Partha Pratim Ray

Assistant Professor • Researcher • IoT, Edge Computing and LLM

Email: parthapratimray1986@gmail.com
GitHub: github.com/ParthaPRay
Google Scholar: scholar.google.co.in/citations?user=ioplfagAAAAJ
Orcid ID: 0000-0003-2306-2792
Mobile: +91-9433668194, +91-7908402850

Professional Summary

Experienced academic with 12+ years of expertise in IoT and Edge Computing, recognized among the world's top 2% scientists by Stanford University (2020–2024). Proven skills in developing scalable IoT and edge solutions for smart systems and healthcare. Currently transitioning to specialize in Large Language Models (LLMs) and Machine Learning (ML), focusing on quantized LLMs and open-source AI applications on edge. Seeking to bridge research and practical innovations in AI, ML, and LLMs for real-world challenges. Currently seeking internships and collaborations to deepen expertise in LLM, and ML with the objective of transitioning into the IT industry along with becoming a versatile educator.

Education

- **Ph.D. in Computer Applications** *Expected: 2026*
  Sikkim University
  Thesis: Enabling Large Language Models on Resource-Constrained Edge: A Multi-Faceted Approach
  GPA: 9.29/10
- **M.Tech. in Electronics and Communication Engineering (Embedded Systems)** *2011*
  Haldia Institute of Technology, MAKAUT
  GPA: 9.60/10
- **B.Tech. in Computer Science and Engineering** *2008*
  B.P. Poddar Institute of Management and Technology, MAKAUT
  GPA: 8.15/10
- **GATE Qualified** in Computer Science (CS) *2009*
  Percentile: 91.62%

Key Skills

Programming Languages:	C, Java, HTML, Python
AI/ML/LLM Frameworks:	**Familiar:** Ollama, scpCy, scikit-learn, Docling, Sentence Transformer
Tools and Platforms:	Raspberry Pi, Arduino, ESP32, FastAPI, Flask, RESTful API, Quantized LLMs
Research & Analytics:	IoT, Edge Computing, Large Language Models, Machine Learning Algorithms
Certifications:	Certified Blockchain Expert-V2, Mastering Digital Twins
Soft Skills:	Mentoring, Research & Analysis, Technical Article Writing & Publication

Key Projects

XML Extraction Visualizer
Developed a XML variable extractor and visualizer
- Parses and extracts variables from any XML file. Performs followsing similarity measures: (i) Jaccard Similarity, (ii) Levenshtein Distance, (iii) Cosine Similarity, (iv) Semantic Similarity (using SentenceTransformer)
- Uses interactive visualization for relationship networks using pyvis.

NLP Pipeline with OpenAI
Developed a NLP pipeline using OpenAI using spaCy, NLTK, wordcloud
- This Natural Language Processing (NLP) Pipeline along with gpt-4o-mini provides a comprehensive solution for analyzing, clustering, and visualizing text data.
- It integrates advanced machine learning techniques with a user-friendly Gradio interface, enabling users to interactively explore results with structured outputs and dynamic visualizations.

Multi-format Document RAG
Developed a RAG system that considers multiple type of documents
- This project demonstrates how to implement Retrieval-Augmented Generation (RAG) using Gradio, LangChain, Docling, Milvus, and HuggingFace.
- The system supports multiple file types, including PDF, Images, HTML, and PPTX, for document conversion, chunking, and question answering.

Book Document Image Mapping Systems
Developed book mapping based on image
- The Document-Image Matching System is an advanced application designed to analyze images, extract keywords, and fetch relevant books using cutting-edge AI capabilities.
- It leverages OpenAI's vision and text APIs, alongside the Open Library API, to deliver accurate and efficient

results.

Weather Forecast Summarizer

Developed NOAA weather forecast summarizer

- The NOAA Weather Summarizer is a Python-based project that retrieves weather forecasts from the National Weather Service (NWS) API, processes the data, and provides detailed and summarized forecasts.
- The application uses NLP, OpenAI, Gradio for a user-friendly web interface and OpenAI for text summarization.

Yolo11 Dog Detection App

Developed do detection fine tuning on stanford dataset

- The YOLO Dog Breed Detection Web App is a powerful and user-friendly tool designed to detect and classify dog breeds within uploaded images using the YOLO (You Only Look Once)-11n object detection framework.
- Leveraging a pre-trained YOLO model with 120 dog breed classes, this web application provides real-time detection results, including bounding boxes, confidence scores, and comprehensive validation metrics.

Sequential Agent

Developed sequential agent using local LLM on resource-constrained edge

- Designed multiple sequential *squad* agents to process user prompts as tasks by allotting some tasks per agent such as 1, 2 , 4 and 8 for agents such as 2, 4, 8, and 16 with collaborative approach.
- It used yaml for designing agents and tasks with agents role, goal, backstory, context, output file directory and other configurations. such as all-minilm:33m, redis, Ollama, qwen2.5:0.5b-instruct, FastAPI ad Quad core Cortex-A72.

Function Call Chaining

Developed function call chaining using local LLM on resource-constrained edge

- Designed multiple functions to process output of one as input to other function in sequential manner with remote API call and local external functions.
- It used Ollama,qwen2.5:0.5b-instruct, llama3.2:1b-instruct-q4_K_M, smollm2:1.7b-instruct-q4_K_M, all-minilm:33m, FastAPI and Quad core ARM v8 64-bit SoC.

Sequential Function Call

Developed function call chaining using local LLM on resource-constrained edge

- Designed multiple functions to process sequential manner with remote API call and local external functions using one-shot and few-shot.
- It used Ollama,qwen2.5:0.5b-instruct, llama3.2:1b-instruct-q4_K_M, all-minilm:33m, FastAPI ad Quad core SoC.

Semantic Static Routing

Developed static routing using local LLM on resource-constrained edge

- Designed multiple static routes to process user prompt for optimal en-routing to optimal LLM calls.
- It used utterance based approach per static route which were compared against sentence transformer such as all-minilm:33m, nomic-embed-text, snowflake-arctic-embed:110m, and mxbai-embed-large with Ollama,qwen2.5:0.5b-instruct, FastAPI and Quad core 64-bit SoC.

Semantic Dynamic Routing

Developed dynamic routing using local LLM on resource-constrained edge

- Designed multiple static routes to process user prompt for optimal en-routing to optimal LLM calls by triggering designated functions using zero-shot, one-shot and few-shot approach with dynamic function schema design.
- It used utterance based approach per static route which were compared against sentence transformer such as all-minilm:33m, nomic-embed-text, snowflake-arctic-embed:110m, and mxbai-embed-large with Ollama, qwen2:0.5b-instruct, FastAPI and ARM v8.

IoT Integration with LLM

Developed *LLMIoT* and *LLMEdge* - frameworks for integrating IoT with localized LLMs using resource constrained edge scenario.

- Designed framework that employed IoT clients to communicate with local RESTful API based LLMs on constrained edge device.
- It used utterance based approach per static route which were compared against sentence transformer such as ESP32, ESP8266 and ARduino MKR1000 with Ollama, qwen2:0.5b, FastAPI and ARM v8 over Wi-Fi.

Linguistic Relativity with ChatGPT

Developed an experiment of hypothesis testing about the applicability of linguistic relativity of various language specific prompts asked to and responded back from ChatGPT 4 mini.

- Designed hypotheses to test whether linguistic relativity is applied on ChatGPT-based multilingual prompt responses.
- It used paraphrase-multilingual-MiniLM-L12-v2 sentence transformer to compare semantic similarity across multiple languages responses. Performed *polarity* comparison across the responses of the languages by using *'polyglot'* package.

API-aware Image and Video Generation Database System

Developed a workflow for generating AI-driven, API-aware images and videos based on text prompts, storing the generated content and logs in databases, and serving the results through a web interface.

- Designed a full-stack project to help user getting image (png format) and videos (mp4 format) in local system using user login data.
- It used DALL-E-3 for image generation, base64 for converting image download in local machine along with the

text prompt to API call RunwayML for 5s video generation and download in local directory. Developed sqlite3, Flask API and sqlalchemy aware databases to manage content and user activity.

Retrieval Augmented Generation (RAG)

Developed a RAG-powered chatbot for legal query resolution based on the Indian Constitution and Indian Penal Code.

- It used five open-source LLMs (Llama3, Mistral, Gemma2, Phi3, and Qwen2) in terms of relevance, faithfulness, context recall, and precision.
- Tools used include LangChain, Ollama, ChromaDB, Streamlit, Python 3.9, and hardware such as an Intel Xeon Processor, 1 TB HDD, and Quadro RTX 5000 GPU.

Professional Experience

Assistant Professor (Senior Scale) *Dec 2020 – Present*

Department of Computer Applications, Sikkim University, India

- Authored over 100 SCI-indexed journal articles, 32 conference articles, 7 books, 3 book chapters, showcasing a strong commitment to research excellence in areas like IoT, Edge Computing, and LLM.
- Supervised numerous master's thesis projects and contributed to student academic growth with innovative projects such as IoT-based applications, blockchain integration, and edge computing systems.
- Played a key role in drafting the syllabus for various programs, including MCA, B.Sc. Computer Science, and vocational courses like B.Voc.
- Taught diverse courses ranging from Discrete Mathematics, Data Structures, and Digital Logic to Cybersecurity under NEP 2020 guidelines.
- Organized numerous workshops and FDPs, such as: (i) Online FDPs on LaTeX and Moodle LMS (ii) Seminars on Blockchain Technology, Green Computing, and Machine Intelligence. Delivered hands-on sessions on IoT, LaTeX, and Raspberry Pi.
- Published science communication articles in Dream 2047 on topics like ChatGPT and historical Indian mathematicians.

Assistant Professor *July 2012 – Dec 2020*

Department of Computer Applications, Sikkim University, India

- Conducted research in IoT and edge computing, leading to innovative publications and patents filing.
- Organized and led workshops and conferences on Blockchain, and IoT technologies.
- Contributed to committees like ICT Policy, Admission Working Committee, and NAAC Criterion Search.
- Facilitated job and training opportunities through active participation in the Training and Placement Committee.
- Served as Organizing Chair for the 6th International Conference on Mathematics and Computing (ICMC) and several TPCs in international-level conferences.
- Contributed to holistic student development through initiatives like Innovation and Entrepreneurship Awareness Workshops and Student Clubs and facilitated student participation in technology-driven activities and competitions.

Assistant Professor *Jan 2012 – June 2012*

Department of Computer Science and Engineering, Surendra Institute of Engineering and Management, Siliguri, India

- Taught undergraduate courses in C Programming and Embedded Systems.
- Developed curriculum and assessment methods for computer science courses.

Metrics (Dynamic)

- **Scholar**: H-index 39, i10-index 68, Citations 9883
- **Journals**: SCI Journals 108, SSCI Journals 1, Scopus Journals 21
- **Conference Proceedings**: 36
- **Book Chapters**: 5
- **Books**: 7
- **Patents Filed**: 7
- **Projects/Dissertations Guided**: 17

Selected Publications

For a complete list of publications, please visit my Google Scholar profile.

Journal Articles

1. **Ray, P. P.**, "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations, and Future Scope," *Internet of Things and Cyber-Physical Systems*, Elsevier, 2023.
2. **Ray, P. P.**, "Benchmarking, Ethical Alignment, and Evaluation Framework for Conversational AI: Advancing Responsible Development of ChatGPT," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2023.
3. **Ray, P. P.**, "A Review on TinyML: State-of-the-art and Prospects," *Journal of King Saud University - Computer and Information Sciences*, Elsevier, 2022.

Projects and Consultancy

- **IoT-Based Plant Temperature Monitoring Systems**, Funded by Sikkim University, 2024–25.
- **Intel IoT Center Setup**, Received Intel Galileo kits from Intel India, 2015–16.

Certifications

- **Certified Blockchain Expert-V2**, Blockchain Council, 2018.
- **Blockchain and Bitcoin Fundamentals**, Udemy, 2018.
- **Mastering Digital Twins**, Coursera, 2019.

Awards & Achievements

- **Best Paper**, LLMIoT: A Framework for Integration of IoT Devices for Localized Large Language Models in the AICTA 2024 at NIT Raipur, 2024.
- **Fellow**, The Institution of Electronics & Telecommunication Engineers (IETE), 2023.
- **World's Top 2% Scientist**, Stanford University, 2020–2024.
- **IEI Young Engineers Award**, The Institution of Engineers (India), 2019–20.
- **Emerald Literati Award**, Highly Commended Paper, 2019.
- **Senior Member**, IEEE, since 2019.
- **Young Scientist Award**, Venus International Foundation, 2017.
- **Bharat Vikas Award**, Institute of Self Reliance, 2017.
- **Best Professor in IT Academic Excellence Award**, ICBM-AMP, 2017.
- **Young Achiever Award**, IEAE, 2018.
- **Academia Liaison Officer** of IEEE CTSoc IoT Technical Committee (2024).

Professional Memberships

- **Fellow**, IETE
- **Senior Member**, IEEE
- **Member**, ACM, CSI, IEI, IETE, IET, ISCA, IACSIT

Additional Information

- **Languages:** Bengali, Hindi, English
- **Hobbies:** Writing Articles and Books, Open-Source Contribution
- **Founder:** Founded Indian Knowledge Forum (IKF) for Dissemination of Ancient and True Indian Knowledge

References and full publication list available upon request.