

Temirlan Dzhoroiev

AI ENGINEER · AGENT ORCHESTRATION & MULTI-AGENT SYSTEMS · RAG · EVALS & MODERATION

AI engineer with 3+ years shipping **production agent systems** — multi-tenant tool-calling agents with checkpointed state and memory, multi-model routing, and the eval and moderation layers that decide what ships — across products used by **10M+ people**. Retrieval depth (hybrid RAG, reranking), real-time 3D engine background (**54M+ game plays**), HCI research (**10 publications, 6 patents**).

12 years in Korea · fluent Korean · Korean workplace experience · no visa sponsorship required

Seoul, South Korea · +82 10-7393-2412 · dzhoroiev1@gmail.com · mastertim.life (portfolio)

linkedin.com/in/dzhoroiev7 · github.com/master-tim

EXPERIENCE

AI Engineer — Redbrick, Seoul

Jul 2024 – Present

- Own the agent orchestration layer in production — **40+ tenants, ~30K requests/day**: tool-calling agents on LangGraph with checkpointed state and per-agent memory, coordinating up to **8 dependent tool calls** per request across retrieval, generation, and sandboxed code execution; checkpoint recovery raised multi-step task success **82% → 96%**.
- Secured a **\$1.2M** government R&D grant with the end-to-end AI game-creation system I built — natural-language prompt to playable 3D game, orchestrating intent parsing, asset generation, scene composition, and logic synthesis.
- Cut hallucination rate **18% → 4%** and lifted answer relevance **+35%** on versioned eval sets by shipping production RAG over **15K+** documents — hybrid dense + BM25 retrieval, query rewriting, cross-encoder reranking.
- Halved cost per answer (**-48%**, 21K → 12K tokens) and dropped latency **3.2s → 0.8s** via rerank-then-budget context assembly and an embedding-similarity semantic cache that removed **~40%** of redundant LLM calls — **~\$2K/month** saved.
- Built and own the moderation and evaluation stack gating every prompt or model change: LangSmith eval datasets (**600+ regression cases**), a **13-category** classifier with **4** severity tiers that escalates to a human rather than silently refusing, fail-closed provider checks, and sub-10ms local guardrails (injection, blocked terms, PII redaction).

3D Frontend Engineer — Redbrick, Seoul

Jul 2023 – Jul 2024

- Scaled a browser-based 3D game engine (Three.js, WebGL, TypeScript) powering a creation platform with **10M+ sign-ups** and **54M+ game plays**; shipped several games on it that surpassed **1M+** combined plays.
- Led modernization of a **50K-line** codebase from Webpack to Vite with a rebuilt render pipeline and replaced physics engine — build times **-60%**, bundles **-35%**, hot reload **8s → under 1s**.
- Refactored the engine into modular, testable packages to unblock **3 teams** working in parallel; added integration tests over the scene graph and asset loading.
- Redesigned onboarding from session-replay analysis and user interviews: 7-day activation **+25%**, Day-30 retention **+15%**.

Embedded Systems & HCI Researcher — DECS Lab, UNIST, Ulsan

Feb 2021 – Feb 2023

- Wrote real-time IoT firmware in C for ARM Cortex-M MCUs driving **50+** networked sensor devices at sub-millisecond latency, powering the data pipeline behind **6** published studies.
- 6** conference papers, **4** journal articles, **6** co-filed patents. First author at IEEE RO-MAN 2023, ICROS 2022, HCI Korea 2022.

SELECTED PROJECTS

Production systems I designed and shipped. Full case studies — architecture, tradeoffs, evaluation — at mastertim.life/work.

Enterprise Agent Platform — Multi-tenant RAG: 5-engine parser fallback, dense + BM25 fused in Milvus, cross-encoder rerank, per-tenant ACLs inside the query. **-48%** tokens/answer, context **51K → 25K** chars. *LangGraph · Milvus · LiteLLM*

Companion Chat Runtime — Companion agent with two-layer memory (Postgres buffer + pgvector recall) behind a model-routing gateway serving **17** models under one alias, with a semantic image cache. Grew revenue **8.8×** in the first month post-launch. *FastAPI · LangChain · pgvector*

Prompt to Playable — Streams typed artifacts into an in-browser Node runtime as they are generated, so the dev server boots while the model is still writing; follow-up edits apply as unified diffs. *Remix · WebContainer · AI SDK*

PDF to Lesson — Pages go to the model as images, not extracted text; activities are schema-validated before render. Every student message passes a **13-category, 4-tier** moderator that escalates self-harm signals to the teacher. *Multimodal · Socket.IO*

EDUCATION

M.S. in Design, Human-Computer Interaction — UNIST

Feb 2023 · GPA 4.0 / 4.3

Thesis: *Human perception of social robot face and color expression using computational emotion models* · Lotte Scholarship

B.S. in Computer Science & Industrial Design — UNIST · Global UNISTAR Silver Scholarship

Feb 2021

TECHNICAL SKILLS

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

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

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

Moderation & Evals: Multi-category classifiers, severity triage & escalation, PII redaction, eval datasets & regression harnesses

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

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