

Mahdi Hasan

AI Training & Evaluation Specialist

Agent benchmark tasks (SWE-bench, Terminal-Bench) · Code review & RLHF ·
Linux/Docker · Bengali-English

Dhaka, Bangladesh · Remote (UTC+6)
business.mahdi.hasan@gmail.com
mahdihasan.space
linkedin.com/in/mahdi-meem
github.com/404mahdi

280+ approved benchmark tasks	3 AI evaluation platforms	2 yrs remote AI evaluation, since 2024	0.91 balanced accuracy, 70B lie detection
---	-------------------------------------	--	---

PROFILE

AI evaluation contractor: **280+ approved** SWE-bench-style and Terminal-Bench tasks for AfterQuery and Snorkel AI, plus RLHF and tool-use evaluation at Outlier. CSE graduate, BRAC University; thesis on lie detection in a 70B LLM.

EXPERIENCE

- **AfterQuery Experts** · Contract Engineer, AI Evaluation 2025 – Present · Remote
 - **130+ approved SWE-bench-style tasks** from real bugs and features in my own private TypeScript, JavaScript and Python repos
 - **120+ approved Terminal-Bench tasks** in ML/AI and compilers & parsers, each checked by behavioral tests
 - Every task ships fail-to-pass tests, a reference fix and a pinned Docker image
 - Hardened tests against shortcuts: hard-coded outputs, monkey-patching and reward forging
- **Snorkel AI** · Benchmark Task Author 2025 – 2026 · Remote
 - **30+ Terminal-Bench 2.0 tasks**: Docker/Linux environments with realistic failures and automated grading
 - Graded a frontier lab's code-review model on real GitHub PRs; wrote rationales used as RLHF data
- **Outlier** · Training Specialist Sep 2024 – Sep 2025 · Remote
 - Ranked, rewrote and critiqued model responses on React, JavaScript and API tasks across RLHF projects
 - Audited an assistant's extension calls (maps, notes and more) for the right call and a grounded answer

RESEARCH & PROJECTS

- Lie detection in a 70B LLM** · Undergraduate thesis · team of 3 · manuscript in preparation
- **0.91 balanced accuracy** detecting a 70B model's lies with no labels, where the benchmark's own probe scores **0.50**
 - Found the published persona method inverts on that dataset (AUROC 0.148); co-designed 10 personas that fix it
 - One zero-shot direction transfers across 6 datasets: **0.916** mean AUROC vs 0.732 for trained probes
 - Ran the evaluation on Llama-3.3-70B with PyTorch and nnsight/NDIF, and wrote the paper
- Leitmotif** · Solo · CSE425 Neural Networks · 2026 [code](#) · [demo](#) · [report](#)
- PyTorch · PyTorch Geometric · BERT · Gradio
- Music tagging and search: a graph neural network over song structure plus BERT over captions
 - **Caught label leakage** (65.6% of tags sat in the captions) and reported the honest **0.468** macro-F1, not the inflated 0.659
- 500 Beginner Python Projects** · Open source · 2024 – present [code](#)
- Python · GitHub Actions
- GitHub Actions that validate, security-scan and auto-merge contributor PRs · **10 stars, 9 forks**, about 8 contributors

TEACHING

- Python Instructor** · Edupy · Python and OOP for beginners Feb 2023 – Oct 2023
- Computer Science Tutor** · Python, data structures & algorithms, math for CS Jul 2024 – Dec 2024

EDUCATION

BRAC University · BSc in Computer Science and Engineering 2022 – 2026 · Dhaka
Coursework: Neural Networks · Artificial Intelligence · Algorithms · Operating Systems · Compilers · Circuits & Electronics · Digital Logic Design · Microprocessors · Linear Algebra · Probability & Statistics

SKILLS

EVALUATION	Benchmark task design	Test harnesses	Difficulty tuning	RLHF ranking	PR review	Tool-use audits
CODE	Python	TypeScript / JavaScript	Bash	C	C++	SQL
SYSTEMS	Linux	Docker	Git	GitHub Actions	pytest	Harbor
ML	PyTorch	PyTorch Geometric	Hugging Face Transformers	nnsight	scikit-learn	pandas
LANGUAGES	Bengali (native)	English (fluent)				