# Ahmed Heakl

Abu Dhabi, UAE

+971585205137 ahmed.heakl@mbzuai.ac.ae ahmed-heakl ahmedheakl Google Scholar

## Education

**Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)** **Aug. 2024 – May. 2026**

*MSc in Computer Vision, GPA 3.77/4.0* *Abu Dhabi, UAE*

- **Courses**: Computer Vision, Probability and Statistics, Object Detection, Vision Language Models
- **Supervisors**: Salman Khan (Primary), Fahad Khan (Secondary)

**Egypt Japan University of Science and Technology (E-JUST)** **Sep. 2019 – Jul. 2024**

*B.Sc. in Computer Science and Engineering, GPA 3.85/4.0 - (top 1/56 students)* *Alexandria, Egypt*

- **Thesis:** LLM-based code-switched Egyptian Arabic–English auto-grading and ASR.

## Work Experience

**Naver (Parameter Lab)** **Jun. 2025 – Now**

*Research Scientist Intern* *Tübingen, Germany*

- Authored Dr.LLM: Dynamic Layer Routing in LLMs, retrofitting frozen LLMs with lightweight per-layer routers (skip/execute/repeat) trained via MCTS-supervised data generation, reducing required training data by **97%**.
- On ARC/DART, improved accuracy by **+3.4%** while **saving 3-11 layers/query**, outperforming adaptive-depth baselines by up to **+7.7%** without changing base model weights.

**Monta AI** **Jul. 2024 – Apr. 2025**

*Applied Scientist (remote)* *California, USA*

- Built an agentic sales-outreach workflow orchestrating prospect discovery, email generation, and multi-step follow-ups using LLMs and external APIs; in internal deployment, 1.8x outreach volume increase and 62% manual time reduction per lead while maintaining reply rates over a multi-week period.
- Co-led design and deployment of an Arabic OCR pipeline for a Saudi government client processing 100k financial documents, covering layout detection, table & text recognition, and chart & image understanding/captioning; trained YOLOv10 on thousands of annotated pages, with internal validation indicating 95.3% mAP@0.5 for layout detection.
- Developed and distributedly trained a 7B VLM stack that converts diagrams and charts into executable Graphviz and Vega-Lite JSON specifications, pretraining on 250k synthetic bilingual (Arabic–English) examples with DDP across 8 GPUs on 2 nodes and reaching 15.3% CER on held-out internal benchmarks.
- Implemented a table recognition component that converts table images into structured formats (XML/Markdown/HTML) using VLMs (e.g., QwenVL, InternVL) and line detection, achieving 93.2% TEDS on an internal dataset of 11k table images.

**And Africa Co., Ltd** **Mar. 2023 – Jul. 2024**

*Software Engineer (remote)* *Tokyo, Japan*

- Adapted a Flutter mobile app for parcel deliveries (ECD) to expand into 5+ countries, implementing key modifications for currency, location, and payment.
- Led the development of Logi-IQ, a web and mobile platform for vehicle route optimization, utilizing MySQL, PHP Laravel, and Vue.js, achieving 90% test coverage.
- Implemented a core route optimization service using Python, optimizing 8 objectives in under 2 seconds for 1000 orders, 3 drivers, and 3 vehicles, and deployed on AWS for robust scalability.

**Scoville Co., Ltd** **Jan. 2023 – Feb. 2023**

*Machine Learning Intern* *Tokyo, Japan*

- Developed and optimized a GAN-based stylometry obfuscation model using GPT-2 and Siamese-ResNet in PyTorch, achieving 0.8s latency for 1,024-token samples with 137M parameters.

## Selected Publications

**SVRPBench: A Realistic Benchmark for Stochastic Vehicle Routing Problem | Link** **May. 2025**

*NeurIPS Datasets and Benchmarks, 2025* *1st Author*

∗ Co-developed SVRPBench, an open-source benchmark of 500+ urban-scale SVRP instances with time-dependent congestion, probabilistic delays, and heterogeneous time windows, plus a full evaluation suite for learning-based solvers.

**GG: An LM Approach for CISC-to-RISC Transpilation with Testing Guarantees | Link** **May. 2025**

*EMNLP 2025 (Findings)* *1st Author*

∗ Introduced Guaranteed Guess (GG), a test-driven CISC-to-RISC transpiler using LLMs, achieving 99.4% functional accuracy on x86→ARMv8 translation and outperforming Rosetta 2 in runtime (1.73×), and memory (2.41×).

**KITAB-Bench: Multi-Domain Arabic OCR and Document Understanding Bench** | Link **Feb. 2025**
*ACL 2025 (Findings)* *1st Author*

* Spearheaded KITAB-Bench, a benchmark with 8,809 Arabic pages across 9 domains and 36 sub-domains, showing VLM-based OCR reduces character error by ∼60% vs. traditional OCR while PDF-to-Markdown peaks at ∼65%.

**LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs** | Link **Jan. 2025**
*ACL 2025 (Findings)* *1st Co-Author*

* Developed a curriculum framework and benchmark (4k+ reasoning steps across 8 categories) for step-by-step visual reasoning, and a multimodal model (LlamaV-o1) that improves over Llava-CoT by 3.8% with 5× faster inference.

## Under review / preprints

**DocAtlas: Multilingual Document Understanding Across 80+ Languages** **Nov. 2025**

* Developed *DocAtlas*, a model-free pipeline producing 360K multilingual OCR pages across 82 languages via differential rendering and RTL synthesis, and introduced a 5.8K-page benchmark showing that *DPO* enables stable cross-lingual transfer with minimal forgetting.

**Mobile-O: Unified Multimodal Understanding and Generation on Mobile Devices** **Nov. 2025**

* Co-architected *Mobile-O*, a 1.6B unified vision–language–diffusion model with a mobile conditioning projector and a quadruplet post-training scheme, enabling real-time on-device multimodal understanding and text-to-image generation with state-of-the-art accuracy and 6–11× faster latency than prior models.

**Dr.LLM: Dynamic Layer Routing in LLMs** | Link **Oct. 2025**

* Introduced **Dr.LLM**, a retrofittable framework enabling frozen LLMs to skip, execute, or repeat layers via MCTS-supervised routing, reducing required training data by **97%** and improving accuracy by **+3.4%p** while saving **3–11 layers/query**.

**How Good are Foundation Models in Step-by-Step Embodied Reasoning?** | Link **Sep. 2025**

* Co-developed **FoMER-Bench**, a large-scale benchmark for step-by-step embodied reasoning across 10 tasks and 8 embodiments, and led evaluations of foundation multimodal models, uncovering systematic failures in long-horizon reasoning and safety-critical behaviors.

**CASS: Nvidia to AMD Transpilation with Data, Models, and Benchmark** | Link **May. 2025**

* Developed CASS, the first large-scale dataset and benchmark (70k samples) enabling source- and assembly-level transpilation between Nvidia CUDA and AMD HIP/RDNA3, and trained domain-specialized LLMs (up to 7B) that outperform GPT-4o and Claude by achieving 95% source and 37.5% assembly accuracy while preserving execution behavior in 85%+ of cases.

**VideoMolmo: Spatio-Temporal Grounding Meets Pointing** | Link **May. 2025**

* Developed VideoMolmo, a spatio-temporal visual grounding model that combines LLM-driven pointing and mask fusion via SAM2; introduced a novel temporal attention module and a bidirectional mask propagation strategy, curated a 72k video-caption dataset with 100k object points, and proposed VPoS-Bench, a challenging benchmark spanning five real-world domains, achieving state-of-the-art performance on video grounding, counting, and reasoning tasks.

## Selected Research Projects

**AIN: The Arabic INclusive Large Multimodal Model** **Report**

* Designed a state-of-the-art Arabic-English multimodal model across multiple scales (2B, 7B, 72B), trained on a diverse dataset of 4M samples (including 900k authentic and 3.7M high-quality translations), achieving a 3% improvement over GPT-4o on CAMEL-Bench and a 17.2% gain in OCR tasks through Qwen2-VL fine-tuning, alongside novel JPEG-based data augmentation for robustness against real-world image quality variations. (+7.5k downloads on huggingface)

**CAMEL-Bench: A Comprehensive Arabic LMM Benchmark** (NAACL 2025 - Findings) **Report**

* Designed a VLMs benchmark with 30k manually verified questions across 8 diverse domains, including medical imaging, video understanding, and cultural-specific knowledge, enhancing model assessment quality and evaluated across open-source and closed-source models.

## Technical Skills

**Languages**: Python (primary), C/C++, CUDA, Go, Java, SQL, JavaScript
**Machine Learning**: PyTorch, VLMs, diffusion models, PEFT/LoRA, RLHF/DPO
**Distributed & Systems**: Linux, HPC & cluster training (Slurm, DeepSpeed, FSDP), CUDA kernels, Docker
**MLOps & Tooling**: Experiment tracking, benchmarking & profiling, CI/CD, AWS, Google Cloud