

Dmitrii Volkov

Current

Member of Technical Staff, Apollo Research, San Francisco, California (June 2026-present)

At Apollo, Dmitrii works on monitoring and controlling AI agents, and carries out red-teaming as part of preparedness for RSI.

Positions held

Cofounder and Head of Research, Palisade Research, Berkeley, California, 2023-2026

Organizer, LaLambda Summer School (formal verification, functional programming, compiler design), 2021 and 2023.

Doctoral student in cybersecurity and formal methods, Purdue University. Withdrew to help start Palisade. 2023.

Operating systems engineer, Kaspersky Lab, 2020-2022

Compilers engineer, JetBrains, S.R.O., 2020

Education

Moscow Polytechnic, Computer Science, 2016-2022

Selected publications

Volkov, D. (2024) Badllama 3: removing safety finetuning from Llama 3 in minutes. arXiv:2407.01376. <https://arxiv.org/abs/2407.01376>

Ryzhenkov, F., Volkov, D., and Ladish, J. (2024) Introducing FoxVox. Palisade Research Blog, 11 July 2024. <https://palisaderesearch.org/blog/foxvox>

Turtayev, R., Petrov, A., Volkov, D., and Volk, D. (2024) Hacking CTFs with Plain Agents. arXiv:2412.02776. <https://arxiv.org/abs/2412.02776>

Krupkina, E., and Volkov, D. (2024) BadGPT-4o: stripping safety finetuning from GPT models. arXiv:2412.05346. <https://arxiv.org/abs/2412.05346>

Bondarenko, A., Volk, D., Volkov, D., and Ladish, J. (2025) Demonstrating specification gaming in reasoning models. arXiv:2502.13295. <https://arxiv.org/abs/2502.13295>

Reworr and Volkov, D. (2025) LLM Agent Honeypot: Monitoring AI Hacking Agents in the Wild. arXiv:2410.13919v2, 10 February 2025. <https://arxiv.org/abs/2410.13919>

Ivanov, I., and Volkov, D. (2025) Resurrecting saturated LLM benchmarks with adversarial encoding. arXiv:2502.06738. <https://arxiv.org/abs/2502.06738>

Volkov, D., Petrov, A., Petukhov, V., Bikmaev, K., and Volk, D. (2025) Biollama: testing biology pre-training risks. Palisade Research, 10 February 2025. <https://palisaderesearch.org/research/biollama>

Petrov, A., and Volkov, D. (2025) Evaluating AI cyber capabilities with crowdsourced elicitation. arXiv:2505.19915. <https://arxiv.org/abs/2505.19915>

Reworr, Petrov, A., and Volkov, D. (2025) Palisade Hacking Cable Technical Report. Palisade Research, 26 August 2025. <https://palisaderesearch.org/assets/reports/hacking-cable-report.pdf>

Bondarenko, A., Ryzenkov, F., Turtayev, R., and Volkov, D. (2025) End-to-end hacking with AI agents. Palisade Research, 12 September 2025. <https://palisaderesearch.org/research/end-to-end-hacking>

Turtayev, R., Fedorova, N., Serikov, O., Koldyba, S., Avagyan, L., and Volkov, D. (2025) Misalignment Bounty: Crowdsourcing AI Agent Misbehavior. arXiv:2510.19738v2, 5 November 2025. <https://arxiv.org/abs/2510.19738>

Reworr, Petrov, A., and Volkov, D. (2025) GPT-5 at CTFs: Case Studies From Top-Tier Cybersecurity Events. arXiv:2511.04860. <https://arxiv.org/abs/2511.04860>

Petrov, A., Koldyba, S., Molchanov, S., Kotov, N., Volkov, D., and Serikov, O. (2026) Technical Report: Shutdown Resistance in Large Language Models, on robots! Palisade Research, February 2026. <https://palisaderesearch.org/research/shutdown-resistance-on-robots>

Air, A., Reworr, Kotov, N., Volkov, D., Steidley, J., and Ladish, J. (2026) Language Models Can Autonomously Hack and Self-Replicate. arXiv:2605.06760. <https://arxiv.org/abs/2605.06760>

Public recognition

Research led or co-led by Volkov has been highlighted by Yoshua Bengio and Dario Amodei, and covered by the Wall Street Journal, Fox News, MIT Technology Review, BBC Newshour, CNBC, and CBS News. His LLM honeypots have captured autonomous hacking attempts worldwide, his model organisms for misalignment have been discussed in TIME and this work and other work at Palisade figured in the 2025 discussions that led to the Singapore Consensus statement on AI. Elon Musk has commented publicly on the team's shutdown-resistance findings.