Name of supervisor: Ali Ramezani-Kebrya (Faculty in ML at IfI, Research Leader at SFF INTEGREAT, PI at TRUST, Senior Researcher at MishMash, NeurIPS and ICML Area Chair)

Preferred background of candidate(s): Solid computation/programming/math background, excited and motivated to publish in top AI venues

Number of available projects: 2

Preferred project period: Flexible

Project:

GPT-3 has been trained on 10,000 V100 GPUs in a cloud data center with unprecedented communication costs and Carbon footprint. Training GPT-4 and their friends including large multimodal models have pushed this number substantially higher—industry sources estimate 20,000–25,000 NVIDIA A100 or possibly H100 and H200 GPUs were used, running simultaneously for weeks to months. The total emissions of a single training Llama 2 family of models and GPT-3 are estimated to be 539 and 552 tons of CO_2 equivalent, respectively, which is not sustainable. The estimated energy consumption per inference for the SotA LLMs today is typically in the range of 0.2–0.35 watt-hours (Wh) per query, depending on model size, context length, system efficiency, and hardware.

To improve inference efficiency,GPTQ (ICLR23 with 2000+ citations and 2000+ Github Stars) have proposed post-training quantization for pre-trained transformers. The Euclidean distance measure of activations under uncompressed and compressed model parameters used in GPTQ is heuristic without any theoretical guarantees. **Our preliminary results have shown that we can indeed significantly improve GPTQ by defining new measures of discrepancy inspired by efficient estimates of mutual information between those activations.** Such efficient implementations to accelerate inference of large-scale transformers backed up with solid theory obtained by our ongoing Integreat and FRIPRO projects Machine Learning in Real World, Embedded Sufficient Statistics, and On the Intersection between Information Theory and Deep Learning.

My MS students work closely with **stellar PhDs and PostDocs and world-class international collaborators**. Some examples:

Røskva Bjørgfinsdóttir and Brage Eilertsen (**first authors of AAAI paper over a summer project**). First job after graduation: somewhere very decent!

Iman Tabrizian (**co-first author of NeurIPS paper**). first job after graduation: full-time engineer at **NVIDIA**.