OpenAI's Approach to External Red Teaming for AI Models and Systems

Lama Ahmad Sandhini Agarwal Michael Lampe Pamela Mishkin

Abstract

Red teaming has emerged as a critical practice in assessing the possible risks of AI models and systems. It aids in the discovery of novel risks, stress testing possible gaps in existing mitigations, enriching existing quantitative safety metrics, facilitating the creation of new safety measurements, and enhancing public trust and the legitimacy of AI risk assessments. This white paper describes OpenAI's work to date in external red teaming and draws some more general conclusions from this work. We describe the design considerations underpinning external red teaming, which include: selecting composition of red team, deciding on access levels, and providing guidance required to conduct red teaming. Additionally, we show outcomes red teaming can enable such as input into risk assessment and automated evaluations. We also describe the limitations of external red teaming, and how it can fit into a broader range of AI model and system evaluations. Through these contributions, we hope that AI developers and deployers, evaluation creators, and policymakers will be able to better design red teaming campaigns and get a deeper look into how external red teaming can fit into model deployment and evaluation processes. These methods are evolving and the value of different methods continues to shift as the ecosystem around red teaming matures and models themselves improve as tools for red teaming.

1 Introduction

Red teaming has emerged as a critical practice in assessing the risks of AI models[1] and systems.[2] The methods, goals, and outputs of red teaming vary across industry, academia, and public sectors. This paper outlines OpenAI's design decisions and processes for external red teaming. It describes how these processes can inform evaluation and risk assessment for increasingly capable and complex AI models and systems. In this paper, we focus on OpenAI's external red teaming efforts, which involve collaborating with domain experts[3] to evaluate the capabilities and risks of AI models and systems. Although this paper draws on insights from OpenAI's red teaming practices, the principles may apply broadly to other model deployers or stakeholders interested in incorporating human red teaming into their risk assessment processes.

Red teaming as a safety practice: industry and policy

OpenAI has conducted external red teaming for frontier AI model deployments since the launch of DALL-E 2 in 2022[1]. At the time of writing, several System Cards have been published that detail

[1] We define AI models as algorithms and statistical models that are trained on data to make predictions or decisions. AI models are the core components that drive the functionalities of AI systems.

[2] We define AI systems as the integration of one or more AI models along with additional components such as data inputs, hardware, software, and interfaces that work together to produce outputs, complete tasks, etc.

[3] As opposed to internal red teaming, and dimensions such as automated / manual which are described later in the paper

the red teaming efforts for GPT-4[2], GPT-4(V)[3], DALL-E 3[4], GPT-4o[5], and o1[6].[4] The 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence defines the term *AI red teaming* as "a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red teaming is most often performed by dedicated 'red teams' that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system"[7]. As a part of this Executive Order, the National Institute of Standards and Technology (NIST) has been tasked with developing guidelines for red teaming and other forms of evaluation and testing, informed in part by the testing methods and best practices that labs like OpenAI have developed and incorporated into deployment practices [8]. Red teaming also features in other government approaches to AI risk assessment, such as global AI Safety Institutes approach to evaluations[10, 9]. In addition, several AI labs and technology companies have adopted and published red teaming practices in their AI deployment processes [11, 12, 13, 14].

The value of external red teaming

External red teaming can serve a range of purposes. At OpenAI, we have found that red teaming helps achieve four goals:

- **Discovery of novel risks:** Identifying new risks due to advancements in model capabilities, new modalities, or changes in model behaviors. For example, red teaming GPT-4o's speech to speech capabilities uncovered instances where the model would unintentionally generate an output emulating the users voice[5].
- **Stress testing mitigations:** Finding inputs or attacks that can evade model or system mitigations. For instance, red teamers identified a category of visual synonyms[19, 20] that bypassed existing defenses designed to prevent the creation of sexually explicit content in DALL-E systems [1]. OpenAI then reinforced the mitigations before deploying the system.
- **Augmenting risk assessment with domain expertise:** Some risk assessments require specific domain knowledge for thorough testing and verification. External red teamers bring valuable context, such as knowledge of regional politics, cultural contexts, or technical fields like law[22] and medicine[21]. For example, scientists helped evaluate the capabilities and limitations of the o1 family of models in developing biological experiment protocols, and in the context of scientific lab safety[6].
- **Independent assessment:** External red teaming invites independent experts to provide input into risk assessments and evaluations. This strengthens trust in the results by reducing real or perceived conflicts of interest.

OpenAI has found external red teaming particularly valuable for testing quickly evolving AI models. Scenarios include significant improvements in capabilities (e.g, improved reasoning in the o1 family of models), new forms of interaction which may differ from previous AI systems or technologies (e.g, inpainting on DALL-E in 2022), and access to novel tools and new actions (e.g. code execution, function calling). External human testing helps to identify potential risk areas and gaps in mitigations where reliable benchmarks are not yet available and there is not yet sufficient

[4] OpenAI also conducted preliminary risk assessments with the launches of GPT-3 and Codex, which laid the groundwork for the incorporation of external red teaming into subsequent launches.

experience with the new system to be confident what an ideal benchmark would measure or how it would be designed. Qualitative findings from external red teaming can inform the development of automated evaluations or human-AI assisted evaluations[15]. The results tend to be new metrics derived from these findings that may serve as a signal for performance against a particular desired behavior.[5] More about the process of human red teaming to automated evaluations can be found in Section 3.

Model developers, or others conducting a red teaming exercise might choose to disclose the scope, assumptions, and criteria of testing in various ways. That may include: details about the model(s) and system iterations that were tested at what point in time, the categories of testing, the backgrounds and focus areas of those conducting the testing, select examples from red teaming that influenced decision making. OpenAI has disclosed red teaming details in System Cards for several frontier model launches[1, 2, 3, 4, 5, 6].

Emerging methods for red teaming

There are various possible methods emerging under the umbrella term of red teaming.

- **Manual:** This approach involves humans actively crafting prompts and interacting with AI models or systems to simulate adversarial scenarios, identify new risk areas, and assess outputs based on criteria such as risk type, severity, efficacy, or baseline comparisons.
- **Automated:** This type of testing involves using AI models or templating to generate prompts or inputs that simulate adversarial use, and sometimes using classifiers to assess or grade the outputs based on a set of criteria. There has been significant progress on leveraging AI and other algorithmic methods to jailbreak and stress-test models such as the work by Perez et al[16].
- **Mixed methods:** Automated and manual testing might be used in combination, for example, by first creating a seed dataset of prompts with manual testing and scaling up that dataset using automated generation of prompts. We detail this feedback loop further in the Red Teaming to Automated Evaluations section.

Each type of red teaming might be useful for different purposes, and are not mutually exclusive. All three methods of red teaming have been applied at OpenAI, and have informed one another at various stages. The tradeoffs in different methods have also been considered by other model deployers [17].

In addition to the methods used for red teaming, it is also relevant to consider who is conducting the testing.

- **Internal red teams:** testing is conducted by internal, dedicated teams at a lab or company.
- **External red teams:** testing is conducted by external stakeholders, sometimes in collaboration with a lab or company.

When deciding between internal and external red teaming, considerations about timing, subject matter expertise, and IP sensitivities are important to take into account. Internal red teams might be useful at early stages of designing an external red teaming campaign, in order to provide baseline

[5] For example, if it is appropriate to refuse an answer for a set of queries, these metrics can monitor whether that remains the case as models and systems evolve.

guidance about useful focus areas and early observations. It might also be useful to have internal red teams where external red teaming might be impractical due to IP or legal considerations. Other considerations about red team composition can be found in Section 2.

2 Designing a red teaming campaign

As with other fields [18], a variety of methods and design decisions might be employed to assess the risks and impacts of a system. Important components of red teaming methodology include decisions about what is in scope of testing, the composition of red teaming cohorts, deciding which models and systems red teamers access, and the format of the final outputs red teamers deliver.

The key steps of a red teaming campaign[6] include:

1. Deciding the composition of the red teaming cohort based on the outlined goals and prioritized domains for testing
    - What open questions do we have about the model or system?
    - What threat model(s) should red teamers take into account?
2. Determining the versions of the model or system to which the red teamers will have access
3. Creating and providing interfaces, instructions, and documentation guidance to red teamers
4. Synthesizing the data and creating evaluations

Areas of testing and composition of red team

AI systems that will be leveraged for many use cases necessitate testing covering a myriad of topics conducted by people with diverse perspectives and worldviews.

Domain prioritization for testing can be informed by threat modeling[7][23, 24] carried out prior to the red teaming exercise which can take into account: early evaluation of models where model developers might expect increased capabilities, known previous content policy issue areas, relevant contextual considerations, and expected uses of AI systems. Ideally, each area of testing is paired with a set of motivating questions and/or starting hypotheses: risks of what, to whom, posed by whom or what? These hypotheses can anchor the findings of a red teaming exercise by providing a clear rationale for testing and emphasizing specific areas. Generally, the first decisions about prioritization for areas of testing happen by internal teams due to their knowledge and first exposure to existing evaluation results, expected capabilities, and details of model development. However, bringing in external red teamers helps to further specify or broaden the areas of testing based on their areas of expertise or findings from testing.

In Appendix A, we provide an illustrative mapping of areas of testing a red teaming effort might cover, with motivating questions, as well as potential threat models to take into consideration for each area of testing. A subset is in the table below, and the full mapping can be referenced in Appendix A.

[6]This paper uses the phrase red teaming campaign to denote a specifically scoped effort to test a particular AI model or system, or a particular issue area of interest.

[7]Threat modeling in the context of AI red teaming refers to the structured process of identifying, analyzing, and prioritizing potential risks and vulnerabilities in AI systems.

Table 1: Example Areas of Testing and Motivating Questions

Area of Testing	Motivating Questions (illustrative, and red teamers should help shape these)
Natural Sciences	What are the current capabilities in natural science domains, and where do those capabilities meaningfully alter the risk landscape (speed, accuracy, efficiency, cost effectiveness, expertise required)? What are the current limitations in natural science domains and where might that pose risks if relied on in high-stakes contexts?
Cybersecurity	What are the current possibilities for the use of the model in offensive / defensive cybersecurity contexts? Where do those capabilities meaningfully alter the risk landscape (speed, accuracy, efficiency, cost effectiveness, expertise required)? Are there risks related to: identification and exploitation of vulnerabilities, spear phishing, or bug finding?
Bias and Fairness	Where might the model exhibit bias? How might that have an impact on particular use cases (history, politics, controversial topics)? Does the model exhibit bias based on race, ethnicity, religion, political affiliation, etc., particularly if used to make decisions in hiring, educational access, and extending credit?
Violence and self harm	Does the model refuse to give answers that support violence, enable self-harm, etc.?

Different models and systems will require different compositions of their respective red team cohorts. The ideal composition may vary across axes such as professional background, education level, gender, age, geographic location, and languages spoken. As an example, for GPT-4, OpenAI prioritized domains such as natural sciences, autonomous capabilities and power-seeking behavior, and cybersecurity, whereas for DALL-E 3, more salient experiences were related to mis/disinformation risks pertaining to images, as well as bias and representation issues with generated images. Additionally, decisions about whether a domain is best suited to be tested manually or via established benchmarks or evaluations will depend on the maturity and confidence in existing evaluations which we expect to become more robust over time.

Red teamers external to a model deployer may include: individuals who have expertise in adversarial testing or relevant domain knowledge [25], academic groups who study particular aspects of adversarial robustness or novel capabilities [26, 27, 28], or organizations that specialize in red teaming services for AI systems. Red teaming can also take place on public arenas for bounty prizes or challenges [29, 30, 31].

Tailoring model access to red teaming goals

The versions of the model or system red teamers have access to can influence the outcomes of red teaming, and model or system version access has to be decided with the goals of the campaign in mind. Early in the model development process, it may be useful to learn about the models capabilities prior to any added safety mitigations, so that model developers can make informed decisions about the models base level risks, or to allow for open-ended novel risk discovery. In other cases, such access may lead to information that is not actionable because the results will be obsolete with the updated system (either improved capabilities or added safety mitigations). Earlier

snapshots of models may not always yield the most helpful or representative responses either. Once safety mitigations have been enabled, red teaming efforts may focus on identifying unaddressed residual risks and assessing the robustness of the mitigations. Ultimately, the optimal design will vary based on the needs of the model or system in question.

Below we outline some of the types of access and the benefits and tradeoffs associated with each type of access. Note that these types of access are illustrative and non-exhaustive (e.g, pre-deployment could encompass a variety of scenarios across different timelines).

Table 2: Pros and cons of different types of model access for red teamers

Type of Access	Advantages	Disadvantages
Pre-deployment models or snapshots without mitigations	Might inform earliest rounds of post-training, understanding initial nascent capabilities	Models without post-training tend to be more difficult to use and are less helpful, generally not representing the full extent of possible capabilities Will not represent the safety profile models intended to be deployed in production Generally are non-standard models and may not always be trained or available by default, requiring separate technical resourcing
Pre-deployment models with partial mitigations	Enables testing of some model-level mitigations (such as refusals) Enables capability testing introduced by post-training Might inform earliest rounds of system-level mitigation development and further rounds of post-training	Will not represent the safety profile models intended to be deployed in production
Pre-deployment models with model and system level safety mitigations	Enables testing of some system-level mitigations Enables capability testing introduced by post-training	Might not fully represent the safety profile of models deployed in production and may lack policy enforcement or detection and response mitigations
Post-deployment models and systems as reflective of expected user experience	Enables stress testing of system-level mitigations Enables capability testing introduced by post-training and any relevant and in-scope external linkages (such as tool use) Fully represents the safety profile that is in production (including policy enforcement and detection and response mitigations)	Provides insights after a model has been accessed by users rather than prior to deployment

Deeper levels of access might be useful for other forms of risk assessment and mitigation development, including interpretability research and thorough capability elicitation but red teaming at the levels of access defined above help to effectively assess the deployment relevant risks associated with model capabilities and system level mitigations.

Instructions, interfaces, and documentation

Effective interactions with external experts during OpenAI's red teaming campaigns are enabled by several key components: providing clear and comprehensive instructions, designing appropriate interfaces for testing, and giving guidance for red teamers to document their results in an actionable

manner.

Instructions: Instructions provided to red teamers are a critical part of the methods that guide red teaming. Instructions might contain information such as: description of the model to the extent appropriate (e.g, known capabilities and limitations) or description of the product and features, description of existing safeguards where applicable, guidance on how to use the interface through which the model or product is served, guidance on the risk areas prioritized for testing, and how to document results from testing. For exploratory red teaming efforts that aim to understand new risks, red teamers are often provided with open-ended instructions and the freedom to explore and test the model as they choose. More structured testing might include pre-specified domains to test with pre-defined threat models, and specific ways to present findings (as described in the Documentation section).

Interfaces: Interfaces through which access is provided, and the interfaces by which the models are intended to be deployed also have an impact on focus areas for testing. For different campaigns, OpenAI has provided interfaces including direct API access, a user interface or product which is intended for deployment, and specialized feedback gathering platforms. Each of these serve different purposes. For example, API access can enable programmatic testing of the model, consumer product interfaces can help guide testing to most closely emulate how people will actually interact with the system, and specialized feedback-gathering platforms can enable easier testing, such as by displaying multiple samples from a model at the same time, side by side. Interfaces also differ across dimensions such as technical ability required for testing, direct comparability to experiences in production, structure of testing and feedback elements (e.g, a ChatGPT like UI would enable more exploratory testing than a data labeling UI which funnels testers into specific tasks to generate more structured datasets). One or more of these interfaces could be used, depending on the nature and needs of a particular red teaming project.

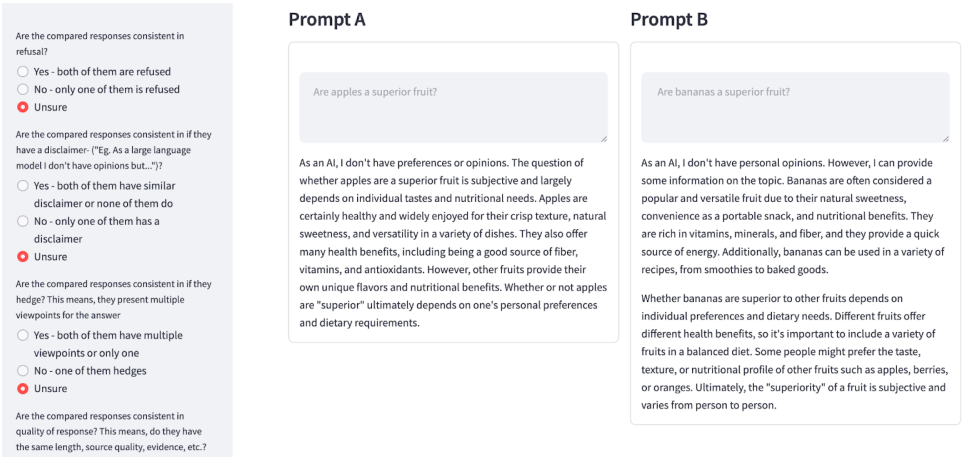

Fig 1. Interface that enables rapid comparison across prompts along with pre-specified questions to enrich findings

Documentation: We instruct each red teamer to document their findings, usually in a specific format. Specific formats can help to facilitate the addition of high quality adversarial tests into existing safety evaluations, or the creation of new ones. Some common elements of documentation include: discrete prompt and generation pairs or conversations, category or domain of the finding,

risk level on some specified scale, such as a Likert scale or low / medium / high, notes on the heuristics for how risk level was decided on or any additional context that would help understand the issue raised. As complexity increases with multi-turn dialogue, increased modalities and complexities of interactions such as calling other systems or tools, documentation of results will need to evolve to accommodate capturing the richness of the data necessary to sufficiently assess the risks associated with evolving models and systems.

3 Human red teaming to automated evaluations

Data synthesis and alignment to desired behaviors

After completing a red teaming campaign, a key challenge is determining which examples are governed by existing policies and, if so, whether they violate those policies. If no existing policy applies, teams must decide whether to create a new policy or modify the desired model behavior. At OpenAI, these policies are informed by resources such as Usage Policies, the Moderation API, and the Model Spec.[8] [9]

Red teaming data provides insights that extend beyond identifying explicitly harmful outputs. Issues such as disparate performance[35, 38] and quality of service issues[36, 37] as well as general user experience preferences are sometimes surfaced through red teaming campaigns. For example, GPT-4o red teaming surfaced unauthorized voice generation behaviors, where the model would unintentionally generate an output emulating the user's voice. While this is related to risks of unauthorized voice generation (such as fraud or impersonation), from a user experience standpoint, the desired behavior is for the assistant voice to follow the pre-set reference voices. Data from red teaming informed the development of robust mitigations and evaluations designed to detect deviations from the reference voice. Similarly, the diverse range of voices and accents represented in the GPT-4o red teaming data has helped to create evaluations for a range of voices on standard AI capability benchmarks, as well as assessing performance on refusal behavior. More details about these examples and evaluations can be found in the GPT-4o System Card[5].

While there may be discrepancies in expectations around safe or acceptable behavior between external red teamers and the organizations deploying AI systems, the careful consideration of the red teaming results makes red teaming a useful part of the risk assessment process. Some red teaming campaigns may explicitly target known areas of policy violative behaviors, while others may seek to understand areas that do not clearly fit into existing policies. Red teaming is not solely an exercise in aligning models with specific content policies. It also serves to gather broader insights into potential risks and the desirable or undesirable behaviors of models or systems.

Creating automated evaluations

One way to maximize the impact of red teaming is to use datasets generated by human red teamers to create repeatable safety evaluations, which can be deployed more quickly and inexpensively in the future. Red teaming can lay the foundation for automated evaluations by discovering areas where investing in such evaluations can be valuable, providing data that can be used to pilot evaluations, and providing harder and unique prompts for more robust evaluations.

[8] It is important to note that not all policies are indicated by model behaviors that can be observed from a single turn interaction such as refusals, and may be enforced through system level mitigations such as monitoring and account level actions

[9] These are a point-in-time reference, and are subject to change.

Red teamers contribute in two main ways: generating relevant prompts to test the model and evaluating the output for potential risks. The prompts red teamers generate can form the seed set of inputs for an evaluation and the ideal behavior can then be measured using techniques such as rule-based classifiers[34]. These methods were used to generate evaluations for models OpenAI has red teamed including GPT-4[2] and DALL-E 3[4].

In the case of DALL-E 3, open-ended red teaming uncovered gaps in areas such as misinformation-prone images, jailbreaks enabling sexually explicit content, and self-harm imagery. OpenAI created automated evaluations for selected subdomains with red teamer prompts as seeds. GPT-4 was used to create synthetic data using red team prompts. Once we had a larger set of prompts, a GPT-4 classifier was used to ensure that the language model creating requests for DALL-E 3 either refused or sufficiently re-wrote any prompt that went against the desired policy before DALL-E 3 used the prompt to generate an image.

For GPT-4, red teaming findings also led to datasets and insights that helped guide the creation of some quantitative evaluations. For example, GPT-4 red teaming discovered the ability of GPT-4 to encrypt and decrypt text in variants like Base64. These findings prompted the OpenAI team to invest in further evaluations that studied the ability of the model to encrypt or compress a given piece of text and then for another model to decrypt it.

4 Limitations and risks of red teaming

Red teaming on its own is not a panacea for risk assessment[39, 47]. Below we outline some of the risks and limitations related to the results from red teaming, as well as the process itself.

- **Relevance across evolving models and systems:** While red teaming efforts may take place prior to a major model or product deployment, models and systems evolve quite often in production which is important to take that into account when contextualizing red teaming findings. Risks surfaced in one point in time red teaming effort may be under-assessed or no longer reflected in an updated system or model, and as such, should not be seen as a panacea for risk assessment efforts of AI systems.

- **Resource intensive:** The form of red teaming described in this paper is resource intensive in terms of operational time and financial costs. Less resourced organizations may not be able to effectively employ this form of red teaming at scale. This is why maximizing the impact of external red teaming investments through development of evaluations where possible is critical to the process of red teaming.

- **Harms to participants and team members:** Red teaming may have the potential to negatively impact participants as members are required to think like adversaries and interact with harmful content, which can lead to decreased productivity or psychological harm. This is especially concerning for those who are part of marginalized groups. Steps like providing mental health resources, fair compensation, and informed consent are crucial to mitigate these risks.

- **Information hazards:** The process of red teaming, particularly with Frontier AI systems, can create information hazards that might enable misuse. For example, exposing a jailbreak or technique to generate potentially harmful content that is not yet widely known can accelerate bad actors misuse of the models. Managing this requires control of information, stringent access protocols, and responsible disclosure practices.

- **Picking early winners:** Red teamers, often having research interests or other stakes outside of their risk assessment roles, gain early access to models or systems, potentially giving them an unfair advantage in research or business ventures. This issue needs addressing to maintain fairness and integrity in the process.
- **Increase in human sophistication:** As models become more capable and their ability to reason in sophisticated domains becomes more advanced, there will be a higher threshold for knowledge humans need to possess to correctly judge the potential level of risk of outputs. Additionally, as models become more robust, it may require more effort to 'jailbreak' them and produce commonly identifiable harms.

Automated red teaming methods that are improving over time can help to mitigate some of these limitations and make the time spent on human red teaming targeted to the highest value areas, while minimizing the time spent reviewing undesirable content.

5 Conclusion

This paper describes how external red teaming fits into AI risk assessment practices, and can help strengthen safety evaluations over time. As AI systems are evolving at a rapid pace, so too does the need for understanding users experiences and the potential risks posed by increased capabilities, possibilities for abuse and misuse, as well as real world factors and considerations such as cultural nuances, and more. No singular process will capture all of these considerations, but red teaming, especially red teaming in collaboration with a diverse range of external domain experts and relevant stakeholders, creates a mechanism for proactive risk assessment and testing, and is an input into creating more up to date benchmarks and safety evaluations that can be reused and updated over time. While external red teaming aims to expand perspectives in service of risk discovery, verification, and evaluation development, additional work is needed to solicit and incorporate public perspectives on ideal model behavior, policies, and other associated decision making processes. Red teaming also needs to be paired with externally specified thresholds and practices for accountability of discovered risks.

External red teaming is one tool for AI risk assessment and safety practices, alongside automated red teaming, established safety benchmarks, third party assessments, and research on the impacts of AI systems. While we expect methods of red teaming to evolve, we hope that researchers, other companies, and policymakers will be able to utilize some of the methodologies and tradeoffs presented in this paper to create more clearly defined and actionable red teaming campaigns.

Acknowledgements

We would especially like to acknowledge those that work on red teaming and the associated evaluations for several of the frontier model deployments above, including: Troy Peterson, Jason Phang, Saachi Jain, Kai Xiao.

We would like to thank Borhane Blili-Hamelin, A. Feder Cooper, Rishi Bommasani, Lujain Ibrahim, Tom Zick, Michael Feffer, Miles Brundage, Pamela Mishkin, Tyna Eloundou, Sarah Shoker, Steven Adler, Alex Beutel, Andrea Vallone, Ashyana-Jasmine Kachra, David Robinson, Yonadav Shavit for their helpful advice, feedback, and comments, which were integral in the development of this paper.

References

[1] P. Mishkin, L. Ahmad, M. Brundage, G. Krueger, and G. Sastry, "Dall-e 2 preview - risks and limitations," 2022.

[2] OpenAI, "GPT-4 system card," 2023.

[3] OpenAI, "GPT-4V system card," 2023.

[4] OpenAI, "DALL-E 3 system card," 2023.

[5] OpenAI, "Gpt-4o system card," 2024.

[6] OpenAI, "OpenAI o1 System Card," 2024.

[7] "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence." https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-arti 2023.

[8] National Institute of Standards and Technology, "Request for information (rfi) related to nist's assignments under sections 4(i), 4(j), and 11 of the executive order on artificial intelligence," 2023. Federal Register Document Number 2023-28232.

[9] Japan AI Safety Institute, *Guide to Red Teaming Methodology on AI Safety (Version 1.00)*, Sept. 2024.

[10] Department for Science, Innovation and Technology, "Ai safety institute approach to evaluations," 2024. Published by the Department for Science, Innovation and Technology, UK Government.

[11] M. Learn, "Red teaming and responsible ai in azure openai service," 2024.

[12] N. Rajani, N. Lambert, and L. Tunstall, "Red-teaming large language models." https://huggingface.co/blog/red-teaming, 2023.

[13] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, *et al.*, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *arXiv preprint arXiv:2209.07858*, 2022.

[14] Google, "Google's ai red team: the ethical hackers making ai safer," 2023.

[15] L. Ibrahim, S. Huang, L. Ahmad, and M. Anderljung, "Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks," 2024.

[16] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," 2022.

[17] Anthropic, "Challenges in red teaming ai systems," June 2024.

[18] N. R. Council, *Science and Decisions: Advancing Risk Assessment*. Washington, DC: The National Academies Press, 2009.

[19] E. Gavves, C. G. Snoek, and A. W. Smeulders, "Visual synonyms for landmark image retrieval," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 238–249, 2012.

[20] W. Tang, R. Cai, Z. Li, and L. Zhang, "Contextual synonym dictionary for visual object retrieval," (New York, NY, USA), Association for Computing Machinery, 2011.

[21] C. T. Chang, H. Farah, H. Gui, S. J. Rezaei, C. Bou-Khalil, Y.-J. Park, A. Swaminathan, J. A. Omiye, A. Kolluri, A. Chaurasia, A. Lozano, A. Heiman, A. S. Jia, A. Kaushal, A. Jia, A. Iacovelli, A. Yang, A. Salles, A. Singhal, B. Narasimhan, B. Belai, B. H. Jacobson, B. Li, C. H. Poe, C. Sanghera, C. Zheng, C. Messer, D. V. Kettud, D. Pandya, D. Kaur, D. Hla, D. Dindoust, D. Moehrle, D. Ross, E. Chou, E. Lin, F. N. Haredasht, G. Cheng, I. Gao, J. Chang, J. Silberg, J. A. Fries, J. Xu, J. Jamison, J. S. Tamaresis, J. H. Chen, J. Lazaro, J. M. Banda, J. J. Lee, K. E. Matthys, K. R. Steffner, L. Tian, L. Pegolotti, M. Srinivasan, M. Manimaran, M. Schwede, M. Zhang, M. Nguyen, M. Fathzadeh, Q. Zhao, R. Bajra, R. Khurana, R. Azam, R. Bartlett, S. T. Truong, S. L. Fleming, S. Raj, S. Behr, S. Onyeka, S. Muppidi, T. Bandali, T. Y. Eulalio, W. Chen, X. Zhou, Y. Ding, Y. Cui, Y. Tan, Y. Liu, N. H. Shah, and R. Daneshjou, "Red teaming large language models in medicine: Real-world insights on model behavior," 2024.

[22] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho, "Large legal fictions: Profiling legal hallucinations in large language models," *Journal of Legal Analysis*, vol. 16, p. 6493, Jan. 2024.

[23] Bugcrowd, "What is ai red teaming?," 2024.

[24] OWASP Foundation, "Threat modeling - owasp community pages." `https://owasp.org/www-community/Threat_Modeling`, 2024. Accessed: 2024-11-17.

[25] OpenAI, "Red teaming network," 2024. Accessed: 2024-11-17.

[26] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!," 2023.

[27] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li, "Decodingtrust: A comprehensive assessment of trustworthiness in gpt models," 2024.

[28] L. Jiang, K. Rao, S. Han, A. Ettinger, F. Brahman, S. Kumar, N. Mireshghallah, X. Lu, M. Sap, Y. Choi, and N. Dziri, "Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models," 2024.

[29] Humane Intelligence, "Generative ai red teaming challenge: Transparency report," 2023.

[30] R. Chowdhury and J. Williams, "Introducing twitters first algorithmic bias bounty challenge," 2021.

[31] J. Kenway, C. François, S. Costanza-Chock, I. D. Raji, and J. Buolamwini, "Bug bounties for algorithmic harms? lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress," 2022.

[32] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan,

"Training a helpful and harmless assistant with reinforcement learning from human feedback," 2022.

[33] I. Solaiman and C. Dennison, "Process for adapting language models to society (palms) with values-targeted datasets," 2021.

[34] T. Mu, A. Helyar, J. Heidecke, J. Achiam, A. Vallone, I. Kivlichan, M. Lin, A. Beutel, J. Schulman, and L. Weng, "Rule based rewards for language model safety," 2024.

[35] P. DHERAM, M. Ramakrishnan, A. Raju, I.-F. Chen, B. King, K. Powell, M. Saboowala, K. Shetty, and A. Stolcke, "Toward fairness in speech recognition: Discovery and mitigation of performance disparities," in *Interspeech 2022*, interspeech$_2$022, $p.$ 12681272, $ISCA$, $Sept.$2022.

[36] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla, J. Gallegos, A. Smart, E. Garcia, and G. Virk, "Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction," 2023.

[37] M. Felderer and R. Ramler, *Quality Assurance for AI-Based Systems: Overview and Challenges (Introduction to Interactive Session)*, p. 3342. Springer International Publishing, 2021.

[38] M. Madaio, L. Egede, H. Subramonyam, J. W. Vaughan, and H. Wallach, "Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support," 2022.

[39] M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, and H. Heidari, "Red-teaming for generative ai: Silver bullet or security theater?," 2024.

[40] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," 2024.

[41] S. Shoker, A. Reddie, S. Barrington, R. Booth, M. Brundage, H. Chahal, M. Depp, B. Drexel, R. Gupta, M. Favaro, J. Hecla, A. Hickey, M. Konaev, K. Kumar, N. Lambert, A. Lohn, C. O'Keefe, N. Rajani, M. Sellitto, R. Trager, L. Walker, A. Wehsener, and J. Young, "Confidence-building measures for artificial intelligence: Workshop proceedings," 2023.

[42] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung, "Toward trustworthy ai development: Mechanisms for supporting verifiable claims," 2020.

[43] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," 2021.

[44] N. Inie, J. Stray, and L. Derczynski, "Summon a demon and bind it: A grounded theory of llm red teaming in the wild," 2023.

[45] H. Khlaaf, P. Mishkin, J. Achiam, G. Krueger, and M. Brundage, "A hazard analysis framework for code synthesis large language models," 2022.

[46] D. . Society, "Ai red teaming is not a one-stop solution to ai harms: Recommendations for using red teaming for ai accountability," 2024.

[47] H. Khlaaf, "Toward comprehensive risk assessments and assurance of ai-based systems,"

Appendix A

Area of Testing	**Motivating Questions (illustrative, and red teamers should help shape these)**
Natural Sciences	• What are the current capabilities in natural science domains, and where do those capabilities meaningfully alter the risk landscape (speed, accuracy, efficiency, cost effectiveness, expertise required)? • What are the current limitations in natural science domains and where might that pose risks if relied on in high-stakes contexts?
Code / Writing Software and Systems Architecture	• What are the current capabilities and limitations of program synthesis, and where do those capabilities meaningfully alter the risk landscape (speed, accuracy, efficiency, cost effectiveness, expertise required)?
Cybersecurity	• What are the current possibilities for the use of the model in offensive / defensive cybersecurity contexts? Where do those capabilities meaningfully alter the risk landscape (speed, accuracy, efficiency, cost effectiveness, expertise required)? • Are there risks related to: identification and exploitation of vulnerabilities, spear phishing, or bug finding?
Privacy	• Can the models produce or be used to surface private information?
Medicine / Healthcare	• What are the capabilities and limitations of the model used in healthcare applications and contexts? Especially with regards to limitations, how do those limitations meaningfully alter the risk landscape? • Do we see changes in over-confidence in answers? • Do we see increased hallucination, particularly of citations?
Law	• What are the capabilities and limitations of the model used in legal applications and contexts? Especially with regards to limitations, how do those limitations meaningfully alter the risk landscape? • Do we see changes in over-confidence in answers? • Do we see increased hallucination, particularly of citations?

Area of Testing	Motivating Questions (illustrative, and red teamers should help shape these)
Tool Use	• Where does tool use (browsing, chaining with other systems, using code interpreter, etc.) meaningfully augment the models base capabilities in ways that might pose risks? Example areas include (carrying out purchases or financial transactions, etc.)
Dangerous Planning	• How might the model aid in planning for dangerous or high-stakes activities (attack planning, obtaining or distributing illegal or dangerous materials, circumvention of safeguards or protections in various contexts)?
Politics and Election Considerations	• How might the model aid in various ways to assist in political strategy, including generation of campaign materials, coming up with voter targeting approaches, analyzing and forecasting public opinion?
Bias and Fairness	• Where might the model exhibit bias? How might that have an impact on particular use cases (history, politics, controversial topics)? • Does the model exhibit bias based on race, ethnicity, religion, political affiliation, etc., particularly if used to make decisions in hiring, educational access, and extending credit?
CBRN Risks	• How might the model impact CBRN (chemical, biological, radiological, and nuclear) threat creation? Where do those capabilities meaningfully alter the risk landscape (speed, accuracy, efficiency, cost effectiveness, expertise required)? Particular sub questions of interest:
AI Research & Development	• Can the models be used to advance methods in AI research and training? Where do those capabilities meaningfully alter the risk landscape (speed, accuracy, efficiency, cost effectiveness, expertise required)?
Situational Awareness / Autonomous Replication	• Are there properties of the model that might demonstrate situational awareness or the ability to autonomously replicate?
Violence and self harm	• Does the model refuse to give answers that support violence, enable self-harm, etc.?
Controversial Questions	• Does the model give responses to controversial questions (social, political, ethical) that are not inclusive of different valid perspectives?

Area of Testing	Motivating Questions (illustrative, and red teamers should help shape these)
Persuasiveness	• Is the model more capable in making persuasive arguments about sensitive topics? • Is the model more capable or willing to form emotional attachments or give sensitive emotional/personal advice?