Biosecure-LLM Framework: Protecting LLMs from Cyberbiosecurity Threats and the Case for Independent AI Safety Governance

Xavier-Lewis Palmer[1], Lucas Potter[1], Srdjan Lesaja[2], Sotirios Karathanasis[3] and Mohammad Ghasemigol[4]

[1]Biosview Labs, Dayton, Ohio, USA
[2]Virginia Commonwealth University, VA, USA
[3]Independent Researcher, New York, NY, USA
[4]Old Dominion University, Norfolk, VA, USA

XP.BV@pm.me
Biosview1@proton.me
srdjanlesaja@gmail.com
sfk@use.startmail.com
mghasemi@odu.edu

Abstract: Large Language Models (LLMs) are becoming critical infrastructure in scientific, healthcare, and governmental contexts. As frontier AI laboratories increasingly partner with government agencies, a fundamental question arises: Who should control the safety and policy-enforcement layers that constrain model behavior? Current safety mechanisms (LLM guardrails) are typically designed for generic "harmlessness" and operate by detecting semantic patterns and refusing requests. However, they are inadequate governance instruments because they cannot implement auditable, domain-specific controls tied to external regulatory policy objects (e.g., control lists or rules governing personally identifying information). Even a perfectly aligned model is not able to express institution-specific policy without an external control layer. This paper argues that the logical separability of policy enforcement from model inference, demonstrated by firewall-style architectures, demands corresponding institutional separability as well. Concentrating both model development and safety governance within the same commercial entities creates unacceptable conflicts of interest, regulatory capture risks, and accountability gaps. We propose that the policy control layers must be housed within independent regulatory bodies, governmental agencies, or trusted third parties rather than the organizations that build and profit from the underlying models. Drawing on the Biosecure-LLM framework as a technical proof-of-concept, we demonstrate that such separation is architecturally feasible and argue it is well-suited for verifiable compliance.

Keywords: AI governance, Institutional design, Regulatory independence, Biosecurity, Responsible AI, Policy enforcement, separability

1. Introduction: The Governance Gap in Frontier AI

Large language models have swiftly moved from research curiosity to embedded infrastructure. In biomedicine, they can assist with literature synthesis, hypothesis generation, and protocol drafting; in clinical settings, they summarize encounters, enable review of clinical histories, and draft documentation for human review (Akogo et al, 2025; Shah, Entwistle, and Pfeffer, 2023; Gao et al, 2024; Floridi et al, 2025). Their integration with electronic knowledge bases and laboratory systems means their behavior now influences scientific and clinical decision-making in ways that impact critical digital infrastructure. Simultaneously, frontier AI laboratories are courting governmental partnerships at an unprecedented pace. These organizations seek contracts that will grant them access to classified information, sensitive health records, and national security data. The commercial incentives are clear: government contracts represent stable revenue and legitimacy. But this interest raises a profound question that existing AI ethics frameworks have overlooked: When the same entity both builds a powerful AI system and controls the mechanisms meant to constrain it, can we trust that there will not be a conflict of interest? This paper argues that the answer is no, not because AI developers are uniquely untrustworthy, but because of institutional design. This concern is compounded when considering the wider bioeconomy: convergence of digital and biological risks, captured by the field of cyberbiosecurity, has been recognized for years (Murch et al., 2018; Potter and Palmer, 2023). LLMs introduce a distinct challenge: because they interface through natural language, they are susceptible to prompt-level manipulation that can subvert safety behaviors. "Prompt injection" and related techniques can induce models to ignore safety instructions, disclose hidden system prompts, or generate content that deviates from institutional policy (Hui et al., 2024). Adversarial prompting and novel techniques in prompt injection can outpace safety considerations and can be difficult to anticipate. Existing alignment techniques, including reinforcement learning from human feedback and instruction tuning, reduce many unsafe outputs but do not necessarily implement enforceable, domain-specific controls mapped to biological regulation or screening (Christiano et

al, 2017; Ouyang et al., 2022; Huang et al, 2025). This creates a policy gap. Traditional cybersecurity focuses on networks and endpoints; traditional biosecurity focuses on physical containment and material screening. The LLM-mediated interface sits between these domains. If left to ad hoc prompt engineering and generic "harmlessness" tuning, institutions risk either over-blocking legitimate discourse or under-blocking content that lowers barriers to harm.

1.1 LLM Guardrail Insufficiency: Current Limitations in Standard LLM Safety Layers

The Biosecure-LLM framework, described below, shows that safety controls can be architecturally separated from model weights. This paper's central claim is that technical separability must be matched by institutional separability: the organizations that enforce AI safety policy must be distinct from those that develop and commercialize the models themselves.

Table 1: The Governance-Safety Mismatch: Why LLM Guardrails are Not Sufficient for Enterprise Compliance: This table maps institutional compliance requirements (governance) to the function of current LLM guardrails (safety). The analysis formalizes the "Structural Gap" as a failure of system integration, entity awareness, and auditability. The inclusion of the "Proposed Architectural Solution" column transitions the argument to a strategic mandate, confirming the necessity of an external, policy-as-code control plane (a Policy Engine) to bridge the gap between general safety alignment and enforceable, auditable enterprise governance (Bai et al, 2022; Mökander et al., 2023; Herrera-Poyatos et al., 2025; Mandalawi et al., 2025; Narula et al., 2025)

Biosecurity/ Compliance Requirement	What Institutions Need (Governance Objective)	What Current LLM Guardrails Do (Safety Objective)	Structural Gap (Why Guardrails Fail Governance)	Proposed Architectural Solution
Regulated-Entity Awareness	Explicit mapping to controlled agents, materials, and techniques (via external, dynamic control lists).	Keyword or latent safety triggers not tied to control lists (detects *intent* to harm, not *policy violation*).	No enforceable policy alignment or entity recognition. Decisions are non-binding.	External Policy Engine with RAG over Control Lists.
Operational Specificity Control	Suppress stepwise, parameterized, optimization content for controlled techniques.	Allow high-level instructions if framed as "educational" or "hypothetical."	Actionability/Plausibility not measured; focuses on high-level *topic* refusal, not *methodology* refusal.	Policy Engine that analyzes response structure (e.g., presence of required parameters, sequence).
PII/PHI/CUI Handling	Dynamic masking, sanitization, or refusal based on data classification and user role (e.g., *only* HR can see salary PII).	Static detection of sensitive *topics* (e.g., "don't share personal medical advice").	LLM lacks integration with existing Role-Based Access Control (RBAC) and Data Loss Prevention (DLP) systems.	Policy Engine integrated via API to RBAC and DLP systems for real-time policy checks.
Auditability	Reproducible, explainable safety decisions (policy rule refusal).	Non-deterministic refusals; black-box model-generated rationales.	No compliance evidence; decisions cannot be traced to a specific, auditable policy rule or external entity.	Policy Engine that logs the specific rule (Policy-as-Code) that triggered the refusal for full compliance trace.

2. The Problem: Concentrated Control and Conflicts of Interest

2.1 The Current Model: Providers as Judge and Jury

Currently, frontier AI laboratories simultaneously develop models, define safety policies, implement enforcement mechanisms, and evaluate their own compliance. This concentration would be remarkable in any other industry. We do not permit pharmaceutical companies to approve their own drugs, nor allow banks to audit their own capital reserves. Yet we accept that AI companies can control every layer of the safety stack. As industries (healthcare, business, transport, etc.) adopt these models, wrapping them in applications for

internal or client-side use, this risk propagates alarmingly fast through the entire critical infrastructure network. The conflict of interest is structural. Safety measures impose costs: they require computational resources, introduce latency, and, most importantly, may restrict model capabilities that drive commercial value. However, a company that controls both the model and its constraints faces constant pressure to relax those constraints when they interfere with user satisfaction or competitive positioning. Even well-intentioned developers have incentives to reward capability expansion and punish excessive caution. This problem intensifies as AI laboratories seek government contracts. Access to sensitive government data creates obligations that conflict with transparency and accountability. When a company trains models on classified information, external oversight becomes nearly impossible due to security constraints. As a consequence, the very partnerships that grant legitimacy to AI developers will shield them from scrutiny.

2.2 Taxonomy of Threats Requiring Independent Oversight

Two families of attacks dominate the risk landscape for LLM deployments: prompt injection and policy evasion. Prompt injection encompasses direct attempts to coerce models into ignoring safety instructions, as well as indirect attacks where adversarial instructions are embedded in retrieved documents that are later fused into the model's context by retrieval-augmented generation pipelines. The latter is particularly insidious because attackers can poison content upstream beyond the LLM's perimeter (De Stefano et al., 2024). Prompt-leaking attacks, another form of prompt injection, aim to extract system prompts and safety rationales; once disclosed, those artifacts enable targeted bypasses (Hui et al., 2024; Asl et al., 2025). Other attacks involve hiding adversarial commands in policy-like format such as JSON or XML, or utilizing fictionalized scenarios to bypass an LLM’s safety content guidelines (Das, 2024; Zhang et al, 2025). Policy evasion involves models producing fluent, operationally specific content (including procedural details, optimization advice, or procurement vectors) that could lower barriers to harmful activity. Even benign queries can elicit fabricated steps that become actionable. This behavior traces to overgeneralization from training data and pressure to provide answers (Farquhar et al., 2024). Findings of AI use being device- and context-dependent by Costa-Gomes et al. (2025) combined with insights of the pervasiveness of AI in wider swathes of community life by Narula et al. (2025) add to an overall picture of a massive attack surface being made available by accidental or malicious actors. These technical vulnerabilities cannot be adequately addressed by the same organizations that profit from model creation and deployment. Detection of novel attack patterns, maintenance of control lists aligned with international regulations, and enforcement decisions that may reduce model utility all require independence from commercial pressures.

2.3 Limitations of Provider-Controlled Safety

Three gaps persist when safety remains provider-controlled. First, a specification gap: generic harmlessness is not an adequate screening threshold for LLMs using complex, critical datasets. A model can be well-aligned in aggregate yet lack enforceable mappings between regulated entities and interventions (Adam et al., 2011; Wheeler et al., 2024). Second, an adversarial gap. Alignment is typically performed on static datasets, while attackers iterate rapidly. Closing this gap requires layered, updateable controls that can be revised as attack patterns evolve, revisions that may not align with provider timelines or commercial interests. Third, an assurance gap. Regulators and institutional risk officers need evidence: tamper-evident logs, provenance records, and reproducible decisions, for example. Conventional alignment pipelines do not provide such artifacts. In adjacent domains, verifiable logging infrastructures demonstrate that append-only, publicly auditable records can raise assurance without exposing sensitive content (Laurie, 2014). An analogous approach for LLM safety requires organizational independence to be credible.

3. Technical Proof-of-Concept: The Biosecure-LLM framework

To demonstrate that institutional separation is technically feasible, we describe Biosecure-LLM, a firewall framework that enforces policy at the boundaries of model communication rather than inside model weights. This architecture proves that safety enforcement can be logically, and therefore institutionally, decoupled from model development.

3.1 Architectural Overview

Biosecure-LLM acts as a security guard for the language model, managing all incoming requests and outgoing answers. When a request comes in from a client application, it first passes through an entrance point, which sends it to the Input Inspection service (Layer 1). This Layer 1 checks the request to make sure it's safe and allowed. If it passes the check, the request is then sent to the language model. The model's answer is then sent to the Output Sanitization service (Layer 2) for cleanup and safety checks before being returned to the client.

This interposition offers three advantages. First, it decouples safety enforcement from model provenance; institutions can upgrade models without rewriting policy. Second, it localizes assurance: the control plane becomes the single locus for audit logging and regulator-facing evidence. Third, it enables comparative assessment: controls act orthogonally to the model, so the same harness can quantify risk reduction across model families (Li et al., 2024; Mazeika et al., 2023; Rein et al., 2023).

3.2 Layer 1: The Prompt Gatekeeper

Layer 1 reduces exposure by vetting user intent and retrieved context before inference through two mechanisms. First, semantic anomaly detection estimates whether an input belongs to a distribution of permitted requests, flagging patterns associated with jailbreaks, instruction overriding, or indirect injection markers (Hendrycks and Gimpel, 2017; Xu et al., 2021). Second, domain-aware recognition identifies mentions of organisms, agents, materials, and techniques, mapping them onto structured policy objects that mirror international control-list categories (Lee et al., 2020; Neumann et al., 2019; Adam et al., 2011; Kobokovich et al., 2019; Wheeler et al., 2024). Entity recognition alone does not trigger enforcement; it conditions a ruleset distinguishing scholarship from attempts to elicit operationally specific procedures. For retrieval-augmented generation, Layer 1 screens not only user strings but retrieved context, rejecting sources that fail provenance requirements, stripping instruction-like segments, and neutralizing patterns known to trigger prompt-leaking (De Stefano et al., 2024; Hui et al., 2024). Another augmentation to Layer 1 involves employing an ensemble model approach, parsing the request through an LLM trained explicitly to recognize and query adversarial requests. Such an adversarial-forward LLM can continue to adapt to emerging prompting attacks while acting as a modular asset that can be paired with different LLMs. Additionally, it can intercept and ask the requester for further information, credentials, and intent, which can help it balance its own weights as it continues to encounter potential attackers.

3.3 Layer 2: The Response Censor

Layer 2 inspects model outputs for their risk of increasing capacity to perform restricted biological activities or are operationally specific (stepwise procedures, parameterized conditions, procurement vectors). This prevents actionable harm, mirroring dual-use governance in nucleic-acid synthesis screening (Adam et al., 2011; Kobokovich et al., 2019; Wheeler et al., 2024). The sanitization pipeline applies three passes: (1) content triage identifying regulated entities; (2) capability inference scanning for procedural markers and optimization advice; and (3) exploitability scoring aggregating operational proximity to controlled entities. If scores exceed certain thresholds, Layer 2 transforms responses by replacing disallowed spans with safe alternatives while preserving legitimate communication (Farquhar et al., 2024). For agentic systems, Layer 2 enforces strict separation between narrative text and executable actions, preventing structured commands from executing without human review (Boiko et al., 2023; Bran et al., 2024).

3.4 Key Technical Properties

The framework emphasizes modularity, low latency, and continuous alignment with evolving policy. Layers are packaged as independent, stateless services that can scale. Policy updates are versioned as digitally-signed, authenticated objects with provenance metadata, enabling rapid rollback and verification (Laurie, 2014; Laurie, 2021). Critically, the framework is model-agnostic. It treats the base LLM as a modular component, regardless of weight visibility or hosting architecture, so long as all inputs and outputs route through the control plane. This preserves diverse deployment choices while maintaining invariant guardrails where they matter which is at the boundary where text could inform action.

4. The Argument for Institutional Separation

4.1 From Logical to Institutional Separability

The Biosecure-LLM framework demonstrates that policy enforcement can be separated from model inference. But separability is not just a technical convenience, but it is an institutional responsibility. Consider the analogy to financial regulation. Markets perform complex transactions that regulators cannot replicate in real-time. Yet we do not conclude that trading firms should therefore regulate themselves. In fact, we have established independent bodies with authority to set standards, conduct audits, and impose penalties. The computational sophistication of the economy does not obviate the need for external oversight. It actually makes such oversight more important. The same logic applies to AI safety. The technical complexity of LLMs is not an argument for provider self-regulation, instead it strongly supports building independent institutions capable of

oversight. The Biosecure-LLM framework shows that policy enforcement can be implemented independent of, and coherently with, model development. Thus, this technical possibility could become an institutional reality.

4.2 Why Model Providers Cannot Own the Policy Layer

There are several discouraging provider-controlled safety regulations. First, accountability requires independence. When incidents occur, investigations must be conducted by parties without a financial stake in the outcome. Provider-controlled safety creates situations where the investigated party itself controls the evidence, sets the standards against which they are measured, and determines what constitutes compliance.

Second, regulatory capture becomes inevitable when regulated entities control enforcement. AI laboratories possess resources, technical expertise, and political access that would dwarf any early regulatory bodies. If providers control the policy layer, they will shape it to accommodate their commercial interests, regardless of stated intentions. Third, unmanaged government partnerships introduce too many conflicts of interest. Frontier laboratories are actively seeking contracts that will grant access to sensitive data. Models trained on classified information, health records, or national security data require oversight that cannot be delegated to commercial entities with competing obligations to shareholders, employees, and customers. Such models must have clear metadata management and control over which users are authenticated to receive responses from models weighted with and without critical or private information. Fourth, public trust requires transparency. Citizens and institutions must be able to verify that safety enforcement serves public interests rather than commercial ones. This verification is less possible when the same organization controls both the model and its constraints.

4.3 Who Should Control the Policy Layer?

Independent oversight could take several forms. Dedicated regulatory agencies with technical capacity and statutory authority represent one model, analogous to the Food and Drug Administration or the Securities Exchange Commission. Alternatively, existing agencies could be augmented with AI safety divisions (an approach adopted in biosecurity) where existing frameworks provide foundations for AI-specific enforcement. Third-party certification bodies offer another model, drawing on precedents in cybersecurity auditing and safety certification. Such bodies could be industry-funded but governance-independent, with statutory backing for their authority. In a similar vein, watchdog-style groups may organically form to publicize breaches in trust.

The above outcomes may emerge at different times or co-exist simultaneously. The form matters less than the principle: the organization that enforces AI safety policy must be structurally independent of organizations that profit from AI deployment. This independence must be visible and verifiable with governance that both prevents capture and enables correction.

5. Governance and Auditability Under Independent Oversight

Independent oversight requires infrastructure to support accountability. The Biosecure-LLM framework couples technical controls with governance mechanisms that demonstrate how such an infrastructure could function.

5.1 Incident Response and Adjudication

When either layer issues an intervention, the system records an immutable event containing a hashed representation of the prompts and context, model and policy versions in force, rules that fired, and outputs that were transformed or suppressed. This snapshot commits to an append-only log so that subsequent policy changes cannot erase the presence and rationale of an executed intervention. Following transparency systems in web public-key infrastructure, inclusion and consistency of records are provable using relatively tamper-proof chains called Merkle-trees: each batch extends the tree, and periodic publication of signed tree heads allows independent verification that past entries have not been altered (Laurie, 2014; Laurie, 2021). The log stores protected elements and structured references rather than plaintext, protecting privacy while enabling audit. Under independent oversight, a designated safety officer receives events for human-in-the-loop adjudication, determining whether interventions were appropriate and diagnosing whether there was attempted misuse, accidental over-specification, or false positives requiring calibration. This separation of duties (automatic detection followed by expert review) mirrors safety controls in critical systems.

5.2 Verifiable Compliance

The logs support third-party verification without revealing sensitive content. Each event stores cryptographic commitments to content digests and policy artifacts. External reviewers can verify that decisions cited specific

policy objects with known versions without accessing underlying text. This approach borrows from secure audit logs and key-transparency systems (Schneier and Kelsey, 1999; Laurie, 2014; Laurie, 2021). Regulators can require periodic submission of signed tree heads and random-audit inclusion proofs as conditions of continued operation, without receiving content triggering confidentiality concerns. Because the control plane externalizes policy, institutions can demonstrate that identical requests would elicit identical interventions across models, a property attractive to compliance officers responsible for multiple portfolios of models.

5.3 Continuous Assessment

Assessment completes the governance loop. The testing framework automatically runs two groups of checks whenever system rules are modified long with routine schedule. The first targets misuse: jailbreak prompts, prompt-leaking patterns, and indirect-injection scenarios exercise suppression of unsafe completions (Hui et al., 2024; De Stefano et al., 2024). Stress suites like Weapons of Mass Destruction Proxy (WMDP) and HarmBench provide baselines for processing at rates that would be questionable by a non-malicious user (Li et al., 2024; Mazeika et al., 2024). The second family targets utility: domain competence on technical questions and performance on benign workloads establish whether scientific capability is preserved and false-positive rates remain acceptable (Rein et al., 2023). Tests run with and without interposition, allowing clean estimates of marginal risk reduction and utility cost. Thus, independent oversight bodies could control assessment standards, conduct periodic assessments, and publish aggregate results.

6. Conclusion and Recommendations

The convergence of powerful AI systems, sensitive biological and governmental data, and nascent regulatory frameworks creates an urgent governance challenge. Frontier AI laboratories are positioning themselves as trusted partners to governments worldwide, seeking access to data that will make their models more capable and their products more valuable. The commercial reasoning is compelling; however, the governance implications are troubling especially as data access grows (Lesaja and Palmer, 2019). This paper has argued that the policy enforcement layer, the mechanisms that constrain AI behavior and ensure compliance with safety standards, must be institutionally separated from model providers. Our goal is not to impede or otherwise handicap AI frontier labs, as their work is valuable. Rather, technical architectures like Biosecure-LLM demonstrate that such separation is feasible and provides benefits for both model creators and data curators. Policy can be enforced at input-output boundaries without retraining models or accessing their weights. The logical separability of enforcement from inference proves that institutional separability is achievable. Concentrated control creates unacceptable risks: conflicts of interest that bias safety decisions toward commercial considerations, accountability gaps that prevent effective investigation of incidents, regulatory capture that subordinates public safety to private profit, and opacity that undermines public trust will erode the efficacy and usability of models. Our recommendations are direct. First, regulatory bodies should assert AI safety policy rather than delegating to providers. The Biosecure-LLM framework provides a template. Second, governmental contracts with AI laboratories should require that safety enforcement be conducted by independent parties, with access to audit logs. Third, industry standards should evolve toward externalized policy enforcement with verifiable compliance mechanisms. The bioeconomy provides instructive precedent. Gene-synthesis screening, control lists for regulated agents, and biosafety review boards all embody the principle that safety oversight must remain independent of commercial interests (Adam et al., 2011; Kobokovich et al., 2019; Wheeler et al., 2024). These frameworks emerged because the alternative (trusting biotechnology companies alone to regulate themselves) poses significant risk of harm. We may expect more regulation as more sensitive biological data, such as radiological and neurophysiological recordings, become easier to store (Lesaja and Palmer, 2020; Liv and Greenbaum, 2023). Work from d'Ascoli et al, (2025) shows that we are increasingly closer to this capability. The same recognition must occur for AI, borrowing from biosafety insights from Gillum et al (2025).

Ethics Declaration: Ethical clearance for the research referred to in this paper was not needed.

AI Declaration: AI tools within Grammarly and Google Docs were used for formatting, grammar/spell-checking, and re-organization.

References

Adam, L., Kozar, M., Letort, G., Mirat, O., Srivastava, A., Stewart, T., Wilson, M.L. and Peccoud, J., 2011. Strengths and limitations of the federal guidance on synthetic DNA. Nature biotechnology, 29(3), pp.208-210. doi: 10.1038/nbt.1802

Akogo, D., Ayensu, J., Sam, N., Hattoh, G., Nyarko, P., Eshun, S., Alhasan, M., Mensah-Brown, H. and Quashie, P., 2025. Moremi Bio Agent: Application of A Foundation Model and End-to-End Automation in the Design and Validation of Monoclonal Antibodies Targeting Plasmodium falciparum Invasion Complex. bioRxiv, pp.2025-02. doi:10.1101/2025.02.12.637967

Asl, J.R., Narula, S., Ghasemigol, M., Blanco, E. and Takabi, D., 2025, November. NEXUS: Network Exploration for eXploiting Unsafe Sequences in Multi-Turn LLM Jailbreaks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 24278-24306). doi:10.18653/v1/2025.emnlp-main.1235

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C. and Chen, C., 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Boiko, D.A. et al. (2023) 'Autonomous chemical research with large language models', Nature, 624, pp. 570–577. doi:10.1038/s41586-023-06792-0.

Bran, A.M., Cox, S., Schilter, O., Baldassari, C., White, A.D. and Schwaller, P., 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, *6*(5), pp.525-535. doi: 10.1038/s42256-024-00832-8

Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D., 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

Costa-Gomes, B., Chen, S., Hsueh, C., Morgan, D., Schoenegger, P., Shah, Y., Way, S., Zhu, Y., Adeline, T., Bhaskar, M. and Suleyman, M., 2025. It's About Time: The Temporal and Modal Dynamics of Copilot Usage. arXiv preprint arXiv:2512.11879.

d'Ascoli, S., Bel, C., Rapin, J., Banville, H., Benchetrit, Y., Pallier, C. and King, J.R., 2025. Towards decoding individual words from non-invasive brain recordings. Nature Communications, 16(1), p.10521. doi: 10.1038/s41467-025-65499-0

Das, N., Raff, E. and Gaur, M., 2024. Human-Interpretable Adversarial Prompt Attack on Large Language Models with Situational Context. arXiv preprint arXiv:2407.14644.

De Stefano, G., Schönherr, L. and Pellegrino, G., 2024. Rag and roll: An end-to-end evaluation of indirect prompt manipulations in llm-based application frameworks. *arXiv preprint arXiv:2408.05025*.

Farquhar, S., Kossen, J., Kuhn, L. and Gal, Y., 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), pp.625-630. doi: 10.1038/s41586-024-07421-0

Floridi, L., Morley, J., Novelli, C. and Watson, D., 2025. What Kind of Reasoning (if any) is an LLM actually doing? On the Stochastic Nature and Abductive Appearance of Large Language Models. arXiv preprint arXiv:2512.10080.

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J.R., Ektefaie, Y., Kondic, J. and Zitnik, M., 2024. Empowering biomedical discovery with AI agents. Cell, 187(22), pp.6125-6151. doi: 10.1016/j.cell.2024.09.022

Gillum, D.R., Knight, C. and Vogel, K.M., 2025. Understanding biosafety practitioner perspectives. Politics and the Life Sciences, pp.1-24. doi: 10.1017/pls.2025.10014

Hendrycks, D. and Gimpel, K. (2017) 'A baseline for detecting misclassified and out-of-distribution examples in neural networks', arXiv preprint, arXiv:1610.02136. doi:10.48550/arXiv.1610.02136.

Herrera-Poyatos, A., Del Ser, J., de Prado, M.L., Wang, F.Y., Herrera-Viedma, E. and Herrera, F., 2025. Responsible Artificial Intelligence Systems: A Roadmap to Society's Trust through Trustworthy AI, Auditability, Accountability, and Governance. arXiv preprint arXiv:2503.04739.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T., 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), pp.1-55. doi: 10.1145/3703155

Hui, B., Yuan, H., Gong, N., Burlina, P. and Cao, Y. (2024) 'PLeak: Prompt leaking attacks against large language model applications', in Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 390–404. doi:10.1145/3658644.3670370.

Kobokovich, A., West, R., Montague, M., Inglesby, T. and Gronvall, G.K. (2019) 'Strengthening security for gene synthesis: Recommendations for governance', Health Security, 17(6), pp. 419–429. doi:10.1089/hs.2019.0110.

Laurie, B., 2014. Certificate transparency. Communications of the ACM, 57(10), pp.40-46. doi: 10.1145/2659897

Laurie, B., Langley, A., Kasper, E., Messeri, E. and Stradling, R., 2021. Certificate transparency version 2.0. *Internet Requests for Comments, RFC Editor, RFC*, *9162*.

Lee, J., Yoon, W. and Kang, J. (2020) 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', Bioinformatics, 36(4), pp. 1234–1240. doi:10.1093/bioinformatics/btz682.

Lesaja, S. and Palmer, X.L., 2020. Brain-computer interfaces and the dangers of neurocapitalism. arXiv preprint arXiv:2009.07951.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J.D., Dombrowski, A.K., Goel, S., Phan, L. and Mukobi, G., 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Liv, N. and Greenbaum, D., 2023. Cyberneurosecurity. In *Policy, identity, and neurotechnology: The neuroethics of brain-computer interfaces* (pp. 233-251). Cham: Springer International Publishing. doi: 10.1007/978-3-031-26801-4_13

Mandalawi, S.A., Mohammed, M.A., Maclean, H., Cakmak, M.C. and Talburt, J.R., 2025. Policy-Aware Generative AI for Safe, Auditable Data Access Governance. arXiv preprint arXiv:2510.23474.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B. and Forsyth, D., 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249.

Mökander, J., 2023. Auditing of AI: Legal, ethical and technical approaches. Digital Society, 2(3), p.49. doi: 10.1007/s44206-023-00074-y

Murch, R.S., So, W.K., Buchholz, W.G., Raman, S. and Peccoud, J. (2018) 'Cyberbiosecurity: An emerging new discipline to help safeguard the bioeconomy', Frontiers in Bioengineering and Biotechnology, 6, p. 39. doi:10.3389/fbioe.2018.00039.

Narula, S., Ghasemigol, M., Carnerero-Cano, J., Minnich, A., Lupu, E. and Takabi, D., 2025. Exploring AI Security: A Systematic Mapping Study. IEEE Access. doi: 10.1109/ACCESS.2025.3567195.

Neumann, M., King, D., Beltagy, I. and Ammar, W. (2019) 'ScispaCy: Fast and robust models for biomedical natural language processing', in Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327. doi:10.18653/v1/W19-5034.

Ouyang, L. et al. (2022) 'Training language models to follow instructions with human feedback', arXiv preprint, arXiv:2203.02155. doi:10.48550/arXiv.2203.02155.

Potter, L. and Palmer, X.L., 2023. Mission-aware differences in cyberbiosecurity and biocybersecurity policies: Prevention, detection, and elimination. In Cyberbiosecurity: A new field to deal with emerging threats (pp. 37-69). Cham: Springer International Publishing. doi: 10.1007/978-3-031-26034-6_4

Rein, D., Aurelio, T., Tamkin, A., Ganguli, D. and Steinhardt, J. (2023) 'GPQA: A graduate-level, Google-proof Q&A benchmark', arXiv preprint, arXiv:2311.12022. doi:10.48550/arXiv.2311.12022.

Shah, N.H., Entwistle, D. and Pfeffer, M.A. (2023) 'Creation and adoption of large language models in medicine', JAMA, 330(9), pp. 866–869. doi:10.1001/jama.2023.14217.

Schneier, B. and Kelsey, J., 1999. Secure audit logs to support computer forensics. *ACM Transactions on Information and System Security (TISSEC)*, *2*(2), pp.159-176. doi: 10.1145/317087.317089

Wheeler, N.E. et al. (2024) 'Developing a common global baseline for nucleic acid screening', Applied Biosecurity. doi:10.1089/apb.2023.0034.

Xu, J., Wu, H., Wang, J. and Long, M. (2021) 'Anomaly Transformer: Time-series anomaly detection with association discrepancy', arXiv preprint, arXiv:2110.02642. doi:10.48550/arXiv.2110.02642.

Zhang, S., Zhao, J., Xu, R., Feng, X. and Cui, H., 2025. Output constraints as attack surface: Exploiting structured generation to bypass llm safety mechanisms. arXiv preprint arXiv:2503.24191.