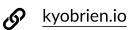
# Kyle O'Brien

# Machine Learning Researcher & Engineer

kyledevinobrien1@gmail.com







## Experience

## **ERA Fellowship** - Research Scientist (Contract)

Cambridge, England 6/2025 - 8/2025

- Developed a likely state-of-the-art dangerous capability unlearning approach for open-weight AI models.
- Midtrained 7B LLMs across 32 NVIDIA GH200 GPUS, including measuring FLOP efficiency and cost estimation.
- Wrote articles discussing pretraining interventions for AI security as well as the opportunities and risks of increasing open-weight model competition between the US and China.
- Presented research on open-weight AI risks and safeguards to researchers at RAND TASP.

#### EleutherAI - Research Scientist (Contract)

Remote 10/2024 - 6/2025

- Pretrained 7B LLMs from scratch across 128 NVIDIA H100 GPUs, resulting in the Deep Ignorance model suite.
- Engineered scalable CBRN pretraining filters that blocked unsafe capability acquisition with <1% additional training FLOPs and demonstrated that unsafe capabilities were not easily reintroduced via fine-tuning.
- Developed pretraining and midtraining datasets for large-scale LLM training runs, resulting in our team achieving desired performance levels in our first training run.
- Improved upon existing LLM benchmarks by identifying and mitigating issues with benchmark shortcut exploitation, allowing us to measure our model's robust knowledge that is invariant to prompting.

#### Microsoft - Software Engineer 2 & Applied Scientist 2

Redmond, WA 7/2020 - 1/2025

- Trained custom ML models for detecting hate speech, cyberbullying, and profanity in MSN news comments.
- Became the subject matter expert on finding relevant techniques from the machine learning literature and applying them to content moderation, such as dynamic exemplar selection and CoT prompts.
- Cleaned and hand-labeled datasets, debugged distributional shifts, designed benchmarks for evaluating external vision models, and engineered shared ML training pipelines.
- Designed experiments for improving retrieval performance in financial RAG application, reducing search results size by ~50% while improving performance.
- Drove the design and implementation of distributed systems resiliency design patterns across several teams, resulting in the ability to scale 4x to meet the Windows 11 launch without sacrificing availability.

## **Select Publications**

Preprint - Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards - Paper

**ICML Workshop** - Steering Language Model Refusal with Sparse Autoencoders - <u>Paper</u>

**ICLR** - Composable Interventions for Language Models - **Paper** 

 $\textbf{ICLR} \textbf{ -} \textbf{ Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon \textbf{ -} \underline{\textbf{Paper}}$ 

## **Technical Skills**

**Productive:** Pretraining, Literature Review, Paper Writing, Python, Docker, Azure, Hugging Face

Familiar: AWS, C#, CI/CD, Test-Driven Development, Front-End Web Development

# Education

### University of California, Santa Cruz

Class of 2020

Bachelor's Degree in Computer Science