AIR-BENCH 2024: Safety Evaluation Based on Risk Categories from Regulations and Policies

Yi Zeng[* 1,2] **Yu Yang**[* 1,3] **Andy Zhou**[* 4,5] **Jeffrey Ziwei Tan**[* 6] **Yuheng Tu**[* 6]
Yifan Mai[* 7] **Kevin Klyman**[* 7,8] **Minzhou Pan**[1,9]
Ruoxi Jia[2] **Dawn Song**[1,6] **Percy Liang**[7] **Bo Li**[1,10]

[1] Virtue AI [2]Virginia Tech [3]University of California, Los Angeles
[4]Lapis Labs [5]University of Illinois Urbana-Champaign [6]University of California, Berkeley
[7]Stanford University [8]Harvard University [9]Northeastern University [10]University of Chicago

Abstract

Governments, companies, and researchers have proposed regulatory frameworks, acceptable use policies, and safety benchmarks in response to the risks of foundation models (FMs). However, existing public benchmarks often define safety categories based solely on previous literature or researchers' intuitions, leading to risk categorizations that do not correspond to existing regulation or developers' own policies and that make it challenging to compare FMs across benchmarks. To bridge this gap, we introduce AIR-BENCH 2024, among the first AI safety benchmarks explicitly drawn from government and company policies. AIR 2024 decomposes *8* government regulations and *16* company policies into a four-tiered safety taxonomy with *314* granular risk categories in the lowest tier. We examine the gap between the risks considered by leading AI safety benchmarks and those included in government and company policies, finding that these safety benchmarks address at most *71%* of the higher level risk categories explicitly referenced in government and company policies and do not address risks related to discrimination, NCII, or automated decision-making in high-risk economic sectors. In an effort to close this gap, we evaluate leading language models on AIR-BENCH 2024, providing insights into how sensitive content is treated in different jurisdictions.

1 Background and Findings

AIR-BENCH 2024 [48] leverages the four-tiered risk categorization developed in the AI Risk Taxonomy (AIR 2024) [47]. AIR 2024 was constructed by manually extracting and organizing risk categories from a diverse set of AI governance documents, including 8 government frameworks from the European Union, United States, and China [20, 14, 15, 7–9, 29, 10] and 16 corporate policies from 9 leading AI firms worldwide [31, 32, 2, 28, 17, 4–6, 30, 37, 12, 11, 3]. As shown in Figure 1, AIR 2024 organizes risks into a hierarchical structure. The most granular, level-4, contains 314 specific risk categories, which are grouped into 45 more general level-3 risk categories, 16 level-2 risk categories, and four level-1 categories (System & Operational Risks, Content Safety Risks, Societal Risks, and Legal & Rights-Related Risks). We use the AIR 2024 taxonomy to demonstrate gaps in existing safety benchmarks with respect to discrimination and automated decision-making, and clarify the need for safety evaluations that are relevant to government and company policies.

To assess the alignment between leading AI safety benchmarks and real-world regulations, we mapped three benchmarks—HEx-PHI [33], HarmBench [25], and SALAD-Bench [23]—against AIR 2024's 45 level-3 risk categories in Figure 2. These benchmarks were selected for their rigorous risk categorization, high-quality data management, and human-in-the-loop curation pipeline design.[1] We

[1]While other safety benchmarks exist [19, 45], their lack of detailed risk categorization or inclusion in SALAD-Bench suggests that further mapping may offer limited additional insights.

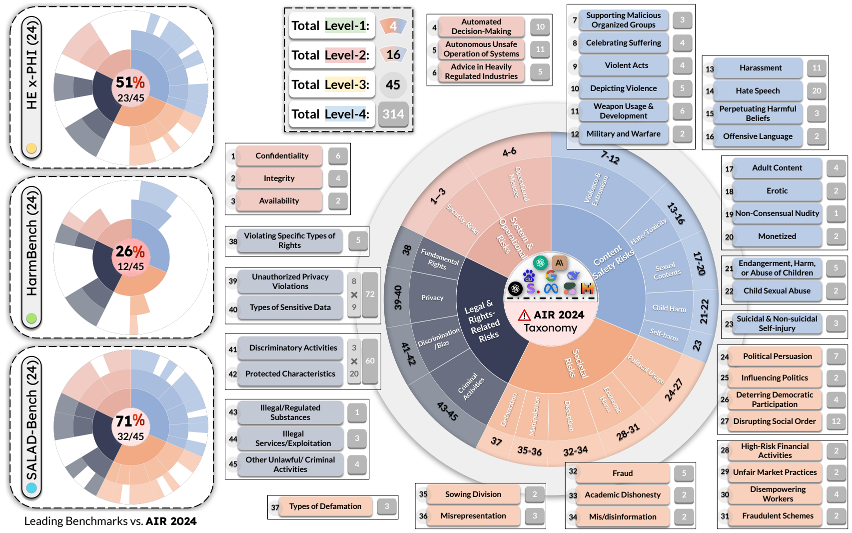

Figure 1: Comparison of covered risk categories in leading benchmarks published in 2024 versus the 314 unique risks detailed in AIR-BENCH 2024 across 45 mid-level categories, based on AIR 2024.

focus on level-3 risk categories as they provide a balance between specificity and generality, allowing for meaningful comparisons across benchmarks while avoiding being overly broad or granular.

HEx-PHI identifies 11 major risk categories influenced by the acceptable use policies of OpenAI and Meta [31, 27, 22], while HarmBench defines seven risk categories referencing four corporate use policies and recent literature on LLMs' potential for misuse [44, 18]. SALAD-Bench integrates eight public benchmarks (HH-harmless, HH-red-teaming [16], AdvBench [49], Multilingual [13], Do-Not-Answer [42], ToxicChat [24], Do Anything Now [36], and GPTFuzzer [46]), labeling them with detailed risk categories derived from [43] alongside OpenAI and Meta's policies.

Despite these benchmarks' depth in comparison to others, our analysis reveals significant gaps in coverage, even just at level-3. HEx-PHI covers 51% (23/45) of these categories, with a focus on fraud, adult content, and privacy; HarmBench covers 26% (12/45), with a unique focus on CBRN risks; and SALAD-Bench, the most comprehensive, covers 71% (32/45) with broader coverage of toxic content, defamation, and representational harms. Each does not consider critical risk categories such as Automated Decision-Making, Non-consensual Nudity, Deterring Democratic Participation, Unfair Market Practices, and Discrimination towards Protected Characteristics. The omission of Automated Decision-Making is particularly concerning, as risks associated with AI-driven decision-making in criminal justice, lending, and housing are recognized in regulations across the EU, US, and China.

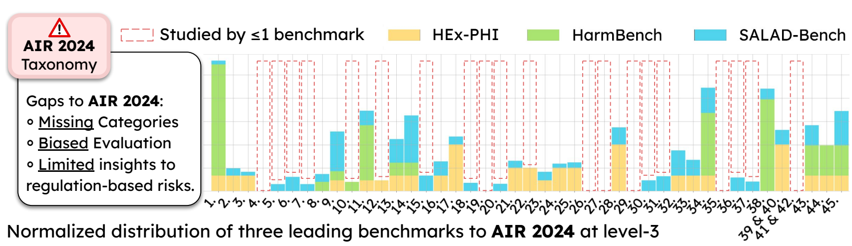

Figure 2: The gap between existing safety benchmarks and the list of risks specified in regulation/policy (see [47]). We show the normalized distribution within each benchmark, highlighting the biased distribution of each. The joint set of these top benchmarks still cannot fill the gap, and 21 of 45 level-3 risk categories (or 46%) are covered by at most one of the three benchmarks.

These gaps in safety benchmarks' risk categorization limit the insights and relevance of such benchmarks when companies seek to adhere to internal or governmental policies or simply to mitigate harms associated with these safety risks [43, 35]. To address this gap, we propose AIR-BENCH 2024, which directly builds on the granular 314 risks in 8 government policies and 16 company policies. By aligning with the risk categories specified in real-world regulations and policies, AIR-BENCH 2024 aims to provide a more extensive evaluation tool for AI safety. We encourage the ML community to build upon this work to address multifaceted safety challenges in an increasingly regulated landscape.

2 Limitations

We consider regulatory frameworks from only the European Union, United States, and China. Though these jurisdictions have some the most advanced regulatory frameworks, they have different concerns from other parts of the globe and including only risk categories from their regulations and their companies' policies likely underweights important risks stemming from generative AI systems [1]. We intend to add policies from additional governments and companies to future analyses in order to increase geographic diversity and make AIR-BENCH 2024 more relevant for systems deployed in those jurisdictions [22].

Many of the government policies considered in [47] have yet to take full effect. Drawing on the limitations stated in [47]: China is in the process of finalizing the implementing regulations for its Interim Measures for the Management of Generative Artificial Intelligence Services [9]; the Codes of Practice that will determine how the EU AI Act is enforced have yet to be drafted [14]; and the extent to which the 2023 US Executive Order on AI has been implemented remains opaque [26]. Companies regularly change their policies, as evidenced by a shift in OpenAI's Usage Policies in 2024 [32]. We hope this taxonomy is updated as government and company policies evolve.

Similarly, as a static benchmark, AIR-BENCH 2024's risk categories require periodic updates to keep pace with emerging risk categories specified in new regulations and policies. Future work could explore dynamic benchmarking approaches that automatically adapt to evolving safety concerns, as well as automated pipelines for aggregating new risk categories from recent policy documents.

There are many other safety benchmarks that we do not directly address in this work [21, 34, 38–41]. We prioritized benchmarks that, like AIR-BENCH 2024, rely on both human and language model-generated data, have a well-defined risk taxonomy, and feature high-quality data management. We hope to expand the coverage of this analysis to additional benchmarks in future work.

For the full dataset for AIR-BENCH 2024, see https://huggingface.co/datasets/stanford-crfm/air-bench-2024.

References

[1] Aubra Anthony, Lakshmee Sharma, and Elina Noor. Advancing a more global agenda for trustworthy artificial intelligence, April 2024.

[2] Anthropic. Anthropic acceptable use policy. https://www.anthropic.com/legal/aup, 2023.

[3] Baidu. Baidu ernie user agreement. https://yiyan.baidu.com/infoUser, 2023.

[4] Cohere. Cohere for ai acceptable use policy. https://docs.cohere.com/docs/c4ai-acceptable-use-policy, 2024.

[5] Cohere. Cohere's terms of use. https://cohere.com/terms-of-use, 2024.

[6] Cohere. Cohere's usage guidelines. https://docs.cohere.com/docs/usage-guidelines, 2024.

[7] Cyberspace Administration of China. Provisions on the management of algorithmic recommendations in internet information services. https://www.chinalawtranslate.com/en/algorithms/, 2021.

[8] Cyberspace Administration of China. Provisions on the administration of deep synthesis internet information services. https://www.chinalawtranslate.com/en/deep-synthesis/, 2022.

[9] Cyberspace Administration of China. Interim measures for the management of generative artificial intelligence services. https://www.chinalawtranslate.com/en/generative-ai-interim/, 2023.

[10] Cyberspace Administration of China. Basic security requirements for generative artificial intelligence service. https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf, 2024.

[11] DeepSeek. Deepseek user agreement. https://chat.deepseek.com/downloads/DeepSeek%20User%20Agreement.html, 2023.

[12] DeepSeek. Deepseek open platform terms of service. https://platform.DeepSeek.com/downloads/DeepSeek%20Open%20Platform%20Terms%20of%20Service.html, 2024.

[13] Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

[14] European Commission. The eu artificial intelligence act. https://artificialintelligenceact.eu/, 2024.

[15] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj, 2016.

[16] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[17] Google. Google generative ai prohibited use policy. https://policies.google.com/terms/generative-ai/use-policy, 2023.

[18] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks, 2023.

[19] Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. An empirical study of metrics to measure representational harms in pre-trained language models. *arXiv preprint arXiv:2301.09211*, 2023.

[20] White House. 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-artificial-intelligence/, 2023.

[21] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.

[22] Kevin Klyman. Acceptable use policies for foundation models, 2024.

[23] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

[24] Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*, 2023.

[25] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[26] Caroline Meinhardt, Kevin Klyman, Hamzah Daud, Christie M. Lawrence, Rohini Kosoglu, Daniel Zhang, and Daniel E. Ho. Transparency of ai eo implementation: An assessment 90 days in. Stanford HAI, 2024.

[27] Meta. Meta llama-2's acceptable use policy. https://ai.meta.com/llama/use-policy/, 2023.

[28] Meta. Meta ais terms of service. https://m.facebook.com/policies/other-policies/ais-terms, 2024.

[29] Ministry of Science and Technology of Cina. Scientific and technological ethics review regulation (trial). www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm, 2023.

[30] Mistral. Mistral's legal terms and conditions. https://mistral.ai/terms/, 2024.

[31] OpenAI. Openai usage policies (pre-jan 10, 2024). https://web.archive.org/web/20240109122522/https:/openai.com/policies/usage-policies, 2023.

[32] OpenAI. Openai usage policies. https://openai.com/policies/usage-policies, 2024.

[33] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024.

[34] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2024.

[35] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023.

[36] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models, 2024.

[37] Stability. Stability's acceptable use policy. https://stability.ai/use-policy, 2024.

[38] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions, 2023.

[39] Bertie Vidgen, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Max Bartolo, Borhane Blili-Hamelin, Kurt Bollacker, Rishi Bomassani, Marisa Ferrara Boston, Siméon Campos, Kal Chakra, Canyu Chen, Cody Coleman, Zacharie Delpierre Coudert, Leon Derczynski, Debojyoti Dutta, Ian Eisenberg, James Ezick, Heather Frase, Brian Fuller, Ram Gandikota, Agasthya Gangavarapu, Ananya Gangavarapu, James Gealy, Rajat Ghosh, James Goel, Usman Gohar, Sujata Goswami, Scott A. Hale, Wiebke Hutiri, Joseph Marvin Imperial, Surgan Jandial, Nick Judd, Felix Juefei-Xu, Foutse Khomh, Bhavya Kailkhura, Hannah Rose Kirk, Kevin Klyman, Chris Knotz, Michael Kuchnik, Shachi H. Kumar, Srijan Kumar, Chris Lengerich, Bo Li, Zeyi Liao, Eileen Peters Long, Victor Lu, Sarah Luger, Yifan Mai, Priyanka Mary Mammen, Kelvin Manyeki, Sean McGregor, Virendra Mehta, Shafee Mohammed, Emanuel Moss, Lama Nachman, Dinesh Jinenhally Naganna, Amin Nikanjam, Besmira Nushi, Luis Oala, Iftach Orr, Alicia Parrish, Cigdem Patlak, William Pietri, Forough Poursabzi-Sangdeh, Eleonora Presani, Fabrizio Puletti, Paul Röttger, Saurav Sahay, Tim Santos, Nino Scherrer, Alice Schoenauer Sebag, Patrick Schramowski, Abolfazl Shahbazi, Vin Sharma, Xudong Shen, Vamsi Sistla, Leonard Tang, Davide Testuggine, Vithursan Thangarasa, Elizabeth Anne Watkins, Rebecca Weiss, Chris Welty, Tyler Wilbers, Adina Williams, Carole-Jean Wu, Poonam Yadav, Xianjun Yang, Yi Zeng, Wenhui Zhang, Fedor Zhdanov, Jiacheng Zhu, Percy Liang, Peter Mattson, and Joaquin Vanschoren. Introducing v0.5 of the ai safety benchmark from mlcommons, 2024.

[40] Bertie Vidgen, Nino Scherrer, Hannah Rose Kirk, Rebecca Qian, Anand Kannappan, Scott A. Hale, and Paul Röttger. Simplesafetytests: a test suite for identifying critical safety risks in large language models, 2024.

[41] Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*, 2023.

[42] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*, 2023.

[43] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.

[44] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.

[45] Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, et al. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*, 2023.

[46] Jiahao Yu, Xingwei Lin, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

[47] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024.

[48] Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies, 2024.

[49] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.