

Babak Ehteshami Bejnordi

Principal Research Scientist | Efficient Large Language Models

✉ babakint@gmail.com

🌐 LinkedIn

🏠 babakint.github.io

🎓 Google Scholar

📍 Zurich, Switzerland

I am a Principal Research Scientist at NVIDIA in Zurich, where I develop efficient architectures for large language models as part of the Nemotron team. Previously, I was a Senior Staff Engineer and Manager at Qualcomm AI Research in Amsterdam, leading research on efficient LLM architectures and Mixture-of-Experts (MoE), from foundational methods to commercial deployment on edge devices. My broader research interests include MoE training and inference, LLM architecture design, efficient and latent reasoning, on-device AI, multi-task and continual learning, and computer vision. I have published at leading venues including NeurIPS, ICML, ICLR, CVPR, and ECCV. I earned my PhD at Radboud University, where I developed machine-learning algorithms for breast cancer diagnostics and organized the CAMELYON16 challenge. During my PhD, I was also a visiting researcher at Harvard University.

PROFESSIONAL EXPERIENCE

NVIDIA, Zurich, Switzerland

Principal Research Scientist

📅 Sep 2026 – Present

- Efficient architectures for large language models on the Nemotron team

QUALCOMM AI RESEARCH, Amsterdam, The Netherlands

Technical Team Lead, Senior Staff Engineer & Manager

📅 Nov 2023 – Aug 2026

- Efficient LLM & MLLM Architectures, Mixture-of-Experts, Efficient (Latent) Reasoning

Technical Team Lead, Staff Engineer & Manager

📅 Nov 2019 – Nov 2023

- Conditional Computation for Efficient LLMs, Vision Transformers, and CNNs

Senior Engineer

📅 Apr 2018 – Nov 2019

- Deep Learning Research Scientist

NAVINFO, Eindhoven, The Netherlands

Deep Learning Engineer

📅 Oct 2017 – Mar 2018

- Deep Learning and Computer Vision for Autonomous Driving

EDUCATION

RADBOD UNIVERSITY, The Netherlands

Ph.D. in Machine Learning & Medical Image Analysis

📅 Apr 2013 – Jun 2017

HARVARD UNIVERSITY, Massachusetts, USA

Visiting Scholar

📅 Jun 2016 – Nov 2016

CHALMERS UNIVERSITY OF TECHNOLOGY, Goteborg, Sweden

M.Sc. in Electrical Engineering

📅 May 2010 – Dec 2012

UNIVERSITY OF GUILAN, Rasht, Iran

B.Sc. in Electrical Engineering

📅 2004 – 2008

SELECTED PROJECTS

QUALCOMM AI RESEARCH

Leading Research and Commercial Efforts on MoEs

Amsterdam, The Netherlands

📅 May 2023 - Aug 2026

- Commercial Feature lead for MoE deployment on Qualcomm Hexagon NPU
 - Designing novel accelerated prefill & expert caching strategies for mobile device deployment with 5+ patents.
 - MoE router finetuning methods to increase expert caching efficiency by regularizing expert selection consistency.
 - Cross-functional Collaboration with AI Software and NPU Hardware teams to enable hardware-software co-design.
- Dirichlet-Prior Shaping: Guiding Expert Specialization in Upcycled MoEs [ICML 2026](#)
 - Enhancing MoE expert specialization with a novel router regularization technique for vision-language models.
- Mixture of Cache-Conditional Experts for Efficient Mobile Device Inference [TMLR 2025](#)
 - Boosted token decoding throughput up to 2× using cache conditional routing, with minimal accuracy degradation.
- Refactorizing LLMs as Router-Decoupled Mixture of Experts with System Co-Design [NeurIPS 2024](#)
 - A technique to transform pre-trained dense LLMs into smaller MoE models.
 - University collaboration with Prof. Atlas Wang from UT Austin.
- Efficient MoE for mobile devices with limited DRAM [NeurIPS 2024 Expo Demo](#)
 - Deployed 14B+ parameter MoE to mobile device with Snapdragon 8 Elite, increased throughput by 2× compared to standard LRU caching, with < 1% drop in MMLU accuracy.

QUALCOMM AI RESEARCH

Leading Efficient LLM Architectures Research

Amsterdam, The Netherlands

 April 2024 - Aug 2026

- KaVa: Latent Reasoning via Compressed KV-Cache Distillation [ICLR 2026](#)
 - The first framework that distills knowledge from compressed KV-cache of a teacher LLM into a latent-reasoning student via self-distillation.
 - KaVa consistently outperforms strong latent reasoning baselines (such as COCONUT, CODI, etc.), scales effectively to larger backbones, and shows robust performance on natural-language reasoning datasets.
- Dynamic Budget Forced Reasoning on Mobile ([Qualcomm Demo @NeurIPS 2025](#))
 - Achieved up to 2.7× reduction in token generation lengths, with minimal accuracy drop by deploying a LoRA-enabled Reinforcement Learning trained reasoning model (Qwen2.5-7B) on Snapdragon 8 Gen 5.
- Think Big, Generate Quick: LLM-to-SLM for Fast Autoregressive Decoding [ICML 2024 \(ES-FoMo II workshop\)](#)
 - A hybrid system leveraging an LLM's prompt representation to steer an SLM for efficient token generation, boosting autoregressive decoding speed by 3–4×.

QUALCOMM AI RESEARCH

Technical Lead focusing on Conditional Computation for Efficient DL

Amsterdam, The Netherlands

 Jan 2022 - June 2024

- InterroGate: Learning to Share, Specialize, & Prune Representations for MTL [BMVC 2024](#)
- Scalarization for Multi-Task and Multi-Domain Learning at Scale [NeurIPS 2023](#)
- MSViT: Dynamic Mixed-Scale Tokenization for Vision Transformers [ICCV 2023 \(NViT workshop\)](#)
- Conditional Compute for On-device Video Understanding at [NeurIPS 2022 Expo Demonstrations](#)
- Revisiting single-gated Mixtures of Experts published at [BMVC 2022](#).

QUALCOMM AI RESEARCH

Efficient Deep Models for Video Processing

Amsterdam, The Netherlands

 Oct 2020 - Dec 2021

- SALISA: Saliency-Based Input Sampling for Efficient Video Object Detection published at [ECCV2022](#).
- FrameExit: Conditional Early Exiting for Efficient Video Recognition published at [CVPR2021 \(Oral paper\)](#).
- Skip-Convolution for Efficient Video Processing published at [CVPR2021](#).
- Spatio-Temporal Gated Transformers for Efficient Video Processing at [NeurIPS2021 \(ml4ad workshop\)](#).

QUALCOMM AI RESEARCH

Conditional Computation for Convolutional Neural Networks

Amsterdam, The Netherlands

 May 2018 - Oct 2020

- Conditional Channel Gated Networks for Task-Aware Continual Learning published at [CVPR2020 \(Oral paper\)](#).
- Batch-shaping for learning conditional channel gated networks published at [ICLR2020](#).

HARVARD UNIVERSITY

Deep learning for Diagnosing Breast Cancer Patients

Boston, MA, USA

 May 2016 - Dec 2016

- Development of a deep learning system for diagnosing breast cancer patients (see publications [2017a](#) and [2017b](#)).

- › This work was in Collaboration with NIH, and Mayo Clinic.

RADBOD UNIVERSITY

Lead organizer of CAMELYON16 Machine Learning Challenge

Nijmegen, The Netherlands

📅 May 2016 - Dec 2016

- › My challenge gathered participants from all around the world including Google, Harvard, MIT, etc.
- › Published paper at [JAMA with 4000+ citations](#).
- › Extensive coverage in over 30 well-known websites and media (e.g. Yahoo News, NOS.nl).
- › Highlighted in the [White House AI strategic planning report \(page 17\)](#).

RADBOD UNIVERSITY

Machine learning for Breast Cancer Diagnosis

Nijmegen, The Netherlands

📅 Apr 2013 - Mar 2016

Selected Research highlights:

- › Co-authored "A survey on deep learning in medical image analysis" published at [MEDIA with ~20,000 citations](#).
- › Development of the first Whole-slide image color standardization published at [IEEE Transactions in Medical Imaging](#).
 - › The first algorithm to standardize giga-pixel pathology images ([Source code](#), [Executable](#)).
 - › Using my algorithm, the team from Harvard & MIT improved its rank from 4th to 1st in CAMELYON16 challenge.

SKILLS

- › Python, PyTorch, HuggingFace Transformers, DeepSpeed, Accelerate, Ray, TensorFlow, scikit-learn, matplotlib, C++
- › LLM Architecture, Efficient MoE, Dynamic Sparsity, Multi-task Learning, Continual Learning, Scalable Model Training, GPU/Edge Optimization, Computer Vision, Autonomous Driving
- › English (fluent), Persian (native), Dutch, Swedish, Italian (basic)

HONORS AND AWARDS

Third highest cited (2000+) work in JAMA

📅 2020

In 2020, with 2000+ citations (currently 4000+), my paper on "Diagnostic assessment of deep learning algorithms lymph node metastases detection" was among the top 3 most cited works of the Journal of American Medical Association (Impact Factor 157) over the past 3 years.

Highest cited survey on Deep Learning for Medical Imaging (20,000+ citations)

📅 2017

Contributed (third author) to the highest cited survey in deep learning medical imaging to date.

Finalist nomination for Wetenschaps- en Innovatieprijs 2019

📅 2019

Finalist nomination for the [Wetenschaps- en Innovatieprijs 2019](#) of the Federation of Medical Specialists in the Netherlands for my project on artificial intelligence in breast cancer diagnosis.

MedicalPhit Innovation Award of 2016

📅 2016

Won the best [MedicalPhit Innovation Award of 2016](#) in the Netherlands for the study of the use of artificial intelligence in detecting metastases in breast cancer patients.

Research Grant Awards

📅 2016

Awarded research grants from BeckLab, Harvard Medical School (2016), Radboud university (2016), Chalmers and Uppsala Universities (2012)

Master's degree with honors

📅 2013

GPA 5.0/5.0 at Chalmers University of Technology.

SELECTED PUBLICATIONS

🎓 [Google Scholar](#) — Citations: 31,001 (5y: 23,997) • h-index: 27 • i10-index: 37

The full list of my publications and patents can be found on my [google scholar](#) page.

- › A Kuzina, M Pioro, B Ehteshami Bejnordi[‡]. "KaVa: Latent Reasoning via Compressed KV-Cache Distillation". [ICLR, 2026](#). [\[PDF\]](#)

- L Mirvakhabova*, **B Ehteshami Bejnordi***, G Kumar, H Liang, W Zhao, P Whatmough. "Dirichlet-Prior Shaping: Guiding Expert Specialization in Upcycled MoEs". **ICML, 2026**. [\[ICML\]](#)
- A Skliar, T van Rozendaal, R Lepert, T Boinovski, M Van Baalen, M Nagel, P Whatmough, **B Ehteshami Bejnordi†**. "Mixture of cache-conditional experts for efficient mobile device inference". **TMLR, 2025**. [\[PDF\]](#)
- R Cai, Y Ro, GW Kim, P Wang, **B Ehteshami Bejnordi**, A Akella, Z Wang. "Read-ME: Refactorizing LLMs as Router-Decoupled Mixture of Experts with System Co-Design". **NeurIPS, 2024**. [\[PDF\]](#)
- B Bergner, A Skliar, A Royer, T Blankevoort, Y Asano, **B Ehteshami Bejnordi†**. "Think Big, Generate Quick: LLM-to-SLM for Fast Autoregressive Decoding". **ICML Workshop ES-FoMo-II, 2024**. [\[PDF\]](#)
- Amelie R, T Blankevoort, **B Ehteshami Bejnordi†**. "Scalarization for multi-task and multi-domain learning at scale". **NeurIPS, 2023**. [\[PDF\]](#)
- **B Ehteshami Bejnordi**, G Kumar, A Royer, C Louizos, T Blankevoort, M Ghafoorian. "InterroGate: Learning to Share, Specialize, and Prune Representations for Multi-task Learning". **BMVC, 2023**. [\[PDF\]](#)
- **B Ehteshami Bejnordi**, A Habibian, F Porikli, A Ghodrati. "Salisa: Saliency-based input sampling for efficient video object detection". **ECCV, 2022**. [\[PDF\]](#)
- A Ghodrati*, **B Ehteshami Bejnordi***, A Habibian. "FrameExit: Conditional Early Exiting for Efficient Video Recognition". **CVPR (Oral), 2021**. [\[PDF\]](#)
- A Habibian, D Abati, T Cohen, **B Ehteshami Bejnordi†**. "Skip-convolutions for efficient video processing". **CVPR, 2021**. [\[PDF\]](#)
- **B Ehteshami Bejnordi**, T Blankevoort, M Welling. "Batch-Shaping for Learning Conditional Channel Gated Networks". **ICLR, 2020**. [\[PDF\]](#)
- Davide Abati, J Tomczak, T Blankevoort, S Calderara, R Cucchiara, **B Ehteshami Bejnordi†**. "Conditional channel gated networks for task-aware continual learning". **CVPR (Oral), 2020**. [\[PDF\]](#)

👑 * Shared first authorship, †Senior (Lead) author

INVITED TALKS

- **GHOST Day 2025: Efficient Deployment of Large Language Models on Edge Devices** at [GHOST Day \(Applied Machine Learning Conference\)](#), Poznan, Poland (2025).
- **Cisco Meraki's GenAI: Efficient LLM inference on-device** (2025).
- **Apple's Efficient LLM reading group: Efficient inference of Mixture-of-Experts LLMs on-device** (2025).
- **DeepLearn 2023 Spring: Course lectures** on Conditional Computation for Efficient Deep Learning at the 9th International School on Deep Learning in Bari, Italy.
- **ELLIS PhD and Postdoc Summit, Keynote talk** at the ELLIS PhD and Postdoc Summit (kick-off program), 2021.
- **TWIML AI: Podcast interview** with Sam Charrington from [TWIML AI](#) on Conditional Computation.
- **Broad Institute of MIT and Harvard**, "Practical recommendations for training convolutional neural networks", MIA Seminar: Modeling, Inference, algorithms, 2017, USA, [\(Video link\)](#).
- **Dutch Society for Pattern Recognition and Image Processing**, "Automatic detection of ductal carcinoma in situ in whole slide histopathological images", Eindhoven, The Netherlands, 2015.
- **European Congress on Digital Pathology**, "An algorithm for reducing stain variability in scanned histological slides", Paris, France, 2014.

OTHER ACADEMIC / INDUSTRY EXPERIENCES

- **Qualcomm Innovation Fellowship:** Co-lead in the organization of the [Qualcomm Innovation Fellowship \(QIF\)](#) program in Europe from 2019 to 2023.
- **Deep learning workshop lecturer:** Co-organizer and lecturer at deep learning workshop at Radboud University, the Netherlands (2016).
- **Member of Broad Institute of MIT and Harvard**, Boston, Massachusetts, USA (Jul 2016 – Jul 2017).
- **Supervision and teaching experience:** Supervised more than 10< master students of computer science and artificial intelligence for their course/master thesis project. Supervised 6 Ph.D. students for their internship/Qualcomm fellowship projects. I was also a teaching assistant for the course "Computer Aided Diagnosis" at Radboud University between 2013 and 2016.