

Suraj Prasad

[LinkedIn](#) | [GitHub](#) | surajprasad8977@gmail.com

Research Interest

I build the infrastructure around LLM post-training: reward functions grounded in execution, rigorous baselines and ablations, and pre-registered go/no-go experiments. I have taken a 7B model through full-parameter RL end-to-end, from data validation to a competitive benchmark result. Across text-to-SQL RL and Hindi-English code-switched ASR fine-tuning, the hard problem turned out to be data and evaluation, not architecture. I also ship production systems end-to-end, from agentic pipelines ranked #1 on independent benchmarks to on-device speech inference at sub-130ms latency.

Research & Shipped Products

Text-to-SQL RL on BIRD-bench | [Link](#) | Full-Parameter RL Pipeline

(Jun'26 – Present)

Built a full-parameter RL pipeline for text-to-SQL, reproduced published 7B SOTA within half a point, and killed two research bets on pre-registered evidence rather than intuition.

- Implemented GRPO plus a custom **MGPO** (max-entropy difficulty-weighting) advantage estimator from a paper, registered into **verl**; trained **OmniSQL-7B** full-parameter on **4×H100** via Modal, cloud-only, no local GPU
- Engineered **FSDP** full-shard training with parameter/optimizer CPU offload to fit full-parameter 7B RL on one node (**41.9GB/53.6GB** per H100, **~233GB** offloaded), colocating the actor, reference model, and vLLM rollout
- Pre-registered a go/no-go gate for a multi-turn RL bet (needed $\geq 5\text{pp}$ lift); measured **-1.04pp**, audited the harness with an oracle upper bound to confirm the negative was real, and killed the direction
- Traced a 3.7pp reproduction gap to its sources (dataset-version + value-linking differences), reproducing OmniSQL's baseline within **0.5pt**; the RL model reached **68.12% maj@8**, $\sim 2\text{pts}$ over the correctly-prompted base
- Built a timeout-safe SQL correctness oracle (multiplicity-aware equivalence checking for join fan-out, NULLs, float/int) and validated all **1,534** dev + full train golds before training on any of it

Srota: Hinglish ASR | [GitHub](#) | [Hugging Face](#) | Open-Source Code-Switched ASR

(May'26 – Present)

Fine-tuned and published an ASR model for Hindi-English code-switched speech from Qwen3-ASR-0.6B; diagnosed and fixed a catastrophic negative-transfer failure through deliberate training-data composition.

- Ran a controlled **3-variant ablation** (HiACC-only, OpenSLR-only, union): the OpenSLR-only specialist regressed **+12.91pp** on conversational speech; the union model reversed this to **-8.88pp** better, a **21.79pp swing**
- Constructed a union training corpus from two heterogeneous Hinglish datasets (5.24h conversational, 89.86h tutorial) via a reproducible, **speaker-disjoint** pipeline (download, JSONL prep, train, evaluate, publish)
- Built **automated WER evaluation** across two held-out test sets (1,036 + 3,132 utterances) plus cohort-level bias checks (adult vs. children speakers, 1.25pp gap, no bias introduced by training)
- Steered model output to stay in natural mixed Devanagari+Latin script (rather than collapsing to monoscript like base) via the training format, achieving a **36% relative WER reduction**
- Trained full-parameter (no LoRA) on **2×H100** via Modal; published 3 models + a dataset to Hugging Face

Srota | [Link](#) | Local-First macOS Voice Dictation

(2026 – Present)

Independently designed, built, and shipped a privacy-first native macOS voice dictation app, with 100+ active users, running entirely on-device at sub-130ms end-to-end latency.

- Architected and shipped a full native macOS app in **Swift, SwiftUI, and AppKit** with global hotkeys, system-wide paste, and floating controls, built solo from zero to production
- Optimized on-device **ASR inference** on Apple Silicon: benchmarked multiple local speech models, reduced latency via buffered audio capture, speculative decoding, and **Metal/CoreML** inference-path optimization
- Built a personalization layer that tracks user vocabulary and correction patterns as a continuous **data feedback loop**, improving per-user transcription accuracy without sending audio to the cloud

Markdrop | [Link](#) | PyPI Python Package

(Dec'24 – Present)

Co-built and co-maintain an open-source PDF-to-Markdown/HTML tool with 210+ stars and 25,000+ downloads.

- Co-engineered a **PDF-to-Markdown** pipeline using Table Transformers, Docling, and XRef IDs for structuring data
- Developed LLM-powered in-place descriptions for non-textual content using 5+ LLM providers

Tests whether vision-language models can follow a webtoon's story across pages; independent, unpublished.

- Designed a 7-task evaluation suite (story generation, panel ordering, VQA, retrieval) over 87 pages
- Found Claude Sonnet 4.6 scores **71.77/100** on narrative quality, 43 points above the closest prior reported result (28.47), while every model tested lands at or near chance on temporal panel ordering

Startup & Industry Experience

AI Engineer, Actioneer (formerly GameRamp) | [Link](#) | Bengaluru

(Dec'25 – Present)

Built SCRIBE, an agent architecture ranked #1 on an independent data-analysis benchmark.

- Built **SCRIBE**, a 3-stage agent (Spec → Executor → Planner) for data-analysis tasks across 4 database systems (PostgreSQL, SQLite, DuckDB, MongoDB), with runtime schema discovery instead of pre-built indexes
- Ranked **#1 on DataAgentBench (81.85% Stratified Pass@1)**, verified via a merged upstream PR
- Separately, built a scalable evaluation pipeline on **BigQuery** processing billions of rows for a deep research agent, with multi-agent orchestration to automate large-scale analysis workflows

LLM Engineer, BharatGen | [Link](#)

(May'25 – Jul'25)

Built retrieval infrastructure for BharatGen's multilingual Indic-LLM initiative.

- Designed a graph-driven indexing system (**LightRAG**) combining entities and relationships for structured retrieval, built to support BharatGen's multilingual document workflows
- Built an agentic **RAG** pipeline with multimodal retrieval and self-evaluation to reduce hallucinations; integrated text and image modalities into a unified storage and retrieval layer

Entrepreneur in Residence, Entrepreneurs First | Bengaluru

(May'25 – Sep'25)

Selected into EF's founder cohort (under 3% acceptance) to build and validate ambitious products.

- Explored and validated startup ideas across agentic AI document parsing and multimodal workflows through rapid prototyping and structured customer discovery, operating as a high-agency solo founder

Publications

SynthPID: P&ID Digitization from Topology-Preserving Synthetic Data | [Link](#) | CVPR Workshop 2026

-
- Introduced SynthPID, a corpus of **665 synthetic P&IDs** seeded from real engineering drawings, and a technique for **steering synthetic data generation** to preserve graph topology
 - Built an automated evaluation pipeline (edge mAP on PIDGraph OPEN100) to **validate synthetic-data fidelity** against a real-data oracle; achieved **63.8±3.1% edge mAP** with **zero real training data**, within 8pp of the oracle

Speech-Synchronized Whiteboard Generation via VLM-Driven Structured Drawing Representations | [Link](#) | CVPR Workshop 2026

-
- Built a dataset of 24 paired Excalidraw demonstrations with narrated audio, each drawing element annotated with **millisecond-precision timestamps** for speech-visual alignment
 - Fine-tuned Qwen2-VL-7B with LoRA to predict stroke sequences aligned to speech; topic-stratified five-fold evaluation showed timestamp conditioning improves temporal alignment over ablated baselines

Federated Cross-Modal Style-Aware Prompt Generation | [Link](#) | AAAI-SAPP 2026

-
- Designed a multi-scale feature extraction and style encoding pipeline with an Injection Block (channel-wise attention) fusing visual and text embeddings, plus a Context Redundancy Penalizing loss
 - Developed a lightweight cross-attention prompt generator that reduces communication overhead in the federated setting; achieved SOTA on 9 benchmarks (+4.3% unseen classes, +1.8% unseen datasets)

Competitions & Hackathons

Real-Time Audio to Indian Sign Language (ISL) Conversion | [Link](#) | SIH 2024

(Aug'24 – Dec'24)

*Bidirectional real-time speech-to-sign and sign-to-speech system; **1st rank nationwide**.*

- Built a real-time audio pipeline (Speech-to-Text → NMT → 3D avatar gesture rendering) and a reverse Sign-to-Text path (PoseNet + LSTM), robust across speakers and signing styles under strict latency constraints

4th rank worldwide; mAP 0.30 vs. prior benchmark 0.25 on nuScenes (1.4M images, 390k LiDAR sweeps).

- Fused camera and LiDAR data in a shared bird’s-eye-view space for robust 3D detection under sensor failure

Research Leadership

- Co-founded an 8-researcher AI lab at IIT Bombay focused on vision-language and multimodal AI; produced multiple peer-reviewed publications (CVPR, AAAI) within one year of founding

Technical Skills

RL / Post-Training	verl, GRPO, custom advantage estimators (MGPO), reward engineering FSDP full-shard + CPU offload, vLLM
Data / Evaluation	WER evaluation (jiwer), dataset curation & ablation design Hugging Face Hub, BigQuery, cohort/bias analysis
Speech / On-Device AI	Whisper, faster-whisper, Qwen3-ASR fine-tuning, CoreML Metal, speculative decoding, streaming inference
ML / LLM	PyTorch, TensorFlow, LoRA fine-tuning, RAG, Agentic AI, VLMs, OpenCV
Programming	Python, Swift, C/C++, SQL, JavaScript, MATLAB, HTML/CSS, LaTeX
Infra / Systems	Modal, Docker, Git, Linux, CI/CD, SwiftUI, AppKit

Education

B.Tech in Mechanical Engineering (CGPA: **8.58/10**)
Minor in Computer Science and Engineering (CGPA: **10/10**)
Relevant Coursework: Data Structures & Algorithms, Machine Learning, Image Processing, Computer Networks, Databases, Linear Algebra, Probability

Scholastic Achievements

- Awarded the Undergraduate Research Award (URA01) from IIT Bombay ('25)
- Attained **99.57** percentile in **JEE-Mains**, outperforming over **1 million** registered candidates ('22)