

Pratham Shah

Ahmedabad, Gujarat, India

[✉ prathamsshah724@gmail.com](mailto:prathamsshah724@gmail.com) · [Phone](#)
[GitHub](#) | [LinkedIn](#) | [Portfolio](#) | [Medium](#) | [LeetCode](#)

About

I like building things that feel fast and considered: lean web products that range from static sites to AI agents, made with real care so the code won't embarrass me later. The rest of my time goes to papers I don't strictly need to read, mostly because I can't help it.

Experience

LLM Intern, Blink Analytics – Ahmedabad, Gujarat

May 2025 to July 2026

- Built a phone-call AI interview agent: which can serve as a reminder agent and screening agent.
- Built bulk candidate processing: recruiters can upload and process large sets of prospects instead of handling them one by one.
- Built payments and interview credits: companies buy plans/credits and those credits control how many AI interviews they can run.
- Shipped ATS and recruiter tools: resume scoring, candidate profiles, job open/close, search/filters, candidate emails, and interview reports (including PDF).
- Improved how quickly interview reports are ready after a session, and cleaned up backend Docker deploys.
- Improved the existing live interview room: better speech recognition, AI agent model routing, noise cancellation, session reliability, and network monitoring.
- Reviewed and helped merge teammate pull requests across company repos.
- Optimized language models using Reinforcement Learning from Human Feedback (RLHF) and Supervised Fine-Tuning (SFT).
- Integrated optimized LLM pipelines into internal recruitment tools with cross-functional teams.

62 merged PRs · 470+ commits across company repos

Projects

Pikd

[\[GitHub\]](#) [\[Live\]](#)

Event photo delivery with face recognition: guests scan once and collect every photo they appear in, in seconds. Next.js + Python backend: each face is detected, embedded, and indexed; a guest selfie is matched via cosine similarity against stored vectors for instant retrieval.

Recursive Reasoning (Small Language Model)

[\[GitHub\]](#) [\[Demo\]](#)

Recursive Transformer SLM for Sudoku that re-processes its own output in shared-weight thinking loops, based on “Less is More: Recursive Reasoning with Tiny Networks,” with a Streamlit dashboard for loop-by-loop visualization. PyTorch model re-processes latent states iteratively rather than adding depth; curriculum training scales 4×4 to 9×9 puzzles.

Reflect Reads

[\[GitHub\]](#)

Full-stack book browsing and ordering: React/Vite + Tailwind frontend, Express REST API, MongoDB, and JWT auth. bcrypt-hashed credentials and middleware guard catalog, cart, and checkout endpoints; order documents persist line items and status through a stateful checkout flow.

Research

Violent HAR Detection

[\[GitHub\]](#)

Fusion model for human activity recognition: YOLO-Pose + LSTM for temporal pose features and RGB frames + MLP for spatial features to detect violent actions. YOLO-Pose extracts 17-keypoint skeleton sequences into an LSTM; a parallel RGB branch passes CNN spatial features through an MLP; late fusion of both embeddings classifies violent actions.

Hazard Surveillance and Maintenance Car

IoT and computer vision system on an ESP-CAM-based robotic car to detect five classes of electrical hazards for maintenance and surveillance. Live frames stream through a lightweight classifier (exposed wiring, damaged insulation, etc.); detected hazards trigger on-device alerts and drive autonomous patrol.

AI Based Prediction for Panic Disorder Detection Using Social and Clinical Factors

[\[IEEE\]](#)

Smart Detection of Passive Smokers using Machine Learning

[\[IEEE\]](#)

Open Source & Writing

Run LLM judge when `-evaluate` is passed – Tracer-Cloud / opensre. Wired the `-evaluate` delivery path to call `run_opensre_llm_judge()` when a rubric is present, instead of skipping evaluation silently. Added regression tests for happy path, skip guards, and error fallback. Closes #2070. [\[PR #2090\]](#)

Attending to all attentions (a try!) (Part 1) – Medium. Part 1 of a 3-part attention series: Bahdanau and Luong attention, scaled dot-product self-attention, multi-head and cross-attention, and $O(n^2)$ quadratic bottleneck—with intuition, math, and PyTorch snippets. [\[Article\]](#)

Education

ME Computer Science, BITS Pilani, Pilani Campus

2026 to 2028

In Progress

B.Tech Computer Engineering, Pandit Deendayal Energy University (PDEU)

2022 to 2026

CGPA 9.562

HBK New High School, Ahmedabad

2020 to 2022

HSC: 82.46%

Technical Skills

Languages Python (Blink, Pikd, Recursive Reasoning), JavaScript/TypeScript (Reflect Reads, Pikd), Java, C, SQL

ML / AI PyTorch, TensorFlow, Hugging Face, LangChain, Transformers, RLHF, SFT, RAG, NLP, CNNs, LSTMs, YOLO-Pose, Face Embeddings

Web / Backend Next.js, React, Vite, Tailwind, Node.js, Express, MongoDB, REST, JWT, FastAPI, Streamlit

Tools / Cloud Docker, Git, GitHub, AWS, Azure, Linux, CI/CD