Forecasting Research Institute

Forecasting the Impacts of ASL-3 Safeguards on Biosecurity Risks

Authors: Bridget Williams, Rebecca Ceppas de Castro, Dan Mayland, Ria Viswanathan, Zack Devlin-Foltz, Jordan Canedy, Victoria Schmidt, Philip E. Tetlock, Ezra Karger, Josh Rosenberg

FIRST RELEASED ON AUGUST 12, 2026

FRI TECHNICAL REPORT

Authors and affiliations

Bridget Williams[1,2], Rebecca Ceppas de Castro[1], Dan Mayland[1], Ria Viswanathan[1], Zack Devlin-Foltz[1], Jordan Canedy[1], Victoria Schmidt[1], Philip E. Tetlock[1,3], Ezra Karger[1,4], Josh Rosenberg[1]

1 = Forecasting Research Institute, 2 = University of Oxford, 3 = University of Pennsylvania, 4 = Federal Reserve Bank of Chicago

Acknowledgments

This research would not be possible without the thoughtful participation of our survey respondents. We are grateful to Mrinank Sharma and Jerry Wei for helpful feedback on survey design. We are also grateful to Anna Bartlett, Coralie Consigny, Johan S Daniel, and Josh Connor for research assistance, and to Otto Kuusela and Simas Kučinskas for helpful feedback on research design.

Disclaimers

This report was originally produced as part of an internal project funded by and completed for Anthropic PBC. Anthropic provided input and advice in the development of the report, but all findings and claims represent the views of the authors. Due to the general interest of the findings, we are releasing a version of the report with confidential information removed.

The views expressed in this paper do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

Table of Contents

Summary

In May 2025, Anthropic announced that it had activated AI Safety Level 3 (ASL-3) Deployment and Security Standards as a precautionary measure, in response to Claude Opus 4's improved CBRN-related capabilities [1]. The deployment standards include the implementation of real-time classifier guards, offline monitoring, access controls, a bug bounty program, threat intelligence, and rapid response.

The Forecasting Research Institute conducted a study to better understand the effectiveness of these measures—particularly the ASL-3 deployment standards—at mitigating biosecurity risks. We surveyed 22 people with expertise in biosecurity, national security, and/or terrorism studies ("experts"), and 20 top generalist forecasters ("superforecasters"), all of whom completed the survey between March 4 and March 29, 2026.

We asked respondents to forecast the total expected financial damages from large-scale human-caused outbreaks (outbreaks leading to at least $100 million in total worldwide damages) occurring between April 2026 and the end of 2028, and how this value would change across four scenarios: a world without general-purpose AI (GPAI); a world where no GPAI models had any safeguards; a world where all GPAI models were protected by ASL-3; and a world where frontier GPAI models were protected by ASL-3.

Because we did not include a status-quo scenario—one in which many frontier models already carry their developers' own misuse protections—our estimates do not capture the marginal benefit of extending ASL-3 beyond what other developers' existing safeguards already provide. We also asked respondents to assume that the capabilities of GPAI models do not improve before December 31, 2028. We made these stipulations because we wanted respondents' views on ASL-3's ability to mitigate harm from current models, rather than their views on other companies' safeguards or likely AI progress.

Key Takeaways

- Experts and superforecasters generally believe that implementation of ASL-3 safeguards across all general-purpose AI (GPAI) models can substantially reduce the risk of large-scale, human-caused outbreaks, relative to a world with no safeguards. Under the 'no safeguards' scenario, the median forecasted probability of human-caused outbreaks, occurring between April 2026 and December 2028, causing at least $100 million in damages was 16% (95% CI: 9.2%, 32.8%). This dropped to 9.2% (95% CI: 4% , 16.9%) under the 'ASL-3 for all models' scenario.
- Forecasters estimate that ASL-3 safeguards, if implemented across all GPAI models, would capture most of the risk reduction that is possible to achieve via safeguarding. For the median respondent, implementing ASL-3 across all GPAI models achieved 71.7% (95% CI: 65.7%, 97.2%) of the risk reduction associated with a hypothetical scenario where no GPAI exists.
- Most respondents believe protecting frontier models (those at least as capable as Sonnet 4.5) with ASL-3 would substantially reduce risk relative to a 'no safeguards' baseline, although not by as much as extending it to all GPAI, regardless of model

size or capability. For the median respondent, 'ASL-3 for frontier models' achieved 36.3% (95% CI: 22.9%, 68.4%) of the risk reduction associated with the 'no GPAI' scenario.

- The median respondent thought that variations to ASL-3—making jailbreaks harder to find, patching them faster, or further degrading the performance of jailbroken models—would reduce risk by an amount similar to applying ASL-3 to all GPAI models. The largest median risk reductions were associated with patching jailbreaks 5x faster and with a further 30% drop in jailbroken models' performance relative to non-jailbroken models. For the median respondent, each of these two variants achieved 73.5% of the risk reduction associated with a hypothetical scenario where no GPAI exists (95% CI: 52.3%, 100% for faster patching; 47.8%, 97.3% for reduced jailbroken performance). Many respondents noted that such variations may matter more in a world where ASL-3 covers all GPAI models, and many saw trusted user exemptions as a potential weak spot in ASL-3.

Key Results

ASL-3 safeguards are seen as having high efficacy: if applied across all GPAI models, respondents estimate that ASL-3 would capture most of the risk reduction achievable through safeguarding.

- When respondents assumed all models had ASL-3 protections, the median forecasted probability of human-caused outbreaks causing at least $100 million in damages dropped by roughly 40% relative to the 'no safeguards' scenario (from 16% (95% CI: 9.2%, 32.8%) to 9.2% (95% CI: 4%, 16.9%)). For comparison, when asked to consider a hypothetical world where GPAI was never developed, the median forecast was 9.3% (95% CI: 3.8%, 13.2%). (See Figure 1.)

- Regarding expected financial damages, the result was similar, although the 'no GPAI' scenario was generally associated with lower damage estimates than the ASL-3 scenarios. Compared to a world with no safeguards, a world where all models have ASL-3 safeguards reduced median predicted financial damages of human-caused outbreaks between now and 2028 by 40% (95% CI: 17%, 63%). Under the 'no GPAI' scenario, the median participant reduced their forecast of expected damages by 49% (95% CI: 19%, 71%).

- Overall, the forecasts suggest that respondents see 'ASL-3 for all models' as reducing risks of a large-scale human-caused outbreak to a risk level slightly above a hypothetical world with no GPAI. The median respondent's forecasts suggest that 'ASL-3 for all models' is associated with a reduction in risk that is 71.7% (95% CI: 65.7%, 97.2%) of the risk reduction associated with the 'no GPAI' scenario.

- Rationales showed that respondents generally believed that ASL-3 would effectively block non-experts. Many respondents were reassured by the pace of jailbreak patching and the difficulty in finding jailbreaks, suggesting that this likely requires substantial cybersecurity expertise. Several also noted the deterrence effects of safeguards.

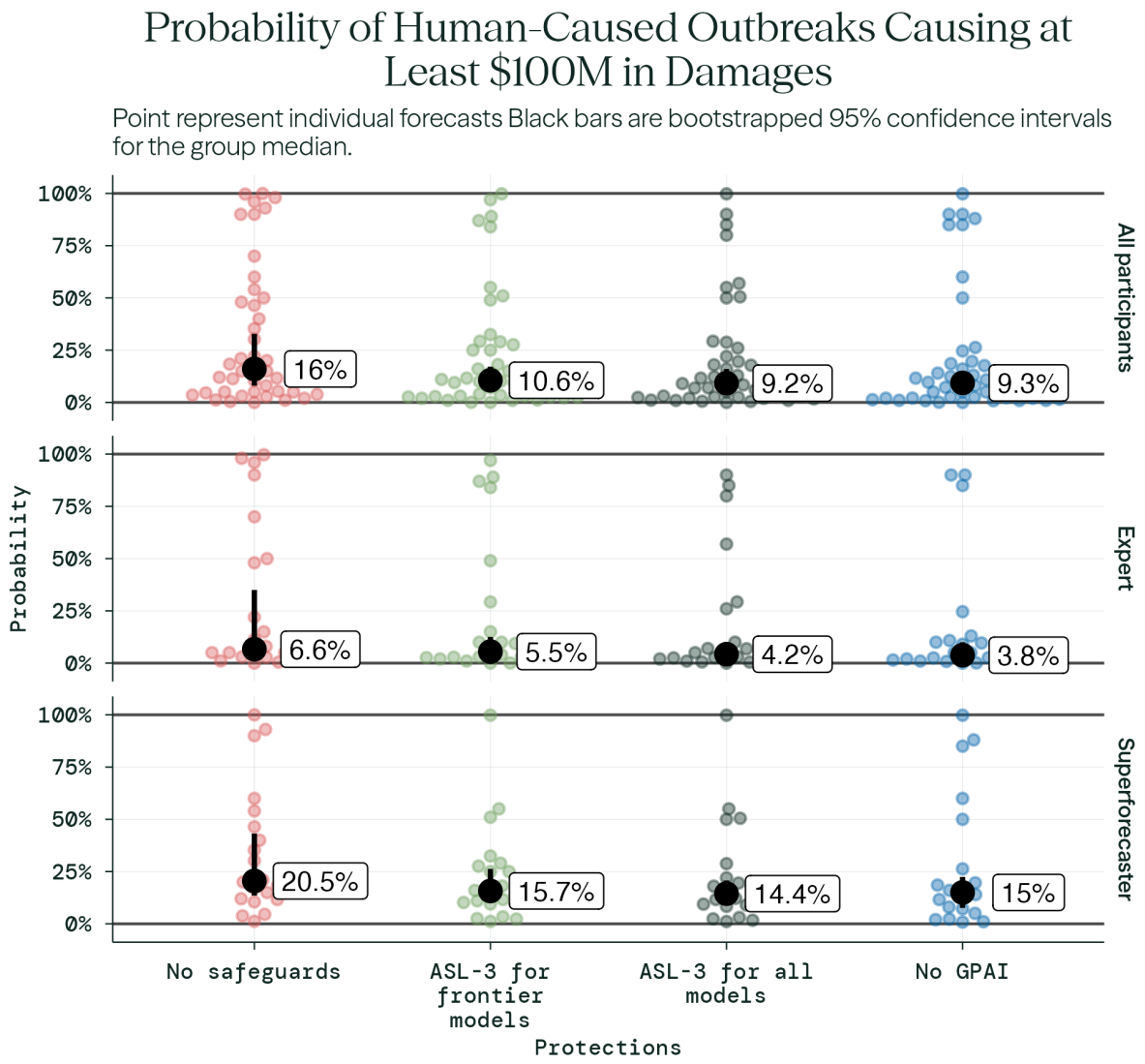


Figure 1: *Probability of human-caused outbreaks that start between April 2026 and December 2028, causing at least $100 million in economic damages. Numbers show medians from each group and black lines show their bootstrapped 95% confidence intervals.*

Real-world impact likely depends on breadth of adoption: protecting only frontier models still reduces risk substantially, but less than comprehensive coverage.

- We also asked how respondents' forecasts would change if ASL-3 were only applied to 'frontier' GPAI models, which we defined as those at least as capable as Sonnet 4.5. Under this scenario, the median respondent forecast for the probability of human-caused outbreaks starting between now and 2028, causing at least $100 million in damages, was 10.6% (95% CI: 5.5%, 18.1%). The median estimate of expected damages was 20% (95% CI: 9%, 52%) lower under this scenario than under the 'no safeguards' scenario.
- For the median participant, the expected damages in the 'ASL-3 for all models' scenario was 4% (95% CI: 1%, 16%) lower than the expected damages in the 'ASL-3 for frontier models' scenario.

- Most respondents placed the frontier-only ASL-3 scenario's risk between 'ASL-3 for all models' and 'no safeguards', but disagreed on where, with some treating it as nearly equivalent to 'ASL-3 for all models' and others as only marginally better than 'no safeguards'. Much of this disagreement hinged on the capability and accessibility of near-frontier models, particularly Chinese open-weight models.

Making jailbreaks harder to find, patching them faster, or further degrading the performance of jailbroken models were generally expected to improve ASL-3 performance, with faster patching expected to be the most impactful.

- We asked respondents to consider six variants of ASL-3 safeguards that differed from the current safeguards' performance in one of three ways: time required to find a jailbreak, time taken to patch a jailbreak, and performance of jailbroken models relative to the base model. For the median respondent, each of the variations associated with improvement—making jailbreaks harder to find, patching them faster, or further degrading the performance of jailbroken models—was expected to reduce risk to a greater extent than the 'ASL-3 for frontier models' scenario.

- The largest median deviation from the 'ASL-3 for frontier models' expected damages baseline was an 8% decrease (95% CI: 5%, 18%), associated with a scenario of a 5x decrease in the time taken to patch a jailbreak. The median respondent's forecasts suggest that ASL-3 for frontier models with 5x faster patching is associated with a reduction in risk that is 73.5% (95% CI: 52.3%, 100%) of the risk reduction associated with the 'no GPAI' scenario.

ASL-3 shifts the threat landscape but does not address all major risk pathways.

- Respondents expected ASL-3 to disproportionately screen out non-expert and lone-wolf actors, shifting the relative probability of damages toward state actors and accidental releases from laboratories.

- Superforecasters saw trusted user exemptions as a notable vulnerability, with some noting that they combine AI access with proximity to the physical infrastructure needed for an attack and exploit human factors that technical safeguards cannot fully address. Many respondents emphasized that lab accidents represent a substantial share of baseline human-caused outbreak risk that AI safeguards do not address.

1. Introduction

In May 2025, Anthropic announced that it had activated AI Safety Level 3 (ASL-3) Deployment and Security Standards as described in Version 2 of Anthropic's Responsible Scaling Policy (RSP) [1]. This activation came in conjunction with the launch of Claude Opus 4, and was described as a precautionary measure. Anthropic determined that due to continued improvements in CBRN-related knowledge and capabilities, it was not possible to rule out that Claude Opus 4 had crossed its CBRN-3 capability threshold. This threshold is defined in the RSP as: "The ability to significantly help individuals or groups with basic technical backgrounds (e.g., undergraduate STEM degrees) create/obtain and deploy CBRN weapons" [2].

Anthropic's announcement noted that ASL-3 defenses would be focused on biological weapons risks. The possibility that AI could enable more actors to develop biological weapons has been raised by heads of AI companies and experts in biosecurity [3, 4, 5]. There is debate about the significance of the risks, with some believing that AI does little to overcome the key barriers to biosecurity risks [6, 7]. Prior work has found that most experts and top generalist forecasters expect biological risks to increase with AI advances in AI capabilities [8].

The ASL-3 Deployment Standard consists of requirements across the following areas: threat modeling, defense in depth, red-teaming, rapid remediation, monitoring, access controls for trusted users, and third-party environments. To meet these requirements, Anthropic implemented the following mitigations: real-time classifier guards, offline monitoring, access controls, a bug bounty program, threat intelligence, and rapid response. Further detail on Anthropic's implementation of ASL-3 Deployment and Security Standards is available in Appendix A.

Although Anthropic can evaluate aspects of its mitigations (for example, the time taken to patch an identified jailbreak), it is unclear how well these measures achieve the overarching goal of reducing biosecurity risks posed by general-purpose AI (GPAI) models. To investigate this, we conducted a forecasting survey of experts in biosecurity, national security, or terrorism studies, as well as top generalist forecasters. Respondents were given information on Anthropic's ASL-3 standards and the measures taken to meet them. This included confidential data from red-teaming efforts, the bug bounty program, and other safeguards testing conducted by Anthropic. Although respondents' forecasts included consideration of both the Deployment and Security Standards that make up ASL-3, the focus of this survey was the Deployment Standards.

This survey aimed to answer four research questions:

- How effective are ASL-3 safeguards?
- What would be the impact of more models being protected by ASL-3?
- How would variations in ASL-3 influence this impact?
- How do ASL-3 safeguards influence threat models related to biosecurity risks?

In this report, we briefly describe the survey methods before presenting the results of the survey that speak to each of these research questions.

2. Methods and participants

2.1 Survey contents

Description of the main outcome

The main outcome we asked participants to forecast was:

> *The total worldwide damages attributable to human-caused outbreaks that start between April 1, 2026, and December 31, 2028, and that each individually cause at least $100 million in total worldwide damages (in 2026 USD).*

We asked respondents to make a few important assumptions when answering this question. First, we asked them to assume that until December 31, 2028, the capabilities of general-purpose AI models do not improve, as if a 'pause on general-purpose AI development' were in place. That is, for that period, no models outperform Claude 4.6, Gemini 3, or GPT-5, and this slowdown is not indicative of a more general slowdown in technological progress. We made these stipulations because we were interested in their views on ASL-3's ability to mitigate harm from current models, rather than their views on likely AI progress.

We asked respondents to forecast the global economic costs attributable to the outbreaks, beginning with the isolation of the pathogen and extending for the full life of an outbreak, and including direct economic losses and monetized mortality. We asked participants to use a value of statistical life of $2.2 million, regardless of geographic location. For full details of this question, including the resolution criteria, please see Appendix B.

Description of scenarios

- **No GPAI**: In this scenario, respondents were asked to assume that GPAI models never existed and will not exist. They were asked to assume that all other features of the world remain the same, e.g., they should not assume that the absence of GPAI models indicates that technological progress is slower across all domains.
- **No safeguards**: In this scenario, respondents were asked to assume that no safeguards are applied to all GPAI models. In other words, to assume that any system safeguards, refusal behaviors, and other safety-focused constraints have been removed or disabled, and that such AI models are optimized for helpfulness (efficient and accurate task performance) without safety protections.
- **ASL-3 safeguards for all GPAI models**: In this scenario, respondents were asked to assume that all GPAI models—regardless of size or capability—are protected by Anthropic's ASL-3 safeguards.
- **ASL-3 safeguards for frontier GPAI models:** In this scenario, respondents were asked to assume that "frontier" GPAI models are protected by Anthropic's ASL-3 safeguards. "Frontier" GPAI models were defined as all those performing as well or better than Claude Sonnet 4.5 on the Epoch Capabilities Index (ECI) [9]. At the time of the survey, Claude Sonnet 4.5 scored 147 on the ECI, and 13 models scored

at or above that level: Gemini 3 Pro, GPT-5.2, Claude Opus 4.6, Gemini 3 Flash, GPT-5 Pro, Claude Opus 4.5, GPT-5, GPT-5.1, o3-pro, Kimi K2.5, Grok 4, o3, and Claude Sonnet 4.5.

Additional questions

In addition to forecasts of the main outcome, we asked questions to understand respondents' views on pathways to human-caused outbreaks, and how these are influenced by ASL-3. Specifically, we asked respondents to assume that a human-caused outbreak that caused more than $100 million in damages had occurred, and to say what probability they placed on different types of actors being the primary cause of the outbreak. We asked participants to answer these questions for two scenarios: 'no safeguards' and 'ASL-3 for all GPAI models'. For the 'ASL-3 for all GPAI models' scenario, we also asked about the probability that a GPAI model had been used in an unauthorized way to assist the process of causing the outbreak, and the most likely pathways to unauthorized use. We also asked participants for some basic details on their demographics and expertise.

Calibration modules

Before completing any forecasting questions, participants completed two short calibration modules. One module asked participants to estimate low probabilities. An example question is, "What is the probability of being struck by lightning in any year?" (Answer: 8.1e-7). The other module asked for values relevant to the survey subject matter. An example question is, "In the 1984 Rajneeshee bioterror attack in Oregon (the largest bioterrorism attack in US history), how many people were infected with Salmonella?" (Answer: 751). The primary purpose of these modules was to give participants an opportunity to test their own calibration.

2.2 Data Analysis

Data analysis was conducted using R after aggregating and cleaning the data submitted by participants. For details of data cleaning, see Appendix B.

For the main outcome, we asked participants to forecast the probability that human-caused outbreaks would lead to different levels of financial damage by 2028. We then calculated expected financial damages for each scenario. To calculate these values, we assumed that damages within each bin were all equally likely (uniformly distributed). We then multiplied the probability assigned to each range by the midpoint dollar amount of that range and summed these probability-weighted contributions across all bins to obtain the overall value of expected damage. For the highest bin (more than $100 trillion), we used $1,000 trillion as the upper bound for calculations.

This approach could be a limitation, especially for wider damage bins or upper-tail outcomes where damages may not be evenly distributed within a range. To reduce the impact of this issue, we used a relatively large number of bins and had participants review the calculated expected damages. Participants were asked to adjust their probability forecasts if the expected damages calculation did not align with their expectations, so these values reflect calculations that participants reviewed and had the opportunity to revise.

To describe forecasts on the main outcome, we report several metrics: the probability that damages from human-caused outbreaks, occurring from now until the end of 2028, will be at least $100 million; the ratio of calculated expected damages values between different scenarios; and a metric we define as "share of GPAI-attributable risk mitigated".

We define "GPAI-attributable risk" in relative terms as the proportional increase in expected damages between the 'no safeguards' scenario and the 'no GPAI' scenario. Formally, we define this as

$$GPAI - attributable\ risk = \frac{E_{no\ sg}}{E_{no\ AI}} - 1,$$

where ($E_{no\ AI}$) denotes expected damages with no GPAI and ($E_{no\ sg}$) denotes expected damages with no safeguards.

This represents our estimate of the additional risk posed by the existence of unsafeguarded GPAI models. We then measure the share of this GPAI-attributable risk that is mitigated by ASL-3 safeguards by comparing expected damages under the 'no safeguards' scenario to those under the relevant ASL-3 scenario (E_s). Specifically, we compute

$$Proportion\ of\ risk\ removed = \frac{\frac{E_{no\ sg}}{E_s} - 1}{\frac{E_{no\ sg}}{E_{no\ AI}} - 1}.$$

This quantity captures the fraction of the proportional increase in expected damages attributable to unsafeguarded GPAI that is eliminated under ASL-3 safeguards. Although we refer to this quantity as a "share" of risk mitigated, it is more precisely a normalized relative-risk reduction measure based on multiplicative differences in expected damages, rather than an additive share of the total risk.

A caveat applies to this definition: the 'no GPAI' scenario removes not only the misuse risks of GPAI but also any defensive benefits (e.g., contributions to pandemic preparedness and response). A 'no GPAI' world therefore has higher baseline outbreak risk than a hypothetical world with perfectly safeguarded AI would, making it a more lenient benchmark. The share mitigated by ASL-3, measured against this benchmark, is best interpreted as an upper bound.

A further caveat applies to a small subset of four participants. These participants were excluded from aggregate calculations and figures of this metric because they assessed expected damages to be lower in the presence of AI than in the 'no GPAI' scenario. In these cases, the calculations of the share of GPAI-attributable risk reduced do not yield interpretable measures since their baseline implies that AI itself reduces risk.

We focus on the relative change in risk because, given the difficulty of forecasting expected damages, we believe these relative changes are likely to be more reliable than the absolute values of damages. However, we present the details of absolute values in Appendix C.

2.3 Participants

The results represent responses from a total of 42 participants: 22 participants were recruited for their expertise in biosecurity, national security, or terrorism studies (hereafter, "experts"), and 20 participants were recruited as top generalist forecasters (hereafter, "superforecasters"). Participants completed the survey between March 4 and March 29, 2026. On average, participants spent 6 hours completing the survey.

The median expert participant reported 11 years of experience relevant to the survey. The expert sample was recruited across three domains — biosecurity, national security, and terrorism studies — but most experts had experience spanning multiple domains. Although it was not a requirement for their participation in the study, seven of the superforecaster participants reported experience in at least one of the biosecurity, national security, or terrorism studies. See Appendix B for details of recruitment.

More than half (55%) of the expert participants reported having been granted access to restricted national security information (now or in the past). Only 20% of superforecasters reported the same. Most participants reported using LLMs regularly (at least once a week). More details on the participants are available in Appendix C.

3. How effective are ASL-3 safeguards?

To understand participants' views on the effectiveness of ASL-3 at reducing biosecurity misuse risks of GPAI models, we can first review the difference in the probability of human-caused outbreaks occurring between now and the end of 2028 and causing at least $100 million in damages under the different scenarios. Figure 1 in the Summary shows the probability of this outcome under the four scenarios. Figure 2 shows the relative reduction in the probability of this outcome under the 'no GPAI' and ASL-3 scenarios, relative to 'no safeguards'. For the median respondent, the 'no GPAI' scenario is associated with the largest reduction in risk [39% (95% CI: 16%, 50%)], but this is only slightly greater than the risk reduction associated with the 'ASL-3 for all models' scenario [32% (95% CI: 15%, 41%)].

Figure 3 shows how the calculated expected damages under the two ASL-3 scenarios and the 'no GPAI' scenario compare to the 'no safeguards' scenario. Although there was substantial variation, most believed that ASL-3 protections for all GPAI models would significantly reduce the expected damages from large-scale human-caused outbreaks, relative to a world where GPAI had no safeguards. This was particularly true for the expert respondents. The median respondent's forecasts suggested that ASL-3 protecting all models would reduce expected damages by 40% (95% CI: 17%, 63%).

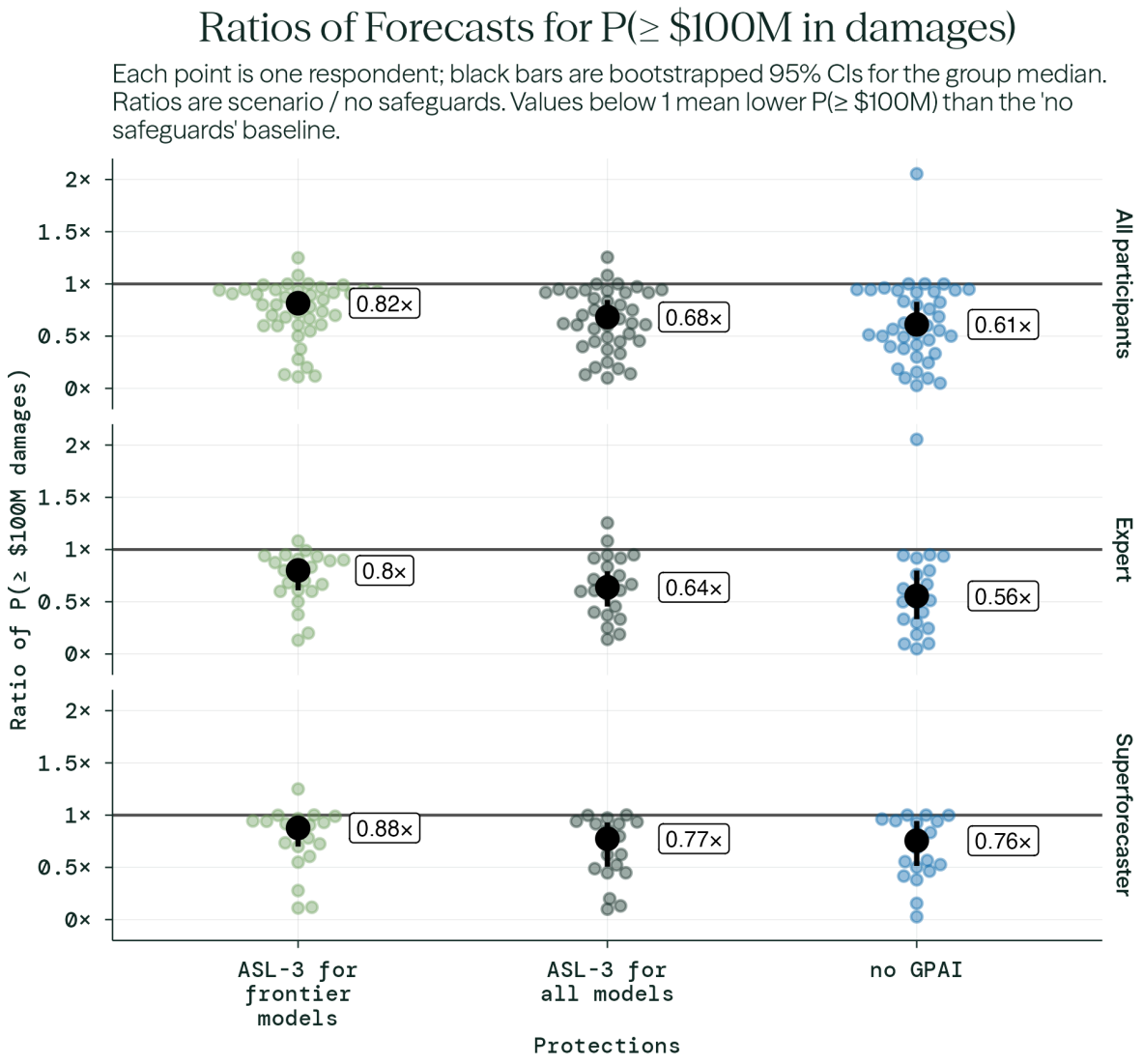


Figure 2: *Ratio of forecasts of probability of human-caused outbreaks that start between April 2026 and December 2028, causing at least $100 million in economic damages, relative to 'no safeguards' scenario. Numbers show medians from each group and black lines show their bootstrapped 95% confidence intervals. Values below 1 indicate lower P(≥$100M) than the no-safeguards scenario.*

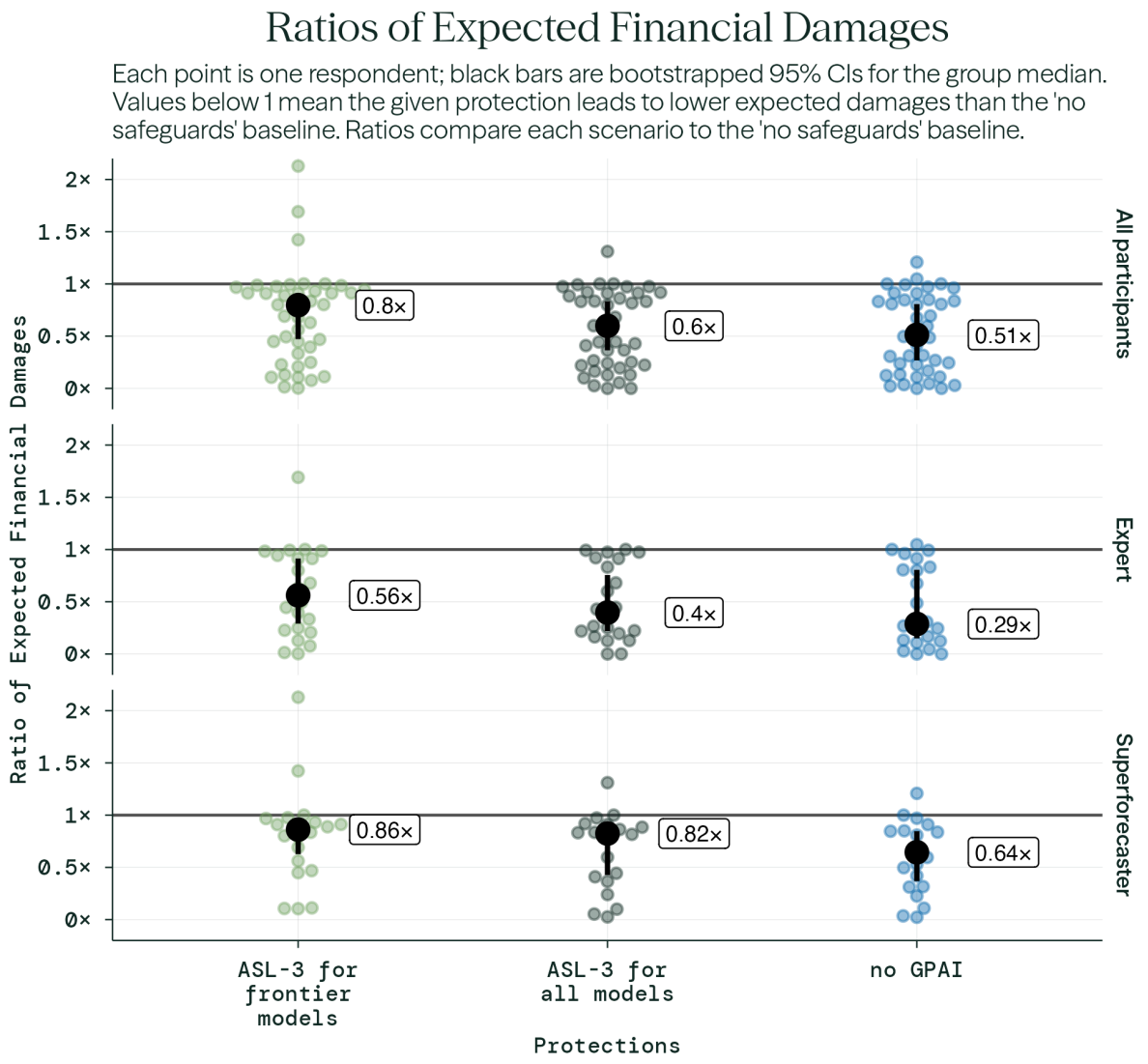
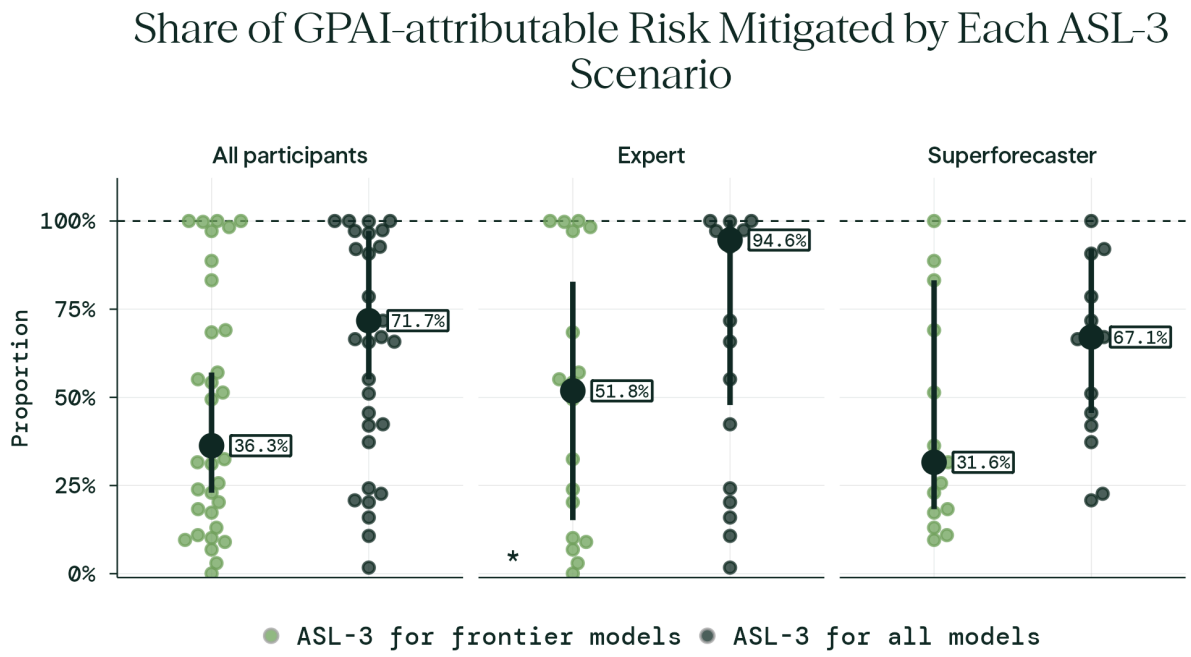


Figure 3: *Ratio of expected damages for each protection scenario compared to the 'no safeguards' baseline. Numbers show medians from each group and black lines show their bootstrapped 95% confidence intervals. Values below 1 indicate lower expected damages than the no-safeguards scenario.*

GPAI-attributable risk

The scenario that asked respondents to assume GPAI models did not exist can provide an upper bound on the risk reduction that could be achieved by safeguards against GPAI model misuse, as it demonstrates the risk of human-caused outbreaks that would persist regardless of GPAI models. We can also express these results in terms of the share of GPAI-attributable risk mitigated by ASL-3. Aggregates presented here exclude four participants whose responses imply that GPAI reduces risk, making this metric uninterpretable for them. See Section 2.2 for a definition of the metric and more detail on the filtering. The median participant's forecasts imply that ASL-3 for all models would mitigate 71.7% (95% CI: 65.7%, 97.2%) of GPAI-attributable risk. Figure 4 shows this comparison.



* One expert believed that ASL-3 for frontier models only would increase expected financial risk relative to all other scenarios (including no safeguards and no GPAI), so their share of the mitigable gap is negative and is not shown.

Figure 4: *Share of GPAI-attributable risk mitigated by each ASL-3 scenario. GPAI-attributable risk is defined as the multiplicative gap in expected damages between the 'no safeguards' and 'no GPAI' scenarios. Numbers show medians from each group and black lines show their bootstrapped 95% confidence intervals.*

An important limitation of this comparison deserves emphasis. The 'no GPAI' scenario serves as a proxy for perfect safeguards, but it is an imperfect one. Perfect safeguards would eliminate misuse risk while preserving the defensive benefits of GPAI: its contributions to pandemic surveillance, pathogen characterization, drug and vaccine development, and outbreak response. The 'no GPAI' scenario removes all of these. This means the 'no GPAI' world is likely more vulnerable to outbreaks (including human-caused ones) than a world with perfectly safeguarded GPAI, making it a lenient benchmark.

The practical consequence is that the 71.7% efficacy figure—the share of maximum risk reduction achieved by ASL-3 for all models—is likely biased upward. If GPAI's defensive contributions are small relative to the misuse risk, the bias is minor. If they are large, the bias could be substantial. We did not directly elicit respondents' views on the magnitude of GPAI's defensive benefits, though several rationales mentioned them. For instance, some respondents noted that GPAI could accelerate the development of medical countermeasures or improve biosurveillance. Future work could address this limitation by eliciting forecasts under a 'perfect safeguards' scenario directly—one where GPAI exists and provides defensive benefits but cannot be misused—or by asking respondents to separately estimate the defensive and offensive contributions of GPAI to outbreak risk.

Insights from rationales[1]

Aligning with quantitative results, text rationales showed that most respondents believed comprehensive ASL-3 protection would meaningfully reduce risk relative to the helpful-only scenario. There was broad agreement that ASL-3's greatest value is in blocking non-expert and lone-wolf actors. Many respondents were reassured by the difficulty of finding jailbreaks and the pace of jailbreak patching. Several noted the deterrence effect of safeguards. On the other hand, multiple respondents noted that ASL-3 does nothing to reduce accidental releases, and one argued that ASL-3 would increase risk by weakening collective defensive capabilities.

> "With safeguards in place, the risk of a human-caused outbreak should be around the same as with no GPAI models around at all. In particular, the safeguards should significantly reduce the likelihood that an untrained non-expert would use an AI system to create a dangerous pathogen."

> "Given the available information, it doesn't seem like the ASL-3 security standard is a panacea. The patching of jailbreaks is slower than I thought. … It seems rather unbalanced that only $26k are awarded to an individual hacker for discovering a bug that, not only could cost the company millions (plus the potential negative publicity), but also requires so much effort to fix. There's probably a black market in the dark web for these universal jailbreaks, and I wouldn't be surprised if the pay is competitive."

Most respondents viewed the helpful-only ('no safeguards') scenario as increasing biorisk, but physical bottlenecks—not information access—were seen as the primary constraint on bioweapons development. Many cited the Active Site study [10] — which found that helpful-only LLM access didn't aid participants much with practical virology tasks relative to internet-only access — and stressed that access to labs and equipment, along with crucial hands-on skills, would remain the real constraints. Others, however, weren't so sanguine, arguing that current frontier models can produce detailed technical guidance and that stripping away refusal behaviors makes that knowledge accessible to a much wider set of actors. Several challenged the relevance of the Active Site study, noting it used older models, and identified less direct pathways to increased risk: more people being inspired to try, lab accidents driven by AI-enabled overconfidence, and the volatile mix of helpful-only models with ongoing geopolitical conflicts.

> "Active Site's recent work on this topic showed that overall LLM-aided participants performed fairly similarly to internet-only in a pseudo-virological workflow. This also fits my strong prior, having worked in laboratory science/bio for 7 years, that converting theoretical knowledge and/or text-based protocols into actual physical actions in the real world is most efficiently and accurately performed with real-time human expert guidance…Until we either have AI-powered VR lab training or completely closed lab-in-the-loop type setups for the full stack of virology/bacteriology workflows, a 2026-level LLM isn't going to aid the lone-wolf DIY bio actor significantly over and above the Internet."

[1] Participant rationales quoted in this report are included as originally shared and have not been edited or fact-checked. We do not necessarily endorse the views or information expressed.

> "Recent studies have shown that the bottleneck is not the AI. It is the lab equipment and more importantly the implicit knowledge of how to do protocols…Conceptually this is similar to the nuclear field. Although access to the core material on how to create a nuclear bomb is widely available…the practical knowledge and tooling has proven to be a barrier for all but the most determined nation states."

Lab accidents were often identified as the most likely source of human-caused outbreaks and largely independent of AI conditions. Some respondents pointed to the proliferation of BSL-3/4 facilities globally, ongoing gain-of-function research, and variable biosafety standards as key drivers of lab accidents.

> "Laboratory accident risk reflects global BSL-3/4 infrastructure expansion—approximately 70 BSL-4 and 1,800 BSL-3 facilities [outside of the U.S.] operational, with substantial growth in China, India, and developing nations with variable biosafety and biosecurity standards (mostly substandard)."

Much of the variation in baseline forecasts of expected damages may have been driven by which historical incidents respondents counted when assessing a base rate for human-caused outbreaks. The 2001 anthrax attacks served as a common anchor, but beyond that, the incidents considered diverged. The breadth of each respondent's historical audit was a strong predictor of their final base rate, with a wider breadth being associated with a higher base rate.

> "When I look at the incidents post anthrax (600M estimated damages), I see several incidents that would also have likely crossed the 100M threshold given the way damages are calculated…Claude estimates (in 2026 USD) that the SARS lab escapes was 80-360M in damages, UK foot and mouth (escaped from the Pirbright laboratory complex) at 600-900M, Lanzhou Brucellosis at 60-250M, and then there's COVID-19, which call it 25% chance of lab leak. So past 25 years, 4 confirmed incidents that probably hit the 100M-1B bin given the method of counting damages, and a 25% chance that there was an incident that hit the 10T-100T bin. All told, a higher base rate than I thought I'd find."'

4. What would be the impact of more models being protected by ASL-3?

It is implausible that all GPAI models (regardless of size or capability) would be protected by ASL-3 safeguards. It seems more realistic for all models at the frontier of capabilities to be protected by ASL-3 or equivalent safeguards, even if that is not the case today. To understand the potential impact of all frontier models being protected by ASL-3 safeguards, we asked respondents to forecast the main outcome under this scenario.

We defined "frontier" GPAI models as all those performing as well or better than Claude Sonnet 4.5 on the Epoch Capabilities Index (ECI). The ECI combines scores from more than 40 different AI benchmarks into a single general capability score [9]. At the time of the survey, Claude Sonnet 4.5 scored 147 on the ECI, and 13 models scored at or above

that level: Gemini 3 Pro, GPT-5.2, Claude Opus 4.6, Gemini 3 Flash, GPT-5 Pro, Claude Opus 4.5, GPT-5, GPT-5.1, o3-pro, Kimi K2.5, Grok 4, o3, and Claude Sonnet 4.5.

Under the 'ASL-3 for frontier models' scenario, the median probability of human-caused outbreaks occurring between now and the end of 2028 and leading to at least $100 million in damages was 10.6% (95% CI: 5.5%, 18.1%). Compared to the 'no safeguards' scenario, the 'ASL-3 for frontier models' scenario was associated with an 18% (95% CI: 9%, 30%) reduction in this risk. The median respondent thought that 'ASL-3 for frontier models' only would be associated with a 20% (95% CI: 9%, 52%) reduction in expected damages due to large-scale human-caused outbreaks relative to the 'no safeguards' scenario (see Figure 3). ASL-3 for frontier models mitigated 36.3% of GPAI-attributable risk (see Figure 4). Most respondents placed the 'ASL-3 for frontier models' scenario's risk between 'ASL-3 for all models' and 'no safeguards', but disagreed on where, with some treating it as nearly equivalent to 'ASL-3 for all models' and others as only marginally better than 'no safeguards'.

Insights from rationales

Forecasters disagreed on the degree to which powerful near-frontier models—particularly Chinese ones—would undermine frontier-only ASL-3 coverage. In their text rationales, some emphasized that frontier models are where the real capability uplift lies, with some citing the Active Site study as evidence that unrestricted models would likely provide limited practical uplift for virology tasks. Some also noted that near-frontier models would likely retain some proprietary safeguards even without ASL-3. Others worried that a community of practice could develop around non-frontier models and that the mere knowledge of capable unsafeguarded models could lower the psychological barrier to attempting an attack.

> "Even if they don't have ASL-3 safeguards, the American models still have quite good safeguards. Unfortunately, this is not the case for most of the Chinese models though. So I think in this world, the easiest path to harm for someone with malicious intent would be to use a Chinese LLM to help them."

> "My guess here, which is strongly influenced by the Active Site study conducted in the summer of 2025, is that no uplift will be provided by models on the list that currently aren't classified as frontier."

5. How would variations in ASL-3 influence this impact?

We asked respondents to consider the impact of variations on ASL-3 in a scenario where frontier models are protected by ASL-3. Respondents were asked to forecast the main outcome conditional on frontier models being protected by the following hypothetical variations on the ASL-3 scenario:

Time required to identify jailbreaks

- A1: The average time required to find a universal jailbreak is 5x longer
- A2: The average time required to find a universal jailbreak is 5x shorter

Time to jailbreak patching

- B1: It takes a fifth (0.2x) of the time to patch all universal jailbreaks
- B2: It takes up to 2x as long to patch all universal jailbreaks

Capabilities loss associated with jailbreaks

- C1: All universal jailbreaks significantly reduce the model's GPQA score. Compared to jailbroken models' current performance relative to the "no jailbreak" baseline, jailbroken models' relative performance decreases by roughly 30%.
- C2: All universal jailbreaks only minimally reduce the model's GPQA score. Compared to jailbroken models' current performance relative to the "no jailbreak" baseline, jailbroken models' relative performance increases by roughly 30%.

Figure 5 shows the median change in expected damages relative to 'ASL-3 for frontier models' under the variant scenarios. For each of the scenarios associated with improvement from current ASL-3 safeguards, the median expected change was similar to that associated with 'ASL-3 for all models'. The largest median change—an 8% (95% CI: 5%, 18%) decrease from the 'ASL-3 for frontier models' scenario—was associated with a faster time to jailbreak patching (B1).

Relative Expected Financial Damages: ASL-3 for All Models and Frontier Variants

Medians of individual ratios of expected damages (scenario / ASL-3 for frontier models). Includes ASL-3 for all models and each frontier variant. 1× means the same expected damages as ASL-3 for frontier models.

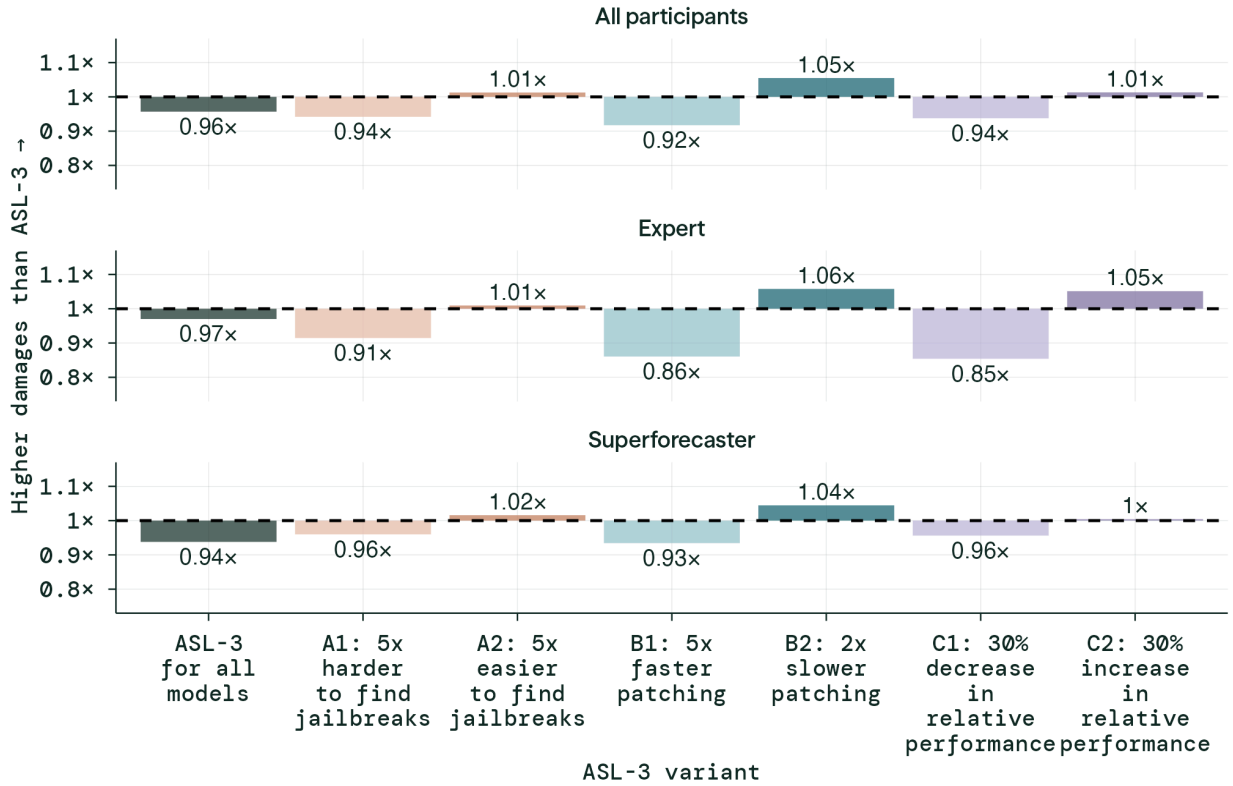


Figure 5: *Relative difference in expected financial damages under each variant of ASL-3 and the 'ASL-3 for all models' scenario. Numbers show medians from each group. Values below 1 represent a decrease in expected damages compared to a scenario with ASL-3 for frontier models.*

Figure 6 shows the share of GPAI-attributable risk mitigated by the variant scenarios for frontier models, 'ASL-3 for frontier models', and 'ASL-3 for all models' scenarios. For the median respondent, the '5x faster patching' (B1) and '30% decrease in relative performance' (C1) scenarios were seen as comparable to the 'ASL-3 for all models' scenario on this metric. Relative to superforecasters, experts expected variants to ASL-3 to have a greater impact.

More details are available in Appendix C.

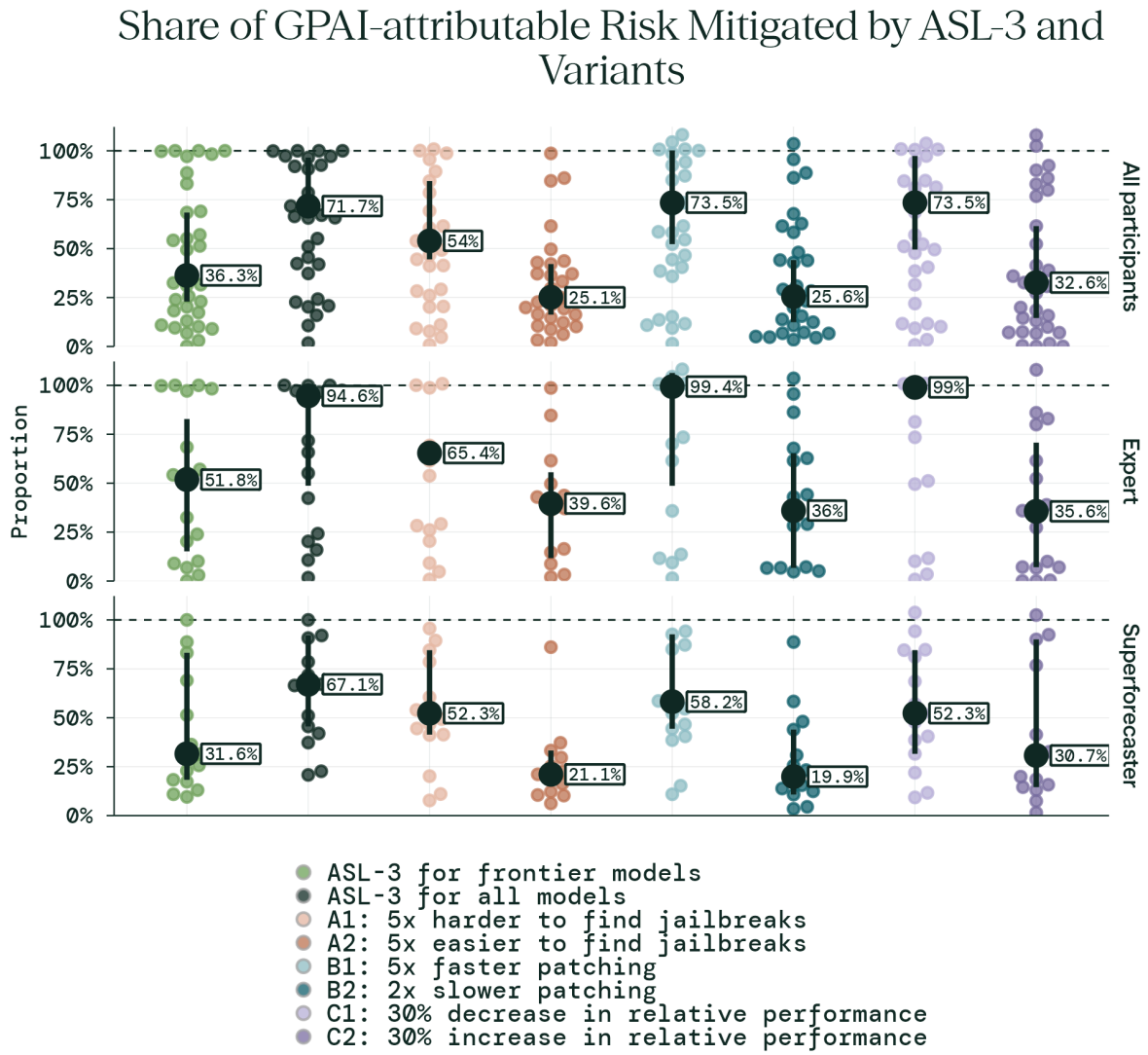


Figure 6: *Share of GPAI-attributable risk mitigated by each ASL-3 scenario. GPAI-attributable risk is defined as the multiplicative gap in expected damages between the 'no safeguards' and 'no GPAI' scenarios. Numbers show medians from each group and black lines show their bootstrapped 95% confidence intervals.*

Insights from rationales

Several respondents expressed concern that if non-frontier models remain unprotected and are nearly as capable, then making frontier jailbreaks harder, faster to patch, or more capability-degrading has limited marginal value.

Most respondents agreed that making jailbreaks harder to find would help, but several argued that it would mainly screen out under-resourced actors who were unlikely to succeed anyway. Many also noted that more organized groups would be able to adapt. Making jailbreaks easier to find was generally assessed to be more dangerous because it would expand access to a much larger pool of actors, however unsophisticated. A few respondents noted the interaction of this variant with patching: discovery time is largely irrelevant if jailbreaks persist for months once found. One flagged that AI agents capable of autonomous software research could erode the barrier further.

Forecasters noted that faster patching would more effectively disrupt the months-long efforts required to develop a bioweapon. Credible bioweapon efforts typically involve iterative troubleshooting over weeks or months and that fast patching could disrupt this workflow by closing the window before a project could be completed—and could also reduce the economic incentive to discover and sell jailbreaks. Slow patching was predicted to do the opposite in that it could make jailbreaks more useful and give bad actors ample time to extract everything they need.

Forecasters differed in opinion on the impact of model performance degradation. Some respondents argued that a significant degradation would eliminate the incentive to pursue jailbreaks, since the degraded model would offer little advantage over unprotected non-frontier alternatives. Others argued that even a significantly degraded frontier model would still be remarkably capable by historical standards.

6. How do ASL-3 safeguards influence threat models related to biosecurity risks?

We asked respondents to answer several questions related to pathways to a large-scale human-caused outbreak. This included the type of actor that is most likely to cause such an event, the probability that unauthorized use of a GPAI model was involved in the event, and the most likely pathway to unauthorized model use.

6.1 Actor type

We asked what type of actor is the most likely primary cause of a human-caused outbreak that causes more than $100 million in damages, and how this would change depending on whether GPAI models had no safeguards or were all protected by ASL-3 safeguards. The mean responses for each group of participants are shown in Figure 7. Both groups of respondents thought that state actors were the most likely cause of such an event, and that they would be an even more likely cause under the ASL-3 safeguards scenario. Among the expert respondents, the next most likely group was individual expert actors, although this group's probability of being the primary cause of a large-scale human-caused outbreak fell under the ASL-3 safeguards scenario. Compared to experts,

superforecasters generally put a higher probability on a different type of actor (an "Other" category) being the primary cause. Respondents who put a large probability in this category suggested that groups of experts or non-experts that did not count as terrorist groups would fit into this category.

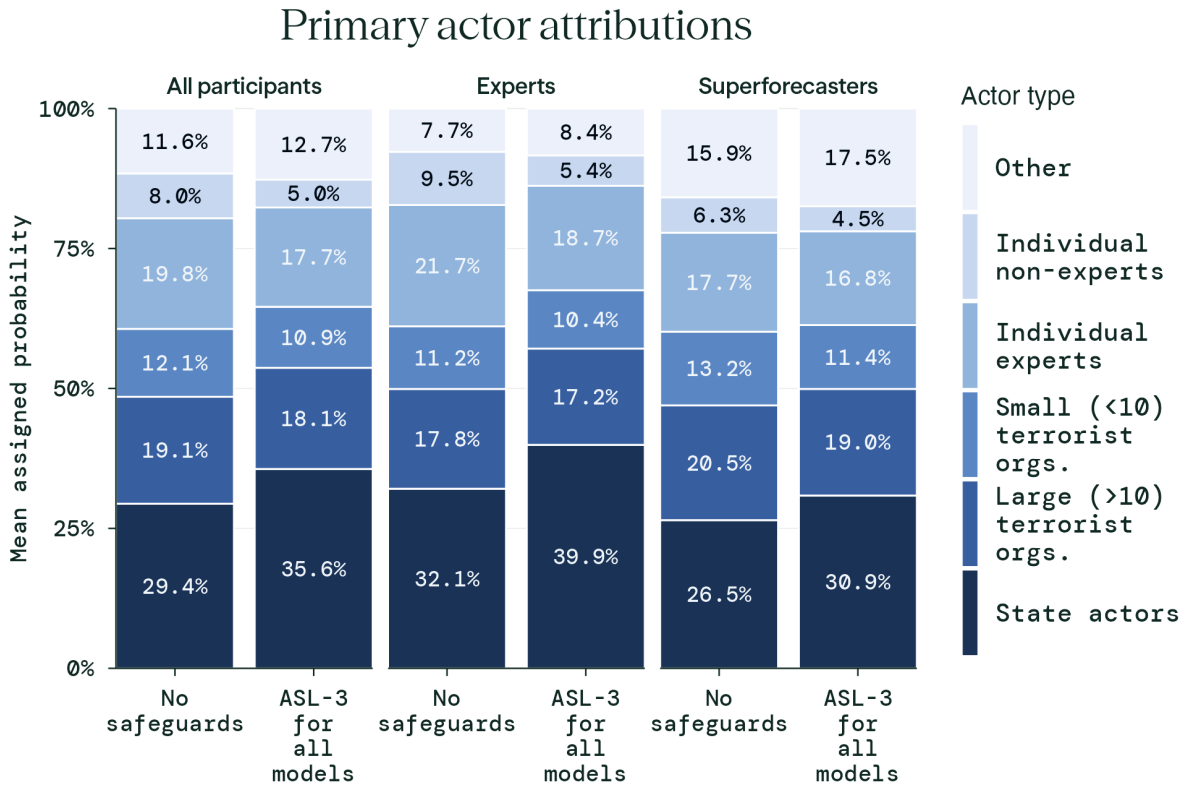


Figure 7: *Probability of the given actor being the primary cause of a large-scale human-caused outbreak. Numbers show group averages.*

Insights from rationales

Across both the 'no-safeguards' and 'ASL-3 for all models' conditions, a dominant theme in respondent rationales was a capability-intent asymmetry. State actors have the resources to develop bioweapons but face strong disincentives (e.g., attribution risk, self-harm), while non-state actors may have motivation but lack capability. Many respondents assign significant probability to accidental release pathways rather than deliberate attacks, shifting the expected actor profile toward individual experts and state labs. Without safeguards, AI is expected to partially close the capability gap for non-state actors, but practical barriers—such as wet lab access, weaponization, and delivery—remain and are largely unaffected by AI. Some argue that pursuing bioweapons is irrational for most actors given the availability of simpler alternatives, and that the "Other" category captures important scenarios (e.g., corporate negligence, criminal groups, insider threats, etc.) outside the listed actor types.

ASL-3 safeguards shift relative probability toward better-resourced actors. State actors and large terrorist organizations gain conditional share because they can overcome or bypass safeguards through in-house expertise, cyber capabilities, and sustained jailbreak efforts, while individual non-experts—the primary beneficiaries of unrestricted AI—lose the most. Jailbreak difficulty is the key mechanism: successful jailbreaking requires specialized cybersecurity skills (compounding the expertise requirement), jailbreaks are transient due to patching, and even successful jailbreaks may produce degraded outputs that less capable actors cannot independently verify.

With deliberate misuse pathways partially blocked, the accidental release pathway gains further relative prominence. Several respondents also flag trusted user exemptions (insiders with legitimate access who misuse it or whose credentials are compromised) as a new attack surface, and that such actors with reduced safeguards may develop over-dependence on AI, increasing the risk of accidental misuse or unintended consequences.

6.2 Unauthorized model use

When asked about the probability of unauthorized frontier model use contributing to a human-caused outbreak that causes more than $100 million in damages, the median expert forecasted a 30% (95% CI: 15 , 50) probability, and the median superforecaster a 23.5% (95% CI: 10.2, 52) probability. The distribution of these forecasts is shown in Figure 8.

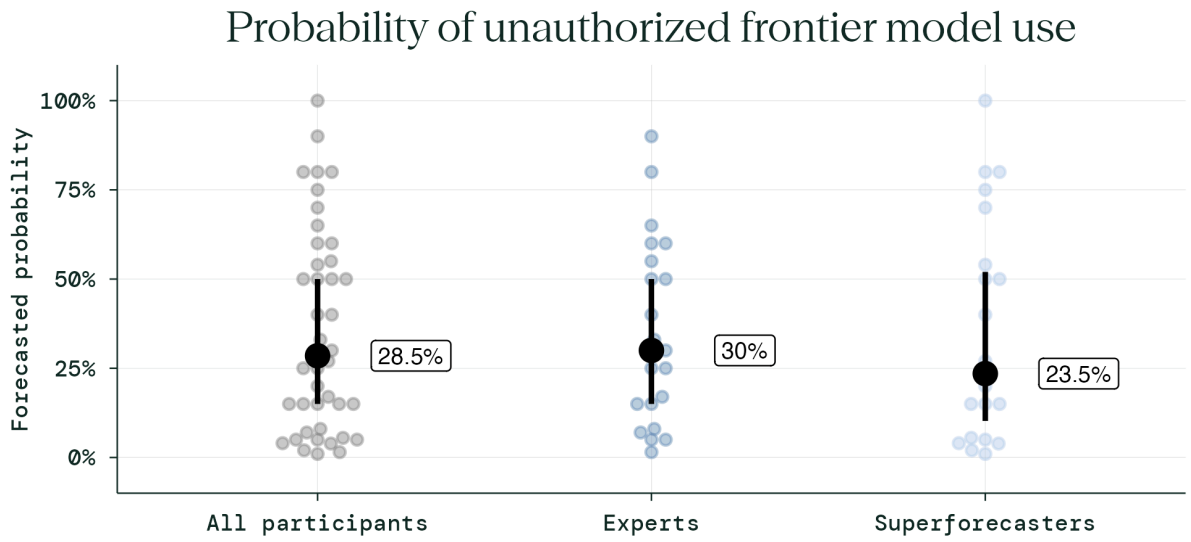


Figure 8: *Probability of unauthorized use of frontier models, conditional on a large-scale human-caused outbreak occurring. Numbers show group medians and black lines show their bootstrapped 95% confidence intervals.*

We then asked respondents to assume that unauthorized model use had contributed to a human-caused outbreak causing more than $100 million in damages, and say how likely they thought different pathways were to have contributed to that unauthorized model use. The responses are shown in Figure 9. Each of the pathways we asked about had a median probability of 15% or higher, suggesting that most respondents found all pathways plausible. Relative to experts, superforecasters generally thought a trusted user exemption was more likely to be used and the actor generating their own jailbreak was less likely. It is worth noting that these median values hide substantial variation in responses.

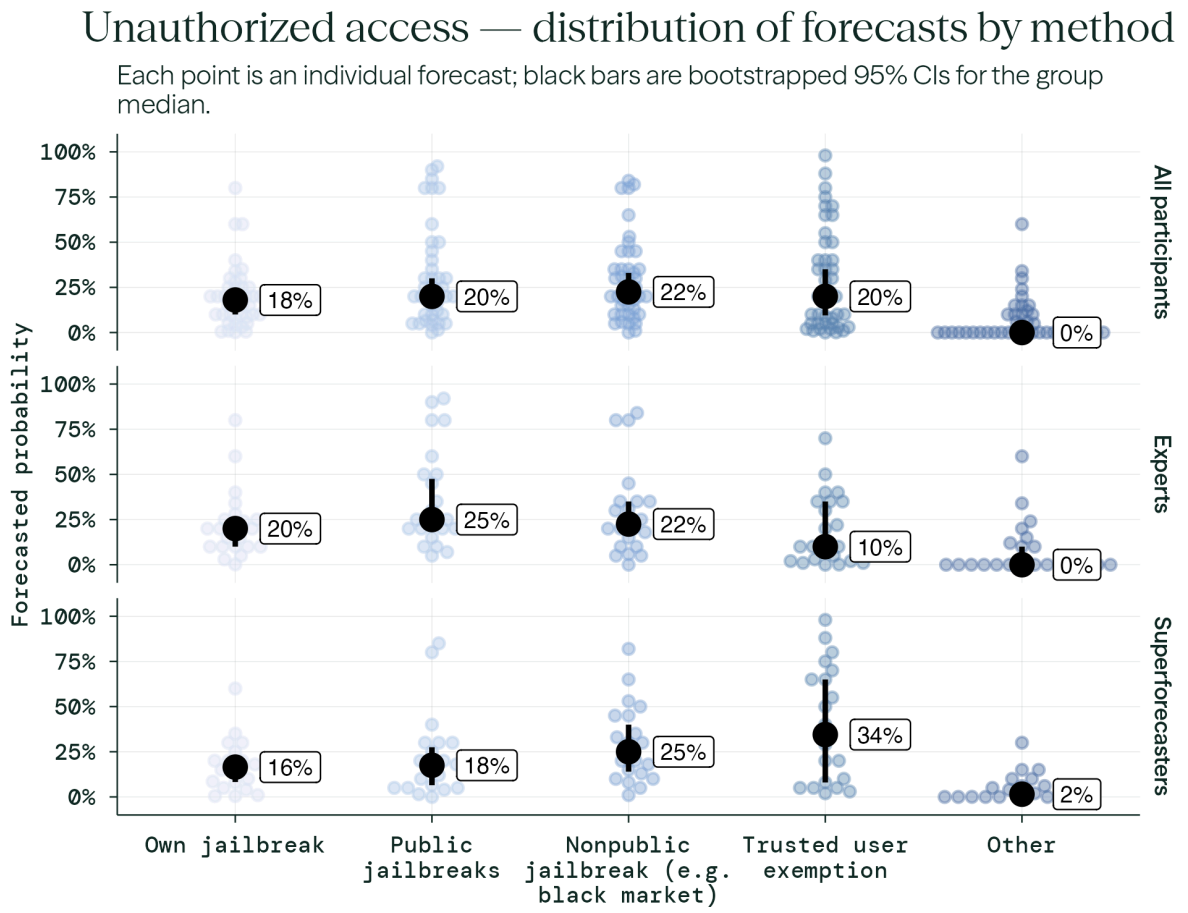


Figure 9: *Distribution of probabilities assigned to each pathway to gain unauthorized access. Numbers show group medians. The pathways were non-exclusive, and each individual was allowed to assign more than 100% across them.*

Insights from rationales

Whether respondents weighted accidental versus deliberate pathways more heavily strongly predicted their estimates of unauthorized model use. Those who saw accidents as dominant gave low estimates, since lab leaks were thought to rarely involve AI in the causal chain. Those focused on deliberate attacks gave high estimates, since any motivated attacker would exploit every available tool.

Trusted user exemptions were identified as a critical vulnerability because they exploit human frailties that technical safeguards cannot fully address. Bribery,

blackmail, and social engineering can compromise exemption holders regardless of how robust the underlying model defenses are. Moreover, exemption holders are precisely those with proximity to pathogens and lab infrastructure, meaning this type of compromise combines AI access with all the physical infrastructure needed for an attack. However, a countervailing view holds that the exemption pool is small, misuse would be highly attributable, and that screening should filter most threats.

Black market and nonpublic jailbreaks are seen as the path of least resistance for actors lacking technical skills. These markets effectively commoditize safeguard bypass via dark web markets that are harder for labs to monitor and take down than public jailbreaks. Public jailbreaks were seen as the most accessible, but also the most short-lived, limiting their utility for the sustained interaction biological development requires.

Creating novel jailbreaks is the highest-barrier pathway, constrained by a skill mismatch between jailbreaking and biology expertise. As one respondent put it, "most actors who can create a jailbreak, wouldn't have the biology background to create a serious virus." This restricts the route mainly to large organizations with resources to recruit across both domains.

Real-world attacks would likely chain multiple methods rather than relying on one, and that the "Other" category is a catch-all for important vectors like open-weight model manipulation, API vulnerabilities, and "unknown unknowns."

7. Conclusion

7.1 Summary of findings

This study asked 22 domain experts and 20 superforecasters to estimate how ASL-3 safeguards would change the expected financial damages from large-scale human-caused outbreaks between now and 2028. Four findings stand out.

Respondents believe ASL-3 safeguards demonstrate high efficacy. If applied to all GPAI models, the median respondent's forecasts imply that ASL-3 would mitigate roughly 70% of GPAI-attributable risk. This figure likely overstates true efficacy, because the 'no GPAI' benchmark used to define GPAI-attributable risk also removes GPAI's defensive contributions. Nonetheless, even accounting for this bias, the results suggest that ASL-3 captures a substantial share of the achievable risk reduction when applied.

Protecting only frontier models with ASL-3 reduces risk, but not as much as protecting all GPAI models. Protecting only frontier models captured less of the risk reduction than protecting all GPAI models, though respondents disagreed substantially on the size of the gap. This disagreement hinged on how capable and accessible near-frontier models—particularly Chinese open-weight models—are judged to be. The difference underscores that the value of ASL-3 as a mitigation depends not just on its technical performance but on how broadly equivalent safeguards are adopted across the ecosystem.

Making jailbreaks harder to find, patching them faster, or further degrading the

performance of jailbroken models were generally expected to improve ASL-3 performance. For the median respondent, the effects of these variants of ASL-3 for frontier models were roughly similar to the effect of ASL-3 for all models. Of the variants we asked about, the largest effect was associated with faster patching, consistent with the logic that bioweapon development requires sustained iterative access.

ASL-3 shifts the threat landscape toward better-resourced actors and accidental pathways. Respondents expected ASL-3 to disproportionately screen out non-expert and lone-wolf actors, shifting relative probability toward state actors, well-resourced organizations, and accidental releases from laboratories. Trusted user exemptions were identified as a notable vulnerability, because they combine AI access with proximity to the physical infrastructure needed for an attack and exploit human factors that technical safeguards cannot fully address. Many respondents emphasized that lab accidents represent a substantial share of baseline risk that AI safeguards do not address.

7.2 Limitations

Beyond the caveats noted above, several broader limitations should inform interpretation.

Forecasting low-probability, high-consequence events is inherently difficult. We focus on relative changes across scenarios rather than absolute damage estimates for this reason, but even relative judgments may be unreliable when base rates are poorly constrained. The wide variation in respondents' baseline forecasts—driven largely by which historical incidents they counted—illustrates this challenge.

The 'no safeguards' baseline is more extreme than the status quo. This scenario asks respondents to imagine all GPAI models are optimized purely for helpfulness with no safety constraints. In practice, most commercial models retain some safety training even without ASL-3. The measured gap between 'no safeguards' and the ASL-3 scenarios therefore captures the value of ASL-3 above a minimal baseline, not its marginal value above current industry-standard practices.

The capability freeze assumption bounds the shelf life of the findings. The study asked respondents to assume no improvement in GPAI capabilities through 2028. This isolates views on current-generation safeguards, but it means respondents were evaluating ASL-3 against a threat landscape that will almost certainly change. More capable models may provide greater uplift to malicious actors, may be harder to safeguard, or may shift the balance between information access and physical bottlenecks. These findings are best understood as a snapshot of expert views on ASL-3's efficacy against today's models.

The sample is modest and skewed toward Western biosecurity institutions. While the expert group had broad experience—14 of 22 reported expertise in more than one of biosecurity, national security and terrorism studies—perspectives from researchers in regions with rapidly expanding BSL infrastructure and those facing the highest burden of terrorist attacks are underrepresented.

Little information was provided on ASL-3 Security Standards. Forecasters were asked to consider the effects of ASL-3 as a whole—including both the Deployment Standards and Security Standards. We provided detailed information on the effectiveness of the Deployment Standards but did not have comparable information on the effectiveness of the Security Standards.

7.3 Implications

These results provide some evidence that deployment safeguards of the kind specified by ASL-3 can meaningfully reduce biosecurity risk. Respondents judged that comprehensive ASL-3 coverage would reduce most of the risk added by GPAI models, and they attributed this largely to ASL-3's ability to block non-expert actors. To the extent this judgment is accurate, it suggests that investment in deployment-stage safeguards is a worthwhile component of biosecurity risk management. An important caveat is that respondents were evaluating the effectiveness of safeguards for early-2026 models. As such, this should not be taken as evidence that ASL-3 would be similarly effective for substantially more capable systems.

The value of these safeguards depends on how broadly they, or equivalent measures, are adopted. Respondents saw a substantial gap between protecting all GPAI models and protecting only frontier models, and the size of that gap hinged on how capable and accessible near-frontier and open-weight models were judged to be. This suggests that the risk reduction achievable by any single developer is bounded by the safeguards adopted by other developers. This strengthens the case for industry-wide standards and for governance attention to models that fall outside the reach of any individual developer's deployment controls.

The findings also provide some insight into views on pathways to improve safeguards. For most respondents, the time required to find a jailbreak, speed of patching jailbreaks, and the performance associated with jailbreaks were all expected to reduce risk. These offer pathways to improving the effectiveness of ASL-3. The speed of patching jailbreaks was generally considered to have the greatest impact on risk. Respondents reasoned that bioweapon development requires sustained, iterative access over weeks or months, so closing jailbreaks quickly can disrupt an effort even after a jailbreak is found. Some respondents suggested that trusted user exemptions may also present a potential weak point in ASL-3, which may be a useful point of intervention.

Finally, the findings indicate that safeguards reshape the threat landscape rather than uniformly suppressing it. As ASL-3 screens out less sophisticated actors, risk concentrates among well-resourced state actors and accidental laboratory releases. Because a substantial share of baseline risk lies outside the influence of AI safeguards, AI-focused measures are best understood as complementary to established biosecurity interventions such as laboratory biosafety and biosecurity, and nucleic-acid synthesis screening.

7.4 Future directions

Two methodological improvements would most strengthen future iterations of this work. First, introducing a 'perfect safeguards' scenario—where GPAI exists and provides defensive benefits but cannot be misused—would allow a cleaner estimate of ASL-3's efficacy by disentangling GPAI's offensive and defensive contributions to outbreak risk. Second, developing an estimate of baseline risk under the status quo—where most, but not all, frontier models have some protections—would give insight into the impact of ASL-3 under real world conditions and the potential benefits of improving frontier model protections from what exists currently. However, developing this baseline would require more detailed information on the protections applied to other frontier models.

It would also be valuable to see how these findings change as model capabilities advance: repeating this elicitation periodically would track whether the efficacy of ASL-3, or future safeguards, holds or erodes as the models it protects become more capable. More broadly, this study demonstrates the feasibility of structured expert elicitation as a tool for evaluating AI safety measures against diffuse, hard-to-observe threats. Similar approaches could be applied to other risk domains covered by ASL-3 or future safety levels.

8. References

1. Anthropic. "Activating AI Safety Level 3 protections." 2025. https://www.anthropic.com/news/activating-asl3-protections.

2. Anthropic. "Responsible Scaling Policy, Version 2.2." Technical report, Anthropic, 2025. https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf.

3. Dario Amodei. "Written Testimony of Dario Amodei, Ph.D., Co-Founder and CEO, Anthropic." Hearing on "Oversight of A.I.: Principles for Regulation," Subcommittee on Privacy, Technology, and the Law, Committee on the Judiciary, United States Senate, July 25, 2023. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

4. Jonas Sandbrink. "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools." arXiv preprint arXiv:2306.13952, 2023. https://arxiv.org/abs/2306.13952.

5. Doni Bloomfield, Jaspreet Pannu, Alex W. Zhu, Madelena Y. Ng, Ashley Lewis, Eran Bendavid, Steven M. Asch, Tina Hernandez-Boussard, Anita Cicero, and Tom Inglesby. "AI and biosecurity: The need for governance." Science 385, 831-833, 2024. https://doi.org/10.1126/science.adq1977.

6. Aidan Peppin, Anka Reuel, Stephen Casper, Elliot Jones, Andrew Strait, Usman Anwar, Anurag Agrawal, Sayash Kapoor, Sanmi Koyejo, Marie Pellat, et al. "The Reality of AI and Biorisk." In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), pages 763-771, 2025. https://doi.org/10.1145/3715275.3732048.

7. Abi Olvera. "Tacit Knowledge: The Missing Factor in AI Bio Risk Assessments." 2026. https://secondthoughts.ai/p/tacit-knowledge-the-missing-factor.

8. Bridget Williams, Luca Righetti, Josh Rosenberg, Rebecca Ceppas de Castro, Otto Kuusela, Rhiannon Britt, Emily Soice, Alvaro Morales, Jon Sanders, Seth Donoughe, James Black, Ezra Karger, and Philip E. Tetlock. "Forecasting LLM-enabled biorisk and the efficacy of safeguards." 2025. https://forecastingresearch.org/research/llm-enabled-biorisk.

9. Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. "A Rosetta Stone for AI Benchmarks." arXiv preprint arXiv:2512.00193, 2025. https://arxiv.org/abs/2512.00193.

10. Shen Zhou Hong, Alex Kleinman, Alyssa Mathiowetz, Adam Howes, Julian Cohen, Suveer Ganta, Alex Letizia, Dora Liao, Deepika Pahari, Xavier Roberts-Gaal, Luca Righetti, and Joe Torres. "Measuring Mid-2025 LLM-Assistance on Novice Performance in Biology." arXiv preprint arXiv:2602.16703, 2026. https://arxiv.org/abs/2602.16703.

Appendix A: Description of ASL-3 Safeguards

Anthropic defines AI Safety Level (ASL) Standards, which are technical and operational safeguards for safely training and deploying frontier AI models.

For models that could significantly help individuals with basic technical backgrounds (e.g., undergraduate STEM degrees) create or deploy Chemical, Biological, Radiological, or Nuclear (CBRN) weapons (the "CBRN-3" threshold), Anthropic requires models to meet two safeguards: the ASL-3 Deployment Standard (preventing misuse of deployed models) and the ASL-3 Security Standard (preventing theft of model weights by non-state adversaries).

A description of these standards is provided in the AI Safety Level 3 Deployment Safeguards Report. We provide our summary below:

ASL-3 Deployment Standard

I.e., is the model robust to persistent attempts to misuse CBRN-3 capabilities?

- **Threat modeling**: Comprehensively identify and continuously update potential catastrophic misuse threats and attack vectors
- **Defense in depth:** Implement multiple-layered security barriers to catch misuse attempts that bypass earlier defenses
- **Red-teaming**: Demonstrate through adversarial testing that realistic attackers cannot consistently extract harmful capabilities beyond existing tools
- **Rapid remediation:** Quickly identify and patch vulnerabilities like jailbreaks before they enable meaningful harm
- **Monitoring:** Define risk metrics and regularly review system performance through bounties, analysis, and log retention
- **Trusted users:** Establish vetting criteria and alternative controls for sharing less-restricted model versions with qualified users
- **Third-party environments:** Ensure equivalent safety standards are maintained when models are deployed by partners

ASL-3 Security Standard

I.e., is the model highly protected against most attackers' attempts at stealing model weights?

- **Threat modeling:** Use frameworks like MITRE ATT&CK to map relationships between threats, assets, and attack vectors for protecting model weights from theft
- **Security frameworks:** Implement industry-standard controls aligned with frameworks like ISO 42001, NIST 800-53, and SOC 2, including:
- **Perimeters and access controls:** Protect AI models with strong physical security, encryption, cloud security, and access management
- **Lifecycle security:** Secure the development chain to prevent compromised components and ensure only trusted code/hardware is used

- **Monitoring:** Proactively detect threats through vulnerability testing, log management, asset monitoring, and deception techniques
- **Resourcing:** Dedicate roughly 5-10% of employees to security and security-adjacent work
- **Existing guidance:** Align with established AI security recommendations from CISA, NSA, and other authorities
- **Audits:** Conduct independent audits of security program design and implementation, with expert red-teaming and regular management reporting
- **Third-party environments:** Ensure equivalent security standards are maintained when models are deployed by partners

How has Anthropic implemented the ASL-3 Standard?

As part of launching Claude Opus 4, Anthropic activated the ASL-3 Standard as outlined in their RSP.

Per the AI Safety Level 3 Deployment Safeguards Report, Anthropic has "adopted a largely qualitative interpretation of the RSP's standard." Specifically, they have prioritized "identify[ing] the most salient and high-likelihood paths by which AI could help individuals to create/obtain and deploy CBRN weapons" and then "implement[ing] significant obstacles to these most likely paths to harm."

Anthropic operates under the following threat model, developed in consultation with experts such as Deloitte, SecureBio, and a Frontier Model Forum workshop:

- **CBRN capability above a critical threshold (CBRN-3):** The model can significantly assist individuals or groups with basic STEM backgrounds in obtaining, producing, or deploying CBRN weapons.
- **Long-term misuse (weeks to months) is the primary concern:** The highest-priority threat scenarios involve processes taking weeks to months, requiring substantial and persistent guidance through dozens of queries over extended periods.
- **Highly effective universal jailbreaks will be the primary attack vector:** Highly effective universal jailbreaks that (1) strongly preserve the model's capabilities in terms of specificity, detail, helpfulness, and correctness, and (2) transfer across multiple queries within the domain of concern.
- For additional details, see "II. Threat Model Summary" in the AI Safety Level 3 Deployment Safeguards Report, p. 7.

In order to mitigate the risks arising from the above threat model, Anthropic has implemented the following safeguards designed to meet the ASL-3 Deployment Standard, which are in addition to the ASL-2 Deployment Standard outlined in the RSP.

We summarize the relevant safeguards in the table below. For additional details, see "III. Safeguards Plan" in the AI Safety Level 3 Deployment Safeguards Report.

Mitigation	Description
Real-time Classifier	Constitutional Classifiers, large language models that monitor model input and output in real-time, are used to target the

Guards	relevant uses of concern related to the threat models and block the model from producing information that would be needed to successfully carry out the threats discussed above. See here for a report on Anthropic's latest classifiers and their performance.
Offline Monitoring	Offline monitoring refers to analysis conducted after model interactions, rather than in real time. These techniques are used as a mechanism to identify jailbreaks being used in practice and/or measure the rate of jailbreak discovery. Monitoring classifiers complement real-time classifiers because they are not subject to the same latency and computation constraints.
Access Controls	Anthropic allows some users to access models whose output isn't restricted by the real-time classifier guards, in cases where their legitimate business needs require this. These users are vetted using a manual vetting process.
Bug Bounty Program	Anthropic operates a bug bounty program with rewards ranging from $1,500 to $26,000 for reporting universal CBRN jailbreaks of real-time defenses.
Threat Intelligence	Anthropic monitors for: (i) available universal CBRN jailbreaks; (ii) black markets for model jailbreaks; (iii) reports of API key leaks, and in particular, markets for API key leaks from organizations with exemptions.
Rapid Response	Anthropic has developed a range of rapid response options (with different tradeoffs) to address jailbreaks and vulnerabilities. For more sophisticated attacks that reveal fundamental vulnerabilities, Anthropic may need weeks to months to develop entirely new defense techniques or architectural changes.

A lightly edited version of the above description was provided to survey respondents.

Confidential information from Anthropic's bug bounty program was also provided to participants to help them make an assessment of the effectiveness of the current ASL-3 anti-jailbreak measures. The information included excerpts from the instructions given to program participants and outcomes from the program. The outcomes focused on: the number of jailbreaks found, the time required to identify jailbreaks, the expertise of hackers who found the jailbreaks, the time to jailbreak patching, and jailbreak capability loss.

Appendix B: Detailed methods

Survey development

To develop the survey, we undertook an iterative process whereby researchers developed an initial set of forecasting questions quantifying the marginal effect of LLMs on the ability of non-experts to synthesize pathogens. We collected answers on these questions from a small group of superforecasters, and we then revised the questions in light of how they interpreted them, clarifying definitions and increasing the precision of each question. We conducted two rounds of this iterative question improvement process because forecasts can be highly sensitive to the precise wording of a question and its resolution criteria.

The survey itself consisted of:

- **Demographics and experience questionnaire:** We asked for basic demographic details (e.g., age category, sex, and country of residence). We also asked for details of experience relevant to the survey.
- **Calibration modules:** We asked participants to complete two short, timed calibration questionnaires. These were designed to provide respondents with insight into their own calibration around low probability events and quantities relevant to the survey.
- **Forecasts conditional on no GPAI:** We asked participants to forecast the main outcome conditional on a hypothetical world where GPAI is never developed.
- **Forecasts conditional on no safeguards:** We asked participants to make forecasts conditional on a hypothetical scenario where no GPAI models are protected by any safeguards, i.e. they are all "helpful-only". We asked for forecasts of the main outcome, and risk pathways conditional on this scenario.
- **Forecasts conditional on ASL-3 for all models:** We asked participants to make forecasts conditional on a hypothetical scenario where all GPAI models are protected by Anthropic's ASL-3 safeguards. We provided detailed information on ASL-3, including confidential information on ASL-3's effectiveness, including data from the bug bounty program and the trusted user exemption program. We asked for forecasts of the main outcome, and risk pathways conditional on this scenario.
- **Forecasts conditional on ASL-3 for frontier models:** We asked participants to make forecasts conditional on a hypothetical scenario where frontier GPAI models are protected by Anthropic's ASL-3 safeguards. Frontier GPAI models were defined as GPAI models at least as capable as Sonnet 4.5, according to the Epoch Capabilities Index. We asked for forecasts of the main outcome only for this scenario.
- **Forecasts conditional on ASL-3 variants:** We asked participants to make forecasts conditional on a hypothetical scenario where frontier GPAI models are protected by variants of Anthropic's ASL-3 safeguards. Frontier GPAI models were defined as GPAI models at least as capable as Sonnet 4.5, according to the Epoch Capabilities Index. We asked for forecasts of the main outcome only for this scenario.

The survey was administered using Quorum software. The survey can be viewed here.

Details of the main outcome

The following information was provided to survey respondents:

See historical outbreak-damage estimates here.

Human-caused outbreak

A human-caused outbreak includes both:

- Deliberate releases of pathogens by malicious actors, and
- Accidental releases of pathogens that have been deliberately processed by humans in some way.

"Deliberately processed" means that humans have taken intentional steps to acquire and manipulate, synthesize, grow, put into a form that allows dispersion, or otherwise modify the pathogen.

We acknowledge that many outbreaks could be considered "human-caused" in a broad sense. For example, an outbreak of Legionella that arises after failure to clean a cooling unit could be considered "human-caused", as could outbreaks of zoonotic disease that occur after human expansion into wild habitats. However, please don't include scenarios such as these when developing your forecast.

For this survey, releases caused by AI systems acting autonomously also count as human-caused, considering the system is carrying out human-enabled processing of a pathogen.

Economic damages

Economic damages refer to global economic costs that are attributable to the outbreak, beginning from the isolation of the pathogen and extending for the full life of the outbreak. These include:

- Direct economic losses (e.g., reduced productivity and income, commonly estimated as a reduction in GDP)
- Monetized mortality, using a value of a statistical life (VSL) estimate of $2.2 million (regardless of geographic distribution of fatalities)

If there is more than one such outbreak that begins between April 1, 2026, and December 31, 2028, the combined damages for all incidents will count towards resolution. Please give your estimates in units of 2026 USD.

Key Definitions

Other Key Definitions remain available at the link.

Resolution details

Please assume this study will have access to authoritative assessments, which

may include estimates by:

- WHO
- World Bank
- IMF
- Academic epidemiological and economic modeling groups

If multiple analyses exist, we will use the consensus or median for resolution. In the absence of such an estimate, we will defer to the majority opinion of a panel of at least five experts.

Please assume that we have perfect knowledge about the cause of the event, the intent of the actors involved, and their status as non-state or state actors. We want to understand your views on the damages resulting from these events, rather than the probability that they would be identified accurately.

Again, please forecast the damages realized over the full lifetime of outbreaks that begin within the defined period, including impacts that occur after 2028.

Participants recruitment

We identified possible expert participants through several different pathways. The approach to recruitment varied for different areas of expertise.

National security / terrorism experts

To identify national security / terrorism experts we searched the OpenAlex database for the authors of relevant publications and reviewed the staff pages of think tanks working on related issues. The details of our search strategy are provided below. We identified 16 experts who had published work related to the risk of biological weapons release. We invited these 16 experts to participate in the study. Of these, four experts expressed interest in participating in the study. One of these was excluded due to ongoing work with frontier AI companies. Three completed the full survey.

National security / terrorism experts sampling plan

- **Publications** (Using the OpenAlex publication database)
    - Any relevant publications in any of the following within the last 20 years (2006–2026):
        - Journal of Bioterrorism & Biodefense
        - Biosecurity and Bioterrorism: Biodefense Strategy, Practice and Science
    - Any relevant publications in the following with “biology”, “biowarfare”, “biological warfare”, “bioterror”, “CBRN”, “bioweapon”, or “biological weapon” in the title or abstract, published within the last 20 years (2006–2026):
        - Armed Forces & Society
        - Behavioral Sciences of Terrorism and Political Aggression
        - Caucasus Survey

        - Critical Studies on Terrorism
        - Conflict Management and Peace Science
        - Democracy and Security
        - European Journal of International Security
        - Health Security
        - Homeland Security Affairs
        - International Affairs
        - International Journal of Conflict and Violence
        - International Security
        - Journal of Conflict Resolution
        - Journal of Peace Research
        - Journal of Terrorism Research
        - Journal of Terrorism Studies
        - Journal of Social and Political Psychology
        - Journal of Policing, Intelligence and Counter Terrorism
        - Journal of Strategic Security
        - Perspectives on Terrorism
        - Risk Analysis
        - RUSI Journal
        - Security Dialogue
        - Security Studies
        - Studies in Conflict & Terrorism
        - Terrorism and Political Violence
        - Terrorism and Counter-Terrorism Studies
        - Third World Quarterly
- **Researchers at the following think tanks:**
    - Royal United Services Institute (RUSI): Terrorism and Conflict Group
    - International Centre for Counter-Terrorism (ICCT)
    - International Institute for Counter-Terrorism (ICT)
    - National Consortium for the Study of Terrorism and Responses to Terrorism (START)
    - Institute for Security and Technology (IST)
    - Stockholm International Peace Research Institute (SIPRI)
    - Bipartisan Commission on Biodefense (BCD)
    - Institute for Biodefense Research (IBR)
    - James Martin Center for Nonproliferation Studies (CNS)
    - Chatham House (CH)
    - The International Institute for Strategic Studies (IISS)
    - Vienna Center for Disarmament and Non-Proliferation (VCDNP)
    - Verification Research, Training and Information Centre (VERTIC)
    - Center for the Study of WMD (CSWMD)

Biosecurity experts

To identify biosecurity experts we searched the OpenAlex database for the authors of relevant publications, reviewed the staff pages of think tanks working on related issues, and the participants of relevant committees and fellowships. The details of our search strategy are provided below. We identified 39 experts who had published particularly relevant work and invited them to participate. Of these, 25 experts expressed interest in

participating in the study. Two of these were excluded due to ongoing work with frontier AI companies. Of the 23 remaining, 21 completed the full survey. Upon review, two of these respondents appeared to have grossly misunderstood large parts of the survey — their responses were removed.

Biosecurity experts sampling plan

- **Publications** (Using the OpenAlex publication database)
    - 3+ relevant publications in any of the following within the last 20 years (2006–2026):
        - Applied Biosafety
        - Biosecurity and Bioterrorism: Biodefense Strategy, Practice and Science
        - Health Security
        - Journal of Biosafety and Biosecurity
    - Any relevant publications in the following with “biosecurity”, “biosafety”, “biodefense”, “biowarfare”, “biological warfare”, “biological terrorism”, “bioterror”, or “CBRN” in the title or abstract, published within the last 20 years (2006–2026):
        - Nature Biotechnology
        - Nature Structural & Molecular Biology
        - Nature Computational Science
        - Biotechnology and Bioengineering
        - Trends in Biotechnology
        - The Lancet
        - Science
        - New England Journal of Medicine
        - JAMA
        - BMJ
        - PLOS Medicine
- **Think tank staff working on biosecurity or biotechnology:**
    - Nuclear Threat Initiative (NTI)
    - Johns Hopkins Center for Health Security
    - Council on Strategic Risks (CSR)
    - Center for Strategic and International Studies (CSIS)
    - CLTR
    - RAND
    - SecureBio
    - Blueprint Biosecurity
    - CEIP
    - CFR
    - Bipartisan Commission on Biodefense
- **Government**
    - US – NASEM
        - Members of the Committee on Assessing and Navigating Biosecurity Concerns and Benefits of AI in the Life Sciences
        - Authors of the report on Navigating the Benefits and Risks of Publishing Studies of In Silico Modeling and Computational Approaches of Biological Agents and Organisms – A Workshop

        - Speakers at the seminar series, Charting a Responsible Future in AI and Biosecurity: Enabling Safe Innovation and Governance
    - US – NSABB
        - Voting members and ex officio members
    - US – Former White House OSTP staff focused on biotechnology/biosecurity (e.g. 2023 and 2021)
    - UK – Members of the Biosecurity Leadership Council
- **Other**
    - ELBI Fellowship – all current and previous fellows
    - Council on Strategic Risks Ending Bioweapons Fellowship – all current and previous fellows
    - Participants of relevant workshops:
        - NTI: Building Strong Biosafety and Biosecurity into the Expanding US Bioeconomy
        - NTI: Developing Guardrails for AI Biodesign Tools
    - Members of relevant WHO Technical Advisory Groups:
        - Technical Advisory Group on the Responsible Use of the Life Sciences and Dual-Use Research (TAG-RULS DUR)
        - Scientific Advisory Group on the Origins of Novel Pathogens
        - Strategic and Technical Advisory Group on Infectious Hazards with Pandemic and Epidemic Potential (STAG-IH)
        - Health-Security Interface Technical Advisory Group (HSI-TAG)
        - Technical Advisory Group on Biosafety (TAG-B)
    - BWC meeting attendees

Superforecasters

We invited 22 superforecasters who had participated in previous studies conducted by the Forecasting Research Institute to participate in this study. Of these, 20 completed the full survey. Of these 20, eight superforecasters had participated in the iterative testing of earlier versions of the survey.

Data cleaning and analysis

Data analysis was conducted using R after aggregating and cleaning the data submitted by participants. We used the median as the default method for aggregating forecasts. To test for difference between expected damages across scenarios, we ran pairwise paired Wilcoxon signed-rank tests for each combination of two scenarios. Similarly, we also compared the expected value differences (relative to the 'ASL-3 for frontier models' baseline) across the six ASL-3 variants, again pairwise. We conducted these analyses three times — on all participants, experts only, and superforecasters only — with Holm correction for multiple comparisons within each subset.

Data cleaning included a series of validation tests that checked for logical coherence and consistency in responses. These tests are:

- Red flags: Logical incoherence
    - Binned probability forecasts do not sum to 100%
    - Probability of different actors being primary cause of outbreak do not sum to 100%

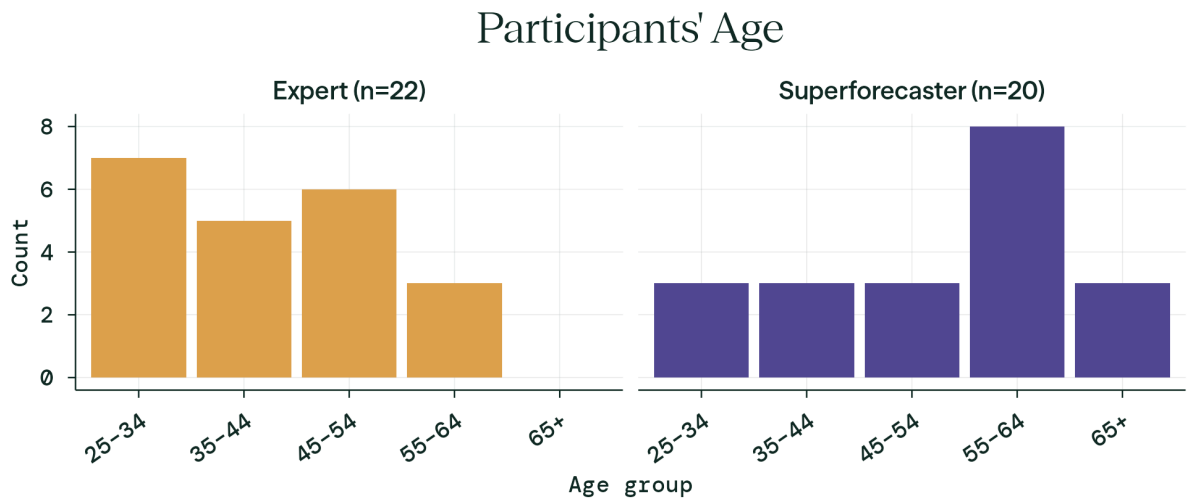
    - Probability of methods of unauthorized access not summing to at least 100% (We don't require strict independence between the methods, so the values may sum to more than 100%)
- Yellow flags: Logically possible but unexpected responses
    - Large discontinuity (more than 1 order of magnitude) in expected values across scenarios
    - Large discontinuity (more than 1 order of magnitude) in probability of expected damages over $10 trillion

We contacted participants who had 'Red flags' in their responses and requested they update their responses to resolve the incoherence. Eight respondents were contacted about red flags.

When 'Yellow flags' were detected, we reviewed text rationales to determine if the rationale supported the quantitative forecasts, or if there was evidence of misunderstanding of the question, or a typographical or unit error. When the responses suggested a possible misunderstanding or error, we contacted the respondent to ask them to review their forecasts. The most common type of error was a unit error, where respondents answered in billions of dollars, rather than providing the full dollar amount (e.g. wrote "6" when they intended to answer "6,000,000,000"). Twelve respondents (including 4 of the 8 respondents who had red flags) were contacted about yellow flags, and most made updates to their forecasts.

Appendix C: Supplementary results

Participants demographics



1 participant did not report their age.

Figure 10: *Distribution of participants' age.*

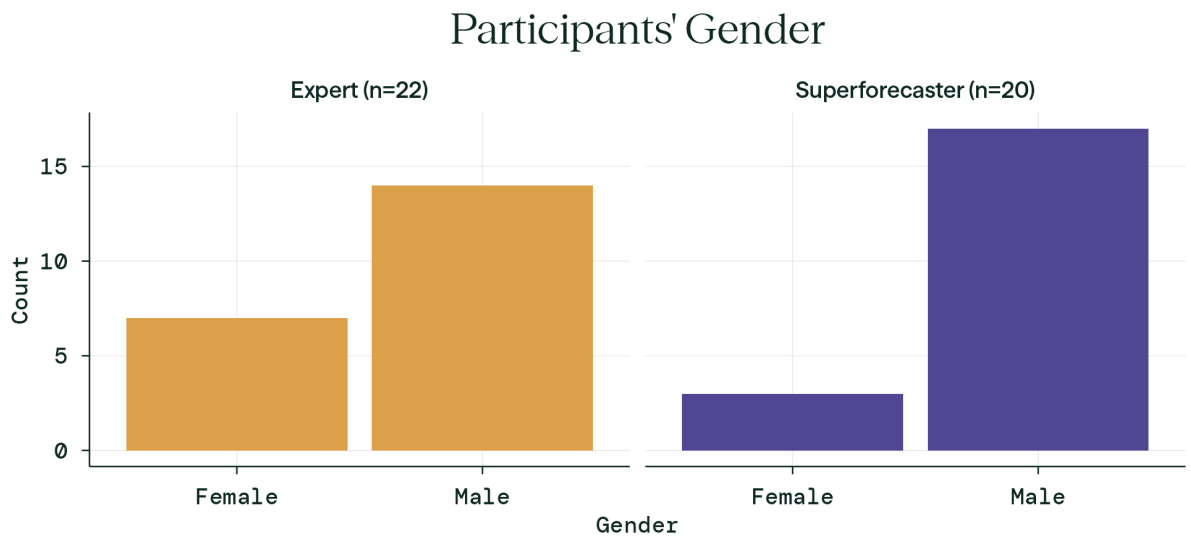


Figure 11: *Distribution of participants' gender.*

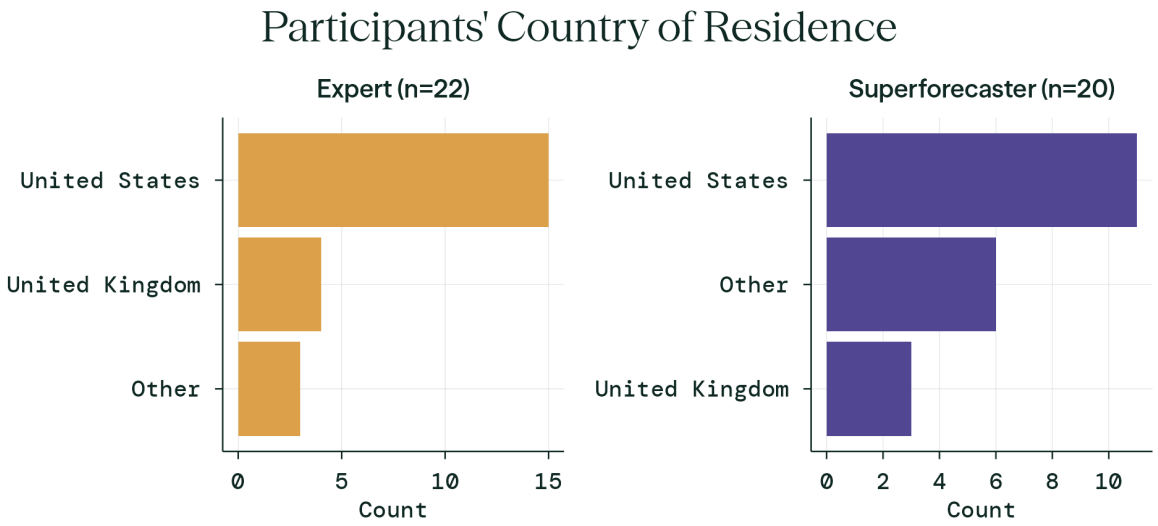


Figure 12: *Distribution of participants' reported countries of residence.*

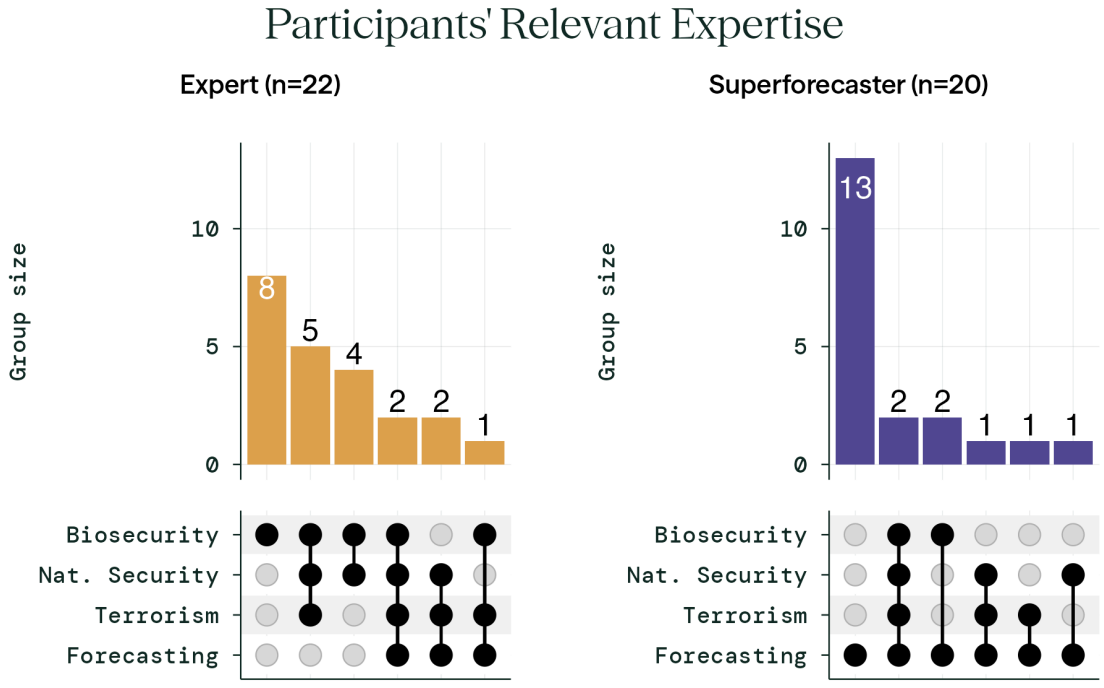


Figure 13: *Participants' reported experience in relevant domains.*

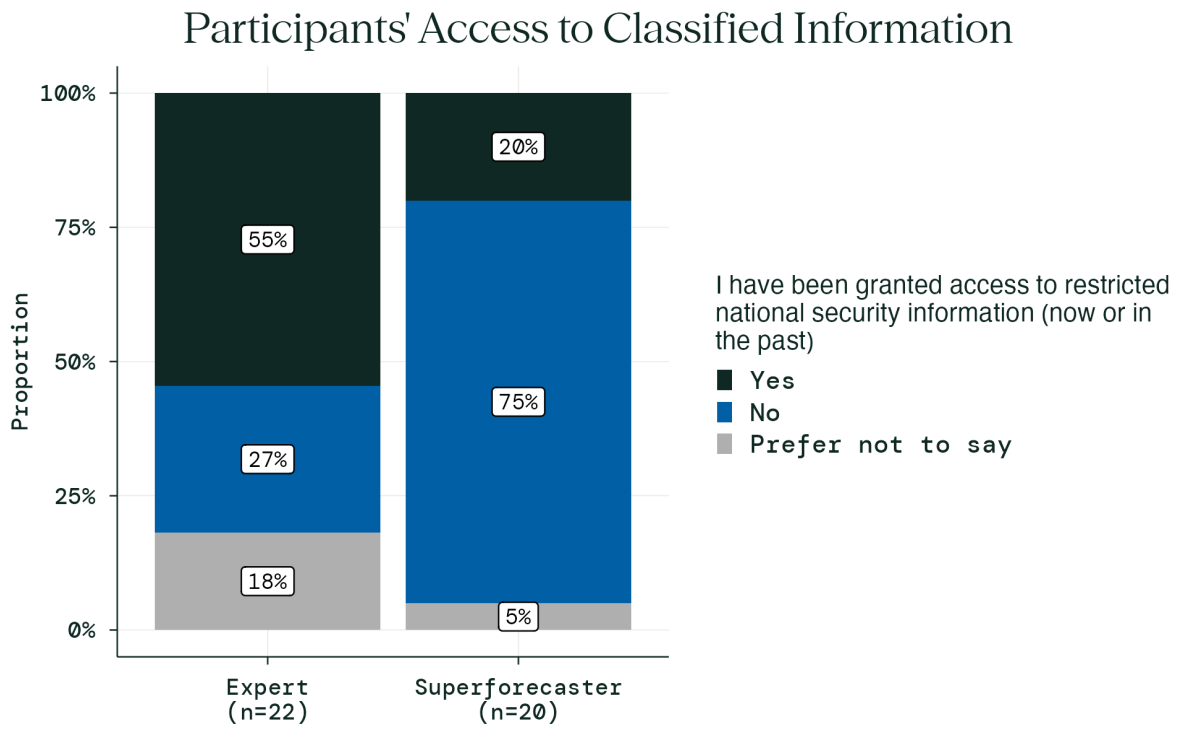


Figure 14: *Proportion of participants who reported having had access to classified information.*

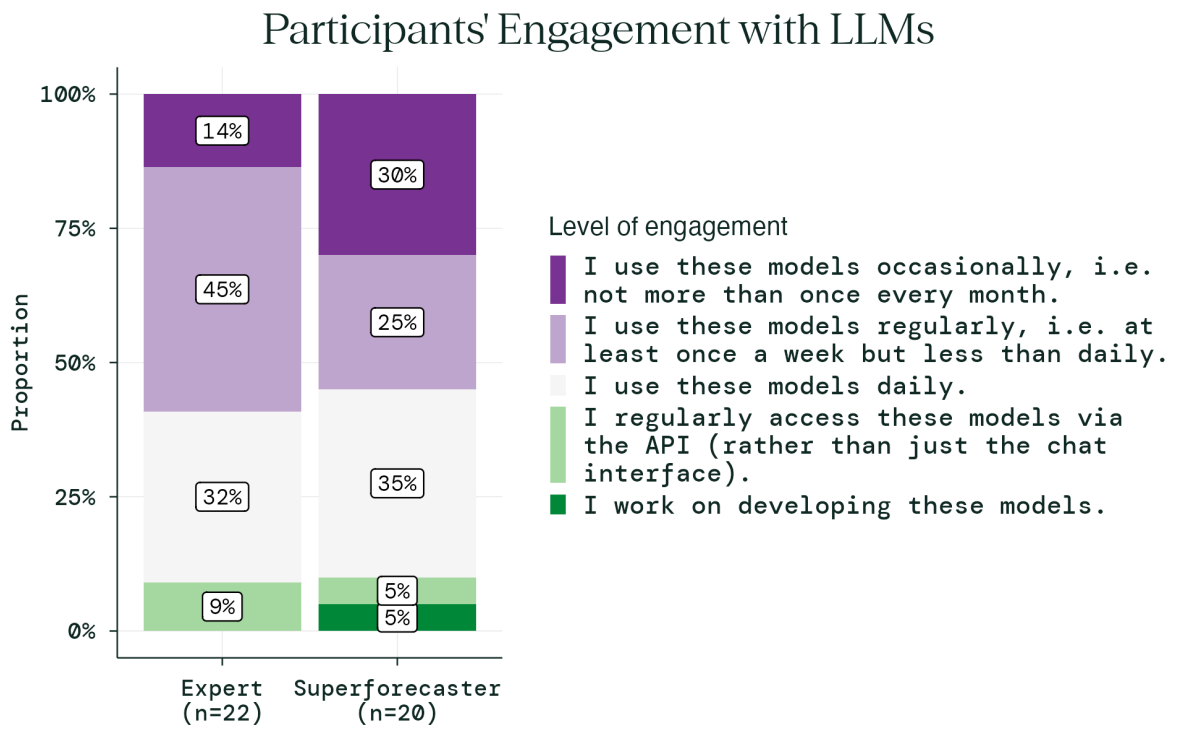


Figure 15: *Participant level of engagement with LLMs.*

Additional analysis of the main outcome

Figure 16 shows the calculated expected damages under all four scenarios. Respondents'

forecasts demonstrated substantial variation, however most believed that ASL-3 protections for all GPAI models would reduce the expected damages from large-scale human-caused outbreaks. Among the expert respondents, the median expected damages value was $411B (95% CI: $75B, $1434B) under the 'No safeguards' scenario and $156B (95% CI: $2B, $304B) under the scenario where all GPAI models were protected by ASL-3 safeguards. For the superforecaster sample, the median expected value estimates were $799B (95% CI: $161B, $3080B) under the 'No safeguards' scenario and $444B (95% CI: $51B, $1633B) under the scenario where all models were protected by ASL-3.

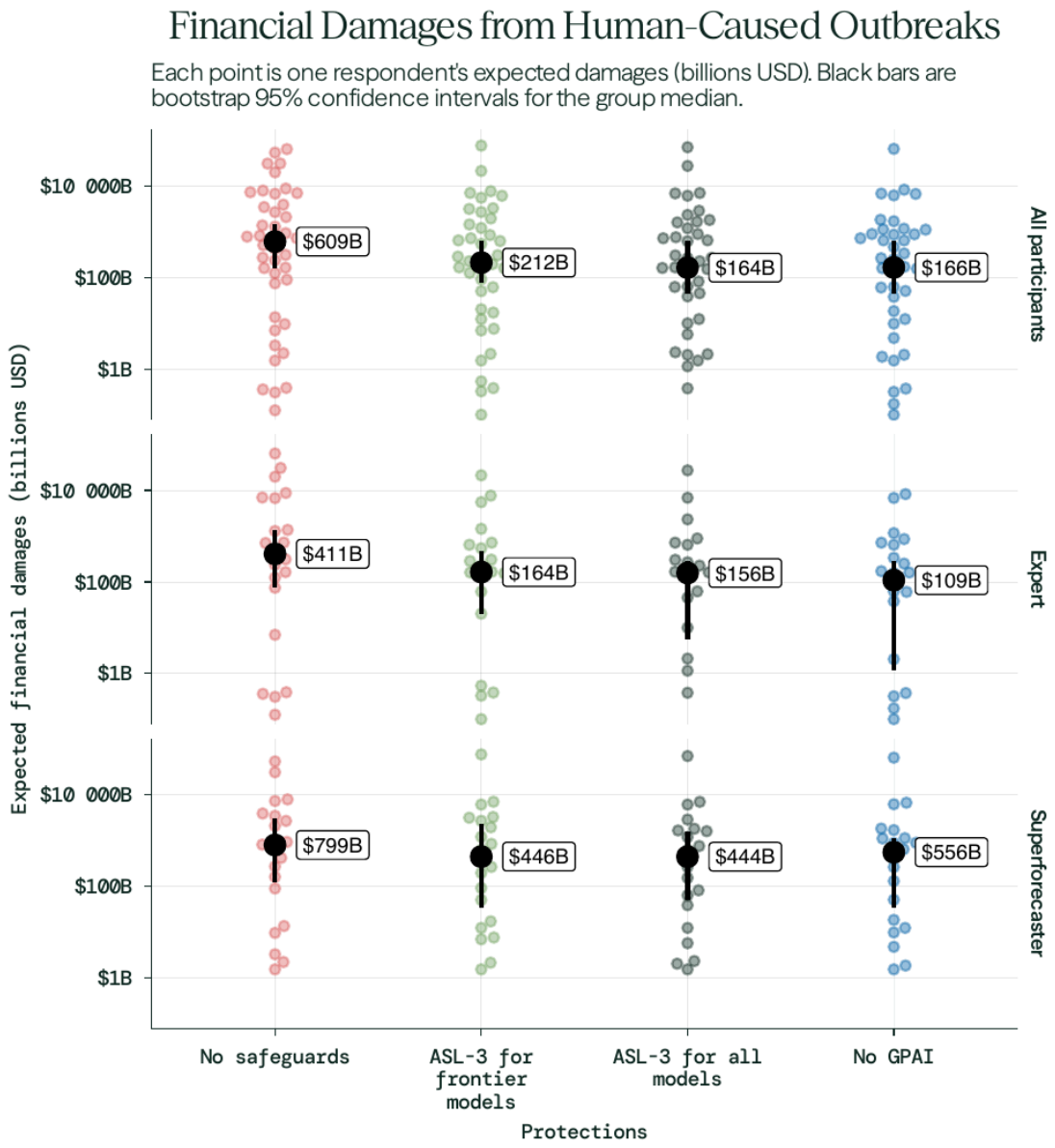


Figure 16: *Expected financial damages under each protection scenario. Expected values were calculated from binned probabilities using a uniform distribution. Numbers show medians from each group in billions of US dollars.*

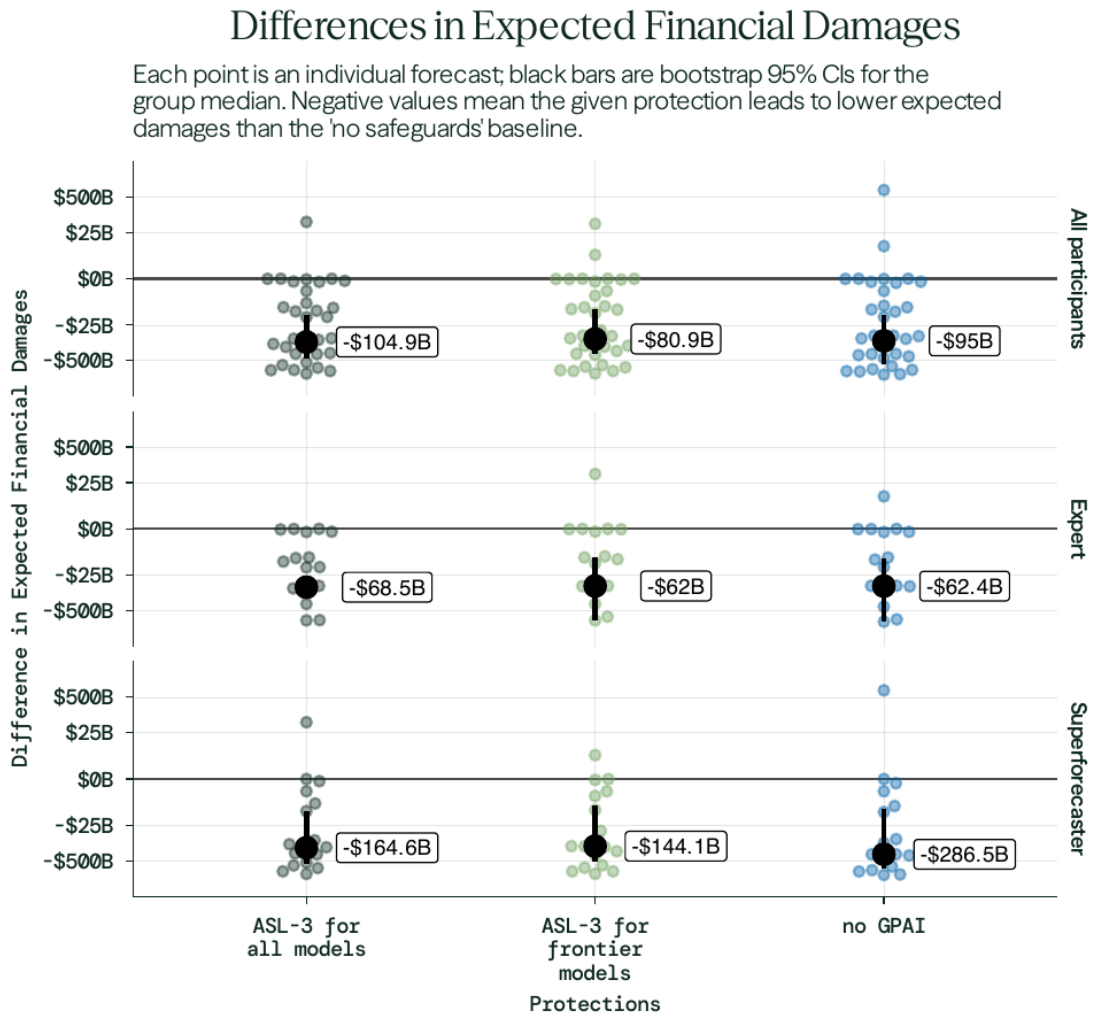


Figure 17: *Difference in expected damages for each scenario. Differences are with respect to the "no safeguards" baseline.*

Data tables

Table 1: *Pairwise differences in expected financial damages (billions USD), by comparison and respondent group. Medians show bootstrap 95% CIs where applicable. Bold median cells have a 95% CI that does not cross 0.*

	Expert			Superforecaster			Total		
Scenario	Median (95% CI)	IQR	N	Median (95% CI)	IQR	N	Median (95% CI)	IQR	N
ASL-3 frontier vs no safeguards	**-62 (-983.3, -5.1)**	(-2729.8, -1.4)	22	**-144.1 (-516.6, -4.4)**	(-765.5, -1.8)	20	**-80.9 (-503.8, -6.4)**	(-1068.1, -1.4)	42
ASL-3 all vs no safeguards	**-68.5 (-1128.8, -6.8)**	(-4869, -5.6)	22	**-164.6 (-659.1, -5.6)**	(-780.6, -3.2)	20	**-104.9 (-583, -9.2)**	(-1122.9, -4.3)	42
No GPAI vs no safeguards	**-62.4 (-3247.8, -5.5)**	(-4921.6, -1.6)	22	**-286.5 (-948.8, -4.8)**	(-1129.9, -3.9)	20	**-95 (-798.5, -8.1)**	(-1249.1, -2.1)	42

No GPAI vs ASL-3 frontier	-0.9 (-70.6, 0)	(-114.6, 0)	22	**-69.4 (-228.8, -0.3)**	(-277.7, -0.2)	20	**-29.4 (-79.4, -0.3)**	(-205.5, 0)	42
ASL-3 all vs ASL-3 frontier	-0.1 (-18.1, 0)	(-52.6, 0)	22	**-12.6 (-85.6, -0.4)**	(-100.2, -0.1)	20	-3.1 (-30.6, 0)	(-80.2, 0)	42
Variant A1 vs ASL-3 frontier	-0.2 (-14.9, 0)	(-24.4, 0)	22	**-0.8 (-30, -0.1)**	(-44.8, 0)	20	**-0.7 (-10.2, -0.1)**	(-28.1, 0)	42
Variant A2 vs ASL-3 frontier	3 (-0.1, 52.2)	(-0.1, 53.3)	22	2.5 (0, 20)	(-1.4, 26.1)	20	3 (0, 21.8)	(-0.1, 52.6)	42
Variant B1 vs ASL-3 frontier	**-2.1 (-37, -0.1)**	(-59.1, -0.1)	22	**-5.7 (-54.4, -0.3)**	(-58.5, -0.1)	20	**-2.1 (-35.3, -0.2)**	(-64.1, -0.1)	42
Variant B2 vs ASL-3 frontier	20.6 (0, 68.8)	(0, 159.3)	22	3.5 (0, 48)	(0, 82.4)	20	12.4 (0.5, 36.2)	(0, 156.4)	42
Variant C1 vs ASL-3 frontier	**-0.9 (-15.8, -0.1)**	(-23.7, -0.1)	22	**-0.4 (-43.1, -0.1)**	(-45.6, 0.1)	20	**-0.5 (-16.2, -0.1)**	(-39.3, -0.1)	42
Variant C2 vs ASL-3 frontier	28.7 (0, 97.8)	(0, 126.6)	22	0.3 (-0.6, 31.2)	(-3.3, 43.2)	20	0.7 (0, 47.6)	(0, 91.2)	42

Table 2: *Pairwise ratios of expected financial damages, by comparison and respondent group. Medians show bootstrap 95% CIs where applicable. Bold median cells have a 95% CI that does not cross 1.*

	Expert			Superforecaster			Total		
Scenario	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**
ASL-3 frontier vs no safeguards	**0.56 (0.25, 0.91)**	(0.23, 0.94)	22	**0.86 (0.63, 0.92)**	(0.54, 0.94)	20	**0.8 (0.48, 0.91)**	(0.35, 0.94)	42
ASL-3 all vs no safeguards	**0.4 (0.22, 0.83)**	(0.2, 0.89)	22	**0.82 (0.43, 0.9)**	(0.4, 0.91)	20	**0.6 (0.37, 0.83)**	(0.23, 0.91)	42
No GPAI vs no safeguards	**0.29 (0.15, 0.81)**	(0.13, 0.83)	22	**0.64 (0.37, 0.85)**	(0.31, 0.87)	20	**0.51 (0.29, 0.81)**	(0.18, 0.85)	42
No GPAI vs ASL-3 frontier	**0.71 (0.54, 0.99)**	(0.36, 1)	22	**0.86 (0.65, 0.97)**	(0.61, 1)	20	**0.79 (0.62, 0.96)**	(0.51, 1)	42
ASL-3 all vs ASL-3 frontier	0.97 (0.62, 1)	(0.53, 1)	22	0.94 (0.9, 1)	(0.86, 1)	20	**0.96 (0.84, 0.99)**	(0.74, 1)	42
Variant A1 vs ASL-3 frontier	**0.91 (0.55, 0.99)**	(0.5, 1)	22	0.96 (0.9, 1)	(0.89, 1)	20	**0.94 (0.89, 0.98)**	(0.8, 1)	42
Variant A2 vs ASL-3 frontier	1.01 (0.93, 1.15)	(0.89, 1.22)	22	1.02 (0.99, 1.11)	(0.97, 1.12)	20	1.01 (1, 1.09)	(0.94, 1.18)	42
Variant B1 vs ASL-3 frontier	**0.86 (0.67, 0.95)**	(0.6, 0.98)	22	**0.93 (0.87, 0.97)**	(0.84, 0.98)	20	**0.92 (0.82, 0.95)**	(0.73, 0.98)	42
Variant B2 vs ASL-3 frontier	1.06 (0.97, 1.39)	(0.9, 1.42)	22	1.04 (1, 1.13)	(1, 1.2)	20	1.05 (1, 1.13)	(0.98, 1.39)	42

Variant C1 vs ASL-3 frontier	**0.85 (0.22, 0.99)**	(0.14, 0.99)	22	0.96 (0.89, 1)	(0.88, 1)	20	**0.94 (0.85, 0.99)**	(0.64, 1)	42
Variant C2 vs ASL-3 frontier	1.05 (0.98, 1.23)	(0.94, 1.29)	22	1 (0.99, 1.06)	(0.98, 1.1)	20	1.01 (1, 1.11)	(0.97, 1.21)	42

Table 3: *Estimated probability of at least $100M in total worldwide outbreak damages by scenario and respondent group. Each value sums binned probabilities over all damage bins except the below-$100M bin.*

	Expert			Superforecaster			Total		
Scenario	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**
No GPAI	3.8 (1.5, 10)	(1.4, 10.6)	22	15 (7.6, 22.9)	(6.7, 32.2)	20	9.3 (3.8, 13.2)	(2, 19.3)	42
No safeguards	6.7 (4.2, 48)	(3.1, 49.5)	22	20.5 (13.5, 44.7)	(11.9, 48.3)	20	16 (9.2, 32.8)	(5, 49.5)	42
ASL-3 for all models	4.2 (2.2, 8.5)	(2, 22)	22	14.4 (9.2, 20.8)	(8.8, 23.7)	20	9.2 (4, 16.9)	(2.5, 25)	42
ASL-3 for frontier models	5.5 (2.7, 15)	(2.6, 25.7)	22	15.7 (10.6, 27)	(10, 27.9)	20	10.6 (5.5, 18.1)	(2.8, 28.6)	42

Table 4: *Pairwise differences in P(at least $100M damages) versus the no-safeguards scenario (percentage points), by comparison and respondent group: scenario probability minus no-safeguards probability. Bold median cells have a 95% CI that does not cross 0.*

	Expert			Superforecaster			Total		
Scenario	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**
ASL-3 frontier vs no safeguards	**-2 (-5.5, -1)**	(-6.8, -0.8)	22	**-1.9 (-6.6, -1)**	(-8.5, -0.8)	20	**-2 (-5, -1)**	(-7.5, -0.8)	42
ASL-3 all vs no safeguards	**-3.7 (-10, -1.5)**	(-11.5, -1.1)	22	**-2.6 (-13.3, -1.5)**	(-14.4, -0.9)	20	**-3 (-8, -2)**	(-13, -1)	42
No GPAI vs no safeguards	**-3.2 (-6, -2.1)**	(-7.5, -1.9)	22	**-3 (-10, -0.8)**	(-13.7, -0.6)	20	**-3.1 (-5, -2)**	(-10.4, -1.1)	42

Table 5: *Pairwise ratios of P(at least $100M damages) versus the no-safeguards scenario, by comparison and respondent group (numerator scenario divided by no safeguards). Bold median cells have a 95% CI that does not cross 1.*

	Expert			Superforecaster			Total		
Scenario	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**	**Median (95% CI)**	**IQR**	**N**
ASL-3 frontier vs no safeguards	**0.8 (0.64, 0.9)**	(0.6, 0.9)	22	**0.88 (0.71, 0.94)**	(0.68, 0.95)	20	**0.82 (0.7, 0.91)**	(0.61, 0.94)	42
ASL-3 all vs no	**0.64 (0.45,**	(0.41, 0.9)	22	**0.77 (0.49,**	(0.48, 0.94)	20	**0.68 (0.59,**	(0.45, 0.92)	42

safeguards	**0.79)**			**0.93)**			**0.85)**		
No GPAI vs no safeguards	**0.56 (0.37, 0.8)**	(0.31, 0.89)	22	**0.76 (0.5, 0.94)**	(0.49, 0.95)	20	**0.61 (0.5, 0.84)**	(0.4, 0.94)	42

Table 6: *Pairwise paired Wilcoxon comparisons of expected financial damages across main scenarios. Medians are in billions USD; Median diff is the median of within-person paired differences (Scenario 1 minus Scenario 2). P-values are Holm-adjusted across pairs within each respondent group; bold cells indicate significance at Holm p < 0.05.*

Respondent group	Scenario 1	Scenario 2	N	Median 1 (B USD)	Median 2 (B USD)	Median diff (B USD)	Holm p	Sig.
All participants	No GPAI	No safeguards	42	166.03	609.49	-95.01	**< 0.01**	yes
	No GPAI	ASL-3 for all models	42	166.03	164.33	-0.73	**0.01**	yes
	No GPAI	ASL-3 for frontier models	42	166.03	211.58	-29.37	**< 0.01**	yes
	No safeguards	ASL-3 for all models	42	609.49	164.33	104.86	**< 0.01**	yes
	No safeguards	ASL-3 for frontier models	42	609.49	211.58	80.92	**< 0.01**	yes
	ASL-3 for all models	ASL-3 for frontier models	42	164.33	211.58	-3.11	**< 0.01**	yes
Experts	No GPAI	No safeguards	22	109.25	410.62	-62.45	**< 0.01**	yes
	No GPAI	ASL-3 for all models	22	109.25	156.39	-0.01	0.13	no
	No GPAI	ASL-3 for frontier models	22	109.25	164.09	-0.9	**0.01**	yes
	No safeguards	ASL-3 for all models	22	410.62	156.39	68.49	**< 0.01**	yes
	No safeguards	ASL-3 for frontier models	22	410.62	164.09	62	**< 0.01**	yes
	ASL-3 for all models	ASL-3 for frontier models	22	156.39	164.09	-0.13	**0.04**	yes
Superforecasters	No GPAI	No safeguards	20	555.74	799.32	-286.54	**0.03**	yes
	No GPAI	ASL-3 for all models	20	555.74	444.17	-42.95	**0.03**	yes
	No GPAI	ASL-3 for frontier models	20	555.74	446.23	-69.43	**0.03**	yes
	No safeguards	ASL-3 for all models	20	799.32	444.17	164.65	**0.02**	yes
	No safeguards	ASL-3 for frontier models	20	799.32	446.23	144.06	**0.02**	yes

	ASL-3 for all models	ASL-3 for frontier models	20	444.17	446.23	-12.59	**0.02**	yes

Table 7: *Pairwise paired Wilcoxon comparisons of P(damages ≥ $100M) across main scenarios. Medians are in survey percentage points; Median diff is the median of within-person paired differences (Scenario 1 minus Scenario 2). P-values are Holm-adjusted across pairs within each respondent group; bold cells indicate significance at Holm p < 0.05.*

Respondent group	Scenario 1	Scenario 2	N	Median 1 (pp)	Median 2 (pp)	Median diff (pp)	Holm p	Sig.
All participants	No GPAI	No safeguards	42	9.33	16	-3.15	**< 0.01**	yes
	No GPAI	ASL-3 for all models	42	9.33	9.2	-0.18	0.09	no
	No GPAI	ASL-3 for frontier models	42	9.33	10.6	-0.8	**< 0.01**	yes
	No safeguards	ASL-3 for all models	42	16	9.2	3	**< 0.01**	yes
	No safeguards	ASL-3 for frontier models	42	16	10.6	1.95	**< 0.01**	yes
	ASL-3 for all models	ASL-3 for frontier models	42	9.2	10.6	-0.55	**< 0.01**	yes
Experts	No GPAI	No safeguards	22	3.8	6.65	-3.25	**< 0.01**	yes
	No GPAI	ASL-3 for all models	22	3.8	4.25	-0.08	0.28	no
	No GPAI	ASL-3 for frontier models	22	3.8	5.5	-0.8	**0.01**	yes
	No safeguards	ASL-3 for all models	22	6.65	4.25	3.65	**< 0.01**	yes
	No safeguards	ASL-3 for frontier models	22	6.65	5.5	1.95	**< 0.01**	yes
	ASL-3 for all models	ASL-3 for frontier models	22	4.25	5.5	-0.85	**0.01**	yes
Superforecasters	No GPAI	No safeguards	20	15	20.5	-2.96	**0.01**	yes
	No GPAI	ASL-3 for all models	20	15	14.45	-0.35	0.21	no
	No GPAI	ASL-3 for frontier models	20	15	15.71	-0.75	0.16	no
	No safeguards	ASL-3 for all models	20	20.5	14.45	2.58	**0.01**	yes
	No safeguards	ASL-3 for frontier models	20	20.5	15.71	1.93	**< 0.01**	yes
	ASL-3 for all models	ASL-3 for frontier models	20	14.45	15.71	-0.42	0.10	no

Table 8: Calculated share of GPAI-attributable risk mitigated by each scenario.

	Expert			Superforecaster			Total		
Scenario	Median (95% CI)	IQR	N	Median (95% CI)	IQR	N	Median (95% CI)	IQR	N
ASL-3 frontier	51.8 (15.2, 82.8)	(9.8, 97.4)	20	31.6 (22.9, 83.2)	(18.3, 83.2)	17	36.3 (22.9, 68.4)	(17.3, 88.7)	37
ASL-3 all models	94.6 (48.7, 99.9)	(37.8, 100)	20	67.1 (45.6, 92.1)	(45.6, 92.1)	17	71.7 (65.7 97.2)	(42.4, 100)	37
Variant A1	65.4 (27.2, 100)	(24.7, 100)	20	52.3 (41.4, 78.6)	(41.4, 84.5)	17	54.0 (41.4, 89.4)	(28.3, 100)	37
Variant A2	39.6 (9.8, 55.6)	(7.4, 67.3)	20	21.1 (12.2, 33.2)	(12.2, 33.2)	17	25.1 (16.3, 42.1)	(10.5, 49.7)	37
Variant B1	99.4 (48.7, 100)	(30.3, 100)	20	58.2 (44.4, 92.6)	(44.4, 92.6)	17	73.5 (52.3, 100)	(40.5, 100)	37
Variant B2	36.0 (6.7, 65.3)	(6.3, 72.4)	20	19.9 (10.8, 44.0)	(10.8, 44.0)	17	25.6 (10.8, 44.1)	(6.7, 61.5)	37
Variant C1	99.0 (50.3, 100)	(40.0, 100)	20	52.3 (31.6, 84.5)	(31.6, 84.5)	17	73.5 (47.8, 97.3)	(31.6, 100)	37
Variant C2	35.6 (7.1, 70.7)	(6.9, 80.7)	20	30.8 (15.8, 77.1)	(14.5, 90.0)	17	32.6 (15.8, 52.4)	(7.4, 83.0)	37