Jacob Dineen

AI, Reasoning & Cognition (ARC) Lab — Arizona State University
jdineen@asu.edu | (480) 603–6994 | jacobdineen.com

RESEARCH INTERESTS

I work on reasoning and alignment in large language models (LLMs), with complementary interests in multi-agent reinforcement learning, controllability, and explainable AI (XAI).

EDUCATION

Arizona State University — Ph.D. Artificial Intelligence[†] GPA: 4.00/4.00 2022–2027
Advisor: Ben Zhou[‡]; Committee: Muhao Chen[§], Chitta Baral[§], Vivek Gupta[§]

University of Virginia — M.Sc. Computer Science GPA: 3.96/4.00 2019–2021
Advisor: Madhav Marathe[‡]

Syracuse University — M.S. Data Science GPA: 4.00/4.00 2017–2018

Grand Canyon University — B.S. Finance & Economics GPA: 3.65/4.00 2012–2015

[†] Expected completion. [‡] Advisor. [§] Committee member.

RESEARCH EXPERIENCE

Graduate Research Assistant, ARC Lab (Arizona State University) 2024–Present

- Research on reasoning, alignment, controllability, and multi-agent RL in LLMs.

Graduate Research Assistant, SEFCOM (Arizona State University) 2022–2024

- Research at the intersection of AI and cybersecurity.

Applied Research, Capital One — Center for Machine Learning (C4ML) 2020–2021

- Built reinforcement learning and agent-based simulations for organizational dynamics.

Research Assistant, Biocomplexity Institute (University of Virginia) 2019–2020

- Worked on graph dynamical systems, cooperative game theory, and behavioral modeling.

PROFESSIONAL EXPERIENCE

Research Engineering Intern, Pareto AI 2025–Present

Machine Learning Engineer, Spring Oaks Capital 2022–2025

Data Scientist, Capital One 2021–2022

Ph.D. Internships (3×), Capital One 2020–2021

Analyst & Business Intelligence, Real World Marketing 2016–2019

Data Scientist, Buffalo Check LLC 2015–2019

Optimization Analyst, Voltari 2012–2015

PUBLICATIONS

Google Scholar Profile

* Equal Contribution, $^{+}$ Corresponding Author / Mentor

Peer-Reviewed Conference Proceedings (C)

C1. **Dineen, Jacob**$^{+}$, Rrv, Aswin, Liu, Qin, Xu, Zhikun, Ye, Xiao, Shen, Ming, Li, Zhaonan, Lu, Shijie, Baral, Chitta, Chen, Muhao, & Zhou, Ben (2025). *QA-LIGN: Aligning LLMs through Constitutionally Decomposed QA*. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20619–20642.

C2. RRV, Aswin, **Dineen, Jacob**, Handa, Divij, Uddin, Md Nayem, Parmar, Mihir, Baral, Chitta, & Zhou, Ben (2025). *ThinkTuning: Instilling Cognitive Reflections without Distillation*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31236–31250.

C3. Xu, Zhikun, Shen, Ming, **Dineen, Jacob**, Li, Zhaonan, Ye, Xiao, Lu, Shijie, RRV, Aswin, Baral, Chitta, & Zhou, Ben (2025). *ToW: Thoughts of Words Improve Reasoning in Large Language Models*. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3057–3075.

C4. **Dineen, Jacob**$^{+}$, Kridel, Donald J., Dolk, Daniel R., & Castillo, David G. (2023). *Unified Explanations in Machine Learning Models: A Perturbation Approach*. In *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS-56)*, pp. 795–804.

C5. **Dineen, Jacob**$^{+}$, Haque, A. S. M. Ahsan-Ul, & Bielskas, Matthew (2021a). *Formal Methods for an Iterated Volunteer's Dilemma*. In *Proceedings of the 14th International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS 2021)*, pp. 81–90.

C6. **Dineen, Jacob**$^{+}$, Haque, A. S. M. Ahsan-Ul, & Bielskas, Matthew (2021b). *Reinforcement Learning for Data Poisoning on Graph Neural Networks*. In *Proceedings of the 14th International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS 2021)*, pp. 141–150. .

C7. Dolk, Daniel R., Kridel, Donald J., **Dineen, Jacob**$^{+}$, & Castillo, David G. (2020). *Model Interpretation and Explainability towards Creating Transparency in Prediction Models*. In *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS-53)*.

Working Papers / Under Review (W)

W1. Li, Zhaonan, Lu, Shijie, Wang, Fei, **Dineen, Jacob**, Ye, Xiao, Xu, Zhikun, Liu, Siyi, Cho, Young Min, Li, Bangzheng, Chang, Daniel, Nguyen, Kenny, Yang, Qizheng, Chen, Muhao, Zhou, Ben (2025). *Unbiased Visual Reasoning with Controlled Visual Inputs*. Pending ICLR 2026.

W2. Li, Zhaonan, Chickering, Kyle R., Li, Bangzheng, **Dineen, Jacob**, Ye, Xiao, Xu, Zhikun, Lu, Shijie, Huang, Yuxi, Shen, Ming, Nguyen, Bach, Pavuluri, Jaya Adithya, Nguyen, Mau Son, Chavan, Sanika, Le, Ngoc Minh Thu, Chen, Muhao, Zhou, Ben (2025).
Visual Analogies: Probing Unified Generation and Reasoning. Pending CVPR 2026.

W3. Liu, Qin, **Dineen, Jacob**, Huang, Yuxi, Zhang, Sheng, Poon, Hoifung, Zhou, Ben, Chen, Muhao (2025). *ArenaBencher: Automatic Benchmark Evolution via Multi-Model Competitive Evaluation*. Pending ICLR 2026.

W4. Ye, Xiao, Li, Zhaonan, **Dineen, Jacob**, Xu, Zhikun, Lu, Shijie, Shen, Ming, Shrivastava, Shaswat, Ahuja, Avneet, Zhou, Ben (2025).
CC-LEARN: Cohort-Based Consistency Learning. Pending ICLR 2026.

W5. Shen, Ming, Xu, Zhikun, **Dineen, Jacob**, Ye, Xiao, Zhou, Ben (2025). *BOW: Bottlenecked Next Word Exploration*. Pending ICLR 2026.

W6. Srinivasan, Adarsh, **Dineen, Jacob**$^{+}$, Afzal, Muhammad Umar, Sarfraz, Muhammad Uzair, Riaz, Irbaz Bin, Zhou, Ben (2025). *RECAP: Transparent Inference-Time Emotion Alignment for Medical Dialogue Systems*. Pending SIGCHI 2026.

W7. Ye, Xiao, **Dineen, Jacob**, Li, Zhaonan, Xu, Zhikun, Chen, Weiyu, Lu, Shijie, Huang, Yuxi, Shen, Ming, Tran, Phu, Yum, Ji-Eun Irene, Khan, Muhammad Ali, Afzal, Muhammad Umar, Riaz, Irbaz Bin, Zhou, Ben (2025). *Evaluating Medical LLMs by Levels of Autonomy: A Survey*. Pending EACL 2026.

SKILLS

OS: Linux (Ubuntu), macOS, Windows

Languages: Python, Rust, C, C++, Java, JavaScript, R, Bash, PRISM, x86-64

Databases: MySQL, PostgreSQL, MongoDB, Snowflake, Redis, SQLite, Redshift

Markup: LaTeX, HTML

ML Libraries: PyTorch, TensorFlow, Keras, JAX, Numpy, Pandas, Polars, Dask, HuggingFace, TRL, vLLM, VeRL, PySpark, NetworkX, DGL, Torch Geometric, BoTorch, SnowparkML

Tools: Git, VSCode, Docker, Kubernetes, Helm, Airflow, AWS (ECR/S3/EKS/CodeBuild), Databricks, OR-Tools, Sigma, Streamlit

SERVICE

Conference Reviewer: HICSS, SBP-BRiMS, NAACL, EMNLP, EACL

Teaching / Tutoring:

- CGCC Calculus & Linear Algebra Tutor (2019)
- Teaching Assistant — ASU CSE 365 / pwn.college (Security)

MISCELLANEOUS

Expert AI Trainer, Pareto AI (2024–Present)

Cybersecurity: Pwn.college green belt (binary exploitation) — pwn.college, user: `jdin`

GRADUATE COURSEWORK

Arizona State University — PhD Computer Science:
Knowledge Representation; Computer Systems Security; Software Security; Planning and Learning Methods in AI; Algorithms; Research/Dissertation Hours.

University of Virginia — MSc Computer Science:
Algorithms; Machine Learning; Computer Vision; Formal Methods; Reinforcement Learning; Graph Mining; Learning Theory (Game Theory); Cloud Computing; Research Hours.

Syracuse University — MS Data Science:
Data Analysis and Decision Making; Business Analytics; Financial Analytics; Marketing Analytics; Ad-

vanced Information Systems; Data Science; Data Warehousing; Text Mining; Scripting for Data Analysis; Information Policy.