Ulysses Sekai Tully Carr

sekaitullycarr@gmail.com | github.com/jammy-eggs | sekaitc.me

Research & Technical Experience

The LLM Data Company (YC x25)

Research Engineer Intern **Sep–Dec 2025**

- Co-authored evaluation rubrics (∼40 criteria per task, 10 domains, 100 tasks) for **Perplexity's DRACO Benchmark**, an open-source benchmark for evaluating frontier deep research agents, now used to score systems from Perplexity, Google DeepMind, and OpenAI.
- Built long-horizon, tool-use, and computer-use evaluation environments for benchmarking frontier models (GPT-4/5, Claude Sonnet/Opus 4.5, Gemini 3 Pro). Designed and reviewed hundreds of complex rubrics across non-verifiable and document-grounded domains including medicine, finance, and law. Contributed technical scoping for evaluation proposals to external labs.
- Designed and implemented an alternative architecture to **GEPA** for reflective prompt optimization in non-verifiable domains, using N candidates per round condensed through a reflection node that carried forward accumulated insights. Integrated this architecture into an internal pipeline for synthetic rubric-generation experiments.
- Expanded infrastructure for synthetic rubric evaluation experiments to align rubrics with task difficulty and calibrate criteria against high-, medium-, and low-quality model outputs.
- Conducted exploratory **Search-R1**-style training experiments with **veRL** on **Modal** to post-train **Qwen2.5-0.5B** for a tool-use agent environment.

Brown University

Capstone Researcher (Advisor: Prof. Stephen Bach) **Jan–May 2025**

- LLM Adaptation for Legal Writing Assessment (**Code**): compared zero-shot behavior, few-shot prompting, and **LoRA** fine-tuning (8-bit quantized, DDP on 4× NVIDIA A6000) for **Mistral-7B-Instruct** on rubric-based feedback for legal memoranda, with manual evaluation against expert professor annotations. Separately adapted **SoftSRV**-style synthetic data generation by training encoder-conditioned MLPs to map **BERT** and **Legal-BERT** embeddings into soft prompts for frozen **Mistral-7B-Instruct**.

Projects

Peven / Peven.jl

Creator **2026–Present**

- Built **Peven**, a **PyPI**-published package for designing environment-grounded LLM evaluations as colored Petri nets. Implemented a Python authoring layer that keeps environment, agent, and tool callbacks in Python while **Peven.jl** executes the net in Julia, handling Petri net scheduling and token semantics. Designed Peven to explore long-horizon and non-verifiable tasks where many rollouts can be evaluated through the same topology and inspected or scored at intermediate checkpoints.

Micrograd.jl

- Reimplemented **micrograd** in **Julia**, building a scalar autograd engine and small neural network library. Benchmarked loop overhead against Karpathy's Python implementation to better understand language-level costs in forward and backward passes.

Work Experience

AlertD

Software Development Intern **Jun–Aug 2025**

- Designed and implemented an agent creation workflow for a public AI agent marketplace.

EY-Parthenon, Software Strategy Group

Summer Associate **Jun–Aug 2024**

Education

Sc.B. Applied Mathematics–Computer Science, Brown University **May 2026**

- GPA: 3.9/4.0.
- Courses: Machine Learning, Artificial Intelligence, Computer Vision, Computational Probability & Statistics, Numerical Optimization, Statistical Inference.

Awards

True Ventures Fellowship, **2025**

- Selected for competitive cohort of student builders and future founders backed by True Ventures.

Undergraduate Teaching & Research Award (UTRA), Brown University **2023**

- Added R as a supported language in EDUC 1230 (Applied Statistics for Education Research) by translating problem sets, solutions, and creating a learning guide.

Watson Institute SPRINT Fellowship, Brown University **2022**

- Competitive research grant. Conducted multilingual research (English, French, Arabic) categorizing global incidents of education under attack; datasets included in the *Education Under Attack 2024* report.

Skills & Languages

Programming Proficient: **Python** Familiar: **Julia**

Libraries Proficient: **PydanticAI**, **NumPy**, **SymPy**, **Hugging Face**
Familiar: **PyTorch**, **veRL**, **Verifiers**

Tools Proficient: **Docker**, **Git**, **uv**, **pytest** Familiar: **Modal**, **Ollama**, **SLURM**

Research Agent evaluation, rubric design, benchmark development, LLM environments, agentic workflows, trajectory evaluation, RL post-training (**GRPO**)

Languages Native: English Fluent: French Proficient: German