Partha Pratim Saha

LinkedIn Github technical.partha@gmail.com

Research Areas:
Natural Language Processing, Large Language Models, Mechanistic & Geometric Interpretability, AI Explainability, AI Safety & Alignment, Cross-cultural Agentic AI, LLM-based Reasoning, Federated & Privacy-Preserving AI

Research Directions:
I develop **information-geometric frameworks** for tracing **how LLMs encode, propagate, and transform** beliefs across layers and how **alignment, fine-tuning, cultural model merging alter that internal geometry**. Three active research threads:

(1) AI Safety, Alignment & Mechanistic Interpretability of LLMs: The **Belief Vector Field (BVF)** models hidden-state trajectories on the Fisher–Rao statistical manifold, extracting spectral torsion norms and geodesic curvature to operationalize **whether alignment modifies internal beliefs or only constrains outputs**. Torsional geometry achieves the highest Cohen's d separation between safe and harmful prompts across Llama-3.2, OLMo-7B, Gemma-3, Qwen, Mistral (11 metrics).

(2) Cross-cultural & Multilingual AI: To investigate epistemic inheritance in merged LLMs using Fisher-Rao Geometry, producing **neural offspring**; emergent cultural nDNA is measured via spectral curvature deviation $\Delta\kappa_\ell$ and thermodynamic length divergence $\Delta\mathcal{L}_\ell$.

(3) Privacy-Preserving FL: Collaborative Federated Learning (CFL) for healthcare: PHI/PII-separated mixed-mode training enabling personalized AI without centralizing sensitive data. Published SpringerNature ICOMP'24.

My **long-term vision**: AI systems aligned not only behaviourally but *geometrically*-with transparent, auditable internal representations.

Education

- **Masters of Technology in Data Science & Engineering(DSE)** — Pilani, India
  Birla Institute of Technology and Science(BITS) Pilani — *Sept 2019 - Aug 2021*
  Final GPA: **9.08/10 Distinction**
  Rank: top 5% out of 600 in DSE, 2022
  Relevant Courses: Natural Language Processing, Machine Learning, Deep Learning, Data Science, Math, Statistics, Data Mining, Big data systems
  Dissertation Research: proposed Collaborative Federated Learning (CFL) cloud-based system separates datasets into public sets and private sets, respectively, based on the removal of Protected Health Information (PHI) and Personally Identifiable Information (PII) for personalized GPT-like systems.

- **Bachelor of Technology in Computer Science & Engineering(CSE)** — Kolkata, India
  West Bengal University of Technology — *August 2006 - July 2010*
  Final GPA: 8.49/10
  Rank: top 5% out of 70 in CSE, 2010
  Relevant Courses: Maths, Statistics, Algorithms, AI, Probability, Data Structures, Programming

Publications

1. [SPINAL] Arion Das, **Partha Pratim Saha**, Aman Chadha, Vinija Jain, Amitava Das: **Scaling-law and Preference Integration in Neural Alignment Layers**: Paper Github: **I provided model experiments, paper writing**

2. [NeurIPS 2025 https://arxiv.org/pdf/2509.00849] Shaina Raza, (Maximus Powers, **PARTHA PRATIM SAHA**), Mahveen Raza, Rizwan Qureshi: **Prompting Away Stereotypes? Evaluating Bias in Text-to-Image Models for Occupations**: **NeurIPS 2025** Workshop: Paper Github: **I provided end-to-end pipeline designing, dataset annotation of all models, paper writing**

3. [Journal Paper **DOI:10.5281/ZENODO.15126395**] Cherishma Kumar Subhasa, Endriyas Zenagebriel, **PARTHA PRATIM SAHA**, Zarah Rezaei, Joseph AKINYEMI : **Enhancing human empathy in conversations using transformer-based models**: published in Sciencematch, Impact Scholar Program 2025: **I was the top contributor & provided all technical ideas and the process-pipeline for this publication**

4. [Conference Paper] [SpringerNature] **PARTHA PRATIM SAHA**, Naresh K. Sehgal, Miad Faezipour : **Collaborative Federated Learning Cloud Based System**: International Conference on Internet Computing & IoT (ICOMP'24) Computer Engineering, & Applied Computing (CSCE), USA [Video] [Short present] [Long present]

Significant Research Projects and Work Experiences

- **WBSCTED - Nalhati Government Polytechnic College, State Govt. of West Bengal** West Bengal, India
  Lecturer in Computer Science and Technology *Dec'21 - Present*

  - **Course Supervisor**: Teaching courses on Machine Learning, Deep Learning, Internet of Things, Python and JAVA programming, Computer Graphics
  - **Act as Project Supervisor**: For the final year major project, I advised 50+ students (during last 4 years) in Machine Learning, Deep Learning, Agentic AI, and NLP projects on Conversational AI, Emotion-based and Empathetic Chatbot development
  - **Administrative Responsibilities**: **Current and Former HoD in CS**, followed **Bloom's taxonomy** and **student-centric innovative pedagogy** and technology integration, held office hours for administrative work, departmental procurements, and academic planning along with the academic council, maintained good inter-and-intra-departmental communication.
  - **(1) Research Project:** ***Belief Vector Field (BVF) of AI Alignment & Safety***:
    - **Aim**: Do alignment procedures modify a models **internal belief geometry** (hidden-state trajectories on the Fisher–Rao manifold) or only constrain outputs? Evaluated across 5 LLM families, 11 metrics.
    - **Finding**: **Torsional geometry** achieves the highest Cohens d separation between safe and harmful prompts (SFT vs. DPO variants, statistically extrapolated).
    - **Models**: LLama-3.2-3B/Instruct, OLMo-3-7B-SFT/DPO, Gemma2-2B-Base/Instruct, Qwen3-4B-SFT-DPO
    - **Datasets**: Litmus, HarmfulQA, CategoricalHarmfulQA, hh-rlhf (Anthropic)
    - **Metrics**: Spectral torsion norm $\|\tau_\ell\|$, geodesic curvature κ_ℓ, nDNA compression ratio, brake onset depth ℓ^*.
  - **(2) Research Project:** ***Semantic Helix of LLMs (nDNA)***:
    - **Aim**: Unify fine-tuning, alignment, distillation, and merging as **measurable deformations of the same depth-wise semantic flow** via spectral curvature κ_ℓ and thermodynamic length $\mathcal{L}_\ell$ (epistemic effort across layers).
    - **Finding**: Provides the first interpretability interface that makes fine-tuning, collapse, merging, alignment, and distillation directly comparable as layer-wise geometric deformations.
    -**Supervisor**: Part of the nDNA: Semantic Helix of Artificial Cognition project under supervision of Prof. Amitava Das
    - **Models**: Llama3-instruct, DeepSeek-R1-Distill-Llama, Qwen-base, tulu3-8b-SFT/DPO, Mistral-instruct
    - **Datasets**: SQuAD, SQuAD-v2, Stanford-Plato, GSM8K
  - **(3) Research Project:** ***Cultural encoding streams in LLMs***:
    - **Aim**: Investigate **epistemic inheritance** in merged LLMs—does model merging produce cultural hybrid vigour or epistemic dominance? Culturally fine-tuned LLMs pairwise-merged via **Fisher-Weighted Averaging (FWA)** using `mergekit`, producing **cultural neural offspring**.
    - **Supervisor**: Part of the nDNA: Semantic Helix of Artificial Cognition project under supervision of Prof. Amitava Das
    - **Finding**: Neural merging produces **non-linear epistemic inheritance**: directional inheritance (one parent dominates), abstract fusion (emergent cross-cultural priors), and representational tension (manifold stretched beyond both parents). [GitHub]
    - **Cultures**: African, Latin American, South Asian, East Asian, Arabic, Indigenous, European, Pacific Islander. **LoRA-finetuned base models**; lower 16 layers frozen post-fusion.
    - **Measurement**: $\Delta\kappa_\ell$ (spectral curvature deviation), $\Delta\mathcal{L}_\ell$ (thermodynamic length divergence), $\Delta_\ell^{inherit}$ (belief vector deviation from mean of parents); FWA preserves curvature-salient weight directions vs. linear interpolation.
  - **(4) Research Project:** ***Neural Robustness learning in Dense Transformers***:
    - **Aim**: Can **formal robustness certificates** for LLMs be derived from data contamination, label noise, and adversarial attacks rather than from input-perturbation bounds alone?
    - **Finding (preliminary)**: DPO-aligned models show **compressed Fisher-norm spectra** vs. SFT counterparts, suggesting alignment induces geometric contraction that doubles as a robustness mechanism; torsion norm outperforms output confidence as a robustness indicator.
    - **Framework**: Using Lifts Lipschitz robustness bounds (Szegedy et al.) and reference GitHub to theoretically guarantee **upto 40% data contamination holds almost $\Rightarrow$ natural robustness**.
    - **Models**: Vision Transformer (ViT), MedSigLIP, CLIP
    - **Datasets**: PathMNIST, DermaMNIST
    - **Evaluation**: based on different loss functions: S-DIV, CCE, GCE, TDPDSCCE

- **Wipro Limited - Conversational Dialog System** Bangalore, India
  Lead Data Scientist *Sept'21 - Nov'21*

  - **Goal**: Develop a conversational chatbot by removing ambiguities in text queries
  - **My contribution**:
    - **Developed a conversational system** Leading and managing a team of 5 members
    - **Implemented 50+ custom intents, entities** and dialog flow with IBM Watson
  - **Impact**: Used by **0.3 million people** worldwide
  - **Tools**: Dialog flow, IBM word alignment models, Postman, NodeJS, IBM Cloud

- **BirlaSoft - Johnson & Johnson(J&J) R&D, Project: Medical Search Engine** New Delhi, India
  Senior Data Scientist *Dec'17 - Aug'19*

  - **Goal**: Develop a medical search engine for queries for healthcare
  - **My contribution**:
    - **Used SciBERT as base transformer model with SciSpacy NLP pipeline** techniques

- **Impact**: Used by **over 0.1 million people** who use Johnson & Johnson products
- **Tools** : Python, Word2Vec, SciSpacy, Fuzzy Search, Flask

- **Indian Institute of Technology(IIT) Kanpur, Research Project:Threat Intelligence System** IIT Kanpur
  Project Engineer in Computer Science Dept., IIT Kanpur *Nov'16 - Jul'17*
  - **Goal**: Develop secure threat management system for academic institute so that that different categories of users can, e.g., instructors to enter grades for their courses, the dean's office to create grade reports for students, administrators can update grades when a grade change request is initiated, and students to check their grades
  - **My contribution**:
    - **Researched on Cyber-Security** to mitigate cyber attacks due to Lack of Integrity, Lack of Confidentiality, and Lack of non-repudiation
    - **Developed Threat Intelligence System** that provides near real time information about threats and threat actors
  - **Impact**: Provided full visibility, situational awareness and actionable threat intelligence software systems
  - **Tools** : Python, Drupal, Django

- **Infosys Technologies Limited, Research Project: Alignment & Cancer Genomics in AI** Chennai, India
  Senior Systems Engineer *Jan'11 - Jul'15*
  - **Goal**: Given DNA sequences in FASTQ file, insertions/deletions/substitutions are needed to convert one sequence into another
  - **My contribution**:
    - **Applied Edit Distance and Needleman Wunsch** algorithms
    - Determined the **minimum no of insertions, deletions and substitutions needed** for the given gene sequences
    - **Identified similarities** between DNA sequences - **Identified top 10 genes that drive Multiple Myeloma(MM)** blood cancer
    - **Implemented 3 research papers** to rank genes that are responsible for MM
  - **Impact**: **Biological hierarchy determination** for new species, **Drug Design**, **Increase in life expectancy**
  - **Tools** : Python, Word2Vec,Numpy, Pandas, Dynamic programming

Skills Summary

- **NLP + AI**: Foundation Models, AI Alignment and Safety, Belief and knowledge editing in LLMs, AI Agents, Conversational Systems, Generative AI, Interpretable and Explainable AI(XAI), Fine-training of Large Language Models from Huggingface, Hybrid and multi-hop RAG
- **LLM**: Worked extensively on DeepSeek, GPT, Llama, Gemma, Mistral, and Qwen models (base / Instruct) for LLM paraphrasing, Self-certainty, model fine-tuning, working on Reasoning in LLMs, NeuroSymbolic AI, Ethics and trustworthiness of LLMs, Model evaluation and benchmarking
- **Deep Learning**: Experience in working on deep learning models such as Autoregressive Generative Models, Neural Network, CNN, RNN, Transformer, GPT
- **Machine Learning and Data Science**: Text and Multimodal Data curation, Regression, Classification, Clustering, Tree-Based Algorithms, Bagging, Boosting
- **Programming Languages**: Python, Java, C, Objective-C
- **Frameworks**: Llama-Index, LangChain, PyTorch, Scikit, Numpy, Pandas, NLTK, SpaCy, PyTorch, TensorFlow, Keras, Flask, NodeJS
- **Tools**: Kubernetes, Docker, GIT, PostgreSQL, MySQL, SQLite
- **Domains**: Cancer Genomics, Bioinformatics and Biomedicine, Real Estate, Finance, and Banking
- **Cloud Platforms**: Worked on Inference systems and deployment, Azure, AWS, IBM Cloud, Watson ML, Amazon EC2

Awards and Achievements

- Got accepted in **SPAR Demo Day, 2025** on AI Safety & alignment research
- Received **5x Google Colab Pro A100/H100 GPU compute units of 300 each** to continue my research experiments on AI Safety & alignment by Neuromatch Academy
- Selected to serve as a **Reviewer** in MTI-LLM @ **NeurIPS 2025**
- Super excited that I am AWS AI & ML Scholars by Udacity [Certificate], 2025
- Got acceptance & received 90% scholarship to Armenian LLM Summer School'25 [Certificate]
- Attended to Duke Machine Learning summer school 2025
- Attended the prestigious Cohere Summer School 2025 [Certificate] and IIIT Hyderabad summer school, India
- My **Proposal** on developing Empathetic & Situationally aware LLMs got **accepted** into Impact Scholar Program
- Got Acceptance in **University of Chicago Data Science Institute (DSI) Summer School'24** and attended AI-Science Research program : Eric and Wendy Schmidt Postdoctoral Fellowship (Schmidt Futures) [Certificate]

- Received MLx **Generative AI Scholarship** Grant in 2024, 2025 from **Oxford ML summer school** [Certificate]
- Got acceptance and will attend Athens Natural Language Processing Summer School 2024
- Got acceptance and attended Neuromatch Academy in **Deep Learning**. [Certificate] [Certificate]
- Accepted to attend **Data Science and Deep learning** and attended **diiP Summer School**, Paris, 2024
- Remotely attended AI Summer School at **New York University(NYU)**, 2022
- Got selected and attended the **Google Developer's program**, Google India, 2019
- Got awarded on AI4 IMPACT from AI Singapore Summer School, Singapore, 2021
- Received **Distinction** in Master Degree program from BITS Pilani, 2022
- **Awarded Udacity Bertelsmann Technology scholarship by Google** for AI Track in ML, 2021

Teaching Experience - Teaching Assistant (TA) at BITS Pilani, Pilani Campus

- **NLP Applications**, AI ML Masters program [Winter 2023]
- **Deep Learning**, AI ML Masters program [Fall 2021]
- **Deep Reinforcement Learning**, AI ML Masters program [Spring 2021]

Responsibilities: My responsibilities were teaching, code demonstrations during webinars, in-class presentations, labs, quizzes, and mid- and end-semester question preparation and evaluations; helping the lead instructor with course content preparation; and student doubt clearance post classes.

Achievement: Awarded teaching contracts and **received** $2,513.11$ **USD as an honorarium** from the Birla Institute of Technology and Science (BITS), Pilani, a deemed university.