Pavel Golikov

Toronto, ON | paulgolikov@gmail.com | pavelgolikov.github.io | linkedin.com/in/pavel-golikov

PROFESSIONAL PROFILE

AI Researcher specializing in machine reasoning and alignment. Blends formal logic with rigorous low-level systems engineering to design highly creative adversarial evaluations. Identified structural vulnerabilities and single-query attention dilution in frontier LLMs. Former Military Intelligence Operator with a Top Secret clearance, bringing a rigorous, threat-modeling mindset to AI security and model capability evaluations.

TECHNICAL SKILLS

Languages: Python, C++, Java, SQL, LaTeX
ML & AI Frameworks: PyTorch, vLLM, HuggingFace, Transformers
Systems & Infrastructure: Linux/Ubuntu Server, AWS, Apache Flink, Distributed GPU Clusters
Core Competencies: Large Language Models (LLMs), Machine Reasoning, AI Alignment, Context Management, Mechanistic Interpretability, Adversarial Evaluations, Threat Modeling, Distributed Systems

FIRST-AUTHOR AI RESEARCH

- **Robust Reasoning Benchmark (RRB)** | *Under Review at a top-tier conference* 2026
  First Author | University of Toronto & Vector Institute *Toronto, ON*
  – Designed a highly creative adversarial evaluation framework leveraging 13 deterministic textual perturbations to decouple an LLM's mechanical deciphering from its underlying mathematical logic.
  – Demonstrated that the well-known phenomenon of attention drift occurs even *within a single query's Chain-of-Thought*, empirically showing that intermediate reasoning steps pollute the dense attention mechanism. This highlights the architectural need for working memory isolation **inside** the model's Chain-of-Thought, therefore identifying the optimal **granularity of reasoning** as a critical open research problem.
  – Engineered custom mechanistic interpretability pipelines in **PyTorch** to extract and analyze causal attention probability matrices across token index boundaries, testing models ranging from 7B to 30B parameters.
  – Identified critical safety-filter vulnerabilities in **Claude 4.6 Opus**, discovering that current alignment strategies penalize abstract character-level reasoning by misclassifying inputs as prompt injections.
  – Deployed large-scale distributed inference experiments using **vLLM** across internal multi-node clusters.
- **Context Management Benchmark** | *Active Research* 2026 – Present
  Lead Researcher *Toronto, ON*
  – Engineering a highly customizable evaluation framework combining LeetCode Python tracing and SymPy mathematical operations to generate arbitrary, programmatic chains of reasoning to allow construction of custom, complex task-subtask solution topologies, outputting both the final textual prompts and the programmatic Python verifiers for every intermediate subtask.
- **Fusing Adds and Shifts for Efficient Dot Products** | *IEEE CAL* 2025
  First Author | Hardware ML Research *Toronto, ON*
  – Proposed and validated a novel algorithmic optimization for dot-product computations, demonstrating a strong foundational understanding of hardware-level ML primitives and efficiency.

ENGINEERING & PROFESSIONAL EXPERIENCE

- **Systems & Infrastructure Engineering (MSc Thesis)** 2020 – 2022
  University of Toronto *Toronto, ON*
  – Engineered a flexible IoT distributed data-streaming framework from scratch, designed to automatically partition computational streaming queries between edge devices and cloud instances.
  – Built the full software stack: programmed Arduino/C++ sensors for real-time biological data collection (EMG/ECG), developed custom socket networking protocols, and deployed cloud infrastructure using **AWS** and **Apache Flink**.
- **Intelligence Operator** 2013 – 2018
  Canadian Armed Forces *Canada*
  – Held a **Top Secret** security clearance, conducting rigorous analysis of classified information streams to produce actionable intelligence reports for command elements.
  – Developed a strong adversarial threat-modeling mindset, emphasizing operational security, rigorous data validation, and the identification of logical vulnerabilities in complex, multi-agent scenarios.
- **Mathematics Teacher** 2012 – 2013
  Blyth Academy *Toronto*
  – Taught foundational mathematics, developing the ability to distill and communicate complex quantitative concepts.

EDUCATION

- **University of Toronto** Toronto, ON
  PhD in Computer Science (Paused to transition to industry) *2022 – Present*
- **University of Toronto** Toronto, ON
  Master of Science (MSc) in Computer Science *2020 – 2022*
- **University of Toronto** Toronto, ON
  Bachelor of Science (BSc) in Mathematics and Philosophy (Formal Logic) *Graduated 2011*