

DR. FATEMEH ZARDBANI

fatemeh.zardbani@gmail.com | (+44) 7356018687 | Edinburgh, UK | [in](#) | [%](#) | [G](#)

PROFESSIONAL EXPERIENCE

Huawei Technologies Research and Development

Senior Software Research Engineer

Jul 2024 - Present

Edinburgh, UK

- Architecting, building & operating large-scale multimodal indexing and retrieval systems, integrating text, image, audio & video embeddings into unified high-dimensional vector spaces using state-of-the-art vector databases and ANN backends (e.g., HNSW, IVF-PQ, ScaNN), enabling low-latency semantic search, cross-modal similarity matching, and downstream Retrieval-Augmented Generation (*RAG*) pipelines with embedding versioning, online/offline re-indexing, and strict latency, throughput & recall KPIs for production-grade Generative AI platforms.
- Designed and deployed Graph-based Retrieval-Augmented Generation (*GraphRAG*) systems leveraging formal knowledge graph modeling and W3C standards (e.g., RDF, OWL, SPARQL), including ontology design, schema alignment, entity recognition, entity linking/alignment, and relationship extraction. Integrated graph-aware context construction and subgraph traversal to augment LLM prompts with structured, high-signal knowledge, enabling deep customization, personalization, and long-term memory for agentic LLM users across multi-hop reasoning, tool invocation, and autonomous workflow execution.
- Designed and implemented a vectorized memory and caching layer for Small Language Models (*SLMs*), leveraging Transformer KV-cache partitioning, prefix/suffix caching, embedding-based context deduplication, and user-specific semantic chunk reuse, with cache-aware decoding and sliding-window/sparse attention to minimize recomputation and reduce latency and cost across multturn, tool-augmented LLM/SLM pipelines.
- Designed and implemented a hardware-agnostic parallelism framework and domain-specific language (*DSL*) for ML workloads—specifically ANN index construction and retrieval; enabling portable data, model, and pipeline parallelism across non-CUDA accelerators and heterogeneous backends via custom IRs, compiler passes, and runtime scheduling for efficient, vendor-neutral execution.
- Designed and operationalized LLM evaluation frameworks including LLM-as-a-Judge, pairwise/listwise ranking, rubric-based and reference-free evaluation, and automatic metrics (e.g., BLEU, ROUGE, BERTScore), with production-grade online/offline pipelines and continuous feedback loops driving prompt optimization, model selection, fine-tuning, and *RLHF/RЛАIF* under strict quality, cost, and latency constraints.

Data-Intensive Systems Lab

PhD Candidate

Sep 2019 - March 2025

Aarhus, Denmark

- Conducted research on adaptive high-dimensional indexing for spatial and vector databases, similarity search under latency-critical, non-stationary workloads, designing query-adaptive, vantage-point-based metric indices with distance reuse and online re-indexing to avoid expensive upfront construction, enabling exact and low-latency retrieval over large-scale, high-dimensional embedding spaces.
- Led algorithmic research on scalable k-means clustering for high-dimensional data, introducing aggressive distance-pruning via tight bounds, multiresolution stepwise evaluation, and triangle-inequality exploitation to drastically reduce centroid distance computations, enabling time- and compute-efficient clustering with strong convergence guarantees on large scientific and embedding datasets.

Harvard University, School of Engineering and Applied Sciences

Research Fellow

Aug 2022 - Feb 2023

Boston, MA, USA

- Designed and implemented adaptive LSM-tree architectures for advanced NoSQL and AI storage engines, with workload-aware read/write optimization via Reinforcement Learning and multi-armed bandits controlling flush, compaction, re-flush, and chunking policies. Deployed in vector databases, feature stores, and online inference backends for large-scale embedding retrieval and real-time ML systems, enabling self-tuning storage with low-latency reads and high-throughput writes under non-stationary workloads.

EDUCATION

Aarhus University

PhD Computer Science

Sharif University of Technology

B.Sc. Computer Engineering

SKILLS & QUALITIES

Programming Languages, Frameworks & Libraries

- C/C++, Python, Java, Bash, Linux, TensorFlow, PyTorch, NLTK, CoreNLP, NumPy, pandas, SciPy, OpenCV, XGBoost, SQL (PostgreSQL, MSSQL, MySQL) & NoSQL (RocksDB, LevelDB, Cassandra, MongoDB, WiredTiger), Apache Spark, Pinecone, LangChain, Git, Docker, OpenCL, Firebase, Keras-RL.

PUBLICATIONS

Incremental Management of Multi-dimensional Data - [Link](#)

F. Zardbani

PhD thesis, Incremental management of multi-dimensional data-low and high dimensional spaces. Includes all work done during the PhD period.

Aarhus University

Updating an Adaptive Spatial Index - [Link](#)

F. Zardbani, K. Lampropoulos, N. Mamoulis, P. Karras

Incremental update mechanisms for adaptive spatial indices enabling low-latency queries under dynamic data updates while preserving adaptive behavior.

ICDE 2025

Benchmarking Adaptive Multidimensional Indices - [Link](#)

K. Lampropoulos, **F. Zardbani**, N. Mamoulis

Comprehensive experimental benchmarking of adaptive multidimensional indexing techniques, analyzing convergence speed, robustness, and performance trade-offs across workloads.

VLDB 2025

Adaptive Indexing of Objects with Spatial Extent - [Link](#)

F. Zardbani, N. Mamoulis, S. Idreos, P. Karras

Workload-driven adaptive indexing for spatial objects with extent, enabling continuous refinement without upfront index construction and improved query efficiency.

VLDB 2023

Adaptive Indexing in High-Dimensional Metric Spaces - [Link](#)

K. Lampropoulos, **F. Zardbani**, N. Mamoulis

Adaptive distance-driven indexing techniques for high-dimensional metric spaces that outperform traditional space partitioning under evolving similarity workloads.

VLDB 2023

Marigold: Efficient k-Means Clustering in High Dimensions - [Link](#)

K.O. Mortensen, **F. Zardbani**, et al

Scalable k-means clustering in high dimensions using aggressive distance pruning via tight bounds, multiresolution evaluation, and triangle-inequality exploitation.

VLDB 2023

Revisiting Multidimensional Adaptive Indexing - [Link](#)

A.H. Jensen, F.A. Lauridsen, **F. Zardbani**, S. Idreos

Theoretical and empirical re-evaluation of multidimensional adaptive indexing, clarifying convergence behavior, strengths, and limitations.

EDBT 2021

Revisiting the Theory and Practice of Database Cracking - [Link](#)

F. Zardbani, P. Afshani, P. Karras

Revisits database cracking from theoretical and systems perspectives, analyzing robustness, performance bounds, and modern workload behavior.

EDBT 2020