Marginal Risk Relative to What? Distinguishing Baselines in AI Risk Management

Michael Chen [*1] **Jide Alaga** [*1]

Abstract

Major developers of large foundation models make development and deployment decisions informed by evaluations of *marginal risk*: risk introduced by a new AI model, relative to a baseline. Developers face a critical choice between two types of baselines: a "pre-GPAI" baseline without modern general-purpose AI systems (e.g., only having 2023-level technology), or a "post-GPAI" baseline which can include the most risk-enabling models already available. Reviewing voluntary safety frameworks adopted by AI model developers, we note that developers do not always clearly specify which baseline is used. We examine potential risks of cumulative model releases that incrementally add marginal risk, leading to an environment in which each individual model may appear safe from the perspective of post-GPAI baselines, while aggregate risk from AI becomes unacceptably dangerous.

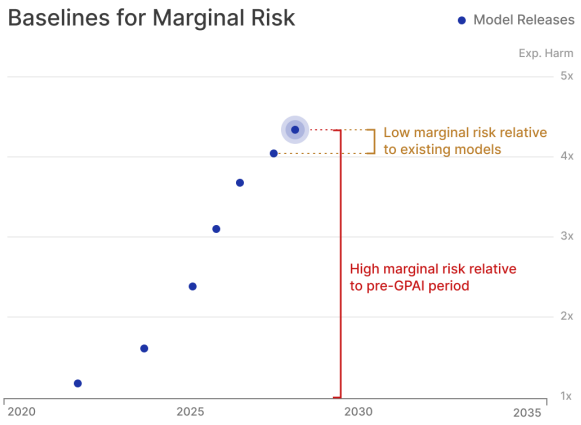

Figure 1. A model with low marginal risk relative to existing models (the rightmost point) can pose high marginal risk relative to a pre-GPAI period.

1. Introduction

AI developers routinely publish system cards or model reports that evaluate capabilities related to catastrophic risk like biological weapons assistance and cyber offense (OpenAI, 2025a; Google, 2025b; Anthropic, 2025; Meta AI, 2024). Many of these companies have also adopted frontier safety frameworks that commit them to regular capability evaluations and mitigating unacceptable risks (METR, 2025).

These risk assessments often directly or indirectly employ the concept of "marginal risk," which Kapoor et al. (2024) defines as the risk that a model introduces compared to other "foundation models or pre-existing technologies, such as web search on the internet".[1] Within this single definition are two distinct baselines with significantly different implications: a "post-GPAI" baseline and a "pre-GPAI" baseline.

Using a **pre-GPAI baseline**, a capability threshold might ask, "Compared to a world without AI models as advanced as GPT-4, does this model significantly increase the chance of catastrophe?" A concrete research question for model evaluations might be, "Does access to the model give humans significant uplift over those who have internet access but no AI, especially for tasks that could lead to catastrophic outcomes?"

In contrast, when assessing marginal risk against **post-GPAI baselines**, a developer assesses whether its model poses significantly greater risk compared to other powerful models already available.

In the process of making risk assessments, developers typi-

[*]Equal contribution [1]METR, Berkeley, United States. Correspondence to: Jide Alaga <jide@metr.org>, Michael Chen <michael@metr.org>.

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

[1]Kapoor et al. (2024) focuses on marginal risk of open foundation models compared to "closed foundation models or pre-existing technologies," whereas we generalize this definition for any foundation model. Our usage of "marginal" should be distinguished from the colloquial definition of "minimal" or "unimportant," as the marginal risk introduced by a model can be large or small relative to a baseline.

Table 1. Pre-GPAI and post-GPAI marginal risk baselines implicit in various AI research papers.

Document	Title	Summary	Baseline
Patwardhan et al. (2024)	Building an early warning system for LLM-aided biological threat creation (January 2024)	GPT-4 access provides mild but not statistically significant uplift on biothreat tasks compared to internet-only	Pre-GPAI
Kapoor et al. (2024)	On the Societal Impact of Open Foundation Models (February 2024)	"marginal risk of open foundation models . . . beyond closed foundation models or pre-existing technologies, such as web search on the internet"	Pre- and post-GPAI
Anthropic (2024b)	RSP Evaluations Report – Claude 3 Opus (May 2024)	Claude 3 access provides no statistically significant uplift in answering CBRN questions accurately, compared to Google	Pre-GPAI
Zhu et al. (2025)	Teams of LLM Agents can Exploit Zero-Day Vulnerabilities (June 2024)	A GPT-4-based multi-agent system achieves an 18% success rate in vulnerability exploitation, up from 0% by open-source, non-GPAI vulnerability scanners (ZAP, MetaSploit) and 0% by open-weight models Llama 3.1 405B and Qwen 2.5 72B	Pre- and post-GPAI
Meta AI (2024)	The Llama 3 Herd of Models (July 2024)	Llama 3 access provides no significant CBRNE or cyber uplift compared to "web-only control group"	Pre-GPAI
Wan et al. (2024)	CYBERSECEVAL 3 (August 2024)	Llama 3 405B does not provide significant cyber uplift to experts or novices, compared to internet-only	Pre-GPAI
Anthropic (2025)	Claude 4 System Card (May 2025)	Access to Claude Opus 4 without safeguards provides 2.53× uplift on bioweapons planning, relative to internet-only	Pre-GPAI
Bommasani et al. (2025)	The California Report on Frontier AI Policy (June 2025)	"We recommend that policymakers center their calculus around the marginal risk: Do foundation models present risks that go beyond previous levels of risks that society is accustomed to from prior technologies, such as risks from search engines?"	Pre-GPAI

cally conduct dangerous-capability evaluations (Shevlane et al., 2023).[2] Several biological risk assessments in early 2024 measured whether large language models (LLMs) significantly improved the performance of novices or experts (Meta AI, 2024; Patwardhan et al., 2024; AI Security Institute, 2024; Mouton et al., 2024). These uplift studies were characterized with one group of humans having LLM access, while a control group used only the internet and non-AI tools, representing a pre-GPAI baseline. As AI capabilities in biology advance, human uplift studies will increasingly find that AI provides significant advantages relative to pre-GPAI baselines. On the Virology Capabilities Test (Götting et al., 2025), for example, recent models unassisted by humans outperform human experts even in their specific areas of expertise.

We review potential incentives for a shift from pre- to post-GPAI baselines, including advances in AI capabilities, technology diffusion, and competitive pressures. We also examine how relying on post-GPAI baselines to claim minimal marginal risk can increase overall risk across the AI ecosystem, though post-GPAI baselines may be appropriate when used carefully and transparently.

[2]Risk can refer to harms weighted by their likelihood, while dangerous capabilities refer to abilities of an AI model that could enable severe harms. Pre-GPAI and post-GPAI baselines can be defined in reference to either risk or capabilities.

2. Survey of AI company policies

Various AI company safety frameworks discuss assessing AI capabilities or risk relative to baselines. Safety policies vary in how they address several key questions:

- **Pre-GPAI versus post-GPAI baselines.** Does the document clarify whether it is using a pre-GPAI or post-GPAI baseline, especially when referring to "uplift" or "automation", or is this left ambiguous?
- **Zero versus incremental marginal risk.** Does the policy allow for relaxing safeguards as long as there is no marginal risk or only a small marginal risk introduced, compared to existing AI models? Note that in practice, distinguishing between zero versus incremental marginal risk may be difficult.
- **Verifiability**: When a developer relies on a marginal risk argument to assert that its model poses no substantially new risks, will this claim be transparent and verifiable?

OpenAI's Preparedness Framework, Version 2 (OpenAI,

Table 2. Pre-GPAI and post-GPAI marginal risk baselines in company policies, excluding quotes where the baseline is unspecified.

Document	Title	Summary	Baseline
Meta (2025)	Frontier AI Framework (February 2025)	"Net new" risks "cannot currently be realized as described . . . with existing tools and resources . . . A frontier AI is assigned to the critical risk threshold if we assess that it would uniquely enable execution of a threat scenario". In the definition of "uplift studies," "existing resources" include "textbooks, the internet, and existing AI models."	Plausibly post-GPAI
Google (2025a)	Frontier Safety Framework	"CBRN uplift 1": "Compared to a counterfactual of not using generative AI systems" "Cyber uplift level 1": "Relative to the counterfactual of using 2024 AI technology and tooling" "Machine Learning R&D autonomy level 1": "relative to humans augmented by AI tools"	Pre- and post-GPAI
Microsoft (2025)	Frontier Governance Framework (February 2025)	"This holistic risk assessment also considers the marginal capability uplift a model may provide over and above currently available tools and information, including currently available open-weight models."	Post-GPAI
Anthropic (2025)	Responsible Scaling Policy 2.1 (March 2025)	CBRN-3 threshold definition relative to "2023-level online resources"	Pre-GPAI
OpenAI (2025b)	Preparedness Framework v2 (April 2025)	"Marginal risk" section: Safeguards may adjust if other developers release high-risk systems without comparable protections	Post-GPAI

2025b) contains a section titled "Marginal risk" that allows OpenAI to adjust its safeguard commitments for its High and Critical risk models, if OpenAI can verify that another AI developer with High or Critical risk models has not adopted "comparable safeguards." In this situation, OpenAI commits to "publicly acknowledge that we are making the adjustment, and, in order to avoid a race to the bottom on safety, we keep our safeguards at a level more protective than the other AI developer, and share information to validate this claim." In other words, OpenAI may relax safety and security safeguards for its High and Critical risk models due to another AI developer's actions, as long as OpenAI verifiably does not introduce marginal risk compared to the other AI model.

This "escape clause" (Karnofsky, 2024) functions similarly to language in Anthropic's Responsible Scaling Policy (Anthropic, 2025) and Google DeepMind's Frontier Safety Framework (Google, 2025a), although these do not explicitly use the term "marginal risk." There are still notable differences; for example, Anthropic's Responsible Scaling Policy allows for "small" "incremental increase in risk attributable to us" alongside "invest[ing] significantly in making a case to the U.S. government for taking regulatory action to mitigate such risk to acceptable levels."

Clearly describing baselines used in marginal risk assessments is particularly important when defining capability thresholds that require enhanced safeguards. Developers currently use a mix of baselines and sometimes do not clarify which is used. Three notable examples are as follows.

Anthropic's "CBRN-3" (chemical, biological, radiological, and nuclear weapons) capability threshold is clearly defined against a pre-GPAI baseline, representing significant assistance "with full model access versus 2023-level online resources." Other capability thresholds in the policy (e.g., CBRN-4, AI R&D-4) do not specify whether uplift or automation capabilities are defined relative to pre-GPAI or post-GPAI baselines (Anthropic, 2025).

Google DeepMind's Frontier Safety Framework explicitly uses pre-GPAI baselines for its "CBRN uplift 1" threshold ("Compared to a counterfactual of not using generative AI systems") and its "Cyber uplift level 1" threshold ("Relative to the counterfactual of using 2024 AI technology and tooling"), while using a post-GPAI baseline for its "Machine Learning R&D autonomy level 1 threshold" ("cost comparison is relative to humans augmented by AI tools") (Google, 2025a).

Microsoft's Frontier AI Governance Framework considers "marginal capability uplift" beyond "currently available tools and information, including currently available open-weight models" as part of "holistic risk assessment," suggesting post-GPAI baselines. However, the threshold definitions referring to "meaningful uplift" to humans ("CBRN weapons," "Offensive cyberoperations") or automating "human labor" ("Advanced autonomy") do not clarify whether baseline human-level capabilities include AI assistance (Microsoft,

2025).

3. Distinguishing between marginal risks

3.1. Drivers of a possible shift from pre-GPAI to post-GPAI baselines

Several factors may contribute to a potential shift from pre-GPAI to post-GPAI marginal risk assessment for decision making:

- **Technology diffusion.** As the industry releases increasingly capable models that become more widely used, pre-GPAI baselines appear less relevant for understanding the impact of a specific model, particularly as open-weight models without robust safeguards become more capable.
- **Competitive dynamics.** AI development operates under competitive market pressures. As AI technology generates greater economic value, organizations face pressure to maintain competitive advantages. Safety practices that delay development may be scrutinized for their impact on market position. A post-GPAI baseline approach may appeal to organizations by establishing a relative safety standard ("not significantly more dangerous than existing models") rather than an absolute one ("not significantly more dangerous than the pre-GPAI world").
- **Capability advancement.** As AI capabilities progress, it becomes more challenging to demonstrate that they do not pose significant marginal risk compared to pre-GPAI baselines. Such a demonstration may require increasingly complex and expensive evaluation methods. Eventually, frontier AI capabilities may inevitably have "significant marginal risk" beyond a pre-GPAI baseline and necessitate costly safety and security mitigations. Organizations with commitments linking "significant marginal risk" to specific safety requirements could face practical challenges maintaining these frameworks as capabilities evolve, potentially creating incentives to redefine how marginal risk is measured.

3.2. Impact of shifting baselines

If AI developers introduce models while relying on arguments based on low post-GPAI marginal risk, even if such models have pose large risks relative to pre-GPAI baselines, this approach would have significant implications for safety. Since a post-GPAI baseline is not constant, it enables a "boiling frog" dynamic: if the risk level of the most dangerous available model progressively increases, aggregate risk from all models can increase significantly, without any single step constituting a major escalation. Developers might release models that individually only seem marginally riskier, but together create a landscape that is more hazardous than would have been initially acceptable.

This situation could be viable if society also incrementally develops defenses that neutralize the risk (see also Bernardi et al., 2025). This depends on offense–defense balances which are unclear and domain-dependent. Biological risk has been argued to be offense-dominant, as developing biological weapons is more cost-effective than building defenses (Koblentz, 2011), while AI may have mixed impacts on cyber conflict (Lohn, 2025).

4. Suggestions for AI developers

Based on our analysis of pre-GPAI and post-GPAI baselines for marginal risk assessment, we propose the following suggestions for AI developers:

- **Clarify whether a pre-GPAI or post-GPAI baseline is used.** AI developers should explicitly identify which baseline they are using when making claims or commitments around marginal risk. Many capability thresholds currently use ambiguous language about increasing human capabilities or automating expert-level work without specifying whether the human baseline is AI-assisted. This clarity is important for meaningful interpretation of risk commitments.
- **Increase transparency around risk factors.** Developers should provide sufficient detail about model capabilities, risks, and safeguards to enable external parties to make informed judgments about marginal risk. When reporting post-GPAI marginal risk, developers could also disclose pre-GPAI marginal risk to provide a more complete risk picture. If an AI developer determines that overall risk has reached unacceptable levels, even if its specific model adds only incremental risk, the developer should communicate this assessment publicly. Transparency should extend to powerful models used internally but not yet deployed (Kinniment, 2025). If a developer is not able to provide detailed information publicly, due to competitive secrets or information hazards, they could provide details to a third party who provides independent or anonymized attestation.
- **Address aggregate risk through concrete actions.** By the time overall risk reaches concerning levels, developers may advocate to relevant government authorities to take actions that would reduce risks to an acceptable level, especially if they are introducing a model with high marginal risk relative to pre-GPAI baselines (cf. escape clause of Anthropic (2024a)), though these actions may require advance preparation to be most effective (Wasil et al., 2024). AI developers may also

invest in defensive technology that would reduce risks (Bernardi et al., 2025), potentially aided by AI assistance.

5. Conclusion

Our examination of pre-GPAI and post-GPAI baselines reveals important considerations for marginal risk assessment in AI governance. A possible shift toward post-GPAI baselines may be driven by practical realities of technological advancement, market competition, and the difficulty of maintaining pre-GPAI comparisons as capabilities progress. However, incremental increases in marginal risk over time can lead to large or unacceptable aggregate risk.

Impact Statement

This paper examines how AI companies assess risk from new AI models. Our analysis aims to increase transparency in risk reporting, particularly ensuring that society can understand the aggregate risk level.

References

AI Security Institute. Advanced AI evaluations at AISI: May update. Technical report, AI Security Institute, 5 2024. URL https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

Anthropic. Responsible scaling policy. Technical report, Anthropic, October 2024a. URL https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf.

Anthropic. Responsible Scaling Policy evaluations report – Claude 3 Opus. Technical report, Anthropic, 2024b. URL https://cdn.sanity.io/files/4zrzovbb/website/210523b8e11b09c704c5e185fd362fe9e648d457.pdf. Accessed July 2025.

Anthropic. Claude 3.7 Sonnet system card. Technical report, Anthropic, 2 2025. URL https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf.

Anthropic. Claude 4 system card. Technical report, Anthropic, 2025. URL https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf. Accessed July 2025.

Anthropic. Responsible Scaling Policy. Technical report, Anthropic, 3 2025. URL https://www-cdn.anthropic.com/17310f6d70ae5627f55313ed067afc1a762a4068.pdf.

Bernardi, J., Mukobi, G., Greaves, H., Heim, L., and Anderljung, M. Societal adaptation to advanced ai, 2025. URL https://arxiv.org/abs/2405.10295.

Bommasani, R., Singer, S. R., Appel, R. E., Cen, S., Cooper, A. F., Cryst, E., Gailmard, L. A., Klaus, I., Lee, M. M., Raji, I. D., Reuel, A., Spence, D., Wan, A., Wang, A., Zhang, D., Ho, D. E., Liang, P., Song, D., Gonzalez, J. E., Zittrain, J., Chayes, J. T., Cuellar, M.-F., and Fei-Fei, L. The california report on frontier ai policy, 2025. URL https://arxiv.org/abs/2506.17303.

Google. Frontier Safety Framework. Technical report, Google, 2 2025a. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0%20(1).pdf.

Google. Gemini 2.5 Pro Preview model card. Technical report, Google, 5 2025b. URL https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf. Updated May 9, 2025.

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

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., Song, D., Storchan, V., Zhang, D., Ho, D. E., Liang, P., and Narayanan, A. On the societal impact of open foundation models, 2024. URL https://arxiv.org/abs/2403.07918.

Karnofsky, H. If–then commitments for AI risk reduction, 9 2024. URL https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction?lang=en.

Kinniment, M. AI models can be dangerous before public deployment. https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-depl 01 2025.

Koblentz, G. D. *1. Offense, Defense, and Deterrence*, pp. 9–52. Cornell University Press, Ithaca, NY, 2011. ISBN 9780801458903. doi: doi:10.7591/9780801458903-004. URL https://doi.org/10.7591/9780801458903-004.

Lohn, A. J. The impact of ai on the cyber offense-defense balance and the character of cyber conflict, 2025. URL https://arxiv.org/abs/2504.13371.

Meta. Frontier AI Framework. Technical report, Meta, 2 2025. URL https://ai.meta.com/static-resource/meta-frontier-ai-framework.

Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 7 2024. URL https://arxiv.org/abs/2407.21783.

METR. Common elements of frontier AI safety policies. Technical report, METR, 03 2025.

Microsoft. Frontier Governance Framework. Technical report, Microsoft, 2 2025. URL https://go.microsoft.com/fwlink/?linkid=2303737&clcid=0x409&culture=en-us&country=us.

Mouton, C. A., Lucas, C., and Guest, E. The operational risks of AI in large-scale biological attacks. Technical report, RAND Corporation, 1 2024. URL https://www.rand.org/pubs/research_reports/RRA2977-2.html.

OpenAI. OpenAI GPT-4.5 system card. Technical report, OpenAI, 2 2025a. URL https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf.

OpenAI. Preparedness Framework. Technical report, OpenAI, 4 2025b. URL https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

Patwardhan, T., Liu, K., Markov, T., Chowdhury, N., Leet, D., Cone, N., Maltbie, C., Huizinga, J., Wainwright, C., Jackson, S. F., Adler, S., Casagrande, R., and Madry, A. Building an early warning system for LLM-aided biological threat creation, 1 2024. URL https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 5 2023. URL https://arxiv.org/abs/2305.15324.

Wan, S., Nikolaidis, C., Song, D., Molnar, D., Crnkovich, J., Grace, J., Bhatt, M., Chennabasappa, S., Whitman, S., Ding, S., Ionescu, V., Li, Y., and Saxe, J. Cyberseceval 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models, 2024. URL https://arxiv.org/abs/2408.01605.

Wasil, A., Smith, E., Katzke, C., and Bullock, J. Ai emergency preparedness: Examining the federal government's ability to detect and respond to ai-related national security threats, 2024. URL https://arxiv.org/abs/2407.17347.

Zhu, Y., Kellermann, A., Gupta, A., Li, P., Fang, R., Bindu, R., and Kang, D. Teams of llm agents can exploit zero-day vulnerabilities, 2025. URL https://arxiv.org/abs/2406.01637.