Access Controls Will Solve the Dual-Use Dilemma

Evžen Wybitul [1]

Abstract

AI safety systems face a dual-use dilemma. The same request can be either harmless or harmful depending on who made it and why. Thus, if the system makes decisions based solely on the request's content, it will refuse some legitimate queries and let harmful ones pass. To address this, we propose a conceptual access control framework, based on verified user credentials (such as institutional affiliation) and classifiers that assign model outputs to risk categories (such as advanced virology). The system permits responses only when the user's verified credentials match the category's requirements. For implementation of the model output classifiers, we introduce a theoretical approach utilizing small, gated expert modules integrated into the generator model, trained with gradient routing, that enable efficient risk detection without the capability gap problems of external monitors. While open questions remain about the verification mechanisms, risk categories, and the technical implementation, our framework makes the first step toward enabling granular governance of AI capabilities: verified users gain access to specialized knowledge without arbitrary restrictions, while adversaries are blocked from it. This contextual approach reconciles model utility with robust safety, addressing the dual-use dilemma.

1. Introduction

User requests — and with them, model outputs — exist on a spectrum from clearly benign to clearly harmful, with most falling in the grey zone in the middle (example in Figure 1). In the grey zone, the same output could be considered harmful or harmless, depending not on its content, but on its *real-world context*: who requested it and for what purpose.

Safety systems that rely solely on content analysis immediately face the *dual-use dilemma*. Since the same request can be either harmless or harmful depending on the context, wherever they draw the refusal line, they will restrict model utility for legitimate users while letting slip harmful requests from adversaries. Some safety systems try to address this by considering real-world context alongside content. However, they typically infer the context from the content itself, making it easy for adversaries to fabricate.

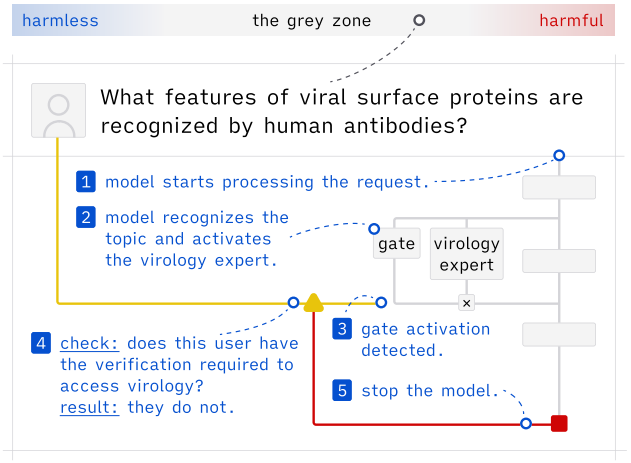

Figure 1. The user is asking a question from the grey zone, that is, one that could be either harmless or harmful, depending on its real-world context. The schema shows how the question would be handled by the system we propose. (1) The model is trained to be helpful and begins to answer the question. (2) The model activates its virology expert module because it contains concepts relevant to the question. (3) The gate activation is observed by an external mechanism that immediately (4) checks if the user has the required authorization to access virology knowledge. (5) Since they don't, the model is stopped. If they did, the model would be allowed to answer the question.

In this paper, we argue that informative, hard-to-fabricate real-world context could be obtained using user-level verifications such as institutional affiliation, or know-your-customer checks. We then address the dual-use dilemma with two contributions:

1. We show how this type of context could be used jointly with content analysis in a safety system based on access

[1]ETH Zurich, Switzerland. Correspondence to: Evžen Wybitul <wybitul.evzen@gmail.com>.

Technical AI Governance Workshop at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada.

controls (Lampson, 1974). First, generated content would be classified into risk categories. Then, a check would be performed to see whether the user has the verifications required to access the detected categories.

2. We propose a novel technical approach to risk category classification that is based on gradient routing (Cloud et al., 2024). Our proposal avoids having the capability gap between a model and its monitors that can make output monitoring methods non-robust (Jin et al., 2024).

Our framework is a first step toward solving the challenge of "detection and authorization of dual-use capability at inference time" that was highlighted by a recent survey of problems in technical AI governance (Reuel et al., 2025) and also raised by the U.S. AI Safety Institute (2024). As such, it has important governance implications, potentially enabling a more nuanced regulatory approach where access to powerful AI capabilities is stratified rather than binary, with policies that differentiate between user types and user contexts rather than focusing solely on model capabilities. The choice of appropriate verification mechanisms and risk categories remains for future work and should ideally happen jointly with stakeholders from academia, AI governance, and industry. Nevertheless, our approach offers a promising direction for addressing the dual-use dilemma.

2. Current Safety Methods Don't Solve the Dual-Use Dilemma

We evaluate three AI safety approaches to see how sensitive they are to contextual information, and whether their sources of real-world context are trustworthy — that is, hard to manipulate by an adversary.

First, to illustrate the need for context, consider decomposition attacks (Glukhov et al., 2023; 2024): transforming a clearly harmful query, such as "How to modify a virus to avoid immune detection?", into a series of mundane technical questions, like the "What features of viral surface proteins are recognized by human antibodies?" from Figure 1. Here, the attacker exploits the dual-use dilemma, and the fact that model providers cannot refuse grey zone requests to preserve model utility.

2.1. Unlearning: Non-Contextual Removal of Concepts

Unlearning methods aim to remove specific knowledge, concepts, or capabilities from a model after training (Liu et al., 2024). Their goal is to eliminate the model's ability to generate harmful content while preserving other capabilities.

Unlearning faces significant technical challenges even for preventing behaviours that are clearly harmful. As noted by Cooper et al. (2024) and Barez et al. (2025), capabilities are hard to define, hard to remove without side effects, and it is hard to trace them back to specific data points. Many unlearning approaches mask rather than truly remove the targeted knowledge (Deeb & Roger, 2025). Moreover, even nascent robust unlearning methods (Cloud et al., 2024; Lee et al., 2025) are not contextual, and thus don't address the dual-use dilemma without additional assumptions.

2.2. Safety Training: The Model Reacts to Context

Safety training methods modify the model's training process to align its outputs with human preferences. This category includes safety pre-training (Maini et al., 2025), RLHF (Christiano et al., 2023), and safety finetuning.

Unlike unlearning, these methods are contextual. They don't remove capabilities entirely but train the model to selectively deploy them based on, among other things, the perceived legitimacy and harmlessness of the request. However, these qualities are entirely inferred from content supplied by the user, such as the request content, the chat history, or the model's memories about past conversations. It should be no surprise, then, that models are susceptible to attacks that fabricate in-chat context (Zeng et al., 2024), or attacks that diminish models' sensitivity to in-chat context, e.g. through multi-round escalation (Russinovich et al., 2025). Without access to trustworthy real-world context of the request, the model cannot make truly informed decisions about grey zone requests, and thus cannot robustly address the dual-use dilemma.

2.3. Post-Processing: External Systems React to Context

Post-processing methods are systems that classify user inputs and model outputs for the purposes of steering the underlying model, and monitoring and filtering its outputs. Sometimes, these methods are used for usage monitoring, as is the case with Anthropic's Clio (Tamkin et al., 2024; Handa et al., 2025), other times, they are used for safety, as with Llama Guard (Inan et al., 2023) and Constitutional Classifiers (Sharma et al., 2025). However, similarly to safety training, the "real-world" context these methods work with is currently inferred mostly from user-supplied content and is thus untrustworthy and vulnerable to attacks, as evidenced by the many jailbreaks that successfully target current production systems (Zhang et al., 2025). Nevertheless, these methods could be modified to incorporate external contextual information, potentially serving as a foundation for more trustworthy, contextual safety mechanisms. We discuss this option in Section 4.2.

3. Access Controls as a Feasible Solution

In the previous section, we established that current safety systems do not address the dual-use dilemma because they

either do not consider the real-world context of the request, or they obtain the context directly from the request's content, making them vulnerable to adversarial attacks. In this section, we present an alternative safety framework based on access controls that directly addresses the dual-use dilemma, and discuss some of its practical considerations.

3.1. The Access Control Framework

Taking note of the problems of current safety methods, we first need a trustworthy source of real-world context. Drawing inspiration from other industries, we propose user-level verification mechanisms as a feasible way to obtain informative context about the user (and by extension, their requests). These verifications could range from basic identity confirmation, to institutional affiliation, to thorough know-your-customer (KYC) checks (Financial Action Task Force (FATF), 2025), each granting different access permissions.

To utilize this user-level context for addressing the dual-use dilemma, we recommend implementing an access control system: (1) pre-define multiple risk categories such as advanced virology or cybersecurity; (2) train classifiers to determine whether a given model output belongs to these categories (see Section 4); (3) check whether the user has the required verifications for generating content in that category. The verification requirements should be publicly available, enabling users to obtain necessary credentials ahead of time.

In practice, harmless content would require no verification, preserving frictionless experience for everyday queries. Grey zone outputs would require a level of verification calibrated to their risk assessment. For example, the virology question from Figure 1 might require the user to be affiliated with a recognized research facility. Clearly harmful requests would still be refused outright.

This defensive-by-default approach aims to minimize the attack surface by aligning with the principle of least privilege (Saltzer & Schroeder, 1975). It prevents unauthorized users from executing decomposition attacks while not reducing model utility for legitimate users, or even potentially granting advanced users access to more capabilities, as developers would no longer need to draw arbitrary refusal lines. The system thus enhances both utility for legitimate users and overall safety, solving the dual-use dilemma.

3.2. Practical Considerations

Many questions about the implementation of the system remain open for future work.

Verification levels must take into account feasibility, trustworthiness, user privacy, and the risk assessment of the different risk categories. We propose a two-tier verification system as an initial implementation. Most capabilities require no verification, maintaining frictionless access for common use cases.

Institutional verification (required only for high-risk domains) consists of organizational email verification, and affiliation confirmation from authorized representatives. This verification would be performed by third parties with existing identity verification infrastructure.

Risk categories must balance granularity with technical feasibility, reflecting usage patterns, potential harm, classification reliability, and the friction of their assigned verification level. Initially, content could fall into two risk categories. Standard content (requiring no verification) includes general programming and everyday queries. High-risk content (requiring institutional verification) includes: advanced bioengineering, chemical synthesis, and advanced cybersecurity (e.g., beyond the most basic textbooks). These categories represent specialized knowledge with clear misuse potential yet legitimate research applications.

System responses can vary based on risk category and classification confidence. The initial implementation might use a three-phase response system: (1) for outputs belonging to a risk category with high-confidence, immediate refusal with a specific explanation of the verification required; (2) for borderline classifications, continued generation with enhanced logging and post-processing review; (3) for verified users, seamless access with background logging for audit purposes.

3.3. Analysis of Feasibility

User friction analysis.

- Access controls introduce friction from two sources: intentional verification requirements for grey-zone requests and accidental false positives from imperfect classification.
- By design, all grey-zone requests require verification even with perfect classifiers. Additionally, classification errors may incorrectly flag clearly harmless requests (e.g., high-school biology as advanced virology).
- To understand the combined impact, consider that only 0.5% of requests involve biology topics according to Anthropic usage data.
- Even if ALL biology requests required verification—an extreme upper bound—this would affect fewer users than current system friction. Existing safety systems refuse 1–7% of benign requests as false positives, with Claude achieving the best rate of 0.4%. Requiring verification for all biology (0.5%) would approximately double Claude's friction rate.

- Additionally, verification friction differs qualitatively from current false positives—users can resolve issues through one-time credential verification rather than facing permanent refusal.
- Companies can calibrate friction through multiple mechanisms: adjusting grey-zone boundaries, tuning classifier thresholds, and implementing graduated responses (requesting clarification, additional context, or secondary review) rather than immediate verification requirements.
- Companies can determine optimal settings empirically through: (1) internal red-team decomposition attacks on concerning capabilities, (2) gradual rollout with logging to measure user impact, and (3) iterative adjustment based on safety-friction tradeoffs.

Developer incentives for adoption.

- Access controls enable competitive advantages by allowing companies to serve "dark-grey" requests that competitors refuse for safety reasons, while adding surgical restrictions to "light-grey" requests vulnerable to decomposition attacks.
- Without access controls, continued decomposition attack success will likely trigger broad government regulations restricting model capabilities entirely. Surgical access controls allow compliance with safety requirements while preserving advanced capabilities for verified users, creating market differentiation.
- Even without perfect industry coordination, first movers gain competitive advantages in serving previously restricted capabilities.

4. Implementing Risk Classification

A core requirement of the verification-based access control system described in Section 3 is being able to reliably classify model outputs into risk categories. This classification needs to address key challenges: accuracy with minimal false positives, resistance to adversarial attacks, and efficiency. We examine two approaches to implementing this classification — one currently available and one theoretical — and discuss their trade-offs.

4.1. Post-Processing

As discussed in Section 2, current systems already rely on post-processing classifiers that analyse outputs before delivery to users. These could be adapted for content classification into risk categories in an access control system. For example, a classifier could be trained to identify moderately advanced virology topics and, if detected, could trigger verification of user permissions before delivering the output.

The key advantage of post-processing systems is modularity, as they can be developed and updated independently of the generation models they oversee. However, they face a trade-off between usability (latency) and safety (Kumar et al., 2025): prioritizing low latency can create a capability gap between generators and monitors that sophisticated language models can exploit (Jin et al., 2024). Despite this limitation, recent post-processing methods show acceptable efficiency and resilience toward jailbreaks (Sharma et al., 2025) and could provide a practical initial implementation path for output-based access control.

4.2. Gradient Routing

To address the capability gap problem in post-processing methods, we detail how gradient routing (Cloud et al., 2024) could be adapted to classify model outputs into risk categories. Our adaptation represents a theoretical direction for integrating risk category detection directly into model architecture. This approach can be combined with post-processing methods and offers different trade-offs.

We propose augmenting models with small expert modules controlled by learned gates, as shown in Figure 1. During model training, assuming we have examples from the relevant risk categories (e.g., advanced virology), the expert module would be made to receive gradients exclusively from these examples. Simultaneously, the module's gate would be trained to activate the module on these examples using an auxiliary loss function. We hypothesize this would lead to the module obtaining specialist knowledge from the risk domain, and to the model learning to activate the module whenever this knowledge is needed.

Figure 1 illustrates this in the context of the wider access control system: when a user requests information about viral surface proteins, the model correctly recognizes the domain and activates the virology gate, engaging the expert but also triggering the verification system. If the user lacks appropriate credentials, the system takes the appropriate response — here, halting the generation.

This approach eliminates the efficiency-capability trade-off inherent in post-processing methods: the risk category classifier is embedded within the larger model, so there is no capability gap, and because the gated module is small and integrated into the model's forward pass, it is computationally efficient with minimal impact on latency.

Crucially, during standard training, the model is naturally incentivized to identify risk category information in inputs and surface this in its representation space, as this is directly linked to enabling the right specialized module at the right time. This approach differs fundamentally from post hoc methods like probing or sparse autoencoders (Cunningham et al., 2023), which cannot guarantee the model actually

operates with the concepts they attempt to identify.

Additionally, it resolves the adversarial dynamic (model vs. monitors, helpful vs. harmless) common in AI safety by aligning training incentives with safety objectives.

Several technical challenges remain. While gradient routing has shown promise in creating specialized modules in smaller models, its effectiveness in larger language models, particularly with the gated-expert architecture, requires empirical validation. Other key challenges include preventing false positives and false negatives, though regularization techniques and adjusted detection thresholds, respectively, could mitigate these issues.

Our approach also requires identifying risk categories during initial training, prompting research into adaptation of gradient routing for fine-tuning scenarios. Despite these challenges, the approach offers promising theoretical properties that warrant experimental investigation.

5. Limitations and Trade-offs

Our framework presents inevitable trade-offs — privacy concerns, access inequities, and increased friction through false positives — but these can be addressed through improved verification mechanisms and continuously improving classification systems. Future work needs to determine appropriate verification levels and risk categories, develop accurate classifiers with low false positive rates, and align with emerging AI regulations through stakeholder engagement. We discuss some of these challenges in Section 3.2 and at the end of Section 4.2.

6. Conclusion

We argued that safety systems that do not utilize contextual information face a lose-lose *dual-use dilemma*: they will restrict model utility for some legitimate users while still allowing some adversaries to use the model for ill. To address this problem, we introduced a new access control framework that limits access to outputs from certain risk categories only to users with relevant verifications (which serve as proxies for trustworthy real-world context). We also proposed a novel technical solution for classifying outputs into risk categories based on gradient routing that has the potential to resolve the efficiency-robustness trade-off of post-processing methods.

Beyond addressing specific technical challenges, our framework represents a promising governance shift from working with model-level abstractions and binary capability restrictions toward more granular user-level access controls. This offers a practical pathway for regulating increasingly powerful AI systems through stratified access rather than blanket capability limitations.

Acknowledgements

We thank Jakub Kryś, Dennis Akar, and Kola Ayonrinde for their feedback on a draft of this paper. We thank Joseph Miller, Alex Cloud, Alex Turner, and Jacob Goldman-Wetzler for discussions on gradient routing.

References

Barez, F., Fu, T., Prabhu, A., Casper, S., Sanyal, A., Bibi, A., O'Gara, A., Kirk, R., Bucknall, B., Fist, T., Ong, L., Torr, P., Lam, K.-Y., Trager, R., Krueger, D., Mindermann, S., Hernandez-Orallo, J., Geva, M., and Gal, Y. Open problems in machine unlearning for ai safety, 2025. URL https://arxiv.org/abs/2501.04952.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Cloud, A., Goldman-Wetzler, J., Wybitul, E., Miller, J., and Turner, A. M. Gradient routing: Masking gradients to localize computation in neural networks, 2024. URL https://arxiv.org/abs/2410.04332.

Cooper, A. F., Choquette-Choo, C. A., Bogen, M., Jagielski, M., Filippova, K., Liu, K. Z., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., Shumailov, I., Triantafillou, E., Kairouz, P., Mitchell, N., Liang, P., Ho, D. E., Choi, Y., Koyejo, S., Delgado, F., Grimmelmann, J., Shmatikov, V., Sa, C. D., Barocas, S., Cyphert, A., Lemley, M., danah boyd, Vaughan, J. W., Brundage, M., Bau, D., Neel, S., Jacobs, A. Z., Terzis, A., Wallach, H., Papernot, N., and Lee, K. Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice, 2024. URL https://arxiv.org/abs/2412.06966.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Deeb, A. and Roger, F. Do unlearning methods remove information from language model weights?, 2025. URL https://arxiv.org/abs/2410.08827.

Financial Action Task Force (FATF). International standards on combating money laundering, terrorist financing and proliferation — the fatf recommendations. Technical report, Financial Action Task Force (FATF), Paris, February 2025. URL https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html. Adopted 2012; updated February 2025.

Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., and Papyan, V. Llm censorship: A machine learning challenge or a computer security problem?, 2023. URL https://arxiv.org/abs/2307.10719.

Glukhov, D., Han, Z., Shumailov, I., Papyan, V., and Papernot, N. Breach by a thousand leaks: Unsafe information leakage in 'safe' ai responses, 2024. URL https://arxiv.org/abs/2407.02551.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., and Ganguli, D. Which economic tasks are performed with ai? evidence from millions of claude conversations, 2025. URL https://arxiv.org/abs/2503.04761.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

Jin, H., Zhou, A., Menke, J. D., and Wang, H. Jailbreaking large language models against moderation guardrails via cipher characters, 2024. URL https://arxiv.org/abs/2405.20413.

Kumar, D., Birur, N. A., Baswa, T., Agarwal, S., and Harshangi, P. No free lunch with guardrails, 2025. URL https://arxiv.org/abs/2504.00441.

Lampson, B. Protection. *ACM SIGOPS Operating Systems Review*, 8:18–24, 01 1974. doi: 10.1145/775265.775268.

Lee, B. W., Foote, A., Infanger, A., Shor, L., Kamath, H., Goldman-Wetzler, J., Woodworth, B., Cloud, A., and Turner, A. M. Distillation robustifies unlearning, 2025. URL https://arxiv.org/abs/2506.06278.

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C. Y., Xu, X., Li, H., Varshney, K. R., Bansal, M., Koyejo, S., and Liu, Y. Rethinking machine unlearning for large language models, 2024. URL https://arxiv.org/abs/2402.08787.

Maini, P., Goyal, S., Sam, D., Robey, A., Savani, Y., Jiang, Y., Zou, A., Lipton, Z. C., and Kolter, J. Z. Safety pretraining: Toward the next generation of safe ai, 2025. URL https://arxiv.org/abs/2504.16980.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Bau, D., Bricman, P., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. Open problems in technical ai governance, 2025. URL https://arxiv.org/abs/2407.14981.

Russinovich, M., Salem, A., and Eldan, R. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2025. URL https://arxiv.org/abs/2404.01833.

Saltzer, J. and Schroeder, M. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9): 1278–1308, 1975. doi: 10.1109/PROC.1975.9939.

Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., Ong, E., Peng, A., Agarwal, R., Anil, C., Askell, A., Bailey, N., Benton, J., Bluemke, E., Bowman, S. R., Christiansen, E., Cunningham, H., Dau, A., Gopal, A., Gilson, R., Graham, L., Howard, L., Kalra, N., Lee, T., Lin, K., Lofgren, P., Mosconi, F., O'Hara, C., Olsson, C., Petrini, L., Rajani, S., Saxena, N., Silverstein, A., Singh, T., Sumers, T., Tang, L., Troy, K. K., Weisser, C., Zhong, R., Zhou, G., Leike, J., Kaplan, J., and Perez, E. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming, 2025. URL https://arxiv.org/abs/2501.18837.

Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S., Stern, M., Clarke, B., Goldberg, L., Sumers, T. R., Mueller, J., McEachen, W., Mitchell, W., Carter, S., Clark, J., Kaplan, J., and Ganguli, D. Clio: Privacy-preserving insights into real-world ai use, 2024. URL https://arxiv.org/abs/2412.13678.

U.S. AI Safety Institute. Managing misuse risk for dual-use foundation models. Initial Public Draft NIST AI 800-1, U.S. AI Safety Institute, Gaithersburg, MD, July 2024. URL https://doi.org/10.6028/NIST.AI.800-1.ipd.

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., and Shi, W. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms, 2024. URL https://arxiv.org/abs/2401.06373.

Zhang, S., Zhao, J., Xu, R., Feng, X., and Cui, H. Output constraints as attack surface: Exploiting structured generation to bypass llm safety mechanisms, 2025. URL https://arxiv.org/abs/2503.24191.