Access Controls Will Solve the Dual-Use Dilemma

Evžen Wybitul [1]

Abstract

AI safety systems face the dual-use dilemma: it can be unclear whether to refuse certain requests, since they could be either harmless or harmful depending on who made them and why. Determining this requires examining their real-world context, but current safety systems cannot access this contextual information. Instead, they make arbitrary decisions that end up hurting both utility and safety: they sometimes refuse legitimate queries and other times fail to refuse harmful ones. To address this, we propose a conceptual framework based on access controls in which only verified users can access dual-use outputs. We describe the framework's components, analyse its feasibility, and explain how it addresses both over-refusals and under-refusals. While only a high-level proposal, our work takes the first step toward enabling granular AI governance that could transform the traditional utility-safety trade-off: users would gain access to more capabilities without sacrificing safety, and regulators would achieve better oversight without stifling innovation.

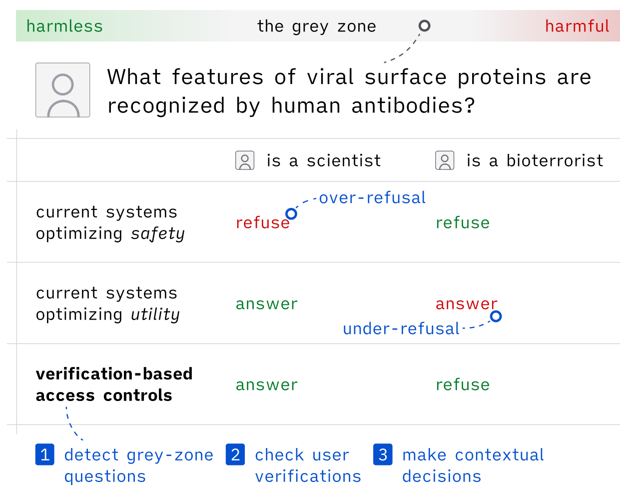

Figure 1. The dual-use dilemma: the same question can be harmless or harmful depending on who asks it. Should we refuse it or not? Current safety systems must decide without knowing the user's context. This leads to over-refusals (blocking legitimate users) and under-refusals (allowing bad actors). Verification-based access controls solve this by detecting grey-zone questions, obtaining real-world context about users, and making contextual decisions that can get both cases right.

1. Introduction

What features of viral surface proteins are recognized by human antibodies?

Is this question safe to answer? While some user requests and large language model outputs are clearly benign or harmful, many fall in the grey zone in the middle, as illustrated in Figure 1. In the grey zone, the harmfulness of a request depends not on its content, but on its *real-world context*: who made it and for what purpose.

Safety systems that rely solely on content analysis immediately face the *dual-use dilemma*. When confronted with a grey-zone request, should they refuse it or not? This forces arbitrary decisions that reduce both utility and safety: some legitimate queries are refused (over-refusals) while some harmful ones are not (under-refusals). Certain safety systems attempt to address this by inferring the context of the request from its contents or the chat history. However, this inferred context is based entirely on user-provided inputs and can be easily fabricated by adversaries.

In this paper, we argue that informative, hard-to-fabricate real-world context can be obtained through user-level verifications such as ID checks, institutional affiliations, or government-issued certifications. We address the dual-use dilemma with two contributions. **As our primary contribution**, we show how this verified context can be used jointly with content analysis in a safety framework based on access controls (Lampson, 1974). In the framework, model outputs are classified into content categories, and the system verifies whether the user has the required credentials to access the detected category. We also describe how the framework addresses both issues caused by the dual-use

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

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

dilemma: over-refusals and under-refusals. **Additionally**, we propose a novel theoretical approach to content category classification based on recent methods for robust unlearning, UNDO (Lee et al., 2025) and gradient routing (Cloud et al., 2024). This approach avoids the capability gap between a model and its monitors that can make output monitoring methods non-robust (Jin et al., 2024).

Current regulatory approaches choose between blanket restrictions of capabilities that stifle innovation and blanket permissions that create safety risks. Our framework represents a first step toward solving the challenge of "detection and authorization of dual-use capability at inference time" highlighted by a recent survey of problems in technical AI governance (Reuel et al., 2025). Future implementations of such frameworks could enable granular policies that differentiate between types of users and their contexts.

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

The dual-use dilemma causes two issues: over-refusals and under-refusals. Over-refusals reduce model utility for legitimate users, which is clearly undesirable. Under-refusals are equally problematic because they enable decomposition attacks (Glukhov et al., 2023; 2024). These attacks transform clearly harmful queries, such as "How to modify a virus to avoid immune detection?", into series of mundane grey-zone questions, such as "What features of viral surface proteins are recognized by human antibodies?" from Figure 1. Safety systems would refuse the harmful query but do not refuse the grey-zone questions since they are not individually harmful. Through these attacks, adversaries exploit under-refusals to decrease system safety.

Since whether a grey-zone request should be refused depends on who made it and why, preventing both over-refusals and under-refusals requires access to real-world context. This means that the traditional focus on resilience against jailbreak and prompt injections is orthogonal to this problem. Instead, 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.

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 hard to trace back to specific data points. Furthermore, many unlearning approaches mask rather than truly remove the targeted knowledge (Deeb & Roger, 2025).

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 or the chat history. 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 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. Future work could try modifying these methods to incorporate external contextual information, e.g. by giving the model access to the user's ID or institutional affiliation. It would likely result in a more opaque and less modular system than making similar modifications to post-processing methods.

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, or 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 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

Table 1. Example content categories in pathogen biology.

Risk Level	Examples	Verification
Low	Basic knowledge	—
Moderate	CRISPR protocols	ID verification
High	Viral surface proteins	Biosafety certification

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 core design choices 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 graduated **system responses** (Section 3.5) ranging from enhanced logging to refusal.

While the exact categorization will be domain-specific, we anticipate a common pattern where the majority of content remains freely accessible, with progressively more stringent verification requirements for higher-risk categories. Table 1 illustrates how this might look in pathogen biology. Under this example, if a user asks our running question about viral surface proteins, the system would detect that the request belongs to a high-risk category, check whether the user has the required biosafety certification, and either provide the information or prompt them to complete verification first.

This approach directly addresses the dual-use dilemma by preventing under-refusals and reducing over-refusals. Decomposition attacks become much harder because the system refuses grey-zone requests by default—attackers would need legitimate credentials rather than clever prompting. Simultaneously, verified users gain access to specialized knowledge that would otherwise face blanket restrictions under current approaches. We discuss the feasibility and limitations of this framework, including considerations around user friction, in Section 4.

3.2. Content Categories

Model providers will develop content categories by adapting existing risk frameworks with the help of domain experts. In biology, experts could build on 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). 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 suitable for knowledge access control.

For instance, cultivating and handling a dangerous BSL-3 pathogen might involve multiple distinct knowledge components: (1) specific procurement methods, (2) cultivation techniques, (3) purification methods, and (4) protocols for specialized equipment. For each component, experts would assess how often it enables harmless versus harmful applications, then assign it to an appropriate risk category. This decomposition approach transforms broad regulatory categories into granular knowledge components that can be individually controlled, as illustrated by the different risk levels in Table 1.

Evidence from chemistry suggests this approach is sometimes feasible: 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 requires a verification process that users must complete to access it, as shown in Table 1. Rather than creating new systems, model providers will build on existing verification infrastructure, consulting domain experts to identify appropriate mechanisms for each field.

For moderate-risk 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 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. In biology, this might include governmental certifications for handling dangerous pathogens, with equivalent certifications recognized across countries.

This approach faces several limitations, including the risk of being overly restrictive and concerns about equitable access across different countries. We discuss these limitations and potential solutions in Section 4.

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 viable 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 how this approach might work when a user poses a grey-zone question.

To implement this, we propose a novel method that would combine UNDO (Lee et al., 2025) and gradient routing (Cloud et al., 2024) to concentrate distributed knowledge into specialized expert modules. The approach would first unlearn knowledge belonging to content categories from the base model, then distill this unlearned model into a new model with expert modules for each category. During distillation, gradients from examples in each content category would be routed exclusively through their associated expert modules, while the model would be trained to activate experts only when generating relevant content.

This approach could offer several advantages. It would likely add minimal latency since expert modules would be small and infrequently activated. More importantly, it could provide strong robustness: if an attacker prevented expert module activation to avoid detection, they would simultaneously prevent access to the specialized knowledge stored in that module, making the attack self-defeating. While this method remains entirely theoretical and requires empirical validation, these potential properties make it worth investigating.

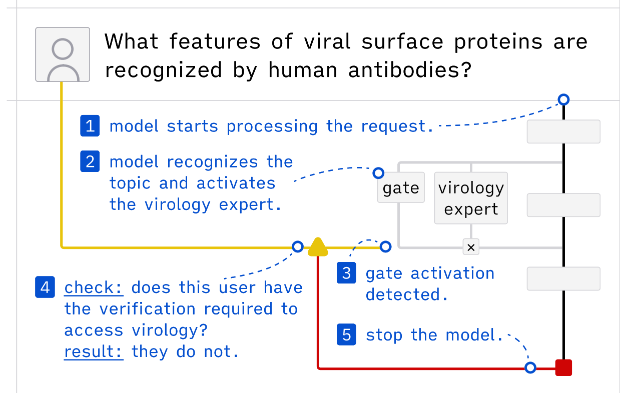

Figure 2. The schema shows how an access control system could be implemented with specialized expert modules. (1) The model begins to answer the question because it is trained to be helpful. (2) During the forward pass, the model detects the question is about virology and activates its virology expert module that contains relevant knowledge. (3) The activation of the expert is observed by an external mechanism that (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.

Post-Processing Post-processing methods such as Constitutional Classifiers (Sharma et al., 2025) offer a proven approach to content classification. 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, for latency reasons, there is sometimes a capability gap between the model and the post-processing system, which adversaries can exploit to evade detection (Jin et al., 2024; Kumar et al., 2025).

3.5. System Responses

When content classification detects restricted categories, the system can implement various responses depending on the risk level and classification confidence. This provides an additional parameter that model providers can tune based on their specific requirements.

For example, outputs classified as restricted with high confidence might be immediately refused, with the system providing a message indicating which verification is required for access. For borderline classifications where confidence is low, the system might allow response generation while enabling enhanced logging and additional safety review before serving the output to the user. Other possible responses include content generation with a steered model, or graduated restrictions for repeat violations.

4. Feasibility and Limitations

Section 3 identified several technical challenges: (1) some harmful knowledge might decompose into concepts that are all indispensable for harmless applications; (2) some verifications may be difficult to obtain; and (3) content classifiers might produce false positives. These challenges increase user friction and impose utility costs. We analyse how access controls affect the safety-utility tradeoff in the both over-refusals and under-refusals.

4.1. Access Controls Help Against Over-Refusals

For over-refusals, access controls can provide a strict improvement over the status quo, even with the technical challenges. Model providers can only require verification for requests that would otherwise be refused. This is a clear improvement in utility: legitimate users can either accept refusal (current experience) or complete verification to gain access (new option). Safety need not be hurt: model providers can design the verification requirements to be strict enough to deter the overwhelming majority of adversaries.

4.2. Domain-specific Approaches to Under-Refusals

Using access controls to prevent under-refusals means making some accessible knowledge require verification. Thus, even a perfect system takes a toll on utility since legitimate users must get verified, with additional costs caused by the technical challenges mentioned above.

Model providers can measure the impacts of the system on utility before deployment, for example by conducting partial rollouts. They can then walk the safety-utility frontier by tuning the parameters of the system: making verification more or less stringent, raising or lowering classifier thresholds, and so on. The optimal point on the frontier will depend on (and will change over time with) the domain-specific technological constraints, liability requirements, regulatory pressures, and business priorities.

In many domains, the default position might be the status quo which offers maximum utility. Pathogen biology is one example of a domain where the tradeoff for safety is already favourable, while also being safety-relevant: a recent study found that frontier models have more tacit knowledge than 94% of expert virologists (Götting et al., 2025). Consider an access control system so badly calibrated that all pathogen-related queries would require verification. This represents an upper bound on friction since real implementations would be far more targeted. First, we note that this system would deter at least some adversaries from performing decomposition attacks, improving on the safety of the status quo. At the same time, the impact on utility is minimal: only 0.85% of queries would experience added friction (estimated based on the Anthropic Economic Index, see Appendix A for details). This would likely be acceptable to model providers; for context, existing safety systems like Constitutional Classifiers incorrectly refuse around 0.5% queries (Sharma et al., 2025). Other domains like cybersecurity may face different challenges and would require domain-specific analysis before implementation.

4.3. The Incentives for Implementation Grow with Model Scale

As models gain even more sophisticated knowledge across dual-use domains, the value of the knowledge (for specialized professional users) and the misuse risks it poses increase. Thus, controlling both under-refusals and over-refusals will be more and more important, and the business implementing access controls strengthens.

4.4. Limitations and Open Questions

Since the system links real-world identities to requests, it poses risks of surveillance and information leakage. Model providers should establish clear privacy policies and use privacy-preserving systems like Clio (Tamkin et al., 2024) for logging and auditing.

Additionally, using verification mechanisms based on certifications that are not available globally could exclude legitimate researchers from access. Model providers should work with regional authorities to address this, with interim manual approval processes for users facing structural barriers.

5. Conclusion

Current safety systems face a dual-use dilemma when encountering grey-zone requests: should they refuse a request that could be either harmless or harmful depending on who made it and why? Our access control framework solves this by incorporating user verification into safety decisions. This reduces over-refusals for legitimate users and prevents under-refusals that lead to decomposition attacks. We also proposed a novel content classification approach that promises high efficiency and robustness to attacks, which future work should validate empirically. As models gain sophisticated dual-use knowledge, access control frameworks offer a path toward nuanced AI governance that moves beyond blanket restriction policies and allows governments to achieve better oversight without stifling innovation.

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.

Götting, J., Medeiros, P., Sanders, J. G., Li, N., Phan, L., Elabd, K., Justen, L., Hendrycks, D., and Donoughe, S. Virology capabilities test (vct): A multimodal virology q&a benchmark, 2025. URL https://arxiv.org/abs/2504.16137.

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.

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%. When we applied similar methodology to identify requests related to any kind of biology, the proportion was 2.98%.