Shreyash Kumar Pandey

+91-8319717414 | shreyash.pandey.katni@gmail.com | shreyash-pandey-katni | shreyash-Pandey-Katni | shreyash.co.in

Experience

Software Development Engineer 2 Jul 2024 – Present
IBM Software Labs *Bengaluru, India*

- Designed AI-powered auto-healing: 3-tier ML cascade (CSS selectors → text embeddings → IBM Granite 3.3 VLM) autonomously self-corrects 5K+ test case locators; 100% Firefox, 83% Chrome accuracy
- Re-architected SAT from monolith to 4 microservices (4–5K req/min, sub-second latency); built intelligent 3-stage executor: direct CSS, embeddings-based search (˜0.85 similarity), VLM fallback
- Led Java 8→17 migration (+30% perf), CI/CD pipelines (Jenkins/Docker/NGINX, +30% build speed, 99% uptime)

Software Engineer 2 Nov 2023 – Jul 2024
Software AG (now IBM) *Bengaluru, India*

- Built Semantic Search Engine using NLP, Knowledge Graphs, and FAISS – won TechInterrupt Hackathon (1st in India, 4th internationally) for intelligent search across enterprise documentation
- Developed AI Chatbot using LangChain and Flask, reducing internal support tickets by 70%
- Created Failure Prediction system using PyTorch and time-series analysis, achieving 99.99% system availability

Software Engineer Aug 2022 – Nov 2023
Software AG *Bengaluru, India*

- Developed enterprise integration platform features (webMethods) using Java, Spring Boot, and RESTful APIs; recognized as 2023 Star Performer

Projects

Phoenix 125M | *Decoder-Only LLM* | HuggingFace 2026

- LLaMA-style 125M decoder-only model trained from scratch on single RTX 3080 Ti; custom tokenizer, pipeline, loop; ˜2B tokens (Wiki, C4, Pile, Indian corpora)
- WinoGrande 0.507 (matches GPT-Neo 125M & OPT-125M trained on 150–300x more data); HellaSwag 0.279, ARC-E 0.358; Apache 2.0 open-source; 150x data efficiency
- **Skills used:** Python, PyTorch, Transformers, LLM Pretraining, Tokenization, Distributed Training, Evaluation Benchmarking, Data Pipelines

Sweta-Hi & Sweta-Kn | *Multilingual Language Models* 2026

- Pretraining 125M-param models for Hindi and Kannada using LLaMA-style architecture; custom tokenizers trained on Sangraha, Samanantar, corpora; end-to-end ML pipeline with async DataLoader and distributed training – near release
- **Skills used:** Python, PyTorch, Multilingual NLP, LLM Pretraining, Custom Tokenizers, Data Engineering, Model Evaluation

LinkedIn Post Swarm | *Agentic Content Automation Pipeline* 2026

- Built a multi-agent pipeline using Claude + Ollama + Playwright + Telegram for end-to-end draft generation, quality review, human approval, and scheduled publishing
- Implemented critic-revision loops, source aggregation, state management, and resilient retry/escalation flows for reliable autonomous operation with human-in-the-loop control
- **Skills used:** Python, Agent Orchestration, Prompt Engineering, Playwright, Telegram Bot API, Automation Workflows, State Management, Reliability Engineering

Rudra | *Autonomous AI Security Orchestrator* 2026

- Designed a multi-agent offensive security architecture with scope guardrails, sandboxed exploit execution, structured event pipelines, and strict layer boundaries
- Implemented security-first controls: input scope validation, network isolation strategy, exploit retry budgets, typed validation, and auditable workflow design for safe autonomous testing
- **Skills used:** Python, AI Security, Red Teaming Workflows, Sandbox Design, Network Isolation, Distributed Systems, Event-Driven Architecture, API Integration

LocalLeads | *AI-Powered Lead Generation and Delivery System* 2026

- Developed an end-to-end backend workflow: Playwright-based business discovery, AI content generation, automated site assembly/deployment, and personalized outreach pipeline
- Integrated operational controls with SQLite state tracking, Telegram approval gates, Netlify deployment automation, and WhatsApp delivery for production-style reliability
- **Skills used:** Python, Backend Development, Playwright, SQLite, Netlify Automation, API Orchestration, Workflow Engineering, Production Operations

Project Scoring Reference (Variant C)

Scoring Dimensions (1–10): AI Alignment, AI Security Depth, Backend Engineering Depth, Recruiter Hook, SEO Relevance.
Primary Weighted Score: $0.30 \times AI + 0.20 \times Security + 0.20 \times Backend + 0.20 \times Hook + 0.10 \times SEO$.
Intent: Reusable master reference for role-specific tailoring (AI-heavy, security-heavy, or backend-heavy variants).

Project	AI	Sec	Backend	Hook	SEO	Weighted
Phoenix 125M	10	4	7	10	9	8.1
Sweta-Hi/Sweta-Kn	9	3	7	8	8	7.1
LinkedIn Post Swarm	8	6	8	9	8	7.8
Rudra	8	10	8	8	9	8.5
LocalLeads	7	5	9	8	8	7.3
AlgoBridge	5	2	8	6	6	5.3

Interpretation Guide:

- **AI-first submissions:** prioritize Phoenix 125M, Sweta-Hi/Sweta-Kn, LinkedIn Post Swarm.
- **AI Security submissions:** prioritize Rudra + Phoenix 125M + LinkedIn Post Swarm.
- **Backend-heavy AI submissions:** prioritize LocalLeads + LinkedIn Post Swarm + Phoenix 125M.
- **Slot strategy:** keep 3 anchors fixed (Phoenix, Sweta, LinkedIn) and switch slot 4 between Rudra and LocalLeads by JD.

Alternative Weight Profiles for fast retargeting:

- **Balanced (default):** 30/20/20/20/10 for AI/Security/Backend/Hook/SEO.
- **Security-first:** 20/30/20/20/10.
- **Backend-first:** 20/20/30/20/10.

Surgical Tailoring Metadata To Maintain Per Project (for focused applications):

- **Project role fit tags:** AI Research, AI Product Engineering, AI Security, Backend Platform, Automation Systems.
- **Proof assets:** GitHub link, demo link, architecture diagram, benchmark/report artifact, and one reproducible command.
- **Ownership clarity:** what was individually built vs team-built, and decision ownership (architecture, implementation, validation).
- **Scale/context fields:** dataset size, request volume, latency targets, retry/error rates, and deployment/runtime environment.
- **Security/compliance fields:** guardrails used, threat model assumptions, sandbox/scope controls, and auditability notes.
- **ATS keyword packs by target role:** maintain separate keyword bundles for SDE2/3 AI, AI Security, and Backend AI roles.
- **Interview hooks:** one advanced system-design discussion point and one failure/recovery story per project.
- **Customization knobs:** preferred 2-bullet and 3-bullet versions per project for short and long resume variants.

Recommended Additional Data To Add For Better Resume Generation:

- **Phoenix/Sweta:** latest checkpoint status, training stability notes, and any new benchmark deltas vs baseline models.
- **LinkedIn Post Swarm:** throughput (posts/cycle), regeneration rate, and publication reliability metrics.
- **Rudra:** concrete validation status of scope guardrails and sandbox controls, plus tested threat scenarios.
- **LocalLeads:** pipeline conversion funnel metrics (discover →approve →deliver), and deploy success rates.
- **Cross-project:** date-stamped changelog of major milestones to keep resume claims current and defensible.

Technical Skills

Languages: Python, Java, Go, SQL, JavaScript
AI/ML: PyTorch, Transformers, LLM Pretraining, Deep Learning, NLP, Computer Vision, Text Embeddings, Knowledge Graphs, LangChain, HuggingFace, FAISS, CNNs, Keras
Backend: Microservices, REST APIs, Spring Boot, Apache Kafka, gRPC, Flask, Node.js, Express.js
DevOps & Tools: Docker, Jenkins, CI/CD, NGINX, Git, Linux
Databases: MySQL, MongoDB, Cassandra, Redis

Education

Bangalore Institute of Technology 2018 – 2022
Bachelor of Engineering, Computer Science *CGPA: 7.2*

Certifications

Neural Networks and Deep Learning – deeplearning.ai | Improving Deep Neural Networks – deeplearning.ai | Introduction to subagents by Antropic Apache Kafka – IBM | Enterprise Design Thinking – IBM