

Aditeya Baral

+1(551)263-8608 | aditeyabaral@nyu.edu | [linkedin.com/in/aditeyabaral](https://www.linkedin.com/in/aditeyabaral) | aditeyabaral.com | [Google Scholar](https://scholar.google.com/citations?user=aditeyabaral)

EDUCATION

New York University, Courant Institute of Mathematical Sciences

New York City, USA

Masters in Computer Science; GPA - 3.78/4.00

Sep 2024 – Present (Expected May 2026)

Concentration: Artificial Intelligence

PES University

Bengaluru, India

Bachelor of Technology in Computer Science & Engineering; GPA - 8.71/10.00

Aug 2018 – May 2022

Specialization: Machine Intelligence & Data Science

RESEARCH EXPERIENCE

Redis

San Francisco, USA

Applied Research Scientist Intern, Redis LangCache; Advisor: Srijith Rajamohan

June 2025 – Dec 2025

- Improved **semantic retrieval** in **Redis LangCache** by building novel **cross-encoder architectures** with **late-interaction attention mechanisms**, yielding a **24% F_1** and **18% precision improvement** over baselines.
- Curated **LangCache-SentencePairs-v1**, a large-scale dataset for supervised fine-tuning of sentence embedding models.
- Fine-tuned and open-sourced LangCache Embed v3** and **LangCache Reranker v1**, two generalist models for semantic retrieval and re-ranking, achieving up to **28% recall increase** and **improving cache-hit quality**.
- Quantified **retriever coverage bottlenecks** and **aggressive vs. conservative reranking effectiveness** by analyzing recall ceilings and reranking movement to **optimize operational trade-offs** and **cache-hit precision**.
- Developed a **comprehensive evaluation framework** for LangCache customers, enabling systematic analysis of achievable cache-hit rates, precision, and recall prior to onboarding.
- Supported **downstream integration** and development of **LMCache** by **building prototypes** and **conducting performance studies** with Redis as an **in-memory KV store**, demonstrating latency and throughput gains.

Computational Intelligence, Vision, and Robotics (CILVR) Lab

New York City, USA

Research Assistant; Advisors: Shauli Ravfogel, Tal Linzen

May 2025 – Present

- Investigating the evolution and generalization of **arithmetic heuristic circuits** in LLMs for **mathematical reasoning** using **circuit analysis**.
- Examining whether **targeted interventions** in LLMs at inference can **alter reasoning strategies** in LLMs.

Computation and Psycholinguistics Lab

New York City, USA

Research Assistant; Advisors: Jackson Petty, Tal Linzen

May 2025 – Present

- Evaluating LLMs on **compositional generalization** and **instruction synthesis** by studying their ability to translate synthetic **Context-Free Grammars (CFGs)** into conforming strings.
- Analyzing model outputs in **few- and zero-shot settings** to assess **grammatical conformity** and uncover **generation strategies** used during translation.

Cisco Systems

Bengaluru, India

Applied AI Engineer, Webex Media Quality Analytics

July 2022 – July 2024

- Instruction fine-tuned** LLMs like Mistral and Llama-2 on-prem to enable **secure** and **cost-effective** AI solutions such as **translation** and **RAG** for engineers and customers, **cutting 3rd party dependency costs by 30%**.
- Led the initiative to build a novel **pre-training algorithm** for conversational data using **PyTorch** and **HuggingFace**, achieving a **40% performance gain** over standard approaches at benchmark fine-tuning tasks.
- Developed the **Webex Contextual Search** engine and **improved searching, ranking, recommendations** and **topic modelling** by **75%** with **<10%** increased overhead latency.
- Integrated OpenAI APIs and **on-prem LLMs** with the **Webex AI Assistant** for **15M+** worldwide users to add **auto-replies, summarisation, querying** and **action-item extraction** to message threads and meeting transcripts.

Intel Corporation

Bengaluru, India

Applied Research Scientist, Intel VSG; Advisors: Anay Majee, Anbumani Subramanian

Aug 2021 – Dec 2021

- Explored **Few-Shot Learning Object Detection (FSOD)** techniques to reduce **catastrophic forgetting** in constrained and heterogenous driving environments.
- Investigated and designed novel **representation learning** and **attention mechanisms** to learn **inter/intra-object relationships** using **PyTorch**.
- Outperformed existing approaches at the time on base and novel classes by **0.2 mAP** and **3 mAP** on the **Few-Shot India Driving Dataset**, a benchmark for FSOD.

Center for Cloud Computing and Big Data

Bengaluru, India

Research Assistant; Advisor: KV Subramaniam

May 2020 – July 2020

- Compiled and used **TailBench** to **simulate and profile** application loads, monitor performance, and analyse results.
- Explored ways to **reduce tail latencies** in latency-critical applications such as translation and image recognition.

SOFTWARE ENGINEERING EXPERIENCE

Cisco Systems

Bengaluru, India

Big Data Engineer, Webex Media Quality Analytics

July 2022 – July 2024

- Developed and deployed streaming jobs in **Scala** and **Flink** to process **1M+ reports/min** and compute **1200+ real-time metrics** from Calls and Meetings.
- Applied *statistical modelling* techniques to investigate and report *media quality insights* to downstream consumers, *reducing errors by 30% and analysis time by 15 hrs/week* per team member.
- Led the development of *real-time (<1 min) auditing pipelines* using **Kafka** and **Python** to ensure *per-minute data consistency* between streaming jobs and **Iceberg** and **Pinot** data stores, *reducing manual effort by >80%*.
- Built graphs and dashboards on the **Webex Media Quality Analytics Dashboard** using **Grafana** and **Kibana** to set up alerts and KPIs for **20,000+** clients and customers.

Big Data Engineering Intern, Webex VideoMesh Analytics

Jan 2022 – June 2022

- Migrated the **Meetings Analytics Engine** from Java and Spark to **Scala** and **Flink** to scale up to **1M+ reports/min** and significantly *improve real-time report generation* by over **40%**.
- Built **VideoMesh Developer APIs** using **Java** and globally rolled them out for **30,000+** enterprises with *customer-facing applications*.

PREPRINTS AND PROJECTS

- [1] **Can LLMs understand Math? – Exploring the Pitfalls in Mathematical Reasoning**
Authors: Tiasa Singha Roy, Aditeya Baral*, Ayush Rajesh Jhaveri, Yusuf Baig*
- [2] **CMLFormer: A Dual Decoder Transformer with Switching Point Learning for Code-Mixed Language Modeling**
Authors: Aditeya Baral, Allen George Ajith, Roshan Nayak, Mrityunjay Abhijeet Bhanja
- [3] **Patch and Control: Steering Behavior of Large Vision-Language Models via Latent Activations**
Authors: Aditeya Baral, Rijul Dahiya, Dilip Venkatesh

PAPERS AND PUBLICATIONS

- [1] **ChatBERT - Multi-task approach to Pre-Training for Structured Conversations**
Webex AI 2023
Authors: Aditeya Baral (Work done as part of Cisco Webex AI Research)
- [2] **CalBERT - Code-mixed Adaptive Language Representations using BERT**
AAAI-MAKE 2022
Authors: Aditeya Baral, Aronya Baksy, Ansh Sarkar, Deeksha D, Ashwini M Joshi
- [3] **Information Maximization to Overcome Catastrophic Forgetting in Few-Shot Object Detection**
Intel VSG Research 2021
Authors: Aditeya Baral, Anay Majee, Anbumani Subramanian
- [4] **MAPLE - MASKing words to generate blackout Poetry using Seq2Seq LEarning**
ACL-ICNLSP 2021
Authors: Aditeya Baral, Himanshu Jain, Deeksha D, Mamatha H R
- [5] **Analysis of Kepler Objects of Interest using ML for Exoplanet Identification**
IEEE CONIT 2021
Authors: Ameya Rajendra Bhamare, Aditeya Baral, Saarthak Agarwal

TEACHING EXPERIENCE

Teaching Assistant, CS322: Big Data

Bengaluru, India

PES University; Professors: Dr. KV Subramaniam, Dr. Prafullata K Auradkar, Animesh Giri

June 2021 – Dec 2021

Designed and graded coursework, assignments and projects, and delivered hands-on sessions on Hadoop and Spark for a class of **600+** enrolled students for the undergraduate Big Data course.

AWARDS AND SCHOLARSHIPS

Webex Analytics Datathon, Cisco Systems	2 nd / 20+ teams	2024
Webex IDEA Hackathon, Cisco Systems	1 st / 300+ teams	2023
Webex Playtime Hackathon, Cisco Systems	Top 20 Globally & Regional Winner / 300+ teams	2023
Undergraduate Researcher Award, PES University	One among 900+ students	2022
Prof. CNR Rao, MRD & DAC Scholarship Awards	Top 20% among 900+ students	2022
Finalist, Intel Technovation, Flipkart, IBM, IISc Hackathons	One among 200+ teams	2022
National newspaper coverage, ToI	Remodelled garbage collection tracking in BLR	2017

SKILLS

- Languages: Python, Scala, Java, C, R, Groovy, Octave, SQL, L^AT_EX
- ML/Stats: PyTorch, Tensorflow, HuggingFace, NLTK, pandas, NumPy, scikit-learn, seaborn, matplotlib, plotly
- Artificial Intelligence Techniques: Representation Learning, Mechanistic Interpretability, Transfer Learning, Language Models
- Big Data/Cloud: Hadoop, Kafka, Zookeeper, Spark, Flink, Iceberg, Pinot, Redis, ELK
- Frameworks/Tools: Git, GitHub, Jenkins, Docker, Kubernetes, Flask, Grafana, PSQL, MongoDB, AWS, Linux

SERVICES AND VOLUNTEERING EXPERIENCE

- Speaker, Guest Lecture on - Building Foundation Models using Transformers
PES University
Bengaluru, India
Sep 2023
Delivered a guest lecture to undergraduate students on the advancements in representation learning techniques for language and highlighted the importance of interdisciplinary research.