

Bharat Singh

AI Systems Architect · AI Engineering Lead

bsbhadwal@gmail.com · +91-9797561011 · Himachal Pradesh, India · github.com/bsbhadwal · bsbhadwal.com

PROFESSIONAL SUMMARY

AI Systems Architect and Engineering Lead with 20+ years of experience directing high-throughput Information Retrieval (IR) architectures, autonomous agent platforms, distributed systems, and knowledge graphs across European institutions and enterprise SaaS. Proven track record leading engineering squads (8–9 developers + QA), executing high-risk core engine modernizations with ~99% production parity, and engineering multi-tenant RAG and search pipelines scaling across 3,500+ commercial websites under sub-50ms p95 latency SLAs.

TECHNICAL EXPERTISE

Agentic AI & Orchestration: LangGraph, Model Context Protocol (FastMCP / SSE), Stateful Multi-Agent Loops, Tool Execution & Function Calling, Supervisor-Worker Pipelines, Deterministic Guardrails.

Information Retrieval & RAG: Multi-Tenant Search Platforms, Hybrid Search (Dense Vector + BM25), Cross-Encoder Neural Rerankers, Reciprocal Rank Fusion (RRF), Local Embedders (BGE, FastEmbed, ONNX Runtime), Quantization (int8, FP16), Vector Stores (MongoDB Atlas Search, Qdrant, pgvector).

Systems & Platform Architecture: Hexagonal Architecture, Protocol-Based Interfaces, High-Throughput Rendering Engines, Distributed Workflow Orchestration (Prefect 3.x), Microservices (FastAPI, gRPC, REST), Parity Verification (~99% Production Parity).

Infrastructure & Operations: Kubernetes (AKS), Docker Containerization, Azure DevOps CI/CD Pipelines, Azure Blob Storage, Linux System Internals, LangSmith Tracing, Telemetry.

Languages & Databases: Python (FastAPI, Pydantic, Prefect, PyTorch), Java, Rust, SQL, MongoDB, Neo4j, GraphDB, RDF/SPARQL.

PROFESSIONAL EXPERIENCE

AI Application Tech Lead (Full-Time) · Milestone Internet

Sep 2025 – Present · Ahmedabad / Remote (India)

- **Core Platform Modernization:** Led the architectural re-engineering of the enterprise content serving engine to asynchronous Python on Kubernetes; achieved ~99% golden rendering parity across 20+ production pods and 3,500+ customer domains, delivering ~\$70,000/year in projected infrastructure cost savings.
- **Autonomous Multi-Agent Platform:** Directed cross-functional engineering squad (8–9 developers + QA) building enterprise multi-agent workflows; designed stateful cyclic graph execution and decoupled tool contracts (FastMCP over SSE/REST), eliminating conversational state loss across complex multi-turn booking and support journeys.
- **In-House Neural Inference & Hybrid Search:** Built and deployed dedicated local inference microservices for neural embeddings and cross-encoder rerankers, replacing third-party API dependencies and achieving sub-50ms p95 retrieval latency SLAs under peak traffic.
- **AI Governance & Automated Telemetry:** Instituted enterprise-grade compliance and observability platform incorporating automated PII scrubbing (HIPAA/GDPR compliance) and unsupervised semantic drift clustering over user query streams to identify failure modes proactively.

Enterprise AI Consultant (Contract) · Milestone Internet

Jun 2025 – Sep 2025 · Remote (India)

- **Multi-Tenant Ingestion & Search Platform:** Architected the foundational multi-tenant document ingestion and vector search platform powering 3,500+ commercial websites; designed resilient DOM scraping, semantic chunking boundaries, and vector indexing pipelines to ingest heterogeneous web catalogs with high availability.
- Engineered configurable discovery, parsing, chunking, and indexing stages, establishing the core retrieval infrastructure for customer experience applications.

AI Specialist Consultant (Contract) · European Patent Office (EPO)2023 – Jan 2025 · *The Hague / Remote (India)*

- Architected multilingual information extraction and semantic search pipelines indexing millions of international patent records.
- Applied Word Sense Disambiguation (WSD) and concept indexing to improve prior-art detection recall and claim classification accuracy.
- Engineered NLP summarization and clustering tools for internal patent examiner tooling.

Knowledge Graph Consultant (Contract) · Knewco Inc.2021 – 2023 · *Rockville, MD / Remote (India)*

- Developed enterprise knowledge graphs using unsupervised relation extraction and semantic inference engines.
- Formulated ontology alignment and concept resolution strategies across large heterogeneous scientific datasets.

AI & Data Modeling Lead (Contract) · EU COVID Task Force2020 – 2021 · *Brussels / Remote (India)*

- Led cross-national health and mobility data schema modeling and data integration pipelines across EU member states.
- Developed interpretable epidemiological machine learning models and visualization pipelines for international policy reporting.

AI Consultant & Scientific Researcher (Contract) · Medical Informatics Research Consortia / NBIC2014 – 2020 · *Netherlands / Remote (India)*

- Advised on semantic knowledge representation, graph databases (RDF, SPARQL), and clinical text mining engines.
- Pioneered graph-based lexical semantic disambiguation models across biomedical ontologies (UMLS, MeSH).

Senior Software Engineer & Scientific Researcher · Erasmus Medical Centre, Rotterdam2010 – 2014 · *Netherlands / Remote (India)*

- Engineered full-stack semantic web architectures, graph stores, and natural language processing pipelines.
- Co-authored peer-reviewed research on clinical entity normalization and data value modeling.

Senior Software Engineer (Full-Time) · Leiden University Medical Centre / NBIC2008 – 2010 · *Leiden, Netherlands*

- Developed semantic web data layers and graph databases integrated with enterprise Java services.
- Built unsupervised text processing pipelines for clinical concept normalization.

Senior Software Engineer (Full-Time) · Quark Inc.2004 – 2008 · *Chandigarh, India*

- Engineered enterprise-scale publishing platforms and high-performance document processing engines in Java.

EDUCATION**Master of Computer Applications (MCA) · Punjab Technical University***India***PEER-REVIEWED PUBLICATIONS****The value of data:** *Nature Genetics*, 43(4), 281–283 (2011). doi:10.1038/ng0411-281**Contribution:** Contributed to semantic data modeling and nanopublication specification implementation for machine-readable assertion networks across scientific corpora.**Using rule-based natural language processing to improve disease normalization in biomedical text:** *JAMIA*, 20(5), 876–881 (2013). doi:10.1136/amiajnl-2012-001173**Contribution:** Designed and implemented rule-based NLP pipelines and dictionary lookup algorithms for normalizing clinical disease mentions against UMLS and MeSH ontologies.**Using an ensemble system to improve concept extraction from clinical records:** *Journal of Biomedical Informatics*, 45(3), 423–428 (2012). doi:10.1016/j.jbi.2012.01.001**Contribution:** Developed ensemble concept extraction pipelines combining machine-learning classifiers with rule-based heuristics to extract clinical entities from unstructured records.