Pranav Ajit Nair

Pre-Doctoral Researcher, Google DeepMind

@ pranavajitnair@google.com   🌐 Homepage   Github   🎓 Google Scholar

Education

Indian Institute of Technology, (BHU) Varanasi 07/2018 - 05/2023
Integrated B.Tech. (Honors) + M.Tech. in Computer Science and Engineering - CGPA: **9.64** / 10

Research Experience

Google DeepMind, *India* 07/2023 - present
Pre-Doctoral Researcher | *Advisors: Dr. Praneeth Netrapalli, Dr. Arun Suggala, Dr. Prateek Jain*

IIT (BHU), Varanasi, *India* 07/2022 - 04/2023
Research Assistant (Master's Thesis) | *Advisor: Prof. Sukomal Pal*

University of Hamburg, *Germany* 05/2021 - 04/2023
Research Intern | *Advisors: Prof. Chris Biemann, Prof. Ricardo Usbeck*

Conference Publications

[1] **CDQuant: Accurate Post-training Weight Quantization of Large Pre-trained Models using Greedy Coordinate Descent** [🔗]
Pranav Ajit Nair, Arun Sai Suggala
Under review at the International Conference on Learning Representations, 2025 [**MLC@NeurIPS'2024**]

[2] **Tandem Transformers for Inference Efficient LLMs** [🔗]
Aishwarya P S, **Pranav Ajit Nair**, Yashas Samaga, Toby Boyd, Sanjiv Kumar, Prateek Jain*, Praneeth Netrapalli*
International Conference on Machine Learning [**ICML'24**]

[3] **Domain Aligned Prefix Averaging for Domain Generalization in Abstractive Summarization** [🔗]
Pranav Ajit Nair, Sukomal Pal, Pradeepika Verma
Findings of the Association for Computational Linguistics [**ACL'23**]

[4] **The Role of Output Vocabulary in T2T LMs for SPARQL Semantic Parsing** [🔗]
Debayan Banerjee*, **Pranav Ajit Nair***, Ricardo Usbeck, Chris Biemann
Findings of the Association for Computational Linguistics [**ACL'23**]

[5] **GETT-QA: Graph Embedding Based T2T Transformer for Knowledge Graph Question Answering** [🔗]
Debayan Banerjee, **Pranav Ajit Nair**, Ricardo Usbeck, Chris Biemann
Extended Semantic Web Conference [**ESWC'2023**]

[6] **Modern Baselines for SPARQL Semantic Parsing** [🔗]
Debayan Banerjee, **Pranav Ajit Nair***, Jivat Neet Kaur*, Ricardo Usbeck, Chris Biemann
ACM SIGIR Conference on Research and Development in Information Retrieval [**SIGIR'22**]

Research Projects

Long-Context Attention
Advisors: Dr. Praneeth Netrapalli, Dr. Arun Suggala, Dr. Prateek Jain

- Developed clustering and approximate logit computation methods to identify the *top-K* keys a query needs to attend to.
- Wrote custom kernels in Pallas to optimize gathers and scatters on TPUs.
- Showed **4**× latency improvements for both prefill processing and per step decode-time on **Gemini Flash** models for long-context attention without any drop in quality. Currently in the pipeline for productionization.

Improving Post Training Quantization
Advisor: Dr. Arun Suggala

- Developed a greedy coordinate descent algorithm to improve post training quantization of LLMs.
- Extended the algorithm to greedily descent over blocks of coordinates, scaling factors, and zero points.
- Improved over GPTQ, (which does not use a greedy strategy, instead cycles over all the coordinates), especially for 2-bits.
- Enhanced several SOTA methods that use GPTQ as a sub-routine. Accepted to MLC@NeurIPS'2024 and under review at ICLR'2025.

Improving Speculative Decoding

Advisors: Dr. Praneeth Netrapalli, Dr. Prateek Jain

- Developed an online distillation strategy to improve drafter quality for speculative decoding.
- For all but the current block of tokens, the drafter attends to projected down representations of the primary model.
- In addition, proposed a routing mechanism to dynamically decide when to fall back to the primary model for verification.
- Obtained **1.36**× speedup over a distilled drafter. Accepted to ICML'2024.

Fast and Efficient Domain Generalization for Abstractive Summarization

Advisor: Prof. Sukomal Pal

- Developed a prefix-merging algorithm to efficiently adapt to previously unseen genres of summarization.
- Trained soft prompts (i.e prefixes) while keeping the backbone model frozen for a given set of genres.
- Generated prefixes for a previously unseen, test time genre by taking a weighted average of the training time prefixes.
- These weights were obtained by measuring the performance of the training time prefixes on a very small set of examples from the test time genre. Improved over several baselines. Accepted to Findings of ACL'2023.

Improving the Output Vocabulary for SPARQL Semantic Parsing

Advisors: Prof. Chris Biemann, Prof. Ricardo Usbeck

- Analyzed the effect of the output vocabulary for SPARQL generation with language models.
- Found that if the SPARQL vocabulary is replaced with a vocabulary more attuned to the LM tokenizer and the pretraining data, the performance on semantic parsing can be significantly improved. Accepted to the Findings of ACL'2023.

Improving Question Answering over Knowledge Graphs

Advisors: Prof. Chris Biemann, Prof. Ricardo Usbeck

- Developed an end-to-end pipeline for question answering over knowledge graphs.
- Trained T5 to generate a skeleton SPARQL query where relation IDs were replaced with their textual labels, and entity IDs were replaced with a concatenation of their textual labels and the first few dimensions of their TransE embeddings.
- Grounded the relations with a BERT reranker, and the entities with ElasticSearch and the generated TransE embeddings.
- Showed improvements over several baselines for questions answering over knowledge graphs. Accepted to ESWC'2023.

Benchmarking Language Models for SPARQL Generation

Advisors: Prof. Chris Biemann, Prof. Ricardo Usbeck

- Found that most SPARQL generation and KGQA methods did not employ language models in their pipelines and relied on RNN-based and traditional semantic parsing-based methods.
- We were among the first ones to benchmark language models such as T5 ans BART on SPARQL generation.
- Showed significant improvements over all existing methods. Accepted to SIGIR'2022.

Selected Honors and Awards

- **Recipient of the DAAD WISE Scholarship:** Received the prestigious DAAD WISE scholarship for a fully funded research internship at the University of Hamburg.
- **Inter IIT Tech Meet:** Awarded silver medal in Bridgei2i's Automatic Headline and Sentiment Generator event at the 9th Inter IIT Tech Meet.

Notable Positions of Responsibility

- **Volunteer** at ICML 2021 (online) and ICLR 2022 (online).
- **Reviewer** at ICLR 2024.
- **TA** for the **Compiler Design** and **Computer Networks** courses at IIT (BHU), Varanasi.
- **General Secretary** of the Social Service Council at IIT (BHU), Varanasi during the academic session 2021-2022. Worked towards improving the living standards and educational awareness among children and adults living in slums.

Key Courses Undertaken

Machine Learning	Artificial Intelligence, Natural Language Processing, Computer Vision, Artificial Intelligence and its Application in Biomedical Engineering
Computer Science	Data Structures and Algorithms, Computer Architecture, Databases, Computer System Organization, Compilers, Computer Networks
Mathematics	Probability and Statistics, Linear Algebra, Mathematical Modeling, Mathematical Methods, Discrete Mathematics, Theory of Computation, Number Theory, Limits and Differential Equations, Theory of Rings and Modules