Gabriel Mukobi

Web: gabrielmukobi.com | Email: gmukobi@cs.stanford.edu | Mobile: 360.525.7299 | GitHub: mukobi | LinkedIn: gabrielmukobi

Summary:

Researcher, engineer, and student passionate about research, governance, and field-building to reduce risks from advanced AI systems. Experienced in machine learning research, software engineering, and leadership in both small-team and large-company environments.

Experience:

Technical AI Safety Research Fellow, Existential Risk Alliance - July 2023-Sept 2023 - Cambridge, UK - erafellowship.org

Led self-directed technical research aimed at reducing risks from advanced language model systems in multi-agent scenarios. Skills: Research, AI safety/alignment/governance, prompt engineering, machine learning.

Gameplay Engineering Intern, Respawn Entertainment - June 2022-Sept 2022 - Remote - ea.com

Engineered core gameplay and AI features as a Software Engineering Intern on the gameplay team of Respawn's unreleased Star Wars first-person shooter title. Skills: Unreal Engine 5, C++.

Gameplay Engineering Intern, Riot Games - June 2021-Sept 2021 - Remote - riotgames.com

Designed and implemented core features as a Software Engineering Intern on the gameplay team of Project L, Riot Games' unreleased fighting game set in the League of Legends universe. Skills: Unreal Engine 4, C++.

Research Programmer Intern and Tools Programmer Intern, Epic Games - June 2020-Jan 2021 - Remote - unrealengine.com

Created deep reinforcement learning samples in Unreal Engine and a plugin to facilitate the use of ML in UE4. Engineered tools to predict LED wall moiré and other issues to improve Unreal virtual production shoots. Skills: RL, Unreal Engine 4, C++, Python.

Google Engineering Practicum Intern, Google Cloud Platform - June 2019-Sept 2019 - Seattle, WA - github.com/knative-portability

Developed several full-stack open-source applications as proof of portability for Knative, an open-source platform for serverless containerized workloads. Skills: Python, Flask, MongoDB, CI/CD, testing, OAuth 2.0, Node.js, Express.js, TypeScript, PostgreSQL.

Selected Projects:

Escalation Risks from Language Models in Military and Diplomatic Decision-Making - Oct 2023-Nov 2023 - Paper Forthcoming, GitHub

First author. Evaluating the risks from autonomous language model decision-makers in escalating international conflicts. Accepted to the MASEC NeurIPS 2023 workshop, for submission to ACL 2024.

Welfare Diplomacy: Benchmarking Language Model Cooperation - June 2023-Sept 2023 - arxiv.org/abs/2310.08901, GitHub

First author. Multi-agent LLM evaluations in a novel general-sum variant of Diplomacy that better incentivizes and measures cooperation. In review at ICLR 2024, accepted to the SoLaR NeurIPS 2023 workshop.

SuperHF: Supervised Iterative Learning from Human Feedback - Jan 2023-Sept 2023 - arxiv.org/abs/2310.16763, GitHub

First author. Alternative to RLHF using supervised learning instead of RL. Accepted to the SoLaR NeurIPS 2023 workshop.

Red Teaming Language Models for Unknown Risks - Oct 2023-Jan 2024- GitHub

Work in progress. Metrics and methods for uncovering qualitatively new harms in language models. For submission to ICML 2024.

Skills:

Artificial Intelligence - software.gabrielmukobi.com/ai

AI safety, NLP, evaluations, AI governance, ML, DL, foundation models, prompt engineering. Languages: Python, PyTorch.

Software Engineering- software.gabrielmukobi.com

Product management, documentation, testing, bug reporting, code review, CS, VCS, GitHub, GitLab. Languages: Python, C++, C#.

Web Development - software.gabrielmukobi.com/web

Full-stack, web design, cloud computing, databases, Docker containerization. Languages: JavaScript, Node.js, Python, HTML.

Game Development - software.gabrielmukobi.com/games

Unreal Engine, Unity, gameplay programming, tools, virtual reality, 3D modelling, computer graphics. Languages: C++, C#, Python.

Education:

Stanford University - M.S. Computer Science - Sept 2023-June 2024, B.S. Computer Science - Sept 2018-June 2023 - GPA: 3.994

Coursework in AI/ML, Computer Graphics, Computer Systems, Algorithms, and Theory. Stanford AI Alignment Founder and President.

Interests:

Photography, digital 3D art, filmmaking, music, animal welfare, video and tabletop gaming, fantasy, and science-fiction.