HumbleBeeAI

E-mail: contact@humblebee.ai
Web: humblebee.ai
Careers: humblebee.ai/careers

UZBEKISTAN
«HUMBLE BEE AI» MChJ
Address: Namangan viloyati, Namangan shahri, Davlatobod tumani, Yuqori Girvon MFY, Afsonalar vodiysi ko'chasi, 6/9-uy

REPUBLIC OF KOREA
Bee Intelligence Global Co., Ltd.
Address: 6 Gyeongin-namgil, Michuhol-gu, Incheon, Korea

NLP / LLM Engineer

Mid-Senior level | Full-time | Onsite | Visa sponsorship available

Position	NLP / LLM Engineer
Seniority	Mid-Senior
Location	Onsite — Incheon (Korea), Tashkent or Namangan (Uzbekistan)
Employment type	Full-time
Reports to	Chief Technology Officer
Team	20–25 people company-wide; day-to-day work with 4–5 engineers
Compensation	Based on experience; agreed at offer stage
Visa sponsorship	Available for eligible candidates
Apply	humblebee.ai/careers · contact@humblebee.ai

About HumbleBeeAI

We specialize in Computer Vision, Generative AI, and Business Analytics. An AI studio with tools, talent, and purpose. We build from our own AI ecosystem to deliver scalable, ethical, and lasting solutions for clients worldwide.

Our offices are located in Incheon (Republic of Korea), and in Tashkent and Namangan (Republic of Uzbekistan). The group operates as «HUMBLE BEE AI» MChJ in Uzbekistan and as Bee Intelligence Global Co., Ltd. in the Republic of Korea.

About the Role

We are looking for an NLP/LLM Engineer to build and improve production-grade AI systems across RAG, agents, evaluation, fine-tuning, document processing, memory, speech, and observability.

You should be comfortable taking an LLM feature from experimentation and dataset preparation through evaluation, deployment, monitoring, and continuous improvement.

Team and Reporting

- You will report directly to the Chief Technology Officer.
- You will work day to day with a group of 4–5 engineers, owning the technical direction and delivery of the LLM workstream.
- The wider company team is 20–25 people across the Korea and Uzbekistan offices.

What You Will Do

- Design and improve production RAG pipelines, retrieval, chunking, embeddings, reranking, and context construction.

- Build multi-step LLM workflows and agents using frameworks such as LangChain, LangGraph, CrewAI, LlamaIndex, or similar.
- Develop query rewriting, intent classification, routing, structured extraction, tool calling, and context-sufficiency checks.
- Improve semantic, keyword, hybrid, filtered, and graph-based retrieval.
- Reduce hallucinations and improve groundedness, factual consistency, and answer quality.
- Build evaluation datasets and automated evaluation pipelines for retrieval and generation.
- Measure retrieval with Recall@K, Precision@K, MRR, and nDCG, and evaluate generation for correctness, relevance, groundedness, safety, latency, and cost.
- Fine-tune and adapt models using SFT, LoRA, QLoRA, PEFT, preference optimization, and distillation where useful.
- Work with OCR, PDF/DOCX processing, speech-to-text, conversational memory, and multilingual NLP.
- Trace and monitor model calls, retrieval, token usage, latency, failures, and cost.
- Run controlled experiments, regression tests, A/B tests, and human evaluations.
- Optimize systems for reliability, concurrency, streaming, caching, retries, and provider failures.

Requirements

Must Have

- 3+ years of experience in NLP, ML, Information Retrieval, or Applied AI.
- Strong Python engineering skills and experience building production services.
- Hands-on experience shipping LLM, RAG, or agent systems to production.
- Strong understanding of transformers, tokenization, embeddings, context windows, prompting, structured outputs, and decoding.
- Strong knowledge of semantic search, hybrid retrieval, metadata filtering, reranking, chunking, and indexing.
- Experience with LangChain, LangGraph, CrewAI, LlamaIndex, or comparable orchestration frameworks.
- Experience with vector or search systems such as Milvus, Pinecone, Qdrant, Weaviate, or Elasticsearch.
- Experience with OpenAI, Gemini, Hugging Face Transformers, and PyTorch.
- Practical knowledge of SFT, LoRA, QLoRA, and PEFT.
- Experience designing automated and human-in-the-loop LLM evaluations.
- Strong understanding of hallucination mitigation, prompt injection, PII and data leakage, and adversarial testing.
- Experience with FastAPI, REST APIs, Docker, Git, testing, and CI/CD.
- Experience with multilingual NLP systems and low-resource language challenges.
- Professional English proficiency.

Nice to Have

- Experience training embedding or reranking models.
- Knowledge of BM25, cross-encoders, bi-encoders, and Reciprocal Rank Fusion.
- Experience with DPO or other preference-alignment methods.
- Experience serving models with vLLM, TGI, Triton, or similar.
- Experience with quantization, batching, KV caching, and GPU inference optimization.

- Familiarity with RAGAS, DeepEval, promptfoo, LangSmith, Langfuse, or custom evaluation systems.
- Experience with OCR and STT evaluation using WER/CER.
- Experience with Graph RAG, knowledge graphs, or synthetic dataset generation.
- Research publications, open-source contributions, or relevant competition experience.
- Experience mentoring or technically guiding a small engineering team.

What We Offer

- End-to-end ownership of real LLM systems: RAG, agents, evaluation, fine-tuning, and deployment.
- Direct collaboration with the CTO, founders, and product leadership.
- High-performance compute, tooling, and infrastructure support.
- Compensation based on experience, visa sponsorship support, and a long-term growth trajectory.

Interview Process

Step 1. Document screening — review of your CV, portfolio, and application form.

Step 2. Online technical interview — Zoom session covering your NLP/LLM experience and system design.

Step 3. In-person interview — final conversation at one of our offices.

How to Apply

Please submit your application through our form, and include the following:

- Your CV or LinkedIn profile.
- Relevant GitHub repositories, projects, or publications.
- A short description of one LLM or NLP system you shipped to production, including how you evaluated and improved it.

Application form: Apply here

Questions: contact@humblebee.ai

HumbleBeeAI is an equal opportunity employer. We consider all qualified applicants without regard to age, gender, nationality, ethnicity, religion, disability, marital status, or any other characteristic protected by applicable law. Hiring decisions are based solely on qualifications, experience, and demonstrated ability to do the job.