Younes Hourri

+1 (819) 448-9285 ⋄ yhourri2@gmail.com ⋄ cs.toronto.edu/~younes/

Education

University of Toronto — **Toronto, ON**
MSc Computer Science — *Sept. 2024 – Dec. 2025*
Research: Pruning techniques and inference acceleration of large language models

McGill University — **Montréal, QC**
BSc Computer Science, Minor Statistics — *Sept. 2020 – May 2024*

Technical Skills

Programming Languages: Python (expert), C/C++

GPU Programming: Triton, CUDA (GEMM)

Machine Learning: PyTorch (expert), scikit-learn

LLM Frameworks & Libraries: vLLM, HuggingFace, Accelerate, Megatron-LM

Optimization Tools: NVIDIA Nsight Systems/Compute

Distributed Systems: SLURM

Publications

Younes Hourri*, M. Mozaffari*, et al. **PATCH: Learnable Tile-Level Pruning of Large Models**. *(under review)*
[Paper] [Code] [Triton Code] (* Equal Contribution)

- Introduced a hybrid tile-level mask that learns to select 2:4 or dense patterns across LLM weights.
- Achieved up to +3.68% accuracy improvement across LLM families and architectures over the state-of-the-art 2:4 pruning method MaskLLM.
- Integrated with the STOICC compiler (built on Triton) to reach 1.18x-1.38× inference speedup on LLaMA-2 7B.

Experience

Analytical Development Intern — **Montréal, QC**
Caisse de dépôt et placement du Québec (CDPQ) — *May 2023 – Aug. 2023*
- Collaborated with internal finance teams to identify process inefficiencies and translate data needs into automated solutions.
- Automated quantitative analyses across 10+ investment portfolios, reducing manual monitoring time.
- Built SQL and Python pipelines to validate financial data, and detect anomalies, reducing manual review time and improving accuracy in internal reporting.

NSERC Research Assistant — **Montréal, QC**
Université de Montréal — *May 2022 – Aug. 2022*
- Implemented and benchmarked embodied agents on the ALFRED benchmark.
- Analyzed performance across observation, planning, and action components in Embodied Instruction Following (EIF).
- Developed a graph-based modeling framework to improve agent task decomposition and reasoning.

Software Developer Intern — **Montréal, QC**
Rockwell Automation — *May 2021 – Aug. 2021*
- Enhanced and debugged legacy software for safety-critical industrial automation systems using C# and .NET.
- Designed and executed system and unit tests to ensure platform stability and compliance.

Invited Talks

Learnable Tile-Level Pruning of Large Models
NVIDIA Research — *Oct. 2025*
- Presented recent research on LLM pruning and compression for efficient inference to research scientists.

Awards & Scholarships

2022: NSERC Undergraduate Student Research Award — Université de Montréal

2020–2024: J.W. McConnell Scholarship — McGill University