SUMIT VERMA

Senior AI Engineer — LLM Systems, Retrieval & Distributed Backend

Kanpur, India (Remote) · +91 8808879249 · imsumi8@gmail.com · linkedin.com/in/imsumi8 · github.com/imsumi8 · imsumi8.tech

SUMMARY

Senior AI Engineer with seven years in software, building production LLM and retrieval systems for a multi-tenant B2B recruitment platform since 2024. Work centres on the parts that decide whether an AI feature survives real users: multi-agent sourcing pipelines, semantic matching over pgvector and a Neo4j knowledge graph, learning-to-rank on human relevance judgments, and generation paths that are fact-checked before anything reaches a customer. Owns services end to end — Python/FastAPI and TypeScript/NestJS backends, Kafka event flows, distributed systems design, and the Kubernetes platform underneath.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, SQL, Cypher

LLM & Agents: RAG pipelines, LangChain, LangGraph, multi-agent orchestration, prompt engineering, structured output, tool/function calling, LLM-as-judge, groundedness evaluation, hallucination guardrails, LiteLLM gateway, Langfuse tracing, OpenAI GPT-4o

ML & Retrieval: pgvector (PostgreSQL vector database), embedding pipelines, semantic and hybrid search, cross-encoder reranking, learning-to-rank (LightGBM LambdaMART), nDCG evaluation, score calibration, PyTorch, Hugging Face Transformers, Sentence-Transformers, ONNX Runtime

Backend & Data: FastAPI, NestJS, Fastify, Node.js, Express, React, REST APIs, GraphQL, WebSockets, Kafka, BullMQ, microservices, event-driven architecture, distributed systems, PostgreSQL, Neo4j, Redis, MongoDB, TypeORM, SQLAlchemy

Platform & Delivery: Docker, Kubernetes, Helm, Argo CD (GitOps), Istio, Kong API Gateway, HashiCorp Vault, GitLab CI/CD, Prometheus, OpenTelemetry, Grafana, multi-tenant SaaS, scalability and system design, code review and mentoring

Computer Vision & Speech: Whisper / Wyoming ASR, VAD, InsightFace (ArcFace), Pyannote diarization, SpeechBrain (ECAPA-TDNN), OpenCV, FFmpeg, liveness detection

PROFESSIONAL EXPERIENCE

Senior AI Engineer Jan 2024 - Present

Techlances Consultancy Services (client: Stafio.ai, B2B SaaS) — Noida, India

Architect and lead developer across two AI product lines on a multi-tenant recruitment platform: an agentic candidate sourcing and outreach system, and an autonomous AI interview platform. Lead a team of three.

Agentic Sourcing, Retrieval & Talent Outreach

- Architected the Virtual Talent Pool sourcing service — a NestJS/Fastify orchestrator directing a fleet of independently deployed Python workers for scraping, NLP extraction, embedding and ranking, coordinated over Kafka and a Redis-backed priority queue, with PostgreSQL and Neo4j behind it.
- Designed the Neo4j knowledge graph, modelling employment as first-class Position nodes (Person → Position → Company) so tenure and career-trajectory questions are answerable directly; co-worker edges are derived by a multi-hop traversal with a temporal overlap predicate, and composite plus full-text indexes keep graph reads inside a sub-100 ms budget.
- Built candidate–job semantic matching on PostgreSQL with pgvector: L2-normalised section embeddings, per-section multi-chunk scoring that takes the best-matching chunk, cross-encoder reranking, and score calibration to spread the narrow raw-cosine band; batched every lookup into a single SQL round-trip.
- Shipped a learning-to-rank layer — LightGBM LambdaMART trained on graded recruiter relevance judgments, evaluated with nDCG@5/@10, with a versioned feature schema, an explicit position-bias feature, and a deterministic linear fallback so ranking degrades instead of failing on cold start.
- Shipped conversational sourcing: a recruiter describes who they want in chat, an LLM asks clarifying questions until a search is launchable, then results refine over judged rounds from thumbs and reason-chip feedback — while routing between the internal pool and external discovery stays invisible to the customer.
- Extracted outreach into a standalone intent-driven service. A declarative registry maps event topic, module and context to an intent with no LLM anywhere in the routing path, which keeps routing testable and auditable; the model's only job is writing the message body.

- Added a grounding judge that checks every generated message against the candidate evidence at temperature 0 on its own model key, fails closed on a timeout or unparseable verdict, and returns the specific unsupported claims so the writer gets one bounded regeneration. Per-intent deterministic templates are the floor when it rejects.
- Made eligibility and suppression per-intent, so the rule that kills cold sourcing outreach no longer blocks an interview reminder, with consent and opt-out as hard blocks no intent can bypass.

Autonomous AI Interview Platform

- Architected a five-microservice platform that runs an interview end to end: the candidate joins the room, an avatar interviewer streams in, speech is transcribed live, an LLM generates adaptive follow-ups, and the avatar answers back.
- Built the real-time transcription service on FastAPI and Wyoming ASR over WebSocket with word-level timestamps and multi-language support, holding sub-second latency so the conversation does not stall.
- Designed the video analysis pipeline, fusing identity matching, blink-based liveness, gaze and posture tracking, emotion recognition, ECAPA-TDNN voice verification and Pyannote diarization into one structured, human-reviewed candidate report.
- Handled the failure modes that decide whether it feels real: LLM timeout fallbacks, predictive pre-loading of the next two or three questions, bias and legal compliance validation on generated questions, and browser-automation crash recovery with session restore from PostgreSQL.

Platform & Delivery

- Deployed on Kubernetes with per-service Helm charts and Argo CD GitOps, promoted through GitLab CI/CD across four environments including a separate UAE production region, fronted by Kong and Istio with secrets in HashiCorp Vault.
- Instrumented with Prometheus, OpenTelemetry, Grafana and Langfuse LLM tracing — including an alert on template-fallback share, which is the signal that the anti-hallucination gate has started rejecting everything.
- Sized model-serving workloads deliberately — memory requests set to hold the embedding model and its working set resident, weights baked into the image to remove cold-start downloads, and probe delays tuned to model warm-up so pods are not marked ready before they are.

Full Stack Developer Jan 2022 - Dec 2023

Routa Digital India — Noida, India

- Built and shipped client-facing product features in React and Node.js, owning them from API design through to the interface on top.
- Worked across the stack on delivery teams — REST services, component architecture, state management and release support.

Senior Full Stack Developer Nov 2020 - Jan 2022

Kenzobhai Consultancy Services Pvt Ltd — Kanpur, India

- Built single-page applications on Vue.js and Laravel with responsive UI and real-time features.
- Shipped native and hybrid mobile applications in Java and Dart.
- Led an agile team, owning sprint planning, code review, delivery timelines and the team's CI/CD workflow.

Software Engineer Jan 2019 - Feb 2020

Beedesigns IT Solution — Kanpur, India

- Built full-stack web applications on Laravel MVC, Livewire, MySQL and a responsive Bootstrap frontend.
- Developed an assessment platform handling real-time data processing across browsers.

EDUCATION

B.E., Computer Science 2016

Dr. A.P.J. Abdul Kalam Technical University