Baturay Saglam

17 Hillhouse Ave, New Haven, CT 06511

baturay.saglam@yale.edu • baturaysaglam.com • Google Scholar • baturaysaglam

Education

Yale University, New Haven, CT — *Aug 2023 – May 2027 (expected)*
Ph.D. in Electrical and Computer Engineering
Advisor: Dionysis Kalogerias
Dissertation: Test-Time Black-Box Control of LLMs for Safety and Alignment

Yale University, New Haven, CT — *2023 – 2026*
M.S. in Electrical Engineering

Bilkent University, Ankara, Turkey — *2020 – 2023*
M.S. in Electrical and Electronics Engineering
Thesis: Novel Deep Reinforcement Learning Algorithms for Continuous Control

Bilkent University, Ankara, Turkey — *2016 – 2020*
B.S. in Electrical and Electronics Engineering

Experience

Cisco Foundation AI, San Francisco, CA — *Oct 2024 – Present*
Ph.D. Intern
Core developer of the Foundation-Sec-8B model family (pre-training, post-training); automated prompt optimization for multi-step LLM pipelines; latent-space interpretability and guardrails.

Robust Intelligence *(acquired by Cisco)*, San Francisco, CA — *Sep 2024 – Oct 2024*
Student Researcher
Model evaluation, LLM robustness and AI safety, prompt-injection firewall development.

Google, New York, NY — *May 2024 – Aug 2024*
Student Researcher, Systems Research Group (SRG)
Out-of-distribution evaluation and long-context generalization.

Turk Telekom, Ankara, Turkey — *May 2022 – Jun 2023*
R&D Operations Assistant Specialist
Reinforcement learning for 5G/6G wireless control and resource allocation.

Publications

Preprints and Working Papers

[P4] Paul Kassianik[*], **Baturay Saglam**[*], Huaibo Zhao, Blaine Nelson, Supriti Vijay, Aman Priyanshu, Amin Karbasi. "FAPO: Fully Automated Prompt Optimization of Multi-Step LLM Pipelines." 2026. *([*]equal contribution)* [arXiv] [code] [blog]

[P3] **Baturay Saglam** and Dionysis Kalogerias. "Self-Improving In-Context Learning." 2026. [arXiv] [code]

[P2] **Baturay Saglam** and Dionysis Kalogerias. "Test-Time Safety Alignment." 2026. [arXiv] [code]

[P1] Supriti Vijay, Aman Priyanshu, Anu Vellore, **Baturay Saglam**, Amin Karbasi. "Think Before You Retrieve: Learning Test-Time Adaptive Search with Small Language Models." 2025. [arXiv]

[W1] **Baturay Saglam** and Dionysis Kalogerias. "Compatible Gradient Approximations for Actor-Critic Algorithms." 2025. *(major revision in progress)* [arXiv] [code]

Technical Reports (released models, benchmarks, and tooling)

[T5] Aman Priyanshu, Supriti Vijay, Kimia Majd, Xuhong He, Fraser Burch, Takahiro Matsumoto, Jianliang He, **Baturay Saglam**, Arthur Goldblatt, Zhuoran Yang, Amin Karbasi. "Vulnerability Localization Benchmark: Measuring Agentic Security Analysis at Repository Scale." 2026. *(contributor)* [report] [code] [blog]
VLoc Bench: two-phase agentic benchmark of 500 real-world vulnerabilities across 290 repositories and 147 CWE categories; 27 language models and four static-analysis tools evaluated.

[T4] Supriti Vijay, Aman Priyanshu, Didier Chapoteau, Arthur Goldblatt, Jianliang He, Kimia Majd, Fraser Burch, **Baturay Saglam**, Takahiro Matsumoto, Zhuoran Yang, Amin Karbasi. "Antares: Foundation Models for Agentic Vulnerability Localization." 2026. *(contributor)* [arXiv] [models] [blog]
Compact open-weight model family (350M–3B) for agentic vulnerability localization; Antares-3B outperforms open-weight models over 200× larger.

[T3] Zhuoran Yang, Ed Li, Jianliang He, Aman Priyanshu, **Baturay Saglam**, . . . , Kojin Oshiba, Yaron Singer, Amin Karbasi. "Llama-3.1-FoundationAI-SecurityLLM-Reasoning-8B Technical Report." 2026. *(core model developer)* [arXiv] [model]
Foundation-Sec-8B-Reasoning: reasoning-tuned variant of Foundation-Sec-8B; improves over Foundation-Sec-8B-Instruct on complex, multi-step cybersecurity tasks.

[T2] Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, **Baturay Saglam**, . . . , Zhuoran Yang, Yaron Singer, Amin Karbasi. "Llama-3.1-FoundationAI-SecurityLLM-8B-Instruct Technical Report." 2025. *(core model developer)* [arXiv] [model]
Foundation-Sec-8B-Instruct: instruction-tuned and RLHF-aligned variant of Foundation-Sec-8B; state-of-the-art on cybersecurity benchmarks and instruction-following tasks versus comparable-size models.

[T1] Paul Kassianik, **Baturay Saglam**, . . . , Omar Santos, Yaron Singer, Amin Karbasi. "Llama-3.1-FoundationAI-SecurityLLM-Base-8B Technical Report." 2025. *(core model developer)* [arXiv] [model]
Foundation-Sec-8B: Cisco's open-weight cybersecurity base model; state-of-the-art on cybersecurity benchmarks versus comparable-size models.

Peer-Reviewed Conference Proceedings

[C8] **Baturay Saglam** and Dionysis Kalogerias. "Test-Time Detoxification without Training or Learning Anything." *ICML*, 2026. *(proceedings forthcoming)* [arXiv] [code]

[C7] **Baturay Saglam**, Paul Kassianik, Blaine Nelson, Sajana Weerawardhena, Yaron Singer, Amin Karbasi. "Large Language Models Encode Semantics and Alignment in Linearly Separable Representations." *International Joint Conference on Natural Language Processing and the Asia-Pacific Chapter of the ACL*, 2025. [paper] [code]

[C6] Jane H. Lee, **Baturay Saglam**, Spyridon Pougkakiotis, Amin Karbasi, Dionysis Kalogerias. "Risk-Averse Constrained Reinforcement Learning with Optimized Certainty Equivalents." *NeurIPS*, 2025. [paper] [code]

[C5] **Baturay Saglam**, Xinyang Hu, Zhuoran Yang, Dionysis Kalogerias, Amin Karbasi. "Learning Task Representations from In-Context Learning." *Findings of the ACL*, 2025. [paper] [code]

[C4] **Baturay Saglam**, Doga Gurgunoglu, Suleyman S. Kozat. "Deep Reinforcement Learning Based Joint Downlink Beamforming and RIS Configuration in RIS-Aided MU-MISO Systems Under Hardware Impairments and Imperfect CSI." *IEEE International Conference on Communications Workshops*, 2023. [paper] [code]

[C3] Dogan C. Cicek, Enes Duran, **Baturay Saglam**, Kagan Kaya, Furkan B. Mutlu, Suleyman S. Kozat. "AWD3: Dynamic Reduction of the Estimation Bias." *IEEE International Conference on Tools with Artificial Intelligence*, 2021. [paper]

[C2] Dogan C. Cicek, Enes Duran, **Baturay Saglam**, Furkan B. Mutlu, Suleyman S. Kozat. "Off-Policy Correction for Deep Deterministic Policy Gradient Algorithms via Batch Prioritized Experience Replay." *IEEE International Conference on Tools with Artificial Intelligence*, 2021. [paper]

[C1] **Baturay Saglam**, Enes Duran, Dogan C. Cicek, Furkan B. Mutlu, Suleyman S. Kozat. "Estimation Error Correction in Deep Reinforcement Learning for Deterministic Actor-Critic Methods." *IEEE International Conference on Tools with Artificial Intelligence*, 2021. [paper]

Journal Articles

[J4] **Baturay Saglam**, Dogan C. Cicek, Furkan B. Mutlu, Suleyman S. Kozat. "Mitigating Off-Policy Bias in Actor-Critic Methods with One-Step Q-learning: A Novel Correction Approach." *Transactions on Machine Learning Research*, 2024. [paper] [code]

[J3] **Baturay Saglam**, Furkan B. Mutlu, Dogan C. Cicek, Suleyman S. Kozat. "Parameter-free Reduction of the Estimation Bias in Deep Reinforcement Learning for Deterministic Policy Gradients." *Neural Processing Letters*, 2024. [paper]

[J2] **Baturay Saglam**, Furkan B. Mutlu, Dogan C. Cicek, Suleyman S. Kozat. "Actor-Prioritized Experience Replay." *Journal of Artificial Intelligence Research*, 2023. [paper] [code]

[J1] **Baturay Saglam** and Suleyman S. Kozat. "Deep Intrinsically Motivated Exploration in Continuous Control." *Machine Learning*, 2023. [paper] [code]

Open-Source Software

FAPO: Fully Automated Prompt Optimization arXiv *2026 – Present*
Co-Creator
Open-source framework for end-to-end optimization of multi-step LLM pipelines (prompts, parameters, and chain structure), driven by step-level failure attribution.

FAITH: Foundation-AI Testing Hub for Cybersecurity *2024 – Present*
Co-Creator
Open-source evaluation harness for benchmarking language-model knowledge and capabilities in cybersecurity.

Latent-Space AI Firewall arXiv *2025*
Creator
LLM guardrail operating on the hidden states of Foundation-Sec-8B-Instruct; outperforms text-level filters (e.g., Llama Guard 3-8B) at detecting malicious inputs and adversarial prompts.

Honors and Awards

Gold Reviewer, ICML *2026*

Ph.D. Fellowship, Yale University *2023*

Ph.D. Fellowship, University of California, San Diego *(declined)* *2023*

Graduate Fellowship, Turk Telekom & Information and Communication Technologies Authority *2022*

Graduate Fellowship, Scientific and Technological Research Council of Turkey (TÜBİTAK) *2021*

Merit-Based Scholarship, Bilkent University *2016, 2020*

Teaching and Academic Service

Program Committee Member, *NeurIPS Workshop on Reliable Machine Learning* *2025*

Graduate Teaching Assistant, *CPSC 1100 — Python Programming for Humanities and Social Sciences, Yale University* *Fall 2025*

Graduate Teaching Assistant, *S&DS 617 — Advances in Large Language Models: Theory and Applications, Yale University* *Spring 2025*

Graduate Teaching Assistant, *CS 115 — Introduction to Programming in Python, Bilkent University* *Spring 2023*

Graduate Teaching Assistant, *EEE 211 — Analog Electronics, EEE 212 — Microprocessors, Bilkent University* *2020 – 2023*

Reviewer, *ICLR, NeurIPS, ICML, TMLR, ACL, Machine Learning, IEEE Robotics and Automation Letters* *2023 – Present*