SoK: Prompt hacking of large language models. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 5392–5401). IEEE.
- Rane, N. L., Tawde, A., Choudhary, S. P., & Rane, J. (2023). Contribution and performance of ChatGPT and other large language models (LLM) for scientific and research advancements: A double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science*, 5(10), 875–899.
- Sarker, I. H. (2024). Generative AI and large language modeling in cybersecurity. In *AI-driven cybersecurity and threat intelligence: Cyber automation, intelligent decision-making and explainability* (pp. 79–99). Springer.
- Yue, C. A., Men, L. R., Mitson, R., Davis, D. Z., & Zhou, A. (2024). Artificial intelligence for internal communication: Strategies, challenges, and implications. *Public Relations Review*, 50(5), 102515.
- Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024). Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *Findings of the Association for Computational Linguistics: ACL 2024*, 10471–10506.