ABHINAV TYAGI

Generative AI Engineer

+91 6398363038 | abhinav_cs_2020@rkgitm.ac.in | linkedin.com/in/abhinav-tyagi-416028244

SUMMARY

Conversational AI Engineer specializing in building LLM-powered chatbots, retrieval systems, and hybrid cloud inference pipelines. Experienced with Gemini API, HuggingFace, embeddings, quantized models, and NLP/NLU workflows. Proven ability to design production-ready conversational systems that reduce latency, cut inference cost, and improve dialogue accuracy.

EDUCATION

B.Tech – Computer Science & Engineering
Raj Kumar Goel Institute of Technology & Management (2020–2024)
CGPA: 6.5 / 10

Senior Secondary (Class XII) – CBSE
Nalanda Public School, Muzaffarnagar (2019–2020)
Percentage: 61%

Secondary (Class X) – CBSE
Nalanda Public School, Muzaffarnagar (2017–2018)
Percentage: 70.04%

Work Experience

Software Developer — TechVimal, Noida Feb 2025 – Present
Built and deployed the company website, improving structure and load performance.
Developed LMS + KPI system for field teams, centralizing performance tracking.
Created KPI Email Automation using Node.js + MongoDB, reducing manual reporting by 60%.
Implemented backend APIs, dashboards, and internal automation workflows.

Java Developer Intern — Webnmobapps, Noida Sep 2022 – Nov 2023
Worked on Java + Spring Boot modules with REST API development.
Enhanced backend logic and debugging for live client modules.
Strengthened fundamentals in OOP, MVC, and microservice patterns.

PROJECTS & RESEARCH

Smart Contextual RAG Chatbot Sep 2025 – Present
- Built a RAG chatbot using FAISS retrieval + Gemini/HF models.
- Added local GPT-2 Q8 fallback, cutting cloud API usage by ~40%.
- Designed light context caching + routing to improve consistency & latency.
- Keywords: RAG, FAISS, embeddings, inference routing.

OmniThinker — Multi-LLM Comparator (Chrome Extension) Nov 2025 – Present
- Automated prompt injection & response capture across ChatGPT, Gemini, DeepSeek using robust content-scripts.
- Cleaned results via noise-removal + dedupe, then analyzed them using Llama (OpenRouter) for unified summaries.
- Implemented parallel processing (Promise.all) → ~3× faster evaluation.
- Keywords: DOM automation, scraping, OpenRouter, LLM comparison.

Psywarp — Cognitive & Behavioral AI Research Aug 2024 – Present
- Proposed a framework modeling emotion drift, deception cues & cognitive behavior.
- Developed core modules: TraitNet, EDM, DDL.
- Research manuscript currently queued under arXiv hold + IJFMR review.
- Keywords: cognitive modeling, affective computing.

SKILLS

AI & NLP:
LLMs, RAG, Embeddings, NLP/NLU, LoRA/QLoRA, Prompt Engineering
Frameworks & Tools:
HuggingFace, LangChain, FAISS, Chroma, Gemini API, OpenRouter
Backend:
Node.js, Express.js, FastAPI, MongoDB, MySQL
Frontend:
HTML, CSS, JavaScript
Other:
Git/GitHub, Postman, Docker (basic), Chrome Extensions (MV3)
Cloud:
GCP (basic), AWS (basic), Render/Vercel deployments