

Temirlan Dzhoroev 테미를란 조로예프

AI 엔지니어 · Agent Orchestration & Multi-Agent Systems · RAG · Evals & Moderation

production agent system을 3년 이상 개발·운영해 온 AI 엔지니어입니다. checkpointed state와 memory를 갖춘 multi-tenant tool-calling agent, multi-model routing, 그리고 실제 배포 여부를 결정하는 eval·moderation 레이어까지 직접 설계했으며, 누적 사용자 **1,000만 명 이상**의 제품에 적용했습니다. hybrid RAG·reranking 기반 retrieval 경험, 실시간 3D 엔진 경험(게임 플레이 **5,400만 회 이상**), HCI 연구 이력(논문 **10편**, 특허 **6건**) 보유.

한국 거주 **12년** · 한국어 자유롭게 구사 · 국내 기업 근무 경험 · 비자 스폰서십 불필요

서울 · +82 10-7393-2412 · dzhoroev1@gmail.com · mastertim.life (포트폴리오) · linkedin.com/in/dzhoroev7 · github.com/master-tim

경력

AI 엔지니어 — Redbrick, 서울

2024.07 – 현재

- production agent orchestration 레이어 총괄 — **tenant 40개 이상, 일 30K 요청** 규모: LangGraph 기반 tool-calling agent에 checkpointed state와 agent별 memory를 적용해 retrieval·generation·sandboxed code execution에 걸쳐 요청당 최대 **8개의 의존적 tool call**을 조율. checkpoint recovery 도입으로 multi-step task 성공률 **82% → 96%** 개선.
- \$1.2M** 규모 정부 R&D 과제 수주 — 자연어 prompt에서 플레이 가능한 3D 게임까지 이어지는 end-to-end AI 게임 제작 시스템을 직접 구축(intent parsing, asset generation, scene composition, logic synthesis를 하나의 워크플로우로 orchestration).
- hallucination 비율 **18% → 4%**, 답변 relevance **+35%** 개선(버전 관리되는 eval set 기준) — 문서 **15K+** 규모 production RAG 구축(hybrid dense + BM25 retrieval, query rewriting, cross-encoder reranking).
- 답변당 비용 **48% 절감**(21K → 12K token), latency **3.2초 → 0.8초** 단축 — rerank-then-budget 방식 context assembly와 embedding similarity 기반 semantic cache로 중복 LLM 호출 **약 40%** 제거, 월 **약 \$2K** 절감.
- 모든 prompt·모델 변경을 검증하는 moderation·evaluation 스택 구축 및 운영: LangSmith eval dataset(**regression case 600건 이상**), **13개 카테고리·4단계** severity classifier(무응답 거부 대신 담당자에게 escalation), fail-closed provider check, **10ms 미만** 로컬 guardrail(injection, 금치어, PII 마스킹).

3D 프론트엔드 엔지니어 — Redbrick, 서울

2023.07 – 2024.07

- 브라우저 기반 3D 게임 엔진(Three.js, WebGL, TypeScript)을 고도화해 가입자 **1,000만 명 이상**, 게임 플레이 **5,400만 회 이상** 규모의 창작 플랫폼을 지원; 해당 엔진으로 출시한 게임들이 누적 **100만 회 이상** 플레이 달성.
- 5만 라인** 규모 코드베이스를 Webpack에서 Vite로 전환하고 렌더 파이프라인 재구축 및 물리 엔진 교체를 주도 — 빌드 시간 **60% 단축**, 번들 크기 **35% 감소**, hot reload **8초 → 1초 미만**.
- 엔진을 모듈 단위 패키지로 리팩터링해 **3개 팀**의 병렬 작업이 가능하도록 개선; scene graph와 asset loading에 대한 integration test 추가.
- 세션 리플레이 분석과 사용자 인터뷰를 바탕으로 온보딩 재설계 — 7일 활성화율 **+25%**, 30일 리텐션 **+15%**.

임베디드 시스템 & HCI 연구원 — DECS Lab, UNIST, 울산

2021.02 – 2023.02

- ARM Cortex-M MCU용 실시간 IoT 펌웨어를 C로 개발해 **50대 이상**의 네트워크 센서 디바이스를 sub-millisecond latency로 제어; 논문 **6편**의 데이터 파이프라인 기반.
- 학회 논문 **6편**, 저널 논문 **4편**, 공동 특허 **6건**. IEEE RO-MAN 2023, ICROS 2022, HCI Korea 2022 제1저자.

주요 프로젝트

직접 설계·배포한 production 시스템. 아키텍처·트레이드오프·평가를 포함한 전체 케이스 스터디는 mastertim.life/work 참고.

Enterprise Agent Platform — Multi-tenant RAG: **5개** parser fallback 체인, Milvus에서 dense + BM25 융합, cross-encoder rerank, 쿼리 내 tenant별 ACL 적용. 답변당 token **48% 절감**, context **51K → 25K**자. *LangGraph · Milvus · LiteLLM*

Companion Chat Runtime — 2단계 memory(Postgres buffer + pgvector recall)를 갖춘 companion agent, 단일 alias로 **17개** 모델을 제공하는 model-routing gateway, semantic image cache 적용. 출시 첫 달 매출 **8.8배** 성장. *FastAPI · LangChain · pgvector*

Prompt to Playable — 생성되는 즉시 typed artifact를 브라우저 내 Node 런타임으로 스트리밍해, 모델이 아직 코드를 작성하는 중에도 dev server가 부팅. 후속 수정은 unified diff로 적용. *Remix · WebContainer · AI SDK*

PDF to Lesson — 페이지를 추출 텍스트가 아닌 이미지로 모델에 전달하고, 학습 활동은 렌더 전에 schema 검증. 모든 학생 메시지는 **13개** 카테고리·4단계 moderator를 거치며 자해 신호는 교사에게 escalation. *Multimodal · Socket.IO*

학력

디자인 석사, Human-Computer Interaction — UNIST

2023.02 · GPA 4.0 / 4.3

학위논문: *계산적 감정 모델을 활용한 소셜 로봇의 얼굴 및 색상 표현에 대한 인간 지각 연구* · 롯데 장학생

컴퓨터공학 & 산업디자인 학사 — UNIST · Global UNISTAR Silver 장학생

2021.02

보유 기술

언어: Python, TypeScript, JavaScript, C/C++, SQL

Agents & Orchestration: LangGraph, LangChain, LangSmith, LiteLLM, Vercel AI SDK, OpenAI / Claude / Gemini API, multi-model routing, ReAct tool loop, checkpointed state & memory, structured output, human-in-the-loop escalation

Retrieval: Milvus, pgvector, ChromaDB, LlamaIndex, PostgreSQL, Redis, BM25 / hybrid search, cross-encoder reranker

Moderation & Evals: 다중 카테고리 classifier, severity 분류 및 escalation, PII 마스킹, eval dataset 및 regression harness

Frontend & 3D: React, Next.js, Three.js, WebGL, Vite, Node.js, Socket.IO

자격증: Agentic AI — DeepLearning.AI (2026) · Advanced RAG with Vector Databases — IBM (2025)