Aditeya Baral

+1(551)263-8608 | aditeyabaral@nyu.edu | linkedin.com/in/aditeyabaral | aditeyabaral.com | Google Scholar

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

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.

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*** 3^{rd} ***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.

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***.

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.

Skills

Languages: Python, Scala, Java, C, R, Groovy, Octave, SQL, LaTeX
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

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