HUNAR BATRA

hunarbatra.com | i@hunarbatra.com | hunar.batra@cs.ox.ac.uk | Linkedin | GitHub | Google Scholar

EDUCATION

University of Oxford – *DPhil (PhD) Computer Science* **[Oct 2023 - Oct 2026]**

Reasoning Models, Reinforcement Learning and Safety | *Supervised by Prof. Ronald Clark (PIXL)*

Distinction in Assessed Work: 85%; Awarded Women in Computer Science Scholarship, Dept. of Computer Science, University of Oxford

University of Oxford – *MSc Advanced Computer Science* **[2021 - 2022]**

Dissertation: "Protein Language Representation Learning to predict SARS-CoV-2 mutational landscape", under Dr. Peter Minary [Overview]

Google Women in Computer Science Scholarship Awardee, Stanford MATS Scholar, GHC Scholar

University of Delhi – *BSc (Hons) Computer Science* **[2017 - 2020]**

8.42 CGPA, First Division Honours with Distinction (Rank 3), Student of the Year (2020), Published 6 research papers

RESEARCH EXPERIENCE [4.5 years research exp. including 2.5 years being full-time]

Oxford Martin School, Technical AI Governance Initiative – Research Assistant **[Nov 2025 – Present]**

Building an automated interpretability agent to detect, surface and control unintended behaviours in models.

VLAA Lab, University of California, Santa Cruz – Visiting Researcher **[March – Aug 2025]**

Built dense RL reward environments with lexicographic multi-objective policy optimisation to improve visual-spatial reasoning in MLLMs

UK AI Security Institute – Research Engineer **[Dec 2024 – Mar 2025]**

Built multi-agent evals for AI-AI exploits via persuasion, jailbreaks & oversight subversion w/ Cooperative AI Foundation (AISI Evals Bounty)

Model Evaluation and Threat Research (METR) – Research Engineer (Contract) **[June – July 2024]**

Designed LLM evals for AI R&D and collected human baselines for these to benchmark language models performance

Anthropic – *Research Engineer (Contract)* **[Sept 2023 – Feb 2024]**

Improved chain of thought transparency in LLMs by mitigating issues of ignored reasoning, sycophancy & biased reasoning via consistency training. Generated synthetic evaluations for detecting hallucination in models. Supervised by Ethan Perez

New York University Center for Data Science (Alignment Research Group) – Visiting Researcher **[June 2023 – June 2024]**

Mechanistic interpretability to decode intermediate encodings for LLaMA-2, trained tuned lenses & mitigate biased CoT reasoning via SFT

Stanford Existential Risks Initiative MATS – *Research Scholar (x2)* **[Nov 2022 – Sept 2023]**

Built research agents & tools for superalignment with process supervision, & worked on improving chain-of-thought faithfulness for LLMs

Oxford Human Centred AI Group, University of Oxford – *Research Intern (EWADA)* **[Nov 2021 – Dec 2022]**

Built decentralised ML apps using SOLID with privacy-preserving ML recommendations under Prof Jun Zhao & **Sir Tim Berners Lee**

Oxford Rhodes AI Lab – *Research Lead* **[May – Oct 2022]**

Leveraged GNNs to predict climate closures equation using symbolic regression in collab with CalTech (CLiMA), MIT & NASA JPL

University of Oxford, Computational Biology Group – *Researcher* **[April – Oct 2022]**

Trained LLMs with SARS-CoV-2 protein sequences to predict COVID-19 mutations with inductive biases from inverse folded AlphaFold2

University of Oxford – *Chatbot Development Research* **[Feb – Aug 2022]**

Developed a Question-Answering language model for the Philosophy Dept to help convey their research work over website & messenger

University of Delhi, Department of Computer Science – *NLP Student Researcher* **[March 2020 – July 2021]**

Researched & developed multiple projects- GPT-3 use-case model extractor, Ensemble ML Fake News detection, GPT-2 Title Generation. COVID-19 News Summariser using transformers, Medical QA bot. Co-authored and published 6 papers in IEEE & Springer Singapore

AI Research Lab, University of Delhi – *Computer Vision Student Researcher* **[June – Sept 2019]**

Built a Computer Vision based Assistive System for Autonomous Vehicles. Compiled Darknet with OpenCV for real-time predictions

WORK EXPERIENCE [1.5 year full-time SWE experience, and 2 years part-time exp.]

Swift Robotics – *Software Engineer, Robotics* **[Aug 2020 – Sept 2021]**

Developed Flask REST API to livestream video processed with Computer Vision techniques (OpenCV, image stitching- KNNs)

Built a React Native app which interacts with ROS melodic nodes to control robot's navigation & visualised LiDAR odometry

HushTech AI – *Co-Founder* **[June 2019 – July 2021]**

Built language model agents including email agents with DIET classifier, & chatbots for messaging platforms (WhatsApp, FB, Slack)

Omdena – *Machine Learning Engineer* *(One of the 28 Global AI experts selected)* **[March – June 2020]**

Applied statistical models: LDA topic modelling, VAR, ARIMA & EDA over COVID-19 policies. Results showcased at UN AI Summit

Impute Inc. – *Mobile Application Development Intern* **[March – June 2019]**

Developed & extensively trained a contextual conversation QA agent for Fluent8 iOS app. Deployed webhooks on Firebase Cloud Function

Inverted Sense – *Chatbot Development Intern* **[Dec 2018 – March 2019]**

Built chatbots using Twilio & developed an in-built shopping cart with up-selling resulting in higher lead conversions & ROAS

RESEARCH PUBLICATIONS | Google Scholar | **650+ Citations**

1. Reinforcement Learning with Simulated Feedback for Physically Grounded Reasoning; **Hunar Batra**, Ronald Clark; under review at ICML 2026
2. Simulated Learning: Self-Evolving Memory for Safe, Efficient and Transparent LLM Adaptation; **Hunar Batra**, Ronald Clark; under review at ICML 2026
3. SpatialThinker: Reinforcing 3D Reasoning in Multimodal LLMs with Spatial Rewards; **Hunar Batra**, Haoqin Tu, Yuanze Lin, Hardy Chen, Cihang Xie, Ronald Clark; under review at ICLR 2026 [**Accepted** to NeurIPS 2025 Workshops on *SPACE in Vision, Language, and Embodied AI* **(Oral)**, *Aligning Reinforcement Learning Experimentalists and Theorists*, *Embodied World Models for Decision Making*, and *Women in Machine Learning* **(Oral)**, Under review at CVPR 2026]

4. Towards Understanding Multimodal Fine-Tuning: A Case Study into Spatial Features; Lachin Naghashyar, **Hunar Batra**, Constantin Venehoff, Ashkan Khakzar, Ronald Clark, Christian Schroeder De Witt, Philip Torr; under review at ICLR 2026 [**Accepted** to NeurIPS 2025 Workshops on *Mechanistic Interpretability*, and *SPACE in Vision, Language, and Embodied AI*, Under review at ICLR 2026]
5. PiXLLaVA: Object-Level Spatial Grounding for Perceptual Reasoning in Multimodal Large Language Models; **Hunar Batra**, Ronald Clark; under review at ICRA 2026
6. Measuring what Matters: Construct Validity in Large Language Model Benchmarks; Accepted to **NeurIPS** 2025
7. Humanity's Last Exam; Contributor Authorship [Link]
8. EVCL: Elastic Variational Continual Learning with Weight Consolidation; **Hunar Batra**, Ronald Clark; ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling [Link]
9. Bias-Augmented Consistency Training Reduces Biased Reasoning in Chain-of-Thought; James Chua, Edward Rees, **Hunar Batra**, Sam Bowman, Julian Michael Ethan Perez, Miles Turpin [Link]
10. Protein Language Representation Learning to predict SARS-CoV-2 Mutational Landscape [MSc Thesis]; **Hunar Batra**, Peter Minary (2022)
11. MUCE - A Multilingual Use Case Model Extractor using GPT-3; Deepali Bajaj, **Hunar Batra** et. al, International Journal of Information Technology (IJIT 2022), Springer [Link]
12. TiGen - Title Generator based on Deep NLP Transformer Model for Scholarly Literature; **Hunar Batra**, Eshika G et. al, 3rd International Conference on Communication, Networks and Computing 2022, Springer [Link]
13. Medbot: Conversational Artificial Intelligence powered Chatbot for delivering Telehealth after COVID-19, IEEE 5th International Conference on Communications and Electronic Systems (ICCES 2020); Urmil Bharti, Deepali Bajaj, **Hunar Batra** et al,. IEEE Xplore [Link]
14. Serverless Deployment of a Voice-Bot for Visually Impaired, International Conference on Applied Soft Computing & Communication Networks (ACN 2020); Deepali Bajaj, Urmil Bharti, **Hunar Batra** et al,. Book Chapter - Springer Singapore [Link]
15. CoVShorts: News Summarization application based on Deep NLP transformers for SARS-CoV-2, **IEEE** 9th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO 2021); **Hunar Batra**, Akansha J, et al. - IEEE [Link]
16. CovFakeBot: a machine learning based chatbot using ensemble learning technique for COVID-19 fake news detection; **Hunar Batra** et. al, International Journal of Artificial Intelligence and Soft Computing 2022 [Link]
17. Solidflix: A decentralised movie social sharing app with privacy-preserving recommendations; **Hunar Batra**, Jun Zhao et. al; (2022)

PROJECTS | Github : github.com/hunarbatra

- **PiXLLaVA**: Interleaved object-centric Vision Language Model Alignment to reduce hallucinations [Link]
- **Visual Hierarchical Reasoning**: Improved acc of GPT-4V by 11% over MMMU by extracting segments + visual attributes [Link]
- **LLaMA-2-7B Tuned Lens**: Trained a tuned lens for LLaMa 2-7B to decoder outputs for intermediate layers [Link]
- **Model written sycophancy evals**: Evals generation using expert oversight guided multiversal dynamics exploration [Link]
- **Elastic Variational Continual Learning**: Combines EWC + VCL to reduce catastrophic forgetting [Link]
- **Scaffold**: Simulates alignment researchers comments on posts/drafts [Link]
- **Alignment Forum Summarisation tool**: Iterative MCTS with tuned expert agent to steer & generate summaries [Slides/Code]
- **GPT++**: Self-learning agent with internet access + episodic memory for reasoning – built before internet access LLMs came out [Link]
- **MuFormer**: Inverted AlphaFold2 for inverse-folding with pLM inductive bias to generate mutational sequences [Link]
- **CoVBERT:** COVID-19 mutation prediction language model [Link]
- **GraphSAGE LSTM & BiLSTM Aggregators**: Merged in PyTorch Geometric Package [Link]
- **HunAI:** DialoGPT DSTC telegram buddy bot

SKILLS

Python, C++, C, Javascript, SQL, App Dev (Native, React Native), Web Dev (React.js, TypeScript, Node.js, Flask, HTML, CSS,)
PyTorch, PyTorch Geometric [Merged PR], TensorFlow, Kubernetes, Google Cloud Platform, AWS, ROS
LLM Frameworks: Inspect AI, AutoGen, LangChain, LangGraph, PlayWright

AWARDS & ACHIEVEMENTS

- **Cosmos Institute AI Research Grant** to work on Interpretable RL reward learning, 2025
- **UK AISI Evals Bounty Recipient**, 2024
- **G-Research PhD Grant**, ICML 2024
- **Dan Kohn Scholarship**, KubeCon + CloudNative AI Conference EU 2024
- **Women in Computer Science Scholarship**, Department of Computer Science, University of Oxford, 2023
- **Long Term Future Fund Grant**, Effective Ventures, 2023
- **Research Scholarship**, Stanford University, Machine Learning Alignment Theory Scholar, 2022 and 2023
- **Google Women in Computer Science Generation Scholarship** EMEA, 2022
- **Grace Hopper Conference Scholarship**, Department of Computer Science, University of Oxford, 2022
- Deep Learning Theory Summer School Scholarship, Simons Institute for Theory of Computing, UC Berkeley, 2022
- **Rank 7**, G-Research Algorithmic Trading Oxbridge Challenge, 2021
- **Student of the Year & Rank 3,** Department of Computer Science, University of Delhi, 2020
- The Mars Generation **24 under 24** Award for Leaders & Innovators in STEM, 2019
- **National Finalist, Smart India Hackathon** Software Edition, (out of 5,000 teams) in India's largest hackathon by MHRD Govt. of India, 2019
- **National Winner**, Summer with Google (out of 20,000 participants), 2018

TEACHING

TA– AI Safety and Alignment, Dept. of Engineering Science, University of Oxford, Michaelmas 2025
TA– Computer Vision, Dept. of Computer Science, University of Oxford, Michaelmas 2025
TA– Machine Learning, Dept. of Computer Science, University of Oxford, Michaelmas 2024
TA– Deep Neural Networks, Dept. of Computer Science, University of Oxford, Michaelmas 2023
TA– Girls Who ML, Introduction to Machine Learning, University of Oxford, Michaelmas 2022

POSITIONS OF RESPONSIBILITY AND ACTIVITIES

- **Reviewer** – ICLR 2025, NeurIPS 2024 MINT Workshop, ICML 2024 Workshop on Agentic Markets 2024, ICML 2024 AI4ABM Workshop
- YCombinator AI Startup School, Summer 2025
- **DPhil Academic Representative –** Department of Computer Science, University of Oxford, 2023-26
- **IT Officer** – Oxford Women in Computer Science, 2023-26
- **Student Entrepreneur** – Oxford University Innovation and Oxford Science Enterprises, Summer 2022
- **Summer Fellow** – Global Leadership Initiative, Oxford Character Project, Summer 2022
- **IT Officer** – Oxford Women in Business, Trinity 2022
- **IT Officer** – Oxford Women in Computer Science, 2021-22
- **MSc Academic Representative** – Department of Computer Science, University of Oxford, 2021-22
- **Solutions Challenge Lead** – Google Developer Student Club Oxford, Michaelmas 2021
- **Oxford Mathematics Admissions Test Marker** – Mathematical Institute, University of Oxford, Michaelmas 2021
- **Lead** – Google Developer Student Club (One of the few students selected globally by Google, 2019-20
- **Mentor** – Google Code-in at TensorFlow, 2019-20
- YCombinator Startup School, Summer 2019

INVITED TALKS & WORKSHOPS

John Hopkins University Dept. of Computer Science Seminar, Dec 2025 – SpatialThinker: Reinforcing 3D Reasoning in MLLMs (Oral)
NeurIPS 2025 Workshop on SPACE in Vision, Language and Embodied AI, Dec 2025 – SpatialThinker: Reinforcing 3D Reasoning in MLLMs (Oral)
NeurIPS 2025 Workshop on Women in Machine Learning, Dec 2025 – SpatialThinker: Reinforcing 3D Reasoning in MLLMs (Oral)
Oxford Internet Institute, June 2024 – Bias-Augmented Consistency Training Reduces Biased Reasoning in Chain-of-Thought
Wolfson Engineering Day, Wolfson College, Oxford, May 2024 – Semantic Visual Tokenization for Vision Language Models
Oxford AI Mini-Conference, Feb 2024 – Large Language Models and AGI Panel
Oxford Women in Computer Science Panel, Nov 2023
Stanford Existential Risks Initiatives MATS Symposium, Feb 2023 – Accelerating Alignment Research via Human-AI Expert Iteration [Slides]
SolidWorld 2022 – Solidflix: A decentralised movie social sharing app with privacy-preserving recommendation [Video] [Slides] [Blog]
GirlsWhoML 2022, University of Oxford – Introduction to Machine Learning (Linear Regression and Logistic Regression) [Slides]
ICML 2022, Oxford Women in Computer Science Virtual Social – Highlighting Women Researchers in Machine Learning
Oxford Computer Science Conference 2022 – Protein Language Modelling to generate de novo SARS-CoV-2 mutations [Slides]
Oxbridge Women in Computer Science Conference 2022 – Protein Language Modelling to generate de novo SARS-CoV-2 mutations [Slides]
NLP Reading Group, University of Oxford (Feb and March 2022) – Presented state-of-the-art work on LLMs
ICRITO 2021, IEEE – Paper Presentation at 9th International Conference on Reliability, Infocom Technologies and Optimization, IEEE
ICCES'20, IEEE – Paper Presentation at 5th International Conference on Communication and Electronic Systems, IEEE
ACN'20, Springer – Paper Presentation at 5th International Conference on Applied Soft Computing and Communication Networks, Springer
Ryerson University (The DMZ, Think Outside the Valley 2020) – Process Automation with Chatbots [Video]
HackOn Hackathon 2020 – Ok Google! Let's build an action for Google Assistant [Video]
SRCC, University of Delhi 2019 – Chatbot Development for Marketing (youngest invited speaker)
Google DevFest New Delhi 2019 – Project showcase, Google Developer Students Club
WHRC 2018 – Ideation Paper Presentation on 'Ingestible Robots' at 15th WONCA World Rural Health Conference 2018