Reyhan Bayraktar

GenAI Engineer — LLM Inference, Evaluation & Agents

Istanbul, Turkey | reyhanbyrktrr@gmail.com | byrayhana.github.io | linkedin.com/in/reyhan-bayraktar
github.com/byrayhana | huggingface.co/byrayhana | Google Scholar

Summary

GenAI Engineer focused on **serving, evaluating and building LLMs and LLM agents**. Self-hosted and benchmarked ∼20 open-weight model families (MoE up to 1T parameters) on H200, GB300 and DGX Spark, and pretrained a 1.7B diffusion language model on the MareNostrum 5 supercomputer. Designed and publicly released **mecellem-legal-bench**, an open benchmark with a rubric-based LLM-as-judge and a public leaderboard of 16 frontier models; built the trajectory-level evaluation and oversight system for a production agent platform, used to benchmark models, gate new MCP tool servers and monitor live agent behaviour. Co-author of five papers on Turkish IR/RAG, hallucination detection, guided decoding and parameter-efficient fine-tuning.

Experience

NewMind AI — **Istanbul, Turkey**
Research Engineer — *Jan 2025 – Present*

- **LLM inference & serving:** self-hosted and benchmarked ∼20 open-weight model families (Qwen3/3.5, Kimi K2, DeepSeek, GLM, Nemotron, Gemma) on **H200**, **GB300** and **DGX Spark** with **vLLM** and **SGLang**; FP8 / NVFP4 / AWQ quantization, tensor/expert parallelism and KV-cache tuning. Reached **1,850 tokens/s** at 100% success on 2×H200; Ray TP16 gave **+33–43%** throughput over multiprocessing for 235B–397B MoE serving. Deployed the company's complete agent platform on-premises on a single 128 GB GB10 workstation.
- **Speculative decoding:** compared EAGLE-3, MTP, DFlash and n-gram drafting across model sizes and hardware; showed a reported 4–5× speed-up to be a benchmarking artefact.
- **Open benchmark (led):** designed and released **mecellem-legal-bench** — 42 Turkish legal practice questions graded on 367 IRAC criteria by an LLM judge with ∼90% agreement with reference verdicts. Maintain its public leaderboard of 16 frontier models (∼2,000 graded attempts) with pass^3 reliability, citation-grounding checks and a release protocol that separates infrastructure failures from model errors. Published the dataset, scorer and leaderboard on Hugging Face and the methodology article on newmind.ai.
- **Agent trajectory evaluation & oversight (lead developer):** designed and built the system that oversees the company's production agent harness at the level of full trajectories — each question is run k times through the harness over MCP tools, every trajectory is traced end to end (Langfuse) and graded step by step: tool selection and call order, deterministic citation grounding, criterion-level LLM-as-judge on the final answer, pass^k, cost and latency; FastAPI, React, ClickHouse, NATS JetStream.
- **Release gating & live oversight:** extended the same trajectory analysis into the release gate for new MCP tool servers — controlled comparison of harness configurations on shared criteria and routing evaluation — and into live oversight of production trajectories: per-server health, before/after analysis around configuration changes, and automated LLM-generated remediation reports. Introduced a rendering-based benchmark for agent-generated web artifacts and a curation pipeline that turns expert-reviewed production conversations into new benchmark items.
- **Agent harness profiling (research):** developed the tracing proxy and call-level profiler used to compare eight agent harnesses under fixed models, a versioned run-profile schema (cost by call role, cache utilisation, tool calls) and its integration with the Harbor evaluation framework; developed an embedding-based runtime monitor for detecting agent misbehaviour.
- **Latency optimisation:** profiled a production six-model fan-out agent workflow from Langfuse traces, identified its critical-path models and replaced them with faster open-weight alternatives on fixed providers.
- **Pretraining on HPC:** pretrained a **1.7B-parameter Turkish diffusion language model** from scratch with multi-node distributed training on the Barcelona Supercomputing Center's **MareNostrum 5** (∼4,400 GPU-hours; SLURM, NCCL/InfiniBand), and built its evaluation harness.
- **Multi-agent RAG:** core engineer of a production multi-agent RAG system for legal and financial documents — orchestrator, up to 16 parallel citation-grounded sub-agents, claim/citation reviewer — with MCP tools and hybrid retrieval; added semantic caching and LLM routing.

Voiser — **Kocaeli, Turkey**
Machine Learning Engineer Intern — *May 2023 – Sep 2023*

- Optimized **YOLOv2-tiny** object detectors for mobile deployment through dataset curation and hyperparameter tuning; contributed to talking-avatar systems (First Order Motion Model).

Auvtech — **Istanbul, Turkey**
Image Processing Team Leader (university robotics team) — *Oct 2021 – Mar 2024*

- Led underwater image enhancement and object-detection pipelines for autonomous underwater vehicles; work published at ICCIDA 2024.

Open Source & Writing

2026: **mecellem-legal-bench: Do models apply Turkish law, or just cite it?** — methodology article (author), newmind.ai/research/mecellem-legal-bench

2026: **mecellem-legal-bench dataset, scorer and leaderboard** — huggingface.co/datasets/newmindai/mecellem-legal-bench

Publications

2026: **RDP LoRA: Geometry-Driven Identification for Parameter-Efficient Adaptation in Large Language Models** — arXiv:2604.19321.

2025: **TurkColBERT: A Benchmark of Dense and Late-Interaction Models for Turkish Information Retrieval** — ACLing 2025, Elsevier Procedia Computer Science.

2025: **Turk-LettuceDetect: Hallucination Detection in Turkish Retrieval-Augmented Generation Systems** — FLLM 2025 (IEEE), Vienna.

2025: **Guided Decoding and Its Critical Role in Retrieval-Augmented Generation** — IEEE SIU 2025; arXiv:2509.06631.

2024: **Automatic Mine Detection for Underwater and Surface Vehicles** — ICCIDA 2024, Springer Smart Innovation, Systems and Technologies, Vol. 387.

Education

Istanbul Sabahattin Zaim University — **Istanbul, Turkey**
B.Sc. in Computer Engineering — *Sep 2020 – Jun 2024*

Skills

Languages: Python, TypeScript, SQL, Bash

Inference: vLLM, SGLang, TensorRT-LLM, ONNX Runtime; FP8 / NVFP4 / AWQ quantization; speculative decoding; tensor/expert parallelism

LLM Evaluation: LLM-as-judge, rubric design, judge reliability, pass@k / pass^k, benchmark design, Langfuse, agent tracing

Agents & RAG: Multi-agent systems, Model Context Protocol (MCP), RAG, hybrid retrieval, embedding & reranking, ColBERT, LLM routing, semantic caching

Training: PyTorch, Transformers, Diffusers, Accelerate, Megatron; pretraining, SFT, LoRA, RL fine-tuning (GSPO)

Infrastructure: Linux, NVIDIA H100/H200/GB300/DGX Spark, HPC (SLURM, NCCL, InfiniBand), Docker, Kubernetes, FastAPI, ClickHouse, PostgreSQL, Qdrant

Languages

Turkish: Native

Arabic: Native

English: Professional working proficiency