

Puneet Bhatt

India | puneetbhatt2000@gmail.com | puneetbhatt.com | +919027282482 | [Github](https://github.com/puneetbhatt) | [Linkedin](https://www.linkedin.com/in/puneetbhatt)

EDUCATION

B.Tech in Information Technology, HNB Garhwal University Aug 2018 – Aug 2022 | Uttarakhand, India

WORK EXPERIENCE

Founding Engineer - AI, [Thine.com](https://thine.com) Jun 2025 – Oct 2025 | Bangalore, India

- Led parametric memory fine-tuning R&D for LLMs to improve recall of customer-specific knowledge.
- Built a pipeline that turned ~25k-token documents into knowledge graphs and 40x-100x synthetic training data for LoRA/SFT.
- Ran 150+ experiments across model (230M to 100B), data, and training setups; improved recall from 5% to 81% on an internal held-out benchmark.

Founding AI Engineer, [Merlin AI](https://merlin.ai) Aug 2022 – Jun 2025 | Bengaluru, India

- #1 hire & founding engineer, built a [chrome extension](#) on a weekend (got 10k+ downloads in a week) leading to the company pivot; scaled it to 20m+ users & 100M+ chats to date.
- Built SOTA (Jun '25; public was 76%, ours 88%) memory system on [LongMemEval](#) based on knowledge graphs & GraphRAG.
- Shipped eval-backed low latency web search (5s to <1s), document RAG (100k+ docs), MCP & chat pipeline used by millions of active users per month.

Software Development Engineer Intern, [PropReturns \(YC 21\)](#) Oct 2021 – Dec 2021 | India

- Integrated a geolocation search showing properties based on location & optimized landing pages for performance.

Full Stack Engineering Intern, [Delego \(SNCA India\)](#) Oct 2020 – Jan 2021 | India

- Built core cloud compliance & corporate action workflows end-to-end, a web-based filesystem, email integrations.

PROJECTS

[Superplan](#), open-source ralph loop alternative

- It's an agent harness available as a plugin for claude code, codex, opencode & anti-gravity.
- It gives the agent a kanban board to work with & internally builds task graphs that let agents execute one by one.

[speechless.sh](#), low latency live in-ear translation app

- Developed a low-latency voice AI app with a server-side speech pipeline for fast turn-taking and responsive conversations.
- Integrated MossFormer, VAD and open-source speech models to improve endpointing and voice quality.

[Puzzled.run](#), multiplayer speedrun puzzle game

- Endless multiplayer puzzle speedrun game where LLM generates personalized puzzles (10 game types) based on your theme, with feedback tracking and self-improvement based on user feedback.

Intwiff, Bike Rental App

- Founded and built the [client](#) app for users to take bikes on rent & a [business](#) app for rental shops to manage their shop & list their bikes for rent.
- Executed all operations with my team of 2 co-founders.
- Contributed by designing & building the front-end for all of our products & making all product-related decisions.

[Tejas Workflow Engine](#), No-code workflow builder for building forms & UI programmatically

SKILLS

- React, Typescript, Django, LLM fine-tuning (LoRA and full), memory systems, RAG pipelines, evaluation and benchmarking, prompt engineering.

ACHIEVEMENTS

- Discovered a privilege escalation vulnerability in the local university's system which exposed 100k+ students' data.