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
Master of Science in Computer Science; GPA – 3.87/4.00 — *Sep 2024 – 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

JPMorganChase — Plano, USA
Associate Machine Learning Engineer, Wholesale Credit Risk and Quant Research — *Aug 2026 – Present*

Redis — San Francisco, USA
Applied Research Scientist Intern, Redis LangCache; Advisor: Srijith Rajamohan — *Jun 2025 – Dec 2025*

- Introduced **Precision–Cache Hit Ratio (P-CHR)** and **Operational Retention Rate (ORR)**, two *cache-aware metrics* for **Redis LangCache**, reframing semantic-cache model selection as a *thresholding problem* rather than a ranking one.
- Curated and open-sourced **LangCache SentencePairs (v1-v3)**, a large-scale dataset family with **1M to 40M** sentence pairs across diverse paraphrase, STS, QA, and adversarial sources for robust semantic-caching fine-tuning.
- Fine-tuned and deployed **LangCache-Embed-v3**, a domain-specific bi-encoder trained with an *ArcFace contrastive objective*, leading all open-source retrievers offline and outperforming larger general-purpose models even without re-ranking.
- Open-sourced the **LangCache ReRanker v1/v2** cross-encoder families with *BCE* and *MNRL* objectives at two training scales, isolating the effect of objective vs. scale on threshold behavior and application-specific tradeoffs.
- Built a comprehensive *evaluation framework* for customer onboarding with a large-scale study of customer data and model baselines across offline and deployment settings, using a two-stage **RedisVL** K-NN retrieval and re-ranking pipeline.
- Demonstrated that the highest-**PR-AUC** models are often the *worst* in deployment, and that re-ranking rarely improves a strong domain retriever, overturning the standard practice in semantic-cache model selection.
- Supported *downstream integration* and development of **LMCache** by *building prototypes* and *benchmarking performance* 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 – May 2026*

- Investigated *arithmetic circuit dynamics* in language models when operators are redefined in-context by analyzing *activation representations* and *attention patterns* across transformer layers.
- Trained tiny (∼3M params) **LLaMA**-style transformers on arithmetic operations under *operator overloading* to test whether transformers reuse circuits across semantically related tasks.
- Developed an *activation-extraction and causal-attribution pipeline* pairing neuron selection with **Activation Patching**, **Direct Logit Attribution**, and **Fourier probes** to separate causally necessary components from merely predictive ones.
- Identified and characterized a *two-tier circuit structure* across a model's depth and breadth, showing transformers organize computation around *semantic operations rather than surface symbols*.

Computation and Psycholinguistics Lab — New York City, USA
Research Assistant; Advisors: Jackson Petty, Tal Linzen — *May 2025 – Aug 2025*

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

Cisco Systems — Bengaluru, India
Big Data and Applied AI Engineer, Webex Media Quality Analytics — *Jul 2022 – Jul 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 third-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 modeling* 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*, *summarization*, *querying* and *action-item extraction* to message threads and meeting transcripts.

- 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 modeling* 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 – Jun 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 Intern, 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 heterogeneous 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 – Jul 2020*

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

Preprints and Projects

[1] **When '+' Means '−': Probing Arithmetic Circuits Under Symbol Redefinition**
Authors: Aditeya Baral, Allen George Ajith, Shauli Ravfogel

[2] **Can LLMs *understand* Math? Exploring the Pitfalls in Mathematical Reasoning**
Authors: Tiasa Singha Roy, Aditeya Baral*, Ayush Rajesh Jhaveri, Yusuf Baig*

[3] **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

[4] **Patch and Control – Steering Behavior of Large Vision-Language Models via Latent Activations**
Authors: Aditeya Baral, Rijul Dahiya, Dilip Venkatesh

Papers and Publications

[1] **Closing the Operational Gap in Semantic Caching**
EMNLP 2026, Industry Track
Authors: Aditeya Baral, Radoslav Ralev*, Iliya Sotirov Zhechev, Srijith Rajamohan, Jen Agarwal*

[2] **Training for Compositional Sensitivity Reduces Dense Retrieval Generalization**
ICLR 2026, Sci4DL
Authors: Radoslav Ralev, Aditeya Baral, Iliya Sotirov Zhechev, Jen Agarwal, Srijith Rajamohan

[3] **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)

[4] **CalBERT – Code-mixed Adaptive Language Representations using BERT**
AAAI 2022, MAKE
Authors: Aditeya Baral, Ansh Sarkar, Aronya Baksy, Deeksha D, Ashwini M Joshi

[5] **Information Maximization to Overcome Catastrophic Forgetting in Few-Shot Object Detection**
Intel VSG 2021
Authors: Aditeya Baral, Anay Majee, Anbumani Subramanian

[6] **MAPLE – MAsking words to generate blackout Poetry using Seq2Seq LEarning**
ACL 2021, ICNLSP
Authors: Aditeya Baral, Himanshu Jain, Deeksha D, Mamatha H R

[7] **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; Faculty: KV Subramaniam, Prafullata K Auradkar, Animesh Giri — *Jun 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	Remodeled garbage collection tracking in BLR	*2017*

Skills

Languages: Python, Scala, Java, C/C++, Groovy, SQL, LaTeX
ML/Stats Libraries: PyTorch, TensorFlow, HuggingFace, WandB, vLLM, FAISS, pandas, NumPy, scikit-learn, matplotlib
AI/ML Techniques: Representation Learning, Mechanistic Interpretability, Transfer Learning, Language Models, RAG
Big Data/Cloud: Hadoop, Kafka, Zookeeper, Spark, Flink, Iceberg, Pinot, Redis, ELK
Frameworks/Tools: Git, GitHub, Jenkins, Docker, Kubernetes, FastAPI, Grafana, PSQL, MongoDB, AWS, Linux

Services and Volunteering Experience

Research Advisor, Summer Internship – Center for Cloud Computing and Big Data — Bengaluru, India
PES University — *Jun 2026 – Aug 2026*
Led the *mechanistic interpretability group*, advising teams on distinct projects that train small transformers to study *grokking*, activation patching, and *circuit-level learning and inference dynamics*.

Speaker, Guest Lecture – Building Foundation Models using Transformers — Bengaluru, India
PES University — *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.