HUZAIFA SHAABAN KABAKIBO

PhD Student | LLM Systems Research

Grünebaumstraße, Paderborn, Germany | huzaifa@mail.uni-Paderborn.de | +49 5251 60-1757

SUMMARY

PhD student in Computer Science at Paderborn University researching Large Language Model (LLM) serving systems and AI infrastructure. My current work focuses on understanding and optimizing LLM serving engines, with particular emphasis on cold-start latency in vLLM and serverless inference. Previously, I developed MoRAGBench, a framework for benchmarking Retrieval-Augmented Generation (RAG) pipelines on mobile devices, and worked on efficient LLM inference for edge devices. My interests lie in scalable AI systems, systems optimization, and hardware-software co-design.

EDUCATION

PhD in Computer Science **Apr 2026 – Present**

Paderborn University

- Research in the Computer Networks Group
- Research focus on Large Language Model serving systems and inference optimization

Masters in Computer Engineering **Oct 2023 – Mar 2026**

Paderborn University

- Focus area in Embedded Systems
- Graduated with distinction (sehr gut, 1.1)

Bachelor Of Science In Computer Engineering **Oct 2019 - Jun 2022**

Lebanese International University

- Graduated with distinguish degree with GPA 3.97/4

PROFESSIONAL EXPERIENCE

Research Assistant, Computer Networks Group, Paderborn University **May 2025 – Mar 2026**

- Built a framework to decompose vLLM initialization into fine-grained steps.
- Analyzed latency contributions across different model configurations.
- Identified cold-start latency as primarily CPU-bound, not GPU-bound.
- Designed a predictive model to estimate latency and guide scheduling.
- Conducted the research that led to the publication "Breaking the Ice: Analyzing Cold Start Latency in vLLM" accepted at MLSys 2026.

Research Assistant, Paderborn Centre for Parallel Computing **May 2024 – May 2025**

- Automate testing and configuration management using Bash scripting and GitLab CI/CD.
- Conduct measurements and performance analysis on FPGA hardware.
- Develop scripts for automating data collection and processing in performance testing.
- Troubleshoot and debug hardware/software issues in FPGA-based workflows.

Full stack developer, BEA Technology Company **Jun 2022 - Oct 2023**

- Developed AI-driven systems for real-time operations with a focus on UI/UX.
- Optimized core systems for better data flow and logical analysis.
- Enhanced database performance, security, and API tokenization.
- Fostered teamwork and ensured compliance with company policies.

PROJECTS

MoRAGBench: A Benchmarking Framework for RAG Pipelines on Mobile Devices **Master Thesis (2026)**

- Designed and implemented the first benchmarking framework for Retrieval-Augmented Generation (RAG) pipelines on Android/mobile devices.
- Supports benchmarking of retrieval quality, indexing performance, and end-to-end LLM inference across multiple mobile hardware backends.
- Provides a modular architecture for evaluating mobile AI systems and reproducible experimentation.
- Open-source implementation: github.com/upb-cn/MoRAGBench

PartInfer: Enabling Efficient LLM Inference on Edge Devices **Oct 2024 – Sep 2025**

- Developed PartInfer, a framework to enable efficient LLM inference on edge devices.
- Categorized neurons into base and task-specific groups using offline profiling.
- Implemented partial loading to fit relevant neurons into memory under constraints.
- Introduced partial computation to accelerate inference by computing only active neurons.

Publications

Breaking the Ice: Analyzing Cold Start Latency in vLLM **2026**
Accepted at MLSys 2026
Huzaifa Shaaban Kabakibo, Animesh Trivedi, Lin Wang

SKILLS

Communication	Time management	Project management
Leadership	Problem solving	Teamwork

Languages

- Arabic: Native
- English: Fluent
- German: B1 Certified