

We're a **stealth-mode AI startup** assembling a **top-notch engineering team** to build and ship **agentic AI products** for internal use.

Hiring **full-time engineers**

Location: Tel Aviv HQ (hybrid)

What you'll do

- Build **agentic workflows** (planning, memory, feedback loops) with success/latency metrics.
- **Fine-tune LLMs** (SFT + preference optimization like DPO/ORPO); help with data curation & evals.
- Run **alignment experiments** (PPO/GRPO, reward models) vs. strong baselines.
- Add **evals & reliability** (groundedness checks, regression tests, small A/Bs).
- **Stay current with** latest advances in **LLM inference optimization** (KV compression, quantization, caching, speculative decoding, paged attention etc.).

You might be a fit if

- MSc/PhD (or BSc **top-of-class**) in CS/EE/Math/Physics; strong ML fundamentals.
- Hands-on **Python + PyTorch**; trained at least one non-toy model.
- Experience with **LLMs\Agentic workflow** (finetuning, prompt tooling, agents).
- You turn papers into **measurable experiments** and communicate clearly.

Nice to have (optional): LangGraph/LangChain, RAG, MCP tool calling, KV-cache tricks, quantization (AWQ/GPTQ/FP8/INT4), PPO/DPO/ORPO, CUDA/Triton.

How to apply:

- Send your **CV + Grade Transcript to:** aigraduate@elmnt.com

Optional: Along with your CV, **send your GitHub (for projects) or scholar page**