Yasmin Moslem, PhD

Efficient NLP and Speech · Dublin, Ireland

hello@yasmin.ie | yasmin.ie | in/yasmin-moslem
ymoslem | hf.co/ymoslem | b4tR-t8AAAAJ

Summary

Hands-on experience building NLP and AI systems across 30+ languages, spanning machine translation, speech, and large language models (LLMs). Recent research focuses on efficient and scalable LLMs, including multi-LLM routing, layer pruning, knowledge distillation, quantization, and efficient fine-tuning. PhD in Computer Science (Adaptive Machine Translation), with retrieval-augmented LLM adaptation methods implemented in production systems. Experience in taking AI research from experimentation and evaluation through to production deployment, with hands-on expertise in PyTorch, Transformers, and vLLM.

Core Skills

- **Efficient inference & training:** multi-LLM routing and cascades, knowledge distillation, iterative layer pruning, model compression, quantization, efficient fine-tuning
- **Modalities:** LLMs, machine translation, speech translation, ASR, TTS / voice cloning, multimodal QA
- **Frameworks & tooling:** PyTorch, Hugging Face Transformers, Accelerate, OpenNMT, vLLM, CTranslate2, ONNX, Faster-Whisper, PEFT (LoRA/QLoRA), bitsandbytes
- **Engineering:** Python, end-to-end system design, data pipelines (collection, filtering, back-translation, synthetic data generation, Pandas, Polars, Scrapy, Airflow), MT/speech evaluation (COMET, BERTScore, chrF++, BLEU), prototyping (FlutterFlow, Flask, FastAPI), and production-level deliverables

Experience

2025 – present

Senior Researcher • ADAPT Centre, School of Computer Science, Trinity College Dublin

• Industry Project: Efficient Foundation Models (sponsored by *Huawei*)

- Built ***CRE-Router***, an open-source two-stage routing and cascading framework for self-hosted LLMs that clusters queries by semantic similarity, routes each cluster to the most cost-effective model under a latency budget measured on the serving GPUs, and escalates answers a quality estimator rejects. On three benchmarks served on A100 GPUs, it matched the accuracy of its strongest model at 20 to 46% lower time per token. (CRE-Router at GitHuub)
- Compressed an 8B multilingual LLM (Aya-Expanse) by 22–44% via iterative layer pruning, retaining up to 99% of translation quality (COMET) and increasing inference throughput by up to 84%, with a further 4.3× serving speedup via vLLM (WMT 2025)
- Halved the size (50% fewer parameters and storage) of an 8B speech translation model (Qwen2-Audio) through iterative layer pruning, 4-bit QLoRA, and knowledge distillation, while retaining 97–100% of the teacher model's quality (IWSLT 2025)
- Developed cross-lingual voice cloning for scientific speech through multi-model distillation, and improved low-resource speech translation using synthetic audio (IWSLT 2024–2026)
- Released AfriNLLB, open compressed translation models for 15 language pairs (30 directions), achieving up to 2.4× faster inference with equal or better quality than the NLLB-200 baseline, plus a curated 1.6M-segment open dataset (AfricaNLP 2026, Spotlight)

• Community leadership: founded the SpeechT mentorship programme (public models, datasets, and a publication); co-organised the IWSLT 2025/2026 shared tasks on Irish-to-English speech translation and cross-lingual voice cloning

2024

Senior NLP Researcher • Bering Lab

• Trained the company's first Arabic translation model from scratch, a 0.5B-parameter model on ∼3B tokens of bilingual data, and delivered higher-quality Korean models for the legal and business domains.

• Led end-to-end development of production MT systems: data collection, filtering, and back-translation, architecture search, quantization, and deployment handover.

Experience (continued)

2022 — **Research Intern** • Microsoft Research
• Researched multilingual and k-nearest-neighbour machine translation (kNN-MT).

2021 – 2022 — **Research Scientist & Architect — NLP** • Big Language Solutions
• Co-built the company's MT capability from the ground up: delivered production translation systems across **10 languages**, from data curation and training to deployment handover.
• Brought state-of-the-art research into production: domain adaptation, robustness, knowledge distillation, handling untranslatables, data augmentation, and low-resource languages.

2020 – 2024 — **PhD Researcher** • ADAPT Centre, School of Computing, Dublin City University
• Developed one of the first approaches to retrieval-augmented LLM adaptation for domain-specific MT (arXiv Jan 2023; EAMT 2023, **300+ citations**), with methods since implemented in commercial MT production systems.
• Published at WMT, EAMT, and AMTA on LLM-based adaptation, terminology integration, and synthetic data generation, including the **Best Presentation Award** at AMTA 2022.
• Funded by Science Foundation Ireland and **Microsoft Research**.

2019 — **NLP Consultant** • African Union
• Built domain-specific neural machine translation models with terminology constraints.

2013 – 2014 — **NLP Engineer (R&D)** • Future Group
• Built domain-specific statistical machine translation engines. Prior background in the translation technology industry (Wordfast).

Education

PhD, Computer Science • Dublin City University
• Thesis: *Language Modelling Approaches to Adaptive Machine Translation*
• Funders: Science Foundation Ireland & Microsoft Research

Research Publications (selection)

* For the full list of publications and citations, please refer to my *Google Scholar profile*.

1. **Y. Moslem** et al., "Cluster, route, escalate: Cascaded framework for cost-aware LLM serving," 2026. arXiv: 2606.27457 [cs.PF]. URL: https://arxiv.org/abs/2606.27457.

2. **Y. Moslem** and J. D. Kelleher, "Dynamic model routing and cascading for efficient LLM inference: A survey," *Transactions on Machine Learning Research (TMLR)*, 2026. URL: https://arxiv.org/abs/2603.04445.

3. **Y. Moslem**, M. H. Al Farouq, and J. D. Kelleher, "Iterative layer pruning for efficient translation inference," in *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, Suzhou, China, Nov. 2025, pp. 1022–1027. URL: https://aclanthology.org/2025.wmt-1.78.

4. **Y. Moslem**, "Efficient speech translation through model compression and knowledge distillation," in *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria, Jul. 2025, pp. 379–388. URL: https://aclanthology.org/2025.iwslt-1.40/.

5. **Y. Moslem**, R. Haque, J. D. Kelleher, and A. Way, "Adaptive Machine Translation with Large Language Models," in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, *300+ citations*, Tampere, Finland, Jun. 2023, pp. 227–237. URL: https://aclanthology.org/2023.eamt-1.22.

6. **Y. Moslem**, R. Haque, J. Kelleher, and A. Way, "Domain-Specific Text Generation for Machine Translation," in *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, *Best Presentation Award*, Orlando, USA, Sep. 2022, pp. 14–30. URL: https://aclanthology.org/2022.amta-research.2.