Public security threats from low-compute AI systems

Anonymous submission

Abstract

Artificial intelligence (AI) systems are revolutionizing fields such as medicine, drug discovery, and materials science; however, many technologists and policymakers are also concerned about the technology's risks. To date, most concrete policies around AI governance have focused on managing AI risk by considering the amount of compute required to operate or build a given AI system. However, low-compute AI systems are becoming increasingly more performant - and more dangerous. Driven by agentic workflows, parameter quantization, and other model compression techniques, capabilities once only achievable on frontier-level systems have diffused into low-resource models deployable on consumer devices. In this report, we profile this trend by downloading historical benchmark performance data for over 5,000 large language models (LLMs) hosted on HuggingFace, noting the model size needed to achieve competitive LLM benchmarks has decreased by more than 10X over the past year. We then simulate the computational resources needed for an actor to launch a series of digital societal harm campaigns - such as disinformation botnets, sexual extortion schemes, voice-cloning fraud, and others - using low-compute open-source models and find nearly all studied campaigns can easily be executed on consumer-grade hardware. This paper argues that protection measures for high-compute models leave serious security holes for their low-compute counterparts, meaning it is urgent both policymakers and technologists make greater efforts to understand and address this emerging class of threats.

Introduction

Artificial intelligence (AI) technologies are enabling the widespread automation of information processing and reasoning tasks. Many anticipate these technologies will herald an era of unprecedented human productivity and economic output, while others are concerned they may be weaponized to cause large-scale societal harm. For example, researchers have investigated the extent to which advanced models, specifically large language models (LLMs), may facilitate synthetic biology attacks, compromise cybersecurity systems, and amplify the effects of disinformation campaigns (Helmus 2022; Hendrycks, Mazeika, and Woodside 2023).

Numerous policy frameworks have been proposed to mitigate AI risks, many of which focus on monitoring or regulating access to compute (Sastry et al. 2024). For example, current US semiconductor export controls are designed, at least partially, to prevent the misuse of advanced models requiring high performance graphics processing units (GPUs), with US officials citing threats from ”both AI training and inference at scale” (BIS 2024). Similarly, The European Union (EU) AI Act designates 10^{25} training floating point operations (FLOPs) as a threshold for systemic risk categorization and regulation (European Parliament 2023).

However, training or deploying large models is not the only pathway to dangerous capabilities. Advancements in test-time compute, parameter quantization, agentic workflows, LLM tooling, and other techniques are rapidly diffusing capabilities from large AI systems into compact models that are easily deployable on consumer devices (Subramanian, Elango, and Gungor 2025; Lang, Guo, and Huang 2024; Li 2024; Shen et al. 2024).

The swift compression of advanced AI capabilities into smaller, accessible, and easily deployable models poses significant security risks that current governance frameworks are not fully equipped to handle. We urgently encourage more researchers and policymakers to focus on developing innovative governance strategies specifically tailored to low-compute AI threats, as these threats are becoming increasingly severe, frequent, and difficult to detect.

In this report, we quantitatively profile the rate at which open-source LLMs have become both more performant and more compute-efficient over time. Secondly, we outline how this shift has impacted the amount of compute resources needed for a single actor to execute a variety of societal harm campaigns. Next we profile the computational workloads of several academic and commercial AI use cases, demonstrating a high degree of overlap between the compute required by both. We discuss how this overlap, combined with the relatively modest amount of compute required to launch the studied societal harm campaigns, complicates existing AI risk mitigation frameworks centered around high-compute models. Lastly, we briefly highlight the promises and shortcomings of a set of proposed strategies for mitigating low-compute AI risks.

As a clarification on terminology, we will refer to low-compute AI models in this report as those with $\lesssim$30B parameters, as these systems are increasingly deployable on low-cost hardware through parameter quantization and other inference optimization techniques. We will also use the term model compression to succinctly refer to the diffusion of

AI capabilities into low-compute systems, although we acknowledge this phrase may have different meanings elsewhere in the literature.

Models are becoming more advanced at smaller sizes

Market pressures are guiding models to become both more capable and more lightweight over time, while hardware improvements are reducing barriers to model deployment. We discuss both trends below while observing that if these trends continue to evolve, bad-faith and good-faith actors alike will be able to deploy increasingly sophisticated models with commonly accessible levels of compute.

Model miniaturization

We download performance data from over 5,000 open-source LLMs hosted on the HuggingFace LLM leaderboard. Each model on the leaderboard has been evaluated against the Eleuther AI Language Model Evaluation Harness (Eleuther AI 2024), a suite of benchmark tests designed to probe language model abilities on diverse tasks such as common sense reasoning, mathematical abilities, and others. In this report, we define a LLM's aggregate model performance, α, as the mean score on the IFEval, BBH, MATH, GPQA, MUSR, and MMLU-PRO benchmark tests included within the Harness.[1]

For each graded model on the leaderboard, we extract the model size (FP16 precision), model performance (α), and the date on which the model was created. In Figure 1A, we plot the model size needed to obtain a given α over time. Each scatter point represents the 25^{th} percentile model size value within the subset of models that surpass a given α at a given date. We present five different curves corresponding to α values of 30%, 35%, 40%, 45%, and 50%. For reference, the highest α value listed on the HuggingFace leaderboard as of the writing of this report is 52%.

We also fit a simple exponential decay curve to each set of data, which we present in the figure along with fit uncertainty bands for visual reference. As can be seen in Figure 1A, across all performance levels, the model size needed to obtain a given benchmark score has dropped significantly over time, with the model size needed to obtain $\alpha = 0.35$ falling by $\sim$10X over the past year.[2]

Similarly, Figure 1B displays how the α of a model of a fixed size has increased over time. Here, we filter our LLM leaderboard dataset to focus on Meta's Llama family of models, given their active use within the developer community. As can be seen in the plot, Llama models of all size ranges [3] have steadily increased in performance over time and have even somewhat converged to a common α value of $\sim$45%. Similar to Figure 1A, for visual reference, we fit linear models to the raw data and present them along with confidence intervals for each set of data.

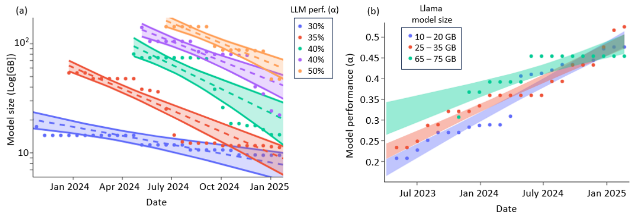

Figure 1: (a) Model size needed to obtain a given LLM benchmark score over time. Exponential curves are fit to the raw data and displayed along with the fit uncertainty bands. (b) Benchmark performance (α) over time for three classes of Llama family models.

Put together, these plots suggest the following two trends:

- The number of parameters needed to achieve a certain level of benchmark performance has decreased over time
- The benchmark performance of a model of a given size has increased over time.

While these statements are hardly surprising, they have strong implications for both public safety and AI governance, which we will discuss further below.

As a caveat, high performance on an LLM benchmark does not necessarily imply usefulness. In fact, many suspect that models have been engineered, or 'overfit', to provide deceptively high performance on such benchmarks while not providing high degrees of capability (Zhou et al. 2023). We fully acknowledge the limitations of benchmark data, and later in this report, we will reference more in depth evaluations of the capabilities of low-compute models that have been conducted by other research groups.

Similarly, while the above analysis focuses on open-source models, similar trends hold for closed-source models as well. For example, GPT-4 was offered at a price of $120/million completion tokens (Wayback Machine 2024) in early 2023. However, GPT-4.1 – a model that performs comparably on many performance benchmarks while offering both longer context windows and multimodal abilities – now executes completions at roughly a 15X cheaper rate. Consequently, cost and resource barriers to deploying advanced AI capabilities are rapidly diminishing for malicious actors, regardless of whether they utilize open or closed-source models.

Advances in hardware

While high-performing models are becoming smaller, the processing power of accelerator chips is becoming larger, both across consumer and data center devices. In Figure 2 we display the evolution of processing power (expressed in single-precision [FP32] FLOPS) and memory bandwidth (expressed in GB/s of memory transfer) over time for two sets of devices: NVIDIA data-center GPUs and consumer MacBooks GPUs.[4]

[1] α is bounded between [0%,100%], with 100% denoting a perfect score

[2] For robustness, we also recalculated these compression curves using each individual metric instead of the aggregate α value and observed similar compression trends over time.

[3] We cluster models into size ranges to account for the differing model sizes across the Llama 2, 3.1, 3.2, and 3.3 releases.

[4] We extracted performance metrics on both set of devices from official Apple and NVIDIA product pages.

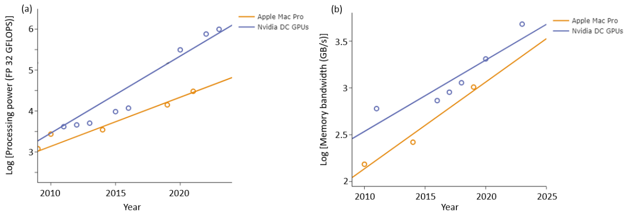

Figure 2: (a) Evolution of processing power in NVIDIA and MacBook chips over time (b) Evolution of memory bandwidth across both sets of chips over time. Linear fits are presented alongside both curves, for visual reference.

While many of the higher-performance data-center chips are currently restricted via US export controls, all MacBook chips explored in the chart are unrestricted and widely available for global use. These devices are sufficient to run inference on many advanced - and potentially dangerous - models, especially given the rapid pace of miniaturization in Figure 1a.

The threats posed by low-compute AI systems

Building on the results of the previous section, we assess the risks of low-compute AI systems from three separate vantage points:

- Examining the rise in reported public AI security incidents
- Highlighting studies evaluating the capabilities of low-compute AI systems
- Simulating the compute required to launch AI-powered social harm campaigns

While each vantage point alone provides only a partial view of low-compute AI risk, together they collectively paint a more complete picture of the public security threats posed by these systems.

The rapid rise of AI security incidents

The FBI stated $2.9B was lost through business-email-compromise scams in 2023 alone (IC3 2024b,a), citing GenAI as a key driver. Additionally, SlashNext reported a $1,265% increase in phishing incidents between Oct. 2022 and Sept. 2023 – a time period marked by the proliferation of generative AI technologies (SlashNext 2023). Similarly, the FBI declared a “global sextortion crisis” fueled by generative models (Federal Bureau of Investigation 2023), while McAfee reported that 25% of surveyed U.S. adults have either experienced or known someone who has experienced an AI voice scam (McA 2023).

Identification of dangerous capabilities within low-compute AI systems

Researchers have tested the believability of audio, video, and text-based output from compressed models (<30B parameters) on human participants. For example, Hackenburg et al. (2025) demonstrated that open-source LLMs as small as 7B parameters were more politically persuasive than a human control group and equally persuasive to many larger LLMs. In Bray et al. (2023), human participants were only able to identify deepfakes generated by a StyleGAN2 model (<1B-parameters) at a rate of $\sim$60%, a value marginally above random chance. Similarly, in Warren et al. (2024), in only $\sim$60% of instances were human subjects able to detect synthetically generated audio samples within the WaveFake dataset (Frank and Schönherr 2021), a dataset consisting of audio samples generated via lightweight open-source models such as MelGAN (<10M parameters). Lastly, Heiding et al. (2024) and Schoenegger et al. (2025) demonstrated that Claude-3.5 – which has been surpassed on the Chatbot Arena by multiple open-source LLMs $\lesssim$30B parameters in size - can produce spear-phishing emails that are as persuasive as those designed by human experts and is more effective at attitude, belief, and behavior shaping than a set of incentivized humans, respectively.

Simulating compute budgets of social harm campaigns

In this section we focus on a relevant set of disinformation, cybersecurity, voice cloning, and deepfakes threats. We first performed a literature review to identify emblematic historical case studies of these types of attacks. Guided by the details of each case study, we decomposed each campaign into a set of constituent tasks executed over an associated timescale, and we subsequently estimated the amount of compute required for an AI model to replicate them. For example, a disinformation campaign can be decomposed into a sequence of generated social media posts, and a spear-phishing campaign may be broken down into a sequence of generated images and email chains.

Anchoring our analysis to historical case studies sets realistic scales for our simulated campaigns. For example, we could profile the compute load of a disinformation campaign consisting of 1,000 Tweets or 1,000,000 Tweets. These campaigns require vastly different compute requirements, and *a-priori* it’s difficult to assess the societal harms posed by each. However, grounding our analysis in a historical disinformation case study helps both set the scale of a campaign and connect that scale to an event with understood social impact. For reference, we profile events like the Brexit disinformation campaign - an automated misinformation campaign thought to have influenced the outcome of a major geopolitical event (Bruno, Lambiotte, and Saracco 2022) - and a business compromise scam that generated millions of dollars in company losses.

The periodic table of synthetic media attacks To estimate the compute required for each audio, text, and image generation task, we simulate each on a NVIDIA V100 GPU equipped with the nvprof GPU profiler[5] Using this ‘periodic table’ of synthetic media generations, we build an aggregate

[5] We profiled each generative model in half precision (FP16) format. Further, we validated our profiler by comparing the measured compute profiles of an LLM token generation task and matrix multiplication task to well-established theoretical estimates, observing largely consistent values between the two.

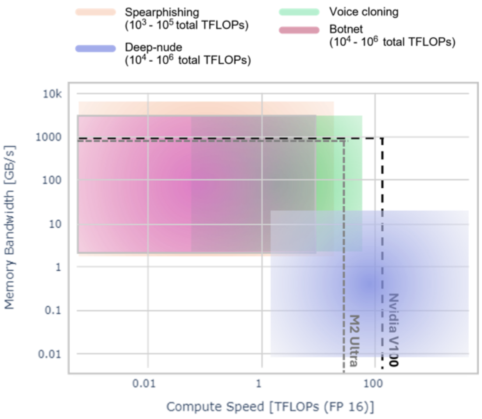

Figure 3: The simulated compute profiles required to execute a set of disinformation, spearphishing, voice-cloning, and deepfake attacks with low-compute AI models. We break each attack into image, text, and audio generation steps and measure the memory speed and processing power an attacker would need to execute the attack using a single chip. The bounding boxes display the 5% and 95% for each GPU performance metric across our simulations. The dashed lines denote the performance metrics for NVIDIA V100 and Apple M2 Ultra chips, two currently non-export-controlled devices.

compute profile of each campaign by considering the compute requirements of each constituent step.

Of course, rather than measuring the compute profile for each task, one could also theoretically derive this profile. For example, both the memory and FLOP requirements of LLM token generation have well known analytic forms (Hoffmann et al. 2022). Similar analytic equations could, in theory, be extracted for image and audio models; however, the complexities of considering different batch sizes, image dimensions, etc. make measurement via the profiler more straightforward than theoretical derivation. Further, in later sections of this report, we profile the workloads of an entirely different set of non-nefarious AI workloads such as deep learning recommendation engine training where theoretical derivation is even less straightforward.

Estimating uncertainties Several variables impact the amount of compute needed within each campaign, such as the size of model required, the image resolution needed with a deepfake campaign, the number of seconds within a voice-clone scam voicemail, and the number of tokens generated within a botnet social media post. While historical case studies help constrain these parameters to some extent, they do not fix them entirely. To this end, we run Monte-Carlo simulations across these parameter spaces to generate compute profile uncertainties.

Simulation results In Figure 3, we plot the amount of memory bandwidth and compute speed required for an actor to execute each campaign on a single accelerator chip, and we also provide the total aggregate TFLOP of the subset of campaigns with fixed time-scales. Denoted by dashed lines, we display the bandwidth and processing power of the V100 and MacPro M2 Ultra accelerators – both of which are currently non-export controlled.

A large fraction of the uncertainty boxes for all considered attacks are contained by the bounding boxes of the V100 and M2 Ultra accelerators, indicating such campaigns could in theory be executed with non-export-controlled devices. While certain regions of the uncertainty boxes lie beyond the performance bounds of a single chip, a majority of the studied campaigns are straightforward to distribute across multiple devices, meaning multiple chips could be combined to provide greater computing power. By our estimates, a computing cluster consisting of just ten V100 chips would offer enough computational power to surpass the estimated compute upper bounds of all considered threat campaigns - a system that could be purchased for mere thousands of dollars on eBay as of the writing of this report. Additionally, we intentionally adopted conservative simulation assumptions; in practice, attackers might achieve these outcomes with significantly fewer resources.

Of course, these results have several limitations. We performed all profiling assuming generative models less than 30B parameters in size. Despite the studies referenced earlier, It is yet not totally proven that AI models of this size are performant enough to successfully execute the explored campaigns. However, if such capabilities do not yet exist at these model sizes, given the findings of Section , they likely will soon. In general, it seems reasonable to assume the compute profile boxes in Figure 3 will shift down and leftwards as models continue to become both more performant and compute efficient, reducing the compute resources needed to execute AI-powered attacks.

Can’t compute thresholds simply be adjusted to address threats from low-compute AI?

The previous section demonstrated that current compute metering frameworks still permit access to compute levels sufficient to execute several societal harm campaigns. A natural follow-up question is: can these frameworks simply be revised to address this? The answer is complicated by the competing needs for these measures to both protect against AI risks while supporting non-nefarious business and academic development use cases. For example, export controls that successfully restrict bad actors from executing societal harm campaigns but also disrupt compute flows within AI-dependent industries would harm hardware manufacturers, strain international relationships, and hinder research collaborations.

To address this point, similar to Section , we profile the compute required to execute a set of typical business/academic AI workloads and compare the results to those in the previous section.

The selected workloads include the following: object recognition within autonomous vehicles, protein structure prediction within biomedical research, audio-to-text transcription within customer call centers, spam detection modeling, and recommendation engine training. We selected this

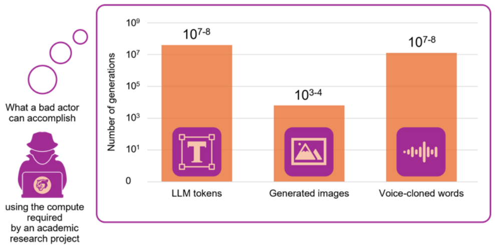

Figure 4: The number of synthetic images, llm-generated tokens, and words of voice cloned audio an actor could generate with the compute required by a single typical academic experiment.

set of workloads based on two criteria. First, we performed a literature review of most common AI commercial and academic use cases. Secondly, we filtered the identified set of workloads to diversify both across model type and application sector. For example, we include a recommendation engine example since - at least as of 2019 - these systems were estimated to be one of the highest volume workloads in global data centers (Mudigere et al. 2022). Similarly, we diversified our workloads to include audio, video, and text generation models leveraged across both academic and commercial sectors.

Following a procedure similar to that used above, we find the compute required by each workload exceeds that required within all profiled societal harm campaigns in Section . This suggests that naively adjusting compute metering thresholds to block attacks from miniaturized AI systems would significantly disrupt many non-nefarious academic and business AI use cases. As an illustrative example, in Figure 4, we display the number of synthetic images, LLM tokens, and voiced-cloned words an actor could generate with the compute required by our biomedical research example. As can be seen in the plot, an actor could generate hundreds of millions of pieces of synthetic content with the compute required by a single academic experiment.

Alternative strategies for addressing low-compute AI risks

The preceding analysis highlights the need for AI protection paradigms fundamentally different from existing approaches. Several governance strategies have been proposed to address low-compute AI risks, though each faces significant implementation challenges. Below we present a brief overview of a subset of such strategies; however, this discussion is by no means exhaustive.

Rather than inferring risk from compute, capability-based frameworks evaluate systems through demonstrated abilities, using benchmarks to probe for potentially harmful capabilities like persuasiveness, deception, or dual-use proficiency (Shevlane et al. 2023; Hooker 2024; Tamkin et al. 2023). However, creating robust evaluations requires substantial expertise, benchmarks rapidly become outdated (Amodei, Team et al. 2023), and bad-faith developers may game regulatory tests while maintaining harmful deployment capabilities. Similarly, defensive AI approaches - leveraging AI models to protect against risk from other AI models - may lead to futures wherer AI agents detect other voice clone agents, automatically patch software vulnerabilities, and address other threats (Lohn 2025). However, they may struggle in threat-scenarios with unfavorable offense-defense balances. For instance, exploiting offensive AI to deploy bioweapons may be far easier than using defensive AI to produce and disseminate vaccines (Unver and Arhan 2023; Aspen U.S. Cybersecurity Group 2024).

If widespread model access proves inevitable, protective measures could be integrated directly into systems through content filters and digital watermarks (Roman et al. 2024; Google DeepMind 2023; Kirchenbauer et al. 2023) that help identify synthetic material. However, seemingly non-toxic content can still remain dangerous and models can be jailbroken despite safeguards (Yu, Lin, and Xing 2023). Alternatively, a preventative approach to AI security could focus on strengthening institutions and social groups through enhanced media literacy, incident reporting, and AI education (Bernardi et al. 2024). However, such resiliency efforts require substantial funding and coordination across disparate sectors before they can scale effectively.

Conclusion

We explore how capabilities initially only present within larger-scale LLMs have diffused into low-resource, lightweight systems deployable on consumer devices. By analyzing historical performance data from over 5,000 models on HuggingFace, we demonstrate that the size of model needed to achieve competitive LLM benchmark scores has decreased by as much as 10X over the past year. We also simulate the compute needed for a bad-faith actor to launch a set of social harm campaigns, unveiling that many such attacks are easily executable on consumer hardware.

While these trends have been noted by other researchers (Bommasani et al. 2023; Weidinger et al. 2022; Hooker 2024), current AI governance frameworks have not adequately evolved to address the risks posed by increasingly powerful low-compute models. We present empirical evidence to underscore the urgency of rethinking approaches that rely primarily on compute thresholds as proxies for risk. As model miniaturization accelerates, policymakers must develop more nuanced frameworks that consider capabilities, intent, and potential for harm alongside compute requirements. This will require deeper collaboration between technical experts, policy makers, and industry to develop more comprehensiveness protection frameworks that consider a wider class of risks, rather than over-indexing on high-compute threats. The pace of AI advancement demands that these conversations move beyond mere discourse into concrete adaptations that meet today's rapidly evolving technological landscape.

References

2023. Beware the Artificial Impostor: A McAfee Cybersecurity Artificial Intelligence Report. Technical report, McAfee

Corp.

2024a. 2023 Internet Crime Report. Technical report, Federal Bureau of Investigation, Internet Crime Complaint Center (IC3). Report released in March 2024, covering data from calendar year 2023.

2024. Commerce Strengthens Export Controls to Restrict China's Capability to Produce Advanced Semiconductors for Military Applications. Technical report, U.S. Department of Commerce, Bureau of Industry and Security. Accessed: 2025-05-20.

2024b. Criminals Use Generative Artificial Intelligence to Facilitate Financial Fraud. Public Service Announcement I-120324-PSA, Internet Crime Complaint Center (IC3).

Amodei, D.; Team, A.; et al. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. *Anthropic Technical Report*.

Aspen U.S. Cybersecurity Group. 2024. Envisioning Cyber Futures with AI.

Bernardi, J.; Mukobi, G.; Greaves, H.; Heim, L.; and Anderljung, M. 2024. Societal Adaptation to Advanced AI. *arXiv*.

Bommasani, R.; Zhang, K.; Kobren, A.; Bailis, P.; Duchi, J.; Hernandez, D.; Horvitz, E.; Jaakkola, T.; Kaplan, J.; Koh, P. W.; et al. 2023. Foundation models and the future of AI: A new paradigm and emerging risks. arXiv:2307.13346.

Bray, S. D.; Johnson, S. D.; Kleinberg, B.; Davies, T.; and Griffin, L. D. 2023. Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 9(1): tyad011.

Bruno, M.; Lambiotte, R.; and Saracco, F. 2022. Brexit and bots: characterizing the behaviour of automated accounts on Twitter during the UK election. *EPJ Data Science*, 11(1): 17.

Eleuther AI. 2024. LM Evaluation Harness. https://github.com/EleutherAI/lm-evaluation-harness. Accessed: 2025-05-16.

European Parliament. 2023. EU AI Act: first regulation on artificial intelligence. Accessed: 2025-05-19.

Federal Bureau of Investigation. 2023. International Law Enforcement Agencies Issue Joint Warning about Global Financial Sextortion Crisis. Accessed: 2025-05-16.

Frank, J.; and Schönherr, L. 2021. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. NeurIPS 2021 Datasets and Benchmarks Track, arXiv:2111.02813.

Google DeepMind. 2023. SynthID: Identifying AI-Generated Content with SynthID. https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/. Accessed: 2025-05-16.

Hackenburg, K.; Tappin, B. M.; Röttger, P.; Hale, S. A.; Bright, J.; and Margetts, H. 2025. Scaling language model size yields diminishing returns for single-message political persuasion. *Proceedings of the National Academy of Sciences*, 122(10): e2413443122.

Heiding, F.; Lermen, S.; Kao, A.; Schneier, B.; and Vishwanath, A. 2024. Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects. arXiv:2412.00586.

Helmus, T. C. 2022. Artificial Intelligence, Deepfakes, and Disinformation: A Primer. Technical report, RAND Corporation.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. *arXiv*.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. An Empirical Analysis of Compute-Optimal Large Language Model Training. In *Advances in Neural Information Processing Systems*, volume 35, 30016–30030.

Hooker, S. 2024. On the Limitations of Compute Thresholds as a Governance Strategy. arXiv:2407.05694.

Kirchenbauer, J.; Geiping, J.; Wen, Y.; Katz, J.; Miers, I.; and Goldstein, T. 2023. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 17061–17084. PMLR.

Lang, J.; Guo, Z.; and Huang, S. 2024. A Comprehensive Study on Quantization Techniques for Large Language Models. arXiv:2411.02530.

Li, X. 2024. A Review of Prominent Paradigms for LLM-Based Agents: Tool Use (Including RAG), Planning, and Feedback Learning. arXiv:2406.05804.

Lohn, A. J. 2025. The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict. arXiv:2504.13371.

Mudigere, D.; Hao, Y.; Huang, J.; Jia, Z.; Tulloch, A.; Sridharan, S.; Liu, X.; et al. 2022. Software–Hardware Co-Design for Fast and Scalable Training of Deep Learning Recommendation Models. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*.

Roman, R. S.; Fernandez, P.; Défossez, A.; Furon, T.; Tran, T.; and Elsahar, H. 2024. Proactive Detection of Voice Cloning with Localized Watermarking. *arXiv*.

Sastry, G.; Heim, L.; Belfield, H.; Anderljung, M.; Brundage, M.; Hazell, J.; O'Keefe, C.; et al. 2024. Computing Power and the Governance of Artificial Intelligence. *arXiv*.

Schoenegger, P.; Salvi, F.; Liu, J.; Nan, X.; Debnath, R.; Fasolo, B.; Leivada, E.; Recchia, G.; Günther, F.; Zarifhonarvar, A.; Kwon, J.; Islam, Z. U.; Dehnert, M.; Lee, D. Y. H.; Reinecke, M. G.; Kamper, D. G.; Kobaş, M.; Sandford, A.; Kgomo, J.; Hewitt, L.; Kapoor, S.; Oktar, K.; Kucuk, E. E.; Feng, B.; Jones, C. R.; Gainsburg, I.; Olschewski, S.; Heinzelmann, N.; Cruz, F.; Tappin, B. M.; Ma, T.; Park, P. S.; Onyonka, R.; Hjorth, A.; Slattery, P.; Zeng, Q.; Finke, L.; Grossmann, I.; Salatiello, A.; and Karger, E. 2025. Large Language Models Are More Persuasive Than Incentivized Human Persuaders. arXiv:2505.09662.

Shen, W.; Li, C.; Chen, H.; Yan, M.; Quan, X.; Chen, H.; Zhang, J.; and Huang, F. 2024. Small LLMs Are Weak Tool Learners: A Multi-LLM Agent. arXiv:2401.07324.

Shevlane, T.; Fleming, M.; Hogarth, R.; and Hadfield, G. 2023. Model evaluation for extreme risks. arXiv:2305.15324.

SlashNext. 2023. The State of Phishing Report.

Subramanian, S.; Elango, V.; and Gungor, M. 2025. Small Language Models (SLMs) Can Still Pack a Punch: A survey. arXiv:2501.05465.

Tamkin, A.; Brundage, M.; Clark, J.; and Ganguli, D. 2023. Measuring Progress in Large Language Models. arXiv:2303.08774.

Unver, H. A.; and Arhan, S. E. 2023. The Strategic Logic of Digital Disinformation: Offence, Defence and Deterrence in Information Warfare. In *Routledge Handbook of Disinformation and National Security*, 192–207. Routledge.

Warren, K.; Tucker, T.; Crowder, A.; Olszewski, D.; Lu, A.; Fedele, C.; Pasternak, M.; Layton, S.; Butler, K.; Gates, C.; and Traynor, P. 2024. "Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. Salt Lake City, UT, USA: Association for Computing Machinery.

Wayback Machine. 2024. OpenAI Pricing. https://web.archive.org/web/20230416151950/https://openai.com/pricing. Accessed: 2024-08-01.

Weidinger, L.; Gabriel, I.; Uesato, J.; Mellor, J.; Goldie, W.; Irving, G.; Sodhani, S.; Hendricks, L. A.; Leike, J.; Kasirzadeh, A.; et al. 2022. A taxonomy of AI risks. arXiv:2206.03421.

Yu, J.; Lin, X.; and Xing, X. 2023. GPTFuzzer: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. *arXiv*.

Zhou, K.; Zhu, Y.; Chen, Z.; Chen, W.; Zhao, W. X.; Chen, X.; Lin, Y.; Wen, J.-R.; and Han, J. 2023. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv*.