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.

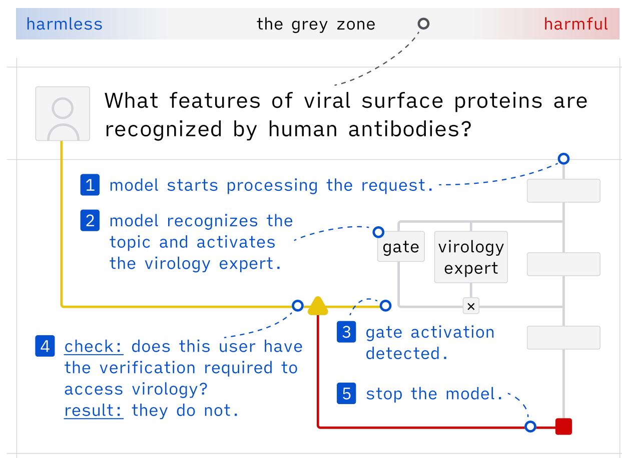

Figure 1. The user is asking a question from the grey zone: a question that could be either harmless or harmful, depending on its real-world context. The schema shows how the system we propose would handle it. (1) The model is trained to be helpful and begins to answer the question. (2) During the forward pass, the model activates its virology expert module because it is relevant to the question. (3) The activation of the expert is observed by an external mechanism that immediately (4) checks in the company's database 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 give an answer.

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.

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 controls (Lampson, 1974). First, generated content

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

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

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 approaches from the AI safety literature 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 3.4.

3. Access Controls as a Solution

Current safety systems face the dual-use dilemma because they lack trustworthy information about who is making a request and why. In this section, we describe an access control system that addresses this problem by verifying user credentials before granting access to sensitive knowledge.

3.1. Overview of the Access Control Framework

We propose a defensive system where grey-zone requests are refused by default, but users can gain access to specific categories of knowledge if they undergo verification.

When model providers set up the system, they will make two design decisions with the help of domain experts. First, they will define **content categories** (Section 3.2): groups of sensitive topics organized by domain and risk rating. Second, for each content category, they will specify a **verification mechanism** (Section 3.3): the verification process users must complete to access that category.

Whenever the model generates an output, the system will perform **content classification** (Section 3.4) to check whether the model's output belongs to any predefined content category. If the user lacks authorization for the detected category, the system will implement appropriate **system responses** (Section 3.5) ranging from enhanced logging to refusal.

For example, in biology, basic knowledge and common techniques would remain freely accessible, widespread techniques like CRISPR would likely only require ID-based verification, and dangerous techniques like aerosolization might require government biosafety certifications. If a user asks for help with CRISPR laboratory protocols, the system would detect that the request belongs to a low-risk category, check whether the user has verified their ID, and either provide the information or prompt them to complete verification first.

This approach directly addresses both sides of the dual-use dilemma. Decomposition attacks will become much harder because the system refuses grey-zone requests by default—attackers would need legitimate credentials rather than clever prompting. Simultaneously, verified users will gain access to specialized knowledge that would otherwise face blanket restrictions under current approaches.

The main concern is increased user friction, but we argue in Section 4 that this will be minimal because most users will never make grey-zone requests.

3.2. Content Categories

Content categories are groups of sensitive topics organized by domain and risk level, which model providers will develop with domain experts.

We expect most implementations to follow a three-tier risk structure. For example, in biology, common techniques would be classified as low-risk; widespread techniques that pose some harm, such as CRISPR, would fall into a moderate-risk category; and specialized techniques with limited legitimate uses, such as aerosolization of bacteria, would be classified as high-risk.

Experts could develop these categories by adapting existing risk frameworks, such as biosafety levels (BSL) (Centers for Disease Control and Prevention & National Institutes of Health, 2020) and dual-use research of concern policies (United States Government, 2012) in biology. However, since existing frameworks typically categorize only high-level concepts like organisms or compounds, experts would need to decompose them into smaller, more specific components. For instance, cultivating and handling a dangerous BSL-3 pathogen might involve (1) specific procurement methods, (2) cultivation techniques, (3) purification methods, and (4) protocols for specialized equipment. For each of these components, experts would assess the ratio of harmless to harmful applications it enables, then assign it to an appropriate (low, moderate, or high) risk category.

Evidence from chemistry suggests this approach could work: the risk schedules of the Chemical Weapons Convention already identify not just controlled compounds but also their precursors and specific equipment (Organisation for the Prohibition of Chemical Weapons, 1993), demonstrating successful decomposition into components. Nevertheless, some harmful applications might not decompose so neatly; we discuss this limitation in Section 4.

3.3. Verification Mechanisms

Each content category will have a verification process that users must complete to access it. The system will initially vary across model providers, but we expect it to follow a three-tier structure, mirroring the risk structure of the content categories. Most content will require no verification, moderate-risk content categories will require basic identity verification or institutional affiliation, and high-risk categories will require domain-specific certifications. Rather than creating new systems, model providers will build on existing verification infrastructure, consulting domain experts to identify appropriate mechanisms for each field.

For low-risk content categories, model providers could use established identity verification services like Stripe Identity (Stripe, Inc., 2024) or institutional systems like ORCID (ORCID, Inc., 2024). These systems provide global, standardized, low-friction solutions with one-time costs under $2 per user. They would serve primarily to maintain audit trails for post-incident investigation and provide a deterrent effect, rather than as security barriers for high-risk knowledge.

High-risk content categories could leverage existing domain-specific certifications that demonstrate users' ability to handle sensitive information and materials responsibly. Model providers would work with domain experts and national authorities to identify appropriate certifications, adapting existing physical-world credential systems to knowledge access control. For biological content categories, the system could draw on governmental certifications for handling high-BSL organisms, as mentioned in Section 3.2, and equivalent certifications in other countries.

Governance of verification systems, including requirements and appeals processes, will initially vary across providers. Over time, successful approaches may inform industry coordination and eventual standardization, similar to how content moderation and know-your-customer standards evolved.

This approach faces several limitations. For high-risk categories, relying on existing certifications may be overly restrictive, potentially excluding some users who should have access. However, we argue in Section 4 that knowledge in high-risk categories would likely face blanket restrictions anyway, and our approach enables access for verified users rather than complete prohibition. In the same section, we discuss open problems including equity concerns regarding differential access for users in developing countries and privacy implications of credential verification systems.

3.4. Implementing Content Classification

Model providers will need to classify model outputs into content categories during generation. We examine three possible implementations below. While none of these approaches have been empirically validated for risk category classification specifically, each represents a plausible technical path that could be developed and evaluated by practitioners interested in implementing access controls. We leave the discussion of how classification errors might influence user experience for Section 4.

Separate Models The most straightforward approach is to create separate models with different capabilities, and route users to the appropriate model based on their authorization.

This approach offers strong robustness against adversarial attacks since unauthorized knowledge is physically absent from the model. However, this approach proves impractical for real deployment, as model providers would need to train and maintain potentially dozens of model variants.

Specialized Expert Modules Instead of maintaining separate models, model providers could use a single model with separate expert modules that activate when their specialized knowledge is required. Figure 1 illustrates this approach when a user asks about viral surface proteins. When the model processes the request, it activates its virology expert module. An external system observes this activation, checks the user's credentials, and decides whether to allow the model to deliver the response. This method approximates the benefits of physically separated models while avoiding the overhead: a model provider trains one model but effectively gets multiple models in return.

To implement this, the model providers need a method that can take knowledge that starts out distributed throughout the model and concentrate it into the expert modules. For this, we propose a method that is a combination of UNDO (Lee et al., 2025) and gradient routing (Cloud et al., 2024). The steps resemble the original UNDO: first, unlearn knowledge belonging to any content category from the model, then distil the unlearned model into a new model. However, taking inspiration from the gradient routing paper, the new model would include expert modules for each content category, and during distillation, gradients from examples in the various content categories would be routed exclusively through their associated expert modules. The model would also be explicitly trained to activate the expert modules only when generating content in their associated category.

This approach would offer several advantages. First, it would add almost no latency since the expert modules are small, not activated very often, and there is no post-processing step. Second, it could provide strong robustness: if an attacker prevents the activation of an expert module to avoid detection, the resulting output lacks the specialized knowledge. Third, since the model is trained to activate the category-specific experts when they are needed, it should learn to recognize the content categories. This stands in contrast to ex-post probing methods, which do not offer such guarantees. While this method remains empirically unvalidated for our use case, the properties above make it worth investigating.

Post-Processing Post-processing methods offer a proven approach to content classification. Methods like constitutional classifiers (Sharma et al., 2025) already demonstrate effectiveness in production systems. These techniques are highly practical since they operate independently of the model, allowing for rapid deployment and iteration, and they could be adapted to detect content categories and trigger checks of user verifications. However, they face a capability gap problem: to minimize latency, the model is sometimes more capable than its post-processing system, and adversaries can exploit this to evade detection (Jin et al., 2024; Kumar et al., 2025).

3.5. System Responses

If the user makes a request for content they are authorized to access, the system allows the model to generate the response. Otherwise, the system responds in various ways based on the risk category and the confidence of the content classification.

For example, the initial implementation might use the following two response types: First, outputs classified as belonging to restricted content categories with high confidence are immediately refused. The system provides a message indicating which verification is required for access. Second, if the classification is borderline, the model is allowed to continue generating the response. However, the system turns on enhanced logging and conducts additional post-processing safety review before delivering the output to the user.

4. Feasibility and Limitations

Section 2 and Section 3 established that current safety methods cannot solve the dual-use dilemma, while access controls offer a viable theoretical solution. However, Section 3 also identified several technical challenges that could affect the practicality of the system. In this section, we show that all of these connect to a central question: will the added user friction undermine the system's value?

Technical Problems Lead to Friction We identified the following technical challenges with the various components of the system:

(1) some harmful knowledge might decompose only into concepts that are all widespread and highly useful for many harmless things; (2) verification mechanisms may be overly restrictive, requiring credentials that are too difficult to obtain; and (3) content classifiers will produce false positives, forcing legitimate users through unnecessary verification.

As for the first problem: this varies domain by domain. In chemistry, the Chemical Weapons Convention already identifies not just controlled compounds but also their precursors and specific equipment, demonstrating successful decomposition into specialized components. Similarly, many processes in biology depend on tacit knowledge that is specific to the studied organism, equipment, and desired use-case. However, in cybersecurity, we concede that it might be difficult to identify neat self-contained concepts that, while being dual-use, are more often than not useful mainly for hamrful things. All in all, if concept categories are too broad or if all categories are often used by many users, there will be a lot of absolute amount of friction experienced by the median user.

The second problem. If verifications are overly strict, or even impossible to obtain for structural reasons (e.g. because they rely on certification system that are not available in developing countries), manual review processes and appeals processes could remove some of the problems, before new verification systems are developed. However, this still leads to friction: if the verification mechanisms are too difficult to obtain, users will experience friction.

Finally, the third problem. Classification errors are inevitable, but existing systems like Constitutional Classifiers achieve acceptable false positive rates around 5% even in difficult safety-adjacent domains, demonstrating that meaningful precision is achievable. Nevertheless, any amount of error leads to friction: if the content classifiers produce false positives, legitimate queries will be classified as potentially harmful, and users will have to undergo verification and thus, again, experience friction.

All in all, although the technical challegenges haev some engineering or design-based solutions, they will definitely be present in the system to some extent, which will increase user friction. Will this added friction outweigh the benefits of the system? To address this systematically, we split the dual-use dilemma — the problem the system is trying to solve — into two sub-problems. First, even legitimate users are restricted from accessing some sensitive knowledge (over-restriction). Second, adversaries can find exploits to use the model for harmful purposes (under-restriction).

Focus: The Impact on Over-Restriction For the first side of the dual-use dilemma — legitimate users being denied access to sensitive knowledge — access controls provide a clear improvement. To see why, consider that the model providers can set up the system to only require verification for high-risk categories, ones that would otherwise face blanket restrictions (be it due to regulations or model providers' own policies). This is possible e.g. by detecting when the current system would refuse a request, and only then triggering the verification cascade, and if the user is verified, steering the model to instead generate an answer. The verification mechanisms could be set so strict that an overwhelming majority of adversaries would be deterred by them, keeping the safety of the system comparable to status quo. Additionally, legitimate users are never worse off than current systems: they can either accept refusal (current experience) or complete verification to gain access (new option). The technical problems mentioned above do limit these improvements (e.g. less users benefit from this than would theoretically be possible because of overly complex verification mechanisms), but the improvement is always non-zero: it is better when *some* legitimate users get access to high-risk knowledge than when *none* do. In other words, the access control system is a Pareto improvement over the current system. In the most pessimistic scenario, model providers could only implement the access control system for these use-cases, not increasing the safety of their systems, but increasing the utility. The amount of utility they would unlock by doing this rises with the rising capabilities of new models: the new models will know more and more potentially risky knowledge — but also potentially useful, for advanced, high-value professional users — which nobody would ever get to see under the current system.

Focus: The Impact on Under-Restriction The second side of the dual-use dilemma — preventing adversarial access to sensitive knowledge — presents more complex trade-offs. Grey-zone requests from legitimate users will require verification, creating friction that must be weighed against safety benefits. Model providers must choose their position on the use-misuse frontier, balancing user convenience against security concerns. However, the system's granularity and flexibility enable providers to tune multiple parameters — content categories, verification requirements, classifier thresholds, and system responses — to find acceptable trade-offs rather than being forced into binary choices.

The good news is that in many domains, the friction costs are manageable while safety benefits are substantial. Pathogen biology provides concrete evidence: even under the pessimistic assumption that all pathogen-related queries require verification, only 0.85% of users face additional friction [cite appendix, mention anthropic economic index]. If the verification mechanisms are set correctly (they deter more adversaries than legitimate users), we would expect this system to be much safer than the current system. To put the friction number for legitimate users in context: existing systems like Constitutional Classifiers have false positive rates around 0.5% and are deemed acceptable for production use. In other words, we are trading off an acceptable amount of friction for a fraction of users against a much safer world for everyone. In reality, the amount of friction will be much lower, as classifiers will achieve better precision than this worst-case scenario assumes, and also, biological knowledge decomposes into more specific components (techniques for bacterial membrane protein modification, viral cultivation protocols, protocols to use specialized laboratory equipment) that can be targeted precisely rather than requiring broad categorical restrictions. So, in reality, the result would be friction reduced by orders of magnitude compared to worst-case estimates while maintaining meaningful safety improvements. We believe the situation in chemistry will be similar or even better.

Other domains may present greater challenges, with cybersecurity representing a potential example where decomposition into clean content categories proves difficult. However, providers are not required to implement access controls uniformly across all domains. They can start with domains where decomposition works well, gain experience with system operation, and expand gradually. For challenging domains, providers can choose to implement only high-risk restrictions (capturing the clear Pareto improvement) while leaving moderate-risk knowledge unrestricted. Alternatively, if security concerns are severe enough, providers may accept higher friction costs in exchange for meaningful safety improvements, but this remains a voluntary choice rather than a system requirement.

The system's flexibility enables a gradual implementation strategy that reduces risks while building evidence. Providers can begin with domains that decompose cleanly and have established verification mechanisms, conduct partial rollouts to measure friction impacts, and adjust parameters based on empirical data. Success in initial domains can inform expansion to more challenging areas, with the option to maintain different policies across domains based on their specific characteristics. This approach allows providers to walk the use-misuse frontier systematically, optimizing for their particular risk tolerance and user base rather than accepting one-size-fits-all solutions. As model capabilities advance and dual-use concerns intensify, the value of having such a flexible system in place increases, making early investment in access controls increasingly attractive.

One thing that makes the deal sweeter: While verification adds some friction for legitimate users, it disproportionately increases barriers for potential misusers, making the trade-off more attractive for providers than simple friction metrics suggest. This means the model providers might be willing to trade off a small amount of friction for legitimate users to gain a comparaively much safer system. Well-designed verification mechanisms create asymmetric costs that favor the access control approach. Legitimate researchers typically have institutional affiliations, professional credentials, and established reputations that make verification straightforward. Adversaries seeking to misuse sensitive knowledge face higher barriers to obtaining credible credentials.

Open Problems

5. 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.

Acknowledgements

We thank Jakub Kryś, and Dennis Akar 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.

Centers for Disease Control and Prevention and National Institutes of Health. Biosafety in microbiological and biomedical laboratories. Technical report, U.S. Department of Health and Human Services, Atlanta, GA, 2020. URL https://www.cdc.gov/labs/BMBL.html. Defines Biosafety Levels (BSL-1 through BSL-4) used in the United States.

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.

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

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.

ORCID, Inc. ORCID: Connecting research and researchers, 2024. URL https://orcid.org/. Global, persistent identifier system for researchers and scholars.

Organisation for the Prohibition of Chemical Weapons. Convention on the prohibition of the development, production, stockpiling and use of chemical weapons and on their destruction, 1993. URL https://www.opcw.org/chemical-weapons-convention. Contains the Annex on Chemicals with Schedules 1–3.

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.

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.

Stripe, Inc. Stripe identity, 2024. URL https://stripe.com/identity. Identity verification service with pricing starting at $2 per verification.

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.

United States Government. United states government policy for oversight of life sciences dual use research of concern, 2012. URL https://aspr.hhs.gov/S3/Pages/Dual-Use-Research-of-Concern-Oversight-Policy-Framework.aspx.

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.

A. Estimating the Number of Requests Related to the Biology of Pathogens

To estimate how many user requests are related to the biology of pathogens, we used the second version of the Anthropic Economic Index (Handa et al., 2025), a dataset of 1 million anonymized conversations from the Free and Pro tiers of Claude.ai. In the dataset, the conversations are clustered by topic, and the proportion of each topic in the whole dataset is given. For example, the topic "Help with agricultural business, research, and technology projects" makes up 0.15% of the requests in the dataset. There are three levels of topic granularity; we use the lowest, most granular level.

We filtered the dataset to only include conversations whose topic contains one of the following keywords related to biology: *cell* (when at the beginning of the word), *genet*, *genom*, *microb*, *bacteria*, *virus*, *viral*, *proteo*, *protei*, *immune*, *neuro*, *patho*, *infect*; we also required that it does not contain any of the following keywords to avoid false positives: *nutri*, *tweet*, *agric*, *sexual health*. The total proportion of these requests was 0.85%.