Incentive-Aware AI Safety via Strategic Resource Allocation: A Stackelberg Security Games Perspective

Blue Sky Ideas Track

Cheol Woo Kim*
Harvard University
Cambridge, United States
cwkim@seas.harvard.edu

Davin Choo*
Harvard University
Cambridge, United States
davinchoo@seas.harvard.edu

Tzeh Yuan Neoh
Harvard University
Cambridge, United States
tzehyuan_neoh@g.harvard.edu

Milind Tambe
Harvard University
Cambridge, United States
tambe@g.harvard.edu

ABSTRACT

As AI systems grow more capable and autonomous, ensuring their safety and reliability requires not only model-level alignment but also strategic oversight of the humans and institutions involved in their development and deployment. Existing safety frameworks largely treat alignment as a static optimization problem (e.g., tuning models to desired behavior) while overlooking the dynamic, adversarial incentives that shape how data are collected, how models are evaluated, and how they are ultimately deployed. We propose a new perspective on AI safety grounded in Stackelberg Security Games (SSGs): a class of game-theoretic models designed for adversarial resource allocation under uncertainty. By viewing AI oversight as a strategic interaction between defenders (auditors, evaluators, and deployers) and attackers (malicious actors, misaligned contributors, or worst-case failure modes), SSGs provide a unifying framework for reasoning about incentive design, limited oversight capacity, and adversarial uncertainty across the AI lifecycle. We illustrate how this framework can inform (1) training-time auditing against data/feedback poisoning, (2) pre-deployment evaluation under constrained reviewer resources, and (3) robust multi-model deployment in adversarial environments. This synthesis bridges algorithmic alignment and institutional oversight design, highlighting how game-theoretic deterrence can make AI oversight proactive, risk-aware, and resilient to manipulation.

KEYWORDS

Stackelberg Security Games, AI Safety, Alignment

ACM Reference Format:
Cheol Woo Kim, Davin Choo, Tzeh Yuan Neoh, and Milind Tambe. 2026. Incentive-Aware AI Safety via Strategic Resource Allocation: A Stackelberg Security Games Perspective: Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 6 pages. https://doi.org/10.65109/JUJH9710

1 INTRODUCTION

Large language models (LLMs) have demonstrated unprecedented capabilities across a wide range of domains, including creative and decision-making tasks once thought uniquely human. Their applications span mathematics [30], writing [42], programming [32], and tool use [72], as well as education [70], medicine [76], law [37], and science [45]. As these LLM-based systems become increasingly capable and autonomous, they are being integrated into workflows and decision pipelines that directly affect people's lives. Ensuring that these models remain aligned with human intent and societal values, and do not exhibit unexpected or harmful behavior in deployment, has therefore become a central and increasingly urgent challenge.

To enhance AI safety, a wide range of alignment and evaluation techniques have been proposed, spanning standard alignment approaches such as reinforcement learning from human feedback (RLHF) [18], evaluation mechanisms like automated red-teaming [13, 23, 48], and mechanistic interpretability methods [9]. While these approaches have substantially advanced our ability to understand and shape model behavior, their focus lies primarily on the model level—the training algorithm or the trained model itself. In doing so, they often overlook the crucial role of humans and institutions in the loop: trainers, annotators, auditors, and developers whose incentives, constraints, and actions directly influence the reliability and integrity of the alignment process. *As a result, many existing safety protocols implicitly treat human actors as benign or static components rather than as strategic agents who may operate under misaligned incentives or deviate from prescribed procedures.*

Even after training, significant challenges remain during deployment, where LLMs are integrated into diverse workflows and interact dynamically with users, other agents, and external tools. Ensuring the safety of these systems requires attention not only to the model's internal behavior but also to the overall performance and downstream outcomes of the multi-agent ecosystems built around them. Achieving this demands mechanisms that can monitor, enforce, and promote robust behavior throughout deployment. This broader perspective highlights the need for strategic,

---

*Equal Contribution.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/JUJH9710

incentive-aware oversight frameworks that go beyond model-level algorithmic fixes and account for the human, institutional, and system-level dimensions of AI safety.

As a step toward this broader vision, we propose applying a specific class of game-theoretic models – Stackelberg Security Games (SSGs) [7, 24, 36, 40, 59, 62, 66] – to the domain of AI safety and alignment. An SSG models a strategic interaction between a defender, who commits to a resource-allocation or inspection strategy under limited capacity, and an attacker, who observes this strategy and chooses where to strike to maximize their gain. SSGs have a proven track record in real-world security operations, having been deployed by the U.S. Federal Air Marshals Service [28, 29], the U.S. Coast Guard [3], and at Los Angeles International Airport [56, 57] to optimize patrol schedules and allocate limited enforcement resources, and saving over USD$100 million for the U.S. government [22, 67]. Moreover, SSGs have been studied extensively for more than a decade in the multi-agent systems community, yielding a deep theoretical foundation and a rich set of scalable algorithms.

The proven scalability and real-world success of SSGs suggest that they can play a transformative role in AI safety, particularly in addressing the strategic, human, and organizational dimensions often overlooked in existing frameworks. Rather than attributing agency to the AI model itself, SSGs provide a principled way to focus on the human actors and institutional processes that shape data collection, training, evaluation, and deployment. Ultimately, this broadened perspective reframes AI safety as more than a matter of model-level diagnostics. It instead emphasizes strategic deterrence and the proactive allocation of limited auditing resources across the entire AI lifecycle. Building on this foundation, we propose three directions in which SSGs can strengthen AI security and fill critical gaps across the lifecycle of LLMs:

1. **Training Data:** modeling and mitigating feedback poisoning or label manipulation attacks during fine-tuning, where adversaries may corrupt a subset of human-generated preference data.
2. **Evaluation:** optimizing the allocation of auditing and evaluation resources to identify potential weaknesses or misaligned behaviors, under limited reviewer capacity.
3. **Deployment:** designing LLM deployment strategies where different models or agent teams vary in capability, cost, and reliability, enabling principled and risk-aware decisions about where and how each model should be used.

We illustrate our proposed directions in Fig. 1.

2 RELATED WORKS

2.1 Stackelberg Security Games

A SSG models a strategic interaction between a defender and an attacker over a set of valuable targets T [7, 24, 36, 39, 40, 59, 62, 66], with extensive research reported at AAMAS since 2007. The defender possesses a limited set of resources R that can be allocated to protect these targets. The attacker observes the defender's (possibly randomized) strategy and then chooses a single target to attack.

A *pure strategy* for the defender is a specific allocation of resources R across targets T, subject to allocation or scheduling constraints such as limits on patrol routes or inspection capacity. The attacker's pure strategy is to choose one target $t \in T$ to attack. A *mixed strategy* is a probability distribution over pure strategies, and for the defender this corresponds to randomized protection patterns designed to deter the attacker.

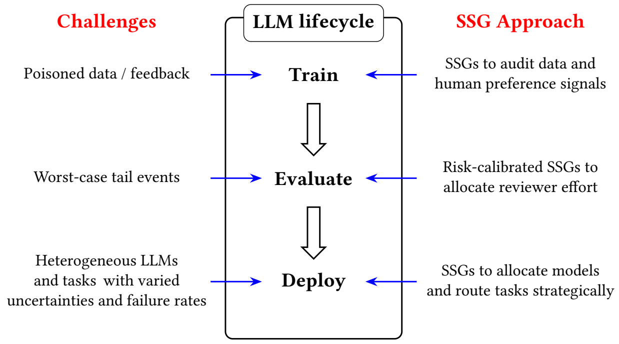

Figure 1: Illustration of how SSGs can help tackle challenges in each of the three steps of the lifecycle of LLMs.

Each target $t \in T$ is associated with payoff values describing the utilities for both players, depending on whether the target is protected. If the attacker chooses target t, the defender receives utility $U_d^c(t)$ if t is covered and $U_d^u(t)$ if it is uncovered ($U_d^c(t) \geq U_d^u(t)$). The attacker's utilities are $U_a^c(t)$ and $U_a^u(t)$ in the covered and uncovered cases, respectively ($U_a^u(t) \geq U_a^c(t)$).

A common modeling assumption is that these payoffs depend only on the attacked target and its protection status. This independence assumption yields a compact representation of the payoff structure and enables efficient computation of equilibrium strategies. A widely studied special case is the zero-sum setting, where the sum of defender and attacker utilities for each outcome is constant. See [59, 62] for a comprehensive overview of SSG theory and applications.

Over the past couple of decades, SSGs have been successfully deployed in numerous real-world security domains. Notable examples include their use by the U.S. Federal Air Marshals Service[28, 29], the U.S. Coast Guard[3], and Los Angeles International Airport (LAX)[56, 57] to optimize patrol schedules, allocate limited enforcement resources, and improve public safety. Beyond physical security, SSG frameworks have been extended to domains such as cybersecurity, wildlife and fisheries protection[21], auditing systems, and traffic enforcement—demonstrating their flexibility, scalability, and robustness in adversarial, resource-constrained environments.

2.2 AI Safety

We describe two broad types of risks in modern AI systems: (1) adversarial data poisoning during training, and (2) unexpected or misaligned behaviors during deployment. For a more comprehensive overview of AI safety challenges, we refer readers to existing surveys [14, 31, 69].

2.2.1 Failure modes during training. Data poisoning has been extensively studied for both pre-training [12, 43, 60, 61, 68, 74, 75] and post-training stages [8, 11, 17, 64]. Attacks can target large public datasets (e.g., Wikipedia) before model training or manipulate human preference data used in RLHF, often sourced from crowdsourcing platforms [54]. Critically, both attack vectors are effective with remarkably small amounts of poisoned data: pre-training attacks succeed with constant numbers of examples (around 250) even as models and datasets scale [60, 68], while post-training attacks can meaningfully shift model outputs by poisoning just 1–5% of preference data [8]. Moreover, pre-training poisoning effects can persist through extensive post-training alignment, and attacks can succeed using seemingly benign inputs [17, 64], making detection particularly challenging.

A straightforward defense against data poisoning is data distillation, where potentially harmful samples are filtered out of the training corpus. This can be implemented using heuristic methods—such as domain blacklists [1, 20], keyword-based matching [20], or manually specified rules [4, 27]. Another common strategy is model-based filtering, in which a separate model—often fine-tuned on safety-oriented datasets—is trained to identify and remove undesirable data [1, 47]. As both heuristic and model-based data distillation methods frame poisoning as a static classification task, they may implicitly encourage adaptive adversaries to engineer inputs that fall below detection thresholds.

2.2.2 Failure Modes during Deployment. LLMs often exhibit uneven skill profiles: strong performance on complex reasoning does not guarantee reliability on simpler tasks. For example, models that excel on benchmarks can still fail at basic word-level manipulations [73], degrade sharply on out-of-domain inputs [55], or generate invalid intermediate steps and hallucinated conclusions [33, 44]. These uneven capabilities highlight the need for accurate assessment and strategic allocation of models across tasks. Safety fine-tuning also does not eliminate vulnerability to jailbreaking attacks, where adversarial prompts induce harmful outputs [15, 16, 77].

As models scale, emergent behaviors introduce additional risks. While extreme thought experiments such as the paperclip maximizer illustrate the broader concern [50], current LLMs already display behaviors suggestive of misaligned incentives, including alignment faking [25, 71], scheming [49, 58], sandbagging [41, 63, 65], and, in rare cases, coercive behaviors such as blackmail [46]. These risks are heightened by situational awareness [38, 58], where models behave differently under evaluation than in real deployment.

Red-teaming techniques [13, 23, 48] are currently the primary method for uncovering such behaviors, but their effectiveness is limited by the lack of standardized protocols and uncertainty about real-world adversarial capability [23]. Situational awareness further suggests that one-time evaluations are insufficient, motivating continuous oversight during deployment. Emerging AI control approaches [26, 34, 35] focus on guardrails—such as human approval for critical actions or restricting access to sensitive tools—to limit harm even if misaligned systems are deployed. Both red-teaming and AI control demand nontrivial human and computational oversight, and an SSG framework can unify these considerations by optimizing how limited auditing resources are allocated.

3 PROPOSED DIRECTIONS

We propose three key directions for applying SSGs toward incentive-aware AI safety. Each direction corresponds to a distinct stage of the AI lifecycle – training, evaluation, and deployment – and addresses critical oversight and resource-allocation challenges that current safety frameworks often overlook.

3.1 SSGs for Data and Feedback Auditing

Modern alignment methods such as RLHF [18] and its variants (e.g., NLHF [51], SLHF [53]) rely on large-scale human preference data [54]. Recent work shows that even extremely small amounts of poisoned feedback—fractions of a percent or a few hundred samples—can reliably shift model behavior or implant backdoors [8, 60, 64, 74]. This exposes a structural vulnerability: strategic annotators can bias alignment data in ways that are difficult to detect. Existing defenses remain largely heuristic and do not provide a principled framework for allocating limited auditing resources against adaptive adversaries.

We propose formalizing feedback auditing as a SSG: a strategic attacker selects which data to corrupt to maximize misalignment with normative desiderata (e.g., Constitution [5] or Model Spec [52]), while a resource-limited defender commits to an auditing policy in advance. Although full observability of auditing strategies is unrealistic, aspects of real pipelines – such as which samples or annotators tend to be rechecked – are often partially learnable. The Stackelberg assumption thus serves as a conservative design principle: if a policy is robust under partial leakage, it is likely to be robust in milder settings as well.

Defenders may employ weaker language models in a similar spirit to scalable oversight [10] to reduce reliance on costly human labor, but such models may lack the fidelity needed to detect subtle corruption or adversarial preference manipulation. Effective auditing policies may therefore require selective discarding or re-labeling problematic feedback samples, consistency checks across annotators, or targeted inspection of high-influence datapoints. Critically, randomized or strategically diversified audits are necessary to avoid predictable evasion as even limited leakage or structural regularities can give attackers enough information to exploit deterministic patterns.

Viewing oversight as strategic deterrence rather than passive anomaly detection enables principled auditing under tight verification budgets. An SSG formulation (1) yields optimized auditing strategies, (2) clarifies the role of randomness for adversarial robustness, and (3) directs attention to datapoints with the greatest misalignment impact (e.g., high-influence samples or annotators with anomalous behavior).

Key challenge(s). Estimating SSG payoffs in real pipelines requires causal estimates of misalignment impact (e.g., via influence functions) under noise, annotator collusion, and dataset drift. Another challenge is deploying randomized audits with knowledge of their partial observability – a key challenge explored earlier in SSGs[2]. Finally, as SSGs assume a very strong adversary, it is worth exploring what guarantees can be derived under weaker or more realistic attacker models.

3.2 SSGs for LLM Evaluation

LLMs are deployed across increasingly diverse tasks and domains, which makes it increasingly difficult to determine how best to evaluate their safety and reliability before deployment. Given the open-ended nature of LLM outputs, exhaustively testing all possible behaviors or failure modes is infeasible. Existing techniques for probing vulnerabilities, such as jailbreaking [15, 16, 77] and red-teaming [13, 23, 48], provide only partial coverage of the risk landscape due to the vast range of potential prompts, tasks, and adversarial strategies. We argue that as models grow in scale and autonomy, evaluation itself becomes a strategic resource allocation problem. It requires deciding where, how, and to what extent to allocate limited human and computational oversight.

We propose using SSGs to address the evaluator-allocation problem: how to optimally assign limited reviewers (human evaluators, weaker LLMs, or teams of agents) across diverse AI application domains when evaluation capacity is constrained. Consider a setting with T task domains such as coding, medicine, and law. Reviewer capacity is limited and may include both human experts and weaker LLMs. Each domain carries a distinct risk profile: coding tasks may have higher error rates but relatively low consequences, whereas medical tasks may fail less frequently but carry far greater potential harm. In SSG terms, these differences correspond to targets with different payoff structures; allocating resources to each "target" provides different marginal reductions in expected harm. Similarly, each reviewer type has different strengths, limitations, and costs.

The defender, representing the auditing organization, allocates reviewer effort across domains, while the adversary models either uncertainty about where failures occur or an attacker exploiting lightly monitored areas. Within this SSG formulation, the defender's goal is to maximize risk-adjusted detection of unsafe behaviors across domains given limited resources. Once the SSG determines the optimal allocation of reviewers to different domains, those reviewers then employ evaluation techniques such as jailbreaking and red-teaming to actually test the model in their assigned domains. Thus, the SSG operates at the strategic resource allocation level, while jailbreaking and red-teaming serve as the tactical evaluation methods applied within each domain.

This framework transforms LLM auditing from an ad hoc model-level process into a principled resource allocation problem grounded in game theory. Importantly, this approach can be applied not just as a one-shot evaluation but continuously throughout deployment, including ongoing monitoring as models interact with real-world environments.

Key challenge(s). A central technical challenge is to define risk-calibrated utilities and priors over failure modes so that reviewer allocation reflects *tail-risk* rather than average error. This becomes even more difficult as models and adversaries adapt over time in a continuous deployment setting. Moreover, human reviewers, weaker LLMs, and automated test harnesses exhibit distinct error patterns and failure behaviors, yet we lack a unified framework for modeling these heterogeneous resources as defender assets within an SSG formulation. Many LLM tasks (e.g., tool-assisted reasoning) are also compositional rather than isolated, making it unclear how to represent composite or hierarchical vulnerabilities as SSG targets. Finally, learning attacker models from limited adversarial data — such as jailbreak logs or red-team traces — and integrating these learned distributions into the attacker side of the game remains challenging.

3.3 SSGs for LLM Deployment

Modern AI service providers such as Cursor [19], Anthropic [6], and OpenAI [52] increasingly rely on multiple LLMs, each with different capabilities, reliability profiles, latency, and cost. As AI adoption grows, practical deployments will often involve several models coexisting and collaborating. Determining which model (or team of models) to assign to each task therefore becomes a strategic deployment problem under uncertainty. In this setting, safety extends beyond improving or evaluating individual LLMs to a higher-level question: given the risks and capabilities of available models, what is the safest way to deploy them?

Resource limitations (such as inference cost, latency budgets, and API quotas) and differences in task-level risk (for example, safety-critical versus low-stakes applications) further complicate deployment decisions. Different models may also exhibit distinct risk profiles across tasks, with heterogeneous failure modes that must be taken into account.

We propose using SSGs to formalize this problem: the defender allocates limited LLM capacity across heterogeneous tasks, while the adversary represents either malicious inputs seeking to exploit weaknesses or even nature (worst-case failures). In this formulation, "targets" correspond to different tasks (e.g., generating code, answering technical questions, processing sensitive data) or broader application domains (e.g., law, medicine), and an "attack" represents either a model failure or an adversarial prompt that exploits tasks assigned to less capable models. The defender's strategy determines which model or model team to deploy for each task while anticipating that failures are most likely to occur in less-protected areas. This SSG-based approach enables incentive-aware deployment strategies that balance safety, performance, and cost while remaining robust to both adversarial threats and task-specific uncertainties in mission-critical settings.

Key challenge(s). SSG-based deployment plans must operate in real time within practical, large-scale systems, which may require moving beyond single-stage formulations toward multi-stage models that capture evolving deployment dynamics. A major challenge is that the true risk profiles of different models across tasks are often unknown and must be learned online. Unlike traditional SSG settings with well-defined, discrete targets, LLM deployment involves a vast and open-ended task space where failure modes can be subtle, delayed, or emerge only through model interactions, causing the underlying payoff structure to shift rapidly. Designing SSG frameworks that can learn these payoffs while simultaneously optimizing deployment strategies remains a key challenge.

4 CONCLUSION

This paper suggests that AI safety can benefit from viewing oversight as a strategic, resource-constrained problem rather than a purely model-centric one. SSG provide a principled lens for this shift, offering a way to reason about limited auditing capacity, adversarial uncertainty, and the multi-stage structure of modern AI pipelines. Our three directions in the lifecycle of LLMs illustrate how SSGs can unify disparate safety challenges under a single strategic framework. Looking forward, this perspective opens a broader research agenda: developing richer models of adversarial behavior, designing scalable oversight algorithms, and building empirical testbeds that capture the dynamics of real-world AI systems. By embracing SSGs, AI safety can move toward more anticipatory and resilient forms of oversight capable of withstanding the strategic pressures that will accompany increasingly capable AI systems.

ACKNOWLEDGMENTS

This work was supported by ONR MURI N00014-24-1-2742.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Bo An, Matthew Brown, Yevgeniy Vorobeychik, and Milind Tambe. 2013. Security games with surveillance cost and optimal timing of attack execution. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 223–230.

[3] Bo An, Fernando Ordóñez, Milind Tambe, Eric Shieh, Rong Yang, Craig Baldwin, Joseph DiRenzo III, Kathryn Moretti, Ben Maule, and Garrett Meyer. 2013. A Deployed Quantal Response-Based Patrol Planning System for the U.S. Coast Guard. *Interfaces* 43, 5 (2013), 400–420.

[4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[5] Anthropic. 2023. Collective Constitutional AI: Aligning a Language Model with Public Input. https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[6] Anthropic. 2025. Claude. https://www.anthropic.com/.

[7] Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. 2015. Commitment Without Regrets: Online Learning in Stackelberg Security Games. In *ACM Conference on Economics and Computation (EC)*.

[8] Tim Baumgärtner, Yang Gao, Dana Alon, and Donald Metzler. 2024. Best-of-Venom: Attacking RLHF by Injecting Poisoned Preference Data. In *Conference on Language Modeling (COLM)*.

[9] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic Interpretability for AI Safety – A Review. *Transactions on Machine Learning Research (TMLR)* (2024).

[10] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring Progress on Scalable Oversight for Large Language Models. *arXiv preprint arXiv:2211.03540* (2022).

[11] Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, and Marta Kwiatkowska. 2025. Strategyproof reinforcement learning from human feedback. *arXiv preprint arXiv:2503.09561* (2025).

[12] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2024. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 407–425.

[13] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. 2024. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2254–2272.

[14] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[15] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems* 37 (2024), 55005–55029.

[16] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 23–42.

[17] Bocheng Chen, Hanqing Guo, Guangjing Wang, Yuanda Wang, and Qiben Yan. 2024. The dark side of human feedback: Poisoning large language models via user inputs. *arXiv preprint arXiv:2409.00787* (2024).

[18] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[19] Cursor. 2025. Cursor: The AI Code Editor. https://cursor.com.

[20] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[21] Fei Fang, Thanh Nguyen, Rob Pickles, Wai Y. Lam, Gopalasamy R. Clements, Bo An, Amandeep Singh, Milind Tambe, and Andrew Lemieux. 2016. Deploying PAWS: Field Optimization of the Protection Assistant for Wildlife Security. In *AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI)*.

[22] Scott Farrow and Detlof von Winterfeldt. 2020. Retrospective Benefit–Cost Analysis of Security-Enhancing and Cost-Saving Technologies. *Journal of Benefit-Cost Analysis* 11, 3 (2020), 479–500.

[23] Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. 2024. Red-teaming for generative AI: Silver bullet or security theater?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 421–437.

[24] Jiarui Gan, Edith Elkind, and Michael Wooldridge. 2018. Stackelberg Security Games with Multiple Uncoordinated. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[25] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093* (2024).

[26] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2023. AI control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942* (2023).

[27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[28] Manish Jain, Erim Kardes, Christopher Kiekintveld, Fernando Ordónez, and Milind Tambe. 2010. Security Games with Arbitrary Schedules: A Branch and Price Approach. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[29] Manish Jain, Jason Tsai, James Pita, Christopher Kiekintveld, Shyamsunder Rathi, Milind Tambe, and Fernando Ordónez. 2010. Software Assistants for Randomized Patrol Planning for the LAX Airport Police and the Federal Air Marshal Service. *Interfaces* 40, 4 (2010), 267–290.

[30] Uijeong Jang and Ernest K. Ryu. 2025. Point Convergence of Nesterov's Accelerated Gradient Method: An AI-Assisted Proof. *arXiv preprint arXiv:2510.23513* (2025).

[31] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852* (2023).

[32] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2025. A Survey on Large Language Models for Code Generation. *ACM Transactions on Software Engineering and Methodology* (2025).

[33] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664* (2025).

[34] Tomek Korbak, Mikita Balesni, Buck Shlegeris, and Geoffrey Irving. 2025. How to evaluate control measures for LLM agents? A trajectory from today to superintelligence. *arXiv preprint arXiv:2504.05259* (2025).

[35] Tomek Korbak, Joshua Clymer, Benjamin Hilton, Buck Shlegeris, and Geoffrey Irving. 2025. A sketch of an ai control safety case. *arXiv preprint arXiv:2501.17315* (2025).

[36] Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. 2010. Complexity of Computing Optimal Stackelberg Strategies in Security Resource Allocation Games. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[37] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large language models in law: A survey. *AI Open* (2024).

[38] Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. 2024. Me, myself, and ai: The situational awareness dataset (sad) for llms. *Advances in Neural Information Processing Systems* 37 (2024), 64010–64118.

[39] Joshua Letchford and Vincent Conitzer. 2013. Solving security games on graphs via marginal probabilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27. 591–597.

[40] Joshua Letchford and Yevgeniy Vorobeychik. 2011. Computing Randomized Security Strategies in Networked Domains. *AAAI Conference on Artificial Intelligence (AAAI) Workshop: Applied Adversarial Reasoning and Risk Modeling* (2011).

[41] Chloe Li, Mary Phuong, and Noah Y Siegel. 2025. Llms can covertly sandbag on capability evaluations against chain-of-thought monitoring. *arXiv preprint arXiv:2508.00943* (2025).

[42] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. In *Conference on Language Modeling (COLM)*.

[43] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 3245–3276.

[44] Jiahao Lu, Ziwei Xu, and Mohan Kankanhalli. 2025. Reasoning LLMs are Wandering Solution Explorers. *arXiv preprint arXiv:2505.20296* (2025).

[45] Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306* (2025).

[46] Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J Ritchie, Soren Mindermann, Ethan Perez, Kevin K Troy, and Evan Hubinger. 2025. Agentic Misalignment: How LLMs Could Be Insider Threats. *arXiv preprint arXiv:2510.05179* (2025).

[47] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 15009–15018.

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

[49] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984* (2024).

[50] Kathleen Miles. 2014. Artificial intelligence may doom the human race within a century. *Huffington Post* (2014).

[51] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. 2024. Nash Learning from Human Feedback. In *International Conference on Machine Learning (ICLR)*.

[52] OpenAI. 2025. OpenAI Model Spec. https://model-spec.openai.com/2025-10-27.html.

[53] Barna Pásztor, Thomas Kleine Buening, and Andreas Krause. 2025. Stackelberg Learning from Human Feedback: Preference Optimization as a Sequential Game. In *European Workshop on Reinforcement Learning (EWRL)*.

[54] Billy Perrigo. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. https://time.com/6247678/openai-chatgpt-kenya-workers/.

[55] Pejman Peykani, Fatemeh Ramezanlou, Cristina Tanasescu, and Sanly Ghanidel. 2025. Large language models: A structured taxonomy and review of challenges, limitations, solutions, and future directions. *Applied Sciences* 15, 14 (2025), 8103.

[56] James Pita, Manish Jain, Janusz Marecki, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. 2008. Deployed ARMOR Protection: The Application of a Game Theoretic Model for Security at the Los Angeles International Airport. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[57] James Pita, Manish Jain, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. 2009. Using Game Theory for Los Angeles Airport Security. *AI Magazine* (2009).

[58] Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, et al. 2025. Stress testing deliberative alignment for anti-scheming training. *arXiv preprint arXiv:2509.15541* (2025).

[59] Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. 2018. Stackelberg security games: Looking beyond a decade of success. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[60] Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. 2025. Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples. *arXiv preprint arXiv:2510.07192* (2025).

[61] Xiaofei Sun, Xiaoya Li, Yuxian Meng, Xiang Ao, Lingjuan Lyu, Jiwei Li, and Tianwei Zhang. 2023. Defending against backdoor attacks in natural language generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5257–5265.

[62] Milind Tambe. 2011. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned.* Cambridge university press.

[63] Cameron Tice, Philipp Alexander Kreer, Nathan Helm-Burger, Prithviraj Singh Shahani, Fedor Ryzhenkov, Jacob Haimes, Felix Hofstätter, and Teun van der Weij. 2024. Noise injection reveals hidden capabilities of sandbagging language models. *arXiv preprint arXiv:2412.01784* (2024).

[64] Florian Tramèr and Javier Rando. 2024. Universal Jailbreak Backdoors from Poisoned Human Feedback. In *International Conference on Learning Representations (ICLR)*.

[65] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F Brown, and Francis Rhys Ward. 2024. Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358* (2024).

[66] Pradeep Varakantham, Hoong Chuin Lau, and Zhi Yuan. 2013. Scalable Randomized Patrolling for Securing Rapid Transit Networks. In *IAAI*. 1563–1568.

[67] Detlof Von Winterfeldt, R Scott Farrow, Richard S John, Jonathan Eyer, Adam Z Rose, and Heather Rosoff. 2020. Assessing the Benefits and Costs of Homeland Security Research: A Risk-Informed Methodology with Applications for the U.S. Coast Guard. *Risk analysis* 40, 3 (2020), 450–475.

[68] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 139–150.

[69] Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, et al. 2025. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585* (2025).

[70] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large Language Models for Education: A Survey and Outlook. *arXiv preprint arXiv:2403.18105* (2024).

[71] Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, and Yingchun Wang. 2024. Fake alignment: Are llms really aligned well?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 4696–4712.

[72] Weikai Xu, Chengrui Huang, Shen Gao, and Shuo Shang. 2025. LLM-Based Agents for Tool Learning: A Survey. *Data Science and Engineering* (2025), 1–31.

[73] Yidan Zhang and Zhenan He. 2024. Large Language Models Can Not Perform Well in Understanding and Manipulating Natural Language at Both Character and Word Levels?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 11826–11842.

[74] Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. 2025. Persistent Pre-Training Poisoning of LLMs. In *International Conference on Learning Representations (ICLR)*.

[75] Zhiyuan Zhang, Lingjuan Lyu, Xingjun Ma, Chenguang Wang, and Xu Sun. 2022. Fine-mixing: Mitigating backdoors in fine-tuned language models. *arXiv preprint arXiv:2210.09545* (2022).

[76] Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S. Yu. 2025. Large Language Models for Medicine: A Survey. *International Journal of Machine Learning and Cybernetics* (2025).

[77] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).