Journal of Technology and Data Science (JTDS)
Vol. 3, No. 2, Desember 2025, hal. 216-228
p-ISSN: 2962-3693 | e-ISSN: 2962-7788

Kajian Konseptual AI Agent On-Chain Berbasis LLM untuk Manajemen Aset DeFi

Ridwan Yusuf[1], Andreas Perdana[1]*, Febri Sugandi[1], Untoro Apsiswanto[1]

[1]Universitas Dharmawacana Metro, Metro, Indonesia

E-mail: ryusuf87@gmail.com; andreas.perdana@dharmawacana.ac.id; fsugandi87@gmail.com; untorolampung@gmail.com

**Penulis Korespondensi (Corresponding Author)*

© 2025. Ridwan Yusuf dkk. Diterbitkan oleh JTDS. Artikel ini bersifat open-access di bawah lisensi Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0).

Info Artikel

Diterima:
25 September 2025
Direvisi:
15 November 2025
Diterbitkan:
25 Desember 2025

Kata Kunci:
AI agent; large language model; smart contract; DeFi; keamanan agen otonom

ABSTRAK

Pada periode 2024–2025, pertemuan antara Large Language Model dan teknologi blockchain melahirkan kelas perangkat lunak baru yang disebut AI Agent on-chain perangkat lunak otonom yang menerjemahkan tujuan pengguna menjadi rangkaian transaksi Decentralized Finance dan menandatanganinya tanpa intervensi manusia. Adopsi yang berlangsung cepat memperlihatkan tiga celah konseptual yang belum terjawab: kompleksitas alur eksekusi multi-langkah pada DeFi, lansekap kerentanan yang khas LLM seperti indirect prompt injection dan jailbreak yang berimplikasi langsung pada aset finansial, dan ketiadaan kerangka tata kelola yang akuntabel pada ekosistem yang terdesentralisasi. Kajian ini menyintesis 34 publikasi terverifikasi dari basis data Zotero ke dalam empat tema yang saling terkait: arsitektur agen dan kolaborasi multi-agen, manajemen kunci pada agentic wallet, lansekap ancaman keamanan, serta peran blockchain sebagai trust layer. Temuan kunci yang muncul adalah bahwa pertahanan yang hanya bertumpu pada lapisan bahasa terbukti tidak memadai; kontrol pada lapisan tindakan melalui kebijakan yang dinamis, isolasi enclave, atau secret sharing multi-pihak menjadi prasyarat ketika konsekuensi finansial bersifat permanen. Kontribusi yang diajukan mencakup taksonomi vektor ancaman per lapis arsitektur, kerangka pertahanan tiga-lapis hibrid, serta agenda riset masa depan dengan lima arah konkret.

Article Info

Received:
September 25, 2025
Revised:
November 15, 2025
Published:
December 25, 2025

Keywords:
AI agent; large language model; smart contract; DeFi; autonomous agent security

ABSTRACT

Over 2024 and 2025, the encounter between Large Language Models and blockchain technology produced a new class of software called on-chain AI Agents autonomous programs that translate user intent into chains of Decentralized Finance transactions and sign them without human intervention. The rapid adoption has exposed three conceptual gaps that remain unresolved: the complexity of multi-step DeFi execution flows, an LLM-specific vulnerability landscape such as indirect prompt injection and jailbreak that directly threatens financial assets, and the absence of an accountable governance framework within decentralised ecosystems. This study synthesises 34 verified publications from a Zotero corpus across four

interconnected themes: agent architecture and multi-agent collaboration, key management within agentic wallets, the security threat landscape, and blockchain's role as a trust layer. The key finding that emerges is that defences relying solely on the language layer prove insufficient; action-layer controls through dynamic policies, enclave isolation, or multi-party secret sharing become preconditions when the financial consequences are permanent. The contributions offered include a per-layer threat taxonomy, a three-layer hybrid defence framework, and a future research agenda with five concrete directions.

1. PENDAHULUAN

Selama beberapa tahun terakhir, dua aliran teknologi yang berkembang independen kecerdasan buatan generatif berbasis Large Language Model dan teknologi blockchain mulai bertemu pada satu titik praktis. Pertemuan itu memunculkan agen otonom yang menggunakan LLM sebagai mesin penalaran untuk menyusun rencana eksekusi dan menandatangani transaksi pada jaringan publik. Pada protokol Decentralized Finance, agen jenis ini bukan lagi konsep teoritis: mereka memindahkan likuiditas, mengatur strategi hasil, dan ikut memberikan suara pada Decentralized Autonomous Organization. Atribut blockchain yang transparan, immutable, dan tanpa perantara secara unik mendukung kebutuhan agen akan identitas yang dapat diverifikasi dan jejak audit yang permanen. Justru atribut yang sama itulah yang menjadi sumber risiko: setiap kesalahan tanda tangan menjadi permanen dan tidak dapat ditarik kembali tanpa cabang konsensus.

Dalam dua tahun terakhir, riset tentang perilaku dan keamanan AI Agent berkembang dengan cepat namun pada jalur yang relatif terpisah. Di sisi arsitektur, Chen [1] menelusuri perpindahan dari kolaborasi yang muncul ad-hoc pada awal generasi multi-agent menuju kecerdasan kolektif yang lebih terstruktur dengan peran eksplisit, dan Kulpa dkk. [2] menerapkan pola yang sama untuk pengambilan keputusan finansial. Pada arah keamanan, bukti empiris bermunculan dengan nada peringatan: Zhan dkk. [3] menunjukkan bahwa pertahanan yang dipublikasikan dapat dipatahkan oleh penyerang yang adaptif, Johnson dkk. [4] memperagakan eksploitasi nyata terhadap agen navigasi web, dan Shahroz dkk. [5] memperlihatkan bahwa serangan prompt yang teroptimasi merambat antar agen pada sistem multi-agent. Pada irisan blockchain, dua karya menonjol: Liu dkk. [6] memadukan Graph Neural Network dengan pengetahuan pakar untuk audit smart contract, sedangkan Mane dkk. [7] memanfaatkan LLM untuk mendeteksi kerentanan firmware Internet of Things. Sintesis lintas-sumber yang menghubungkan ketiga arah inilah yang masih jarang dilakukan.

Pemetaan terhadap literatur memperlihatkan tiga celah substantif yang saling berkait. Yang pertama adalah pemisahan epistemik: studi keamanan LLM dan studi keamanan blockchain hidup sebagai dua percakapan yang tidak banyak saling membaca, sehingga rumusan tentang bagaimana ancaman LLM diterjemahkan menjadi kerugian on-chain belum tersedia secara sistematis. Celah kedua bersifat arsitektural: mitigasi dominan terhadap prompt injection dan jailbreak sebagaimana dilaporkan Fathima dkk. [8], Lan dkk. [9], Naik dkk. [10], dan Choudhary dkk. [11] masih terkonsentrasi pada lapisan bahasa, sementara karakter kerugian on-chain bersifat permanen sehingga membutuhkan kontrol pada lapisan tindakan. Celah ketiga berkaitan dengan tata kelola: praktisi yang merancang agentic wallet, prosedur audit, atau kebijakan kepatuhan tidak menemukan kerangka rujukan yang menyatukan ketiga aspek tersebut dalam satu peta yang kohesif. Ketiga celah ini bersifat saling memperkuat menyelesaikan satu tanpa yang lain akan menghasilkan solusi parsial.

Berangkat dari ketiga celah tersebut, kajian ini mengarahkan pertanyaan pada tiga aspek. Pertama, pemetaan arsitektur AI Agent on-chain ke dalam lapisan-lapisan fungsional yang memadukan literatur multi-agent LLM dengan literatur blockchain (RQ1). Kedua, identifikasi ancaman keamanan paling kritis pada agen yang mengelola aset DeFi, beserta bagaimana mitigasi pada lapisan bahasa dan lapisan tindakan saling melengkapi (RQ2). Ketiga, perumusan agenda riset yang dapat menyatukan langkah pengembang LLM, peneliti keamanan blockchain, dan regulator pasar finansial (RQ3). Tiga kontribusi spesifik yang ditawarkan adalah: kerangka empat-lapis arsitektur agen on-chain yang berfungsi sebagai bahasa rujukan bersama; taksonomi ancaman yang memetakan vektor serangan ke permukaan eksploitasi pada masing-masing lapis; dan kerangka pertahanan tiga-lapis hibrid yang menggabungkan kontrol bahasa, kebijakan dinamis, dan isolasi kunci. Susunan naskah berikutnya: Bagian 2 menyusun landasan teoritis, Bagian 3 menjelaskan metode kajian, Bagian 4 menyajikan hasil sintesis, dan Bagian 5 menutup dengan agenda penelitian.

2. KAJIAN PUSTAKA

Lima tema membentuk landasan teoritis kajian: perkembangan sistem multi-agent berbasis LLM, peta ancaman prompt injection dan jailbreak, mekanisme pertahanan dari lapisan bahasa hingga lapisan tindakan, ekosistem blockchain dan keamanan smart contract, serta peran blockchain sebagai trust layer. Tabel I memberikan pemetaan posisi kajian terhadap empat publikasi rujukan utama.

2.1 Sistem Multi-Agent Berbasis Large Language Model

Selama dua tahun terakhir, sistem multi-agent berbasis LLM bergerak dari konsep penelitian menuju penerapan operasional. Chen [1] mendokumentasikan transisi dari kolaborasi emergent yang muncul tanpa koordinasi formal menuju kecerdasan kolektif dengan peran yang lebih eksplisit dan protokol pertukaran pesan yang lebih disiplin. Pada domain finansial, Kulpa dkk. [2] menerapkan pola serupa dengan modul perencanaan, riset, dan eksekusi yang terkoordinasi untuk pengambilan keputusan. Bukti tambahan muncul dari domain teknik: Ghafarollahi dan Buehler [12] menggabungkan Graph Neural Network dengan LLM untuk merancang alloy melalui sistem multi-agent, memperlihatkan kemampuan agen yang terkoordinasi melampaui model tunggal. Pada konteks lain, Lan dkk. [13] memperkenalkan AutoQual sebagai agen LLM untuk discovery fitur interpretable, Teplyi dan Dosyn [14] memakai agen LLM dengan prompt panduan untuk pembelajaran ontologi end-to-end, Kakkar [15] memperluas penerapan ke supervisi Robotic Process Automation, dan Hazime dkk. [16] mengevaluasi kemampuan agen LLM menyelesaikan multi-arm bandit problem. Mukherjee [17] menambahkan dimensi keamanan dengan memakai agentic AI berbasis LLM untuk mendeteksi pola pada perangkat IoT. Apa yang dapat ditarik dari konvergensi temuan ini? Agen LLM bekerja efektif ketika tujuan dirumuskan secara sempit, ruang tindakan terbatas, dan tools yang relevan tersedia kondisi yang juga diharapkan pada agen on-chain pengelola DeFi.

2.2 Lansekap Ancaman Prompt Injection dan Jailbreak

Dua kelas serangan menyita perhatian utama dalam keamanan LLM: prompt injection dan jailbreak. Pada varian indirect-nya, prompt injection beroperasi melalui konten eksternal yang dibaca agen halaman web, email, respons API yang kemudian menyusupkan instruksi pembajak alur eksekusi. Bukti efektivitas serangan ini hadir dengan nada peringatan: Zhan dkk. [3] memperlihatkan bahwa pertahanan yang ada dapat dipatahkan dengan serangan adaptif yang dirancang khusus, dan Johnson dkk. [4] memperagakan eksploitasi nyata pada agen navigasi web. Pada sisi pertahanan, Wang dkk. [18] mengusulkan AgentVigil sebagai red-teaming black-box otomatis, Jia dkk. [19] memperkenalkan Task Shield yang menegakkan task alignment, dan An dkk. [20] mengusulkan IPIGuard berbasis graf ketergantungan tool. Sisi ofensif tidak kalah produktif: Hu dkk. [21] menunjukkan bahwa fusi prompt multi-agent dapat mengotomasi jailbreak, Wang dkk. [22] memakai model difusi untuk memanipulasi

prompt, dan Deng dkk. [23] mengembangkan MASTERKEY untuk jailbreak otomatis pada chatbot LLM. Yang paling mengkhawatirkan adalah EchoLeak eksploitasi zero-click prompt injection pertama yang dikonfirmasi pada sistem produksi sebagaimana dilaporkan Reddy dan Gujral [24]. Pada konteks multi-agent, temuan Shahroz dkk. [5] sangat penting: serangan dapat merambat antar agen, sehingga kompromi pada satu agen menulari yang lain. Untuk agen on-chain yang membaca data eksternal atau berinteraksi dengan agen lain, permukaan serangan ini sangat luas dan tidak bisa diabaikan.

2.3 Mekanisme Pertahanan: dari Lapisan Bahasa ke Lapisan Tindakan

Pertahanan terhadap serangan LLM berkembang dalam dua jalur yang saling melengkapi. Jalur pertama bekerja pada lapisan bahasa: Lan dkk. [9] mengkaji deteksi prompt injection pada aplikasi terintegrasi LLM, Naik dkk. [10] menelaah penanganan output yang tidak aman, Fathima dkk. [8] mengusulkan kerangka deteksi dan mitigasi serangan prompt adversarial, dan Choudhary dkk. [11] secara kritis menunjukkan keterbatasan deteksi yang sendirinya bergantung pada LLM. Sebagai pelengkap, Yang dkk. [25] mengembangkan ShellBox honeypot interaktif berbasis LLM yang dirancang untuk mempelajari pola serangan. Panterino dan Fellington [26] menambahkan moving target defense yang mengubah konfigurasi prompt secara dinamis untuk menyulitkan serangan terarah. Jalur kedua bertindak pada lapisan tindakan: alih-alih mengandalkan interpretasi bahasa yang rentan, sistem dirancang sehingga agen yang berhasil dipaksa tetap memiliki kemampuan tindakan yang sempit. Logika ini diterapkan Kavian dkk. [27] pada domain kode dan memberi inspirasi langsung pada operasi on-chain: kebijakan pengeluaran yang ketat, daftar putih alamat, dan otorisasi multi-pihak harus diberlakukan di lapisan eksekusi transaksi. Pemisahan kontrol ini menjadi fondasi kerangka tiga-lapis pertahanan yang dijabarkan pada Bagian 4.

2.4 Blockchain, Smart Contract, dan Ekosistem Web3

Blockchain memberikan substrat eksekusi yang transparan tetapi immutable bagi agen otonom sifat yang membuatnya menarik sekaligus berbahaya. Liu dkk. [6] menggabungkan Graph Neural Network dengan pengetahuan pakar untuk mendeteksi kerentanan smart contract; pendekatan ini relevan untuk audit kontrak yang dipanggil agen pada deployment baru. Wu dkk. [28] menunjukkan bagaimana deteksi phishing pada Ethereum dapat dilakukan dengan network embedding pada graf transaksi memberi sinyal bahwa anomali pada pola pemanggilan kontrak dapat dijadikan basis sistem deteksi ancaman real-time. Pada lapisan keamanan perangkat keras, Arcenegui dkk. [29] mengikat perangkat fisik ke smart Non-Fungible Token melalui Physically Unclonable Function preseden yang secara konseptual paralel dengan agentic wallet berbasis Trusted Execution Environment dimana kunci agen tidak pernah meninggalkan enclave. Karandikar dkk. [30] menunjukkan kematangan smart contract pada infrastruktur energi komunitas melalui penggunaan token fungible dan non-fungible secara komplementer. Pada sisi pasar, Ante [31] memetakan keterkaitan pasar Non-Fungible Token dengan Bitcoin dan Ethereum, sementara Wang dkk. [32] mendokumentasikan perilaku gelembung pada NFT dan DeFi pengingat penting bahwa volatilitas tinggi adalah kondisi normal yang menjadi konteks operasional bagi setiap agen yang berdagang.

2.5 Blockchain sebagai Trust Layer Ekonomi Agen

Posisi blockchain sebagai trust layer bertumpu pada tiga atribut yang saling menguatkan: identitas yang dapat diverifikasi, jejak audit permanen, dan eksekusi tanpa perantara. Wilson dkk. [33] memetakan stakeholder ekonomi digital berbasis NFT sebagai contoh konkret bagaimana atribut tersebut menstrukturkan ekosistem yang kompleks. Ritterbusch dan Teichmann [34] menyusun definisi sistematis Metaverse yang memposisikan Web3, blockchain, dan AI sebagai komponen yang saling berkait. Sintesis kedua karya ini menyiratkan bahwa agen otonom membutuhkan infrastruktur identitas dan reputasi on-chain agar dapat berinteraksi dengan agen lain bukan sekadar dengan manusia secara

akuntabel. Kesenjangan yang terlihat: belum tersedia kerangka standar yang menyatukan ketiga atribut tersebut menjadi protokol lengkap. Berbagai usulan ERC mutakhir mencoba mengisi celah ini, namun pada saat kajian ini ditulis, adopsi massalnya belum terbentuk. Implikasi praktis yang ditarik: pertahanan agen harus dirancang dengan asumsi konservatif tidak menggantungkan keamanan pada protokol yang belum matang, melainkan pada kombinasi kontrol berlapis yang masih berlaku jika satu atau dua lapis gagal.

TABEL I
PERBANDINGAN EMPAT STUDI RUJUKAN UTAMA TERKAIT AI AGENT ON-CHAIN

Studi	Fokus Utama	Pendekatan	Domain	Kontribusi
Chen [1]	Multi-agent LLM	Sintesis literatur	AI umum	Taksonomi kolaborasi multi-agent
Zhan dkk. [3]	Pertahanan prompt injection	Eksperimental	LLM Agent	Bukti adaptive attacks dapat menembus pertahanan
Liu dkk. [6]	Deteksi kerentanan smart contract	GNN + ekspert	Blockchain	Akurasi deteksi melebihi rule-based
Mane dkk. [7]	Vulnerability detection berbasis LLM	Pipeline EMBA + Ghidra + LLM	IoT firmware	Otomasi audit kode tertutup
Kajian ini	Arsitektur, ancaman, pertahanan agen on-chain DeFi	Sintesis konseptual	AI Agent + Blockchain	Kerangka 4-lapis arsitektur, taksonomi ancaman, pertahanan 3-lapis

3. METODOLOGI KAJIAN

Kajian ini berbentuk tinjauan konseptual yang menyintesis publikasi mutakhir dari basis data Zotero personal peneliti. Tiga elemen metodologis dijelaskan pada bagian ini: protokol pencarian literatur, kriteria seleksi dan verifikasi, serta skema ekstraksi dan analisis. Alur tahapan tersebut diilustrasikan pada Gbr. 1.



Gbr. 1. Alur kajian konseptual: pencarian, seleksi, ekstraksi, dan sintesis.

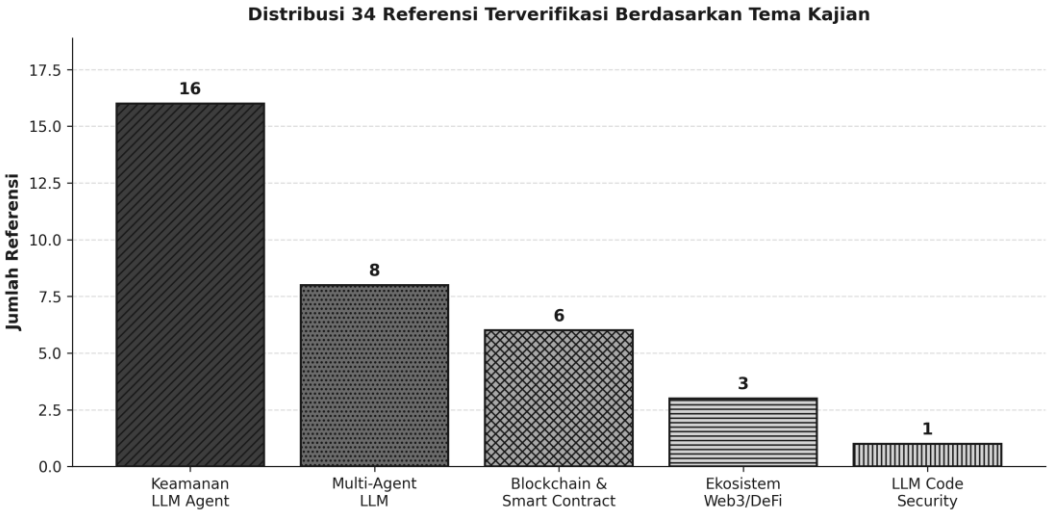
3.1 Sumber Data dan Protokol Pencarian

Basis data tunggal yang dipakai adalah Zotero personal peneliti yang telah dikurasi selama beberapa tahun pada bidang sistem informasi, AI, dan blockchain. Pencarian memakai kombinasi kata kunci pada bidang judul yang dikelompokkan menjadi empat klaster tematik. Klaster arsitektur agen mencakup 'agent', 'autonomous', 'multi-agent', dan 'large language model'. Klaster keamanan mencakup

'prompt injection', 'jailbreak', 'LLM security', dan 'adversarial'. Klaster blockchain mencakup 'smart contract', 'Ethereum', 'blockchain', dan 'DeFi'. Klaster infrastruktur mencakup 'TEE', 'multi-party computation', dan 'zero-knowledge'. Kriteria inklusi yang diberlakukan: artikel jurnal atau prosiding terindeks yang diterbitkan antara 2021 dan 2025, memiliki Digital Object Identifier valid, dan tersedia berkas PDF yang dapat dibuka di basis data Zotero. Kriteria eksklusi diberlakukan untuk artikel yang topiknya berada di luar kerangka kajian misalnya klasterisasi kabur untuk akreditasi atau data mining medis yang muncul karena tumpang-tindih kosakata.

3.2 Seleksi dan Verifikasi Referensi

Seleksi dilakukan melalui tiga lapis penyaringan yang memperketat tahap demi tahap. Lapis pertama menerapkan ekspresi reguler pada judul untuk memberi skor relevansi yang berbobot. Lapis kedua memverifikasi keberadaan berkas PDF pada basis data Zotero artikel tanpa berkas PDF dieksklusikan demi menjaga ketertelusuran konten yang dirujuk. Lapis ketiga membaca abstrak untuk memastikan relevansi konseptual; hanya yang lulus lapis ini yang menjadi rujukan kajian. Tiga puluh empat referensi memenuhi seluruh kriteria, dengan distribusi tema sebagai berikut: enam belas pada keamanan LLM Agent, delapan pada sistem multi-agent dan agen otomatis, sembilan pada blockchain dan smart contract, serta tiga pada konteks ekosistem digital terdesentralisasi. Distribusi ini menggambarkan dengan baik dominasi diskursus keamanan agen pada literatur terbaru, dan sekaligus memperlihatkan tipisnya jembatan konseptual antara studi LLM dan studi blockchain. Visualisasi distribusi ditampilkan pada Gbr. 2.



Gbr. 2. Distribusi 34 referensi terverifikasi berdasarkan tema kajian.

3.3 Ekstraksi Data dan Skema Analisis

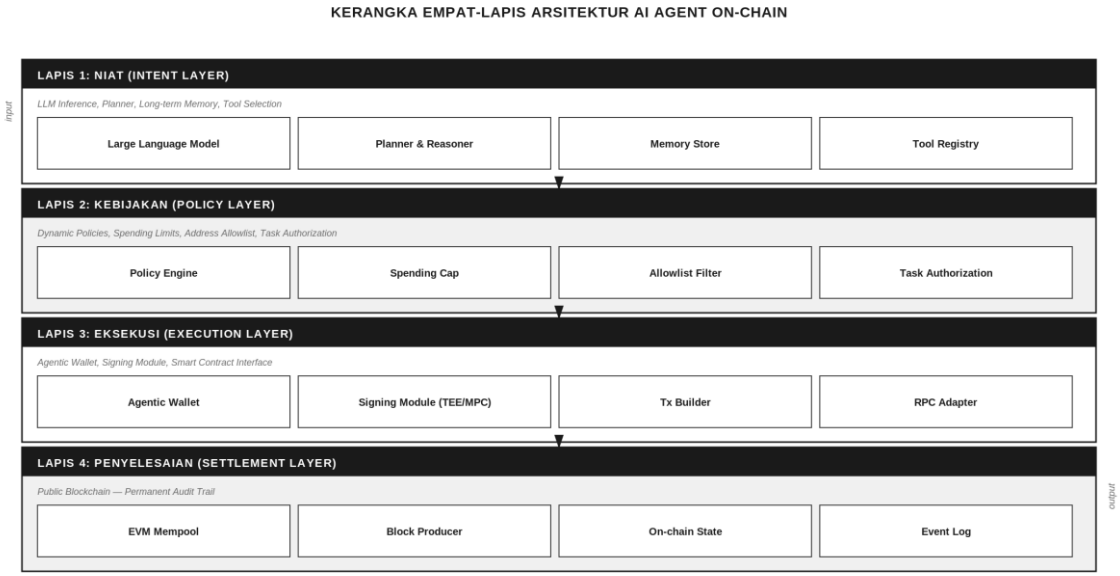
Untuk masing-masing referensi terpilih, peneliti mengekstrak metadata standar (penulis, tahun, judul, sumber, DOI) bersama tiga elemen analitis: klaim utama atau temuan kunci yang dilaporkan; metode atau pendekatan yang digunakan; dan keterbatasan yang diakui penulis. Informasi tersebut disusun dalam matriks sintesis yang menjadi fondasi empat artefak konseptual yang dihasilkan kajian: kerangka empat-lapis arsitektur AI Agent on-chain (Gbr. 3), taksonomi vektor ancaman (Gbr. 4), kerangka pertahanan tiga-lapis (Gbr. 5), dan agenda riset (Gbr. 6). Validitas konstruk diperiksa melalui konsistensi pemetaan: setiap elemen kerangka dirujuk minimal dua referensi independen, dan setiap kategori taksonomi ancaman wajib memiliki rujukan empiris yang spesifik. Keterbatasan pendekatan ini diakui secara terbuka pada Bagian 4.5.

4. ANALISIS DAN DISKUSI

Empat artefak analitis disajikan dalam bagian ini sebagai hasil sintesis: kerangka arsitektur empat-lapis, taksonomi ancaman, kerangka pertahanan, dan diskusi posisi blockchain sebagai trust layer. Setiap artefak diakhiri pengakuan keterbatasan yang memungkinkan penelitian lanjutan menguji konstruksi yang diajukan.

4.1 Kerangka Empat-Lapis Arsitektur AI Agent On-Chain

Sintesis terhadap literatur multi-agent LLM [1], [2], [12]–[17] dan literatur smart contract [6], [28]–[30] menghasilkan kerangka empat-lapis arsitektur AI Agent on-chain seperti dipaparkan pada Gbr. 3. Lapis pertama, Lapis Niat, berisi LLM untuk pemahaman bahasa, modul perencanaan, dan memori jangka panjang. Lapis kedua, Lapis Kebijakan, menampung mesin kebijakan dinamis, batasan pengeluaran, dan daftar putih alamat. Lapis ketiga, Lapis Eksekusi, berisi agentic wallet, mekanisme tanda tangan transaksi, serta antarmuka ke smart contract. Lapis keempat, Lapis Penyelesaian, adalah blockchain publik yang menyediakan jejak audit permanen. Setiap transaksi menempuh keempat lapis secara berurutan; kegagalan validasi pada lapis manapun membatalkan transaksi sebelum mencapai tahap tanda tangan. Pemisahan kontrol ini memberi properti penting: kompromi pada satu lapis tidak otomatis berarti kerugian finansial selama lapis-lapis berikutnya tetap utuh. Perlu diakui bahwa kerangka empat lapis ini merupakan abstraksi implementasi nyata sering memadukan beberapa lapis dalam satu komponen atau memecah satu lapis menjadi sub-modul. Sebagai abstraksi rujukan, kerangka ini dimaksudkan untuk menjadi bahasa bersama, bukan cetak biru implementasi tunggal.



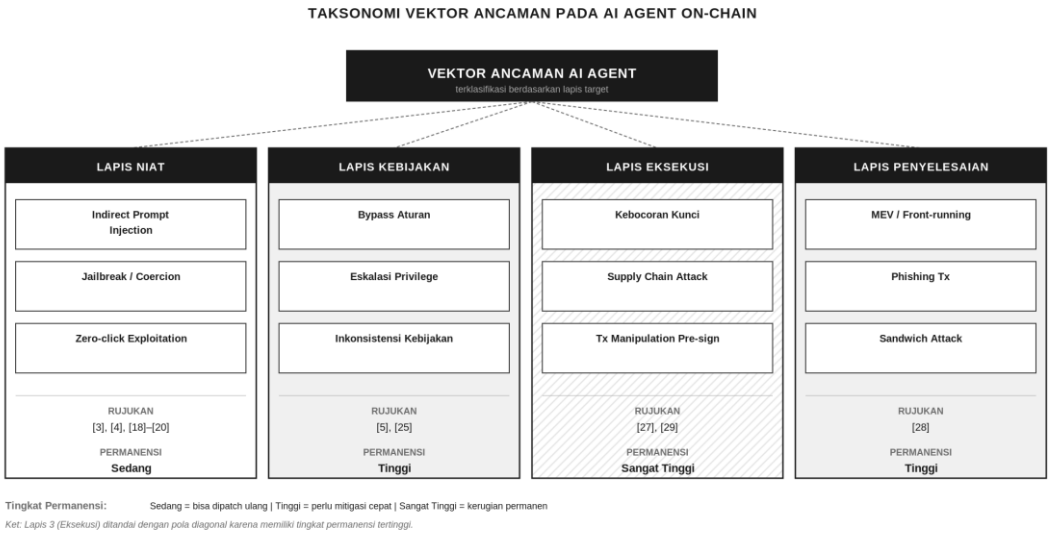
Gbr. 3. Kerangka empat-lapis arsitektur AI Agent on-chain.

4.2 Taksonomi Ancaman: Empat Vektor Eksploitasi

Taksonomi yang ditampilkan pada Gbr. 4 mengelompokkan vektor eksploitasi ke dalam empat kategori sesuai lapis yang dijadikan sasaran. Kategori pertama menyerang Lapis Niat dan mencakup indirect prompt injection [3], [4], [18]–[20], jailbreak [21]–[23], serta zero-click exploitation [24]. Kategori kedua menyerang Lapis Kebijakan: penyusupan kebijakan via prompt yang menuntun agen melewati batas, atau eksploitasi kontradiksi antar aturan yang membuka eskalasi privilege. Kategori ketiga menyerang Lapis Eksekusi: kebocoran kunci dari proses penandatanganan, manipulasi input transaksi sebelum disubmit, atau supply chain attacks pada library yang dipakai agen. Kategori keempat menyerang Lapis Penyelesaian: pemilihan blok yang menguntungkan penyerang (Maximal Extractable

Journal of Technology and Data Science (JTDS)
Vol. 3, No. 2, Desember 2025, hal. 216-228
p-ISSN: 2962-3693 | e-ISSN: 2962-7788

Value), front-running, dan phishing pada level transaksi sebagaimana didokumentasikan Wu dkk. [28]. Tabel II merangkum keempat kategori beserta rujukan empirisnya. Hal yang perlu disoroti adalah perbedaan tingkat permanensi kerugian antar kategori: serangan pada Lapis Niat dapat dipatch ulang ketika model di-update, sedangkan serangan pada Lapis Eksekusi yang berhasil menyebabkan transfer dana adalah permanen observasi inilah yang melatari penekanan pertahanan pada Lapis Eksekusi pada bagian berikutnya.



Gbr. 4. Taksonomi vektor ancaman pada AI Agent on-chain.

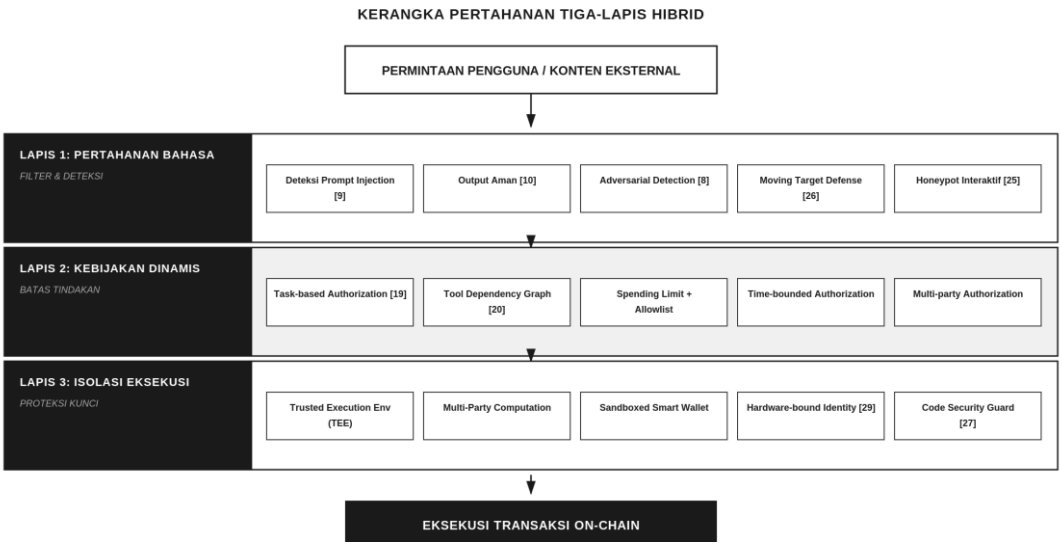
TABEL II
TAKSONOMI VEKTOR ANCAMAN PADA AI AGENT ON-CHAIN

Lapisan Target	Vektor Ancaman	Permukaan Eksploitasi	Rujukan Empiris	Tingkat Permanensi
Niat	Indirect prompt injection	Konten eksternal yang dibaca agen	[3], [4], [18]–[20]	Sedang
Niat	Jailbreak / coercion	Prompt manipulasi langsung	[21], [22], [23]	Sedang
Kebijakan	Bypass kebijakan	Kontradiksi aturan, eskalasi privilege	[5], [25]	Tinggi
Eksekusi	Kebocoran kunci / supply chain	Library agen, lingkungan tanda tangan	[27], [29]	Sangat Tinggi
Penyelesaian	MEV, front-running, phishing tx	Mempool publik, alamat penerima	[28]	Tinggi

4.3 Kerangka Pertahanan Tiga-Lapis Hibrid

Berlandaskan prinsip bahwa pertahanan tunggal tidak mencukupi untuk lansekap ancaman yang sudah dijabarkan, sintesis menghasilkan kerangka tiga-lapis hibrid sebagaimana ditampilkan pada Gbr. 5. Lapis pertama, Pertahanan Bahasa, merangkum mekanisme deteksi prompt injection [9], [10], penanganan output yang aman [10], moving target defense [26], dan honeypot interaktif [25]. Lapis kedua, Kebijakan Dinamis, mengaplikasikan otorisasi berbasis tugas, batasan pengeluaran berdurasi

terbatas, dan daftar putih alamat pendekatan ini selaras dengan prinsip Task Shield [19] dan IPIGuard [20]. Lapis ketiga, Isolasi Eksekusi, melindungi kunci privat agen melalui Trusted Execution Environment, Multi-Party Computation, atau Sandboxed Smart Wallet, secara konsep paralel dengan pengikatan fisik perangkat ke token dalam Arcenegui dkk. [29]. Ketiga lapis ini saling menutup celah: kompromi pada satu lapis tidak otomatis berarti kerugian finansial selama lapis berikutnya bertahan. Perlu dicatat bahwa kerangka ini bukan jaminan absolut penyerang yang berhasil menembus ketiga lapis tetap dapat menyebabkan kerugian permanen, dan tidak ada bukti empiris yang dipublikasikan sampai saat ini bahwa konfigurasi tiga lapis selalu lebih unggul dari konfigurasi dua lapis yang sangat ketat. Klaim efektivitas kerangka ini bersifat argumentatif berdasarkan konsistensi dengan pola pertahanan berlapis yang berhasil pada konteks kode [27], dan menanti verifikasi empiris di lapangan.



Gbr. 5. Kerangka pertahanan tiga-lapis hibrid untuk AI Agent on-chain.

4.4 Blockchain sebagai Trust Layer: Implikasi untuk Ekonomi Agen

Blockchain berfungsi sebagai trust layer yang memungkinkan agen otonom membangun reputasi dan saling memverifikasi tanpa perantara terpusat. Karandikar dkk. [30] memperlihatkan bagaimana token fungible dan non-fungible dapat dipakai sebagai primitif identitas dan akses pada infrastruktur energi komunitas; pola ini dapat diadaptasi untuk identitas agen otonom. Dua karya melengkapi pemandangan ini: studi pasar Ante [31] dan Wang dkk. [32] memperlihatkan bahwa volatilitas tinggi adalah kondisi normal pada DeFi sehingga agen wajib memiliki guardrails finansial yang ketat; sementara studi ekosistem Wilson dkk. [33] dan definisi sistematis Ritterbusch dan Teichmann [34] menggarisbawahi pentingnya interoperabilitas antar platform. Tiga implikasi praktis dapat ditarik untuk desain agen on-chain. Pertama, identitas agen harus tahan terhadap impersonasi, baik melalui tanda tangan kriptografis berbasis kunci yang terlindung enclave maupun mekanisme reputasi on-chain yang sulit dipalsukan. Kedua, riwayat transaksi harus dapat di-audit oleh pihak ketiga sehingga akuntabilitas tetap berlaku ketika sengketa muncul. Ketiga, kerangka kepatuhan harus dapat menempel pada protokol tanpa membutuhkan perubahan konsensus jaringan sebab perubahan konsensus berbiaya politis tinggi pada blockchain publik.

4.5 Keterbatasan Kajian

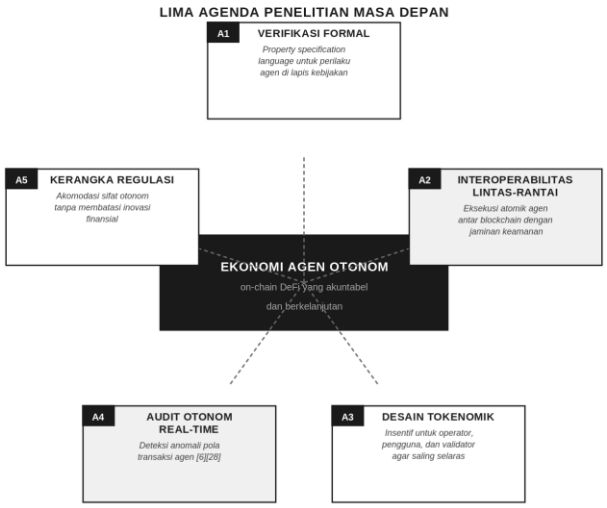
Empat keterbatasan kajian ini perlu disampaikan secara jujur. Pertama, sumber referensi terbatas pada basis data Zotero personal peneliti meski telah dikurasi selama beberapa tahun, basis data ini tidak mengklaim kelengkapan literatur internasional, terutama untuk publikasi yang sangat baru pada akhir

2025 yang kemungkinan belum sempat terindeks. Kedua, kajian bersifat konseptual sehingga klaim tentang efektivitas kerangka pertahanan tiga-lapis perlu diuji secara empiris pada implementasi nyata; klaim yang dibuat di sini bersifat argumentatif berdasarkan konsistensi dengan literatur tentang pertahanan berlapis. Ketiga, literatur tentang protokol baru seperti ERC mutakhir, x402, dan agentic wallet komersial belum tercakup dalam basis data akademik standar pada saat kajian ditulis sehingga rujukan teknisnya tidak dapat dimasukkan; praktisi yang membaca kajian ini perlu melengkapi dengan dokumentasi resmi protokol terkait saat akan diadopsi. Keempat keterbatasan yang paling fundamental penyelesaian permanen pada blockchain berarti satu kesalahan tanda tangan dapat menghapus efek puluhan pertahanan yang berhasil sebelumnya. Kerangka pertahanan tiga-lapis yang diusulkan menurunkan probabilitas kompromi total, namun tidak menghilangkannya.

5. KESIMPULAN

Sintesis terhadap 34 publikasi mutakhir yang berhasil dilakukan kajian ini memperlihatkan tiga simpulan saling berkait. Pertama, arsitektur AI Agent on-chain dapat dipetakan ke dalam empat lapis fungsional Niat, Kebijakan, Eksekusi, dan Penyelesaian di mana setiap lapis mengemban tanggung jawab spesifik dan menghadirkan permukaan serangan yang berbeda (RQ1). Kedua, lansekap ancaman terbagi ke dalam empat kategori yang masing-masing menyerang lapis tertentu: indirect prompt injection dan jailbreak menyerang lapis Niat, bypass kebijakan menyerang lapis Kebijakan, kebocoran kunci dan supply chain attack menyerang lapis Eksekusi, serta MEV dan front-running menyerang lapis Penyelesaian; mitigasi yang efektif menuntut kombinasi pertahanan bahasa, kebijakan dinamis, dan isolasi eksekusi yang saling melengkapi (RQ2). Ketiga, agenda riset yang terkonkretkan disusun dalam lima arah yang mencakup verifikasi formal, interoperabilitas, tokenomik, audit otonom, dan kerangka regulasi (RQ3).

Lima arah riset masa depan yang direkomendasikan diuraikan sebagai berikut. Pengembangan metode verifikasi formal khusus untuk perilaku agen otonom menjadi prioritas pertama, dengan property specification language yang dapat dipakai pada lapis Kebijakan. Arah kedua adalah mekanisme interoperabilitas lintas-rantai yang memberikan jaminan keamanan pada eksekusi atomik antar blockchain. Arah ketiga menyoroti desain insentif tokenomik yang menyelaraskan kepentingan operator agen, pengguna akhir, dan validator jaringan agar tidak ada pihak yang mendapatkan keuntungan dengan merugikan yang lain. Arah keempat membangun sistem deteksi ancaman otonom yang memanfaatkan teknik pada Liu dkk. [6] dan Wu dkk. [28] untuk mengaudit transaksi agen secara real-time. Arah kelima merumuskan kerangka regulasi yang mengakomodasi sifat otonom agen tanpa membatasi inovasi; isu ini menuntut kolaborasi antara pengembang, regulator, dan komunitas. Operasionalisasi kelima agenda hanya dapat berhasil jika dilakukan secara lintas-disiplin: tidak ada satu komunitas yang sendirinya cukup untuk membangun ekonomi agen on-chain yang akuntabel dan berkelanjutan.



Gbr. 6. Lima arah agenda penelitian masa depan untuk ekonomi agen otonom.

UCAPAN TERIMA KASIH

Penghargaan disampaikan kepada Universitas Dharmawacana Metro yang menyediakan dukungan akademik selama proses penyusunan kajian. Apresiasi juga diberikan kepada komunitas riset terbuka di bidang keamanan LLM dan blockchain publikasi ilmiah merekalah yang memungkinkan sintesis konseptual ini terlaksana.

KONTRIBUSI PENULIS

Ridwan Yusuf bertanggung jawab pada konseptualisasi kerangka kajian, perumusan pertanyaan riset, dan penulisan draf awal Bagian 1–2. Andreas Perdana menangani perancangan kerangka empat-lapis arsitektur, perumusan taksonomi ancaman dan kerangka pertahanan tiga-lapis, korespondensi naskah, dan revisi akhir. Febri Sugandi mengurus kurasi basis data referensi, ekstraksi metadata, validasi PDF terverifikasi, dan penyusunan matriks sintesis. Untoro Apsiswanto melakukan review konsistensi kajian pustaka, penyusunan agenda penelitian masa depan, dan penyempurnaan diskusi serta kesimpulan.

PERNYATAAN KONFLIK KEPENTINGAN

Penulis menyatakan tidak terdapat konflik kepentingan finansial maupun non-finansial yang dapat memengaruhi objektivitas kajian ini. Pemilihan referensi disandarkan pada kriteria akademik dan ketersediaan PDF di basis data Zotero, bukan pada afiliasi vendor atau kepentingan komersial pihak manapun.

REFERENSI

[1] F. Chen, "Multi-Agent LLM Systems: From Emergent Collaboration to Structured Collective Intelligence," 2025.

[2] V. Kulpa et al., "Prompt Engineering in Finance: An LLM-Based Multi-Agent Architecture for Decision Support," 2025.

[3] Q. Zhan, R. Fang, H. S. Panchal, dan D. Kang, "Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents," 2025.

[4] S. Johnson, V. Pham, dan T. Le, "The Dangers of Indirect Prompt Injection Attacks on LLM-based Autonomous Web Navigation Agents: A Demonstration," 2025.

[5] R. Shahroz, Z. Tan, S. Yun, C. Fleming, T. Chen et al., "Agents Under Siege: Breaking Pragmatic Multi-Agent LLM Systems with Optimized Prompt Attacks," 2025.

[6] Z. Liu et al., "Combining Graph Neural Networks With Expert Knowledge for Smart Contract Vulnerability Detection," IEEE Transactions on Knowledge and Data Engineering, 2023, doi: 10.1109/TKDE.2021.3095196.

[7] S. Mane, J. Bhortake, V. Wankhade, dan F. Kazi, "Automated IoT Firmware Vulnerability Detection Using Large Language Models," Preprints, 2025, doi: 10.20944/preprints202510.0166.v1.

[8] S. Fathima et al., "LLM Adversarial Prompt Attack Detection and Mitigation Engine: A Novel Framework for Securing Generative AI Systems," 2025.

[9] Q. Lan, A. Kaul, dan S. Jones, "Prompt Injection Detection in LLM Integrated Applications," 2025.

[10] D. Naik, N. Naik, dan I. Naik, "Insecure Output Handling in Large Language Models (LLMs) and Approaches to Enhance Output Security, Including Prevention of LLM-Based Attacks," 2025.

[11] S. Choudhary, D. Anshumaan, N. Palumbo, dan S. Jha, "How Not to Detect Prompt Injections with an LLM," 2025.

[12] A. Ghafarollahi dan M. J. Buehler, "Rapid and automated alloy design with graph neural network-powered large language model-driven multi-agent AI," 2025.

[13] X. Lan, J. Feng, Y. Liu, X. Shi, dan Y. Li, "AutoQual: An LLM Agent for Automated Discovery of Interpretable Features for Review Quality Assessment," 2025.

[14] Y. Teplyi dan D. Dosyn, "Prompt-guided LLM agent for end-to-end ontology learning," 2025.

[15] S. Kakkar, "Prompt-Robotics: Leveraging Large Language Model (LLM) Agents to Enhance Robotic Process Automation (RPA) Supervision and Efficiency," 2025.

[16] A. Hazime et al., "Evaluation of LLM Powered Agentic AI for Solving Multi-Arm Bandit Problems," 2025.

[17] A. Mukherjee, "Use of Agentic AI with LLM and Prompt Engineering and State-of-the-Art Machine Learning Algorithm to Detect Patterns in IoT Devices," 2025.

[18] Z. Wang et al., "AGENTVIGIL: Automatic Black-Box Red-teaming for Indirect Prompt Injection against LLM Agents," 2025.

[19] F. Jia, T. Wu, X. Qin, dan A. Squicciarini, "The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents," 2025.

[20] H. An, J. Zhang, T. Du, C. Zhou, dan Q. Li, "IPIGuard: A Novel Tool Dependency Graph-Based Defense Against Indirect Prompt Injection in LLM Agents," 2025.

[21] J. Hu, J. Zhang, dan Z. Li, "Breaking Barriers: Multi-Agent Prompt Fusion for Automated LLM Jailbreaks," 2025.

[22] H. Wang et al., "DiffusionAttacker: Diffusion-Driven Prompt Manipulation for LLM Jailbreak," 2025.

[23] G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, dan Z. Liu, "MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots," 2024.

[24] P. Reddy dan A. S. Gujral, "EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System," 2025.

[25] G. Yang, Z. Sun, dan Y. Wang, "ShellBox: Adversarially Enhanced LLM-Interactive Honeypot Framework," 2025.

[26] S. Panterino dan M. Fellington, "Dynamic Moving Target Defense for Mitigating Targeted LLM Prompt Injection," 2024.

[27] A. Kavian, M. M. P. Kallehbasti, S. Kazemi, dan F. Khansari, "LLM Security Guard for Code," 2024.

[28] J. Wu et al., "Who Are the Phishers? Phishing Scam Detection on Ethereum via Network Embedding," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, doi: 10.1109/TSMC.2020.3016821.

[29] J. Arcenegui, R. Arjona, R. Roman, dan I. Baturone, "Secure Combination of IoT and Blockchain by Physically Binding IoT Devices to Smart Non-Fungible Tokens Using PUFs," Sensors, vol. 21, no. 9, hal. 3119, 2021, doi: 10.3390/s21093119.

[30]N. Karandikar, A. Chakravorty, dan C. Rong, "Blockchain Based Transaction System with Fungible and Non-Fungible Tokens for a Community-Based Energy Infrastructure," Sensors, vol. 21, no. 11, hal. 3822, 2021, doi: 10.3390/s21113822.

[31]L. Ante, "The Non-Fungible Token (NFT) Market and Its Relationship with Bitcoin and Ethereum," FinTech, 2022, doi: 10.3390/fintech1030017.

[32]Y. Wang, B. M. Lucey, S. A. Vigne, dan L. Yarovaya, "Bubbles all the way down? Detecting and date-stamping bubble behaviours in NFT and DeFi markets," Journal of Chinese Economic and Business Studies, 2022, doi: 10.1080/14765284.2022.2138161.

[33]K. B. Wilson, A. Karg, dan H. Ghaderi, "Prospecting Non-Fungible Tokens in the Digital Economy: Stakeholders and Ecosystem, Risk and Opportunity," Business Horizons, 2022, doi: 10.1016/j.bushor.2021.10.007.

[34]G. D. Ritterbusch dan M. R. Teichmann, "Defining the Metaverse: A Systematic Literature Review," IEEE Access, 2023, doi: 10.1109/ACCESS.2023.3241809.