Research Associate Statement

I am a final-year B.Tech student in AI and ML at NMIMS MPSTME, and I am applying to Precog as a Research Associate for a six-month internship in my final semester, from December 2026 to May 2027. I like the parts of machine learning that sit inside the model and close to the hardware: what an LLM computes, where inference time, memory, and energy go, and how to run a good model on modest infrastructure. My B.Tech thesis and my capstone project both come from that question.

With two teammates and under the guidance of Prof. Rajesh Verma, I am comparing the two 4-bit post-training quantization formats, FP4 and NF4, on energy use as well as accuracy. Our harness runs Mistral-7B in FP16, FP4, and NF4 on an NVIDIA RTX PRO 6000. It records MMLU accuracy, WikiText-2 perplexity, hidden-state similarity to the FP16 model, throughput, and GPU power read through NVML, idle-subtracted and scaled for PUE, over five trials in randomized order. The first results cut against the assumption that smaller means greener. 4-bit reduces peak GPU memory from 14.5 GB to about 4.6 GB and raises throughput from 71.1 to about 90 tokens per second, yet a full MMLU run uses about 5% more energy than FP16 (0.104 and 0.105 kWh against 0.0994). NF4 beats FP4 on accuracy (0.491 against 0.487), perplexity, and hidden-state similarity, and it uses slightly more energy. We have run only Mistral-7B so far, with Phi-3, Gemma 3, and Qwen 2.5 7B next, and we are writing the results up as a paper.

My capstone project follows from that result. I am writing my own matrix multiplication kernel in Python Triton, aimed at reducing GPU power draw and energy consumption. Matrix multiplication dominates LLM inference cost, so a saving there would carry into every model I measured in the thesis. I plan to test the kernel with the same NVML-based measurement, so its energy numbers can be compared directly with the thesis results.

That is why Precog fits. Your Vocabulary Compression project cuts memory use and raises throughput during pretraining, a saving like the ones I am measuring after training. The Small Language Models work on on-device function calling puts inference efficiency next to agentic behavior. Representation Surgery would let me study LLM internals by editing them, and MACA, your work on distilling LLMs into efficient retrievers, is close to the retrieval systems I have built. I would also like to try RL-based optimization on small open models. Precog describes its aim as measuring AI systems, and I would apply that to efficiency: what a compression or kernel change does to speed, energy, and quality.

The role asks for deep learning skill, a base in data structures, databases, and programming, and the ability to prototype research ideas. I built a physics-based spatiotemporal ConvLSTM that forecasts PM2.5 over India, with a WindWarp module that simulates pollutant transport through differentiable grid sampling. It reached 89.25% accuracy and ranked first in the pollution forecasting theme at ANRF AISEHack 2026. I am building Koala, a vector search library, from scratch, writing the indexing and nearest-neighbour algorithms myself. I work in Python, C++, MySQL, and SQL Server, and at my data analysis internship at Bhilwara Hulchul (May to August 2026) I used SQL and Python to clean and analyse app-usage data. My prototypes include a fully offline multimodal RAG system built on FAISS, Whisper, and CLIP, which reached 73.87% on more than 2,000 documents and 500 audio clips, and FloatChat, a natural-language interface for Argo oceanographic data that cut manual report time by a factor of four.

Apart from this thesis, my research experience is limited to projects and a hackathon. Over six months I would like to take one well-scoped problem, such as reducing the energy cost of running a compressed small open model, prototype it, and measure it carefully enough that the result could go into a paper. I am available for the full semester, on site in Hyderabad.

Kartik Soni