Filip Szatkowski | PhD student at WUT working on inference efficiency

Email: fmszatkowski@gmail.com | **Webpage** | **Google Scholar** | **LinkedIn** | **GitHub**

I am a PhD student at Warsaw University of Technology, with research interests in inference efficiency, adaptive computation, and continual learning. I have interned at Amazon and Samsung, and collaborated with researchers from Sapienza University of Rome, Computer Vision Center Barcelona, and Jagiellonian University. My work has been published at top-tier venues including ICML and NeurIPS. I am also active in the Polish ML community through the ML in PL organization. In my free time, I enjoy bouldering, playing guitar, reading, traveling, and learning languages.

EDUCATION

PhD Student at Warsaw University of Technology and IDEAS NCBR (ELLIS Unit Warsaw) 2021 - now
Researching inference efficiency, adaptive computation, and continual learning. Recipient of the NCN Preludium Grant (2025) for early-career researchers, funded by the Polish National Science Centre. Supervisor: prof. Tomasz Trzciński.

Visiting PhD student at Sapienza University, Rome Oct - Nov 2023
Working on adaptive computation algorithms with prof. Simone Scardapane, which resulted in a NeurIPS paper.

MSc in Computer Science, Warsaw University of Technology 2015 - 2020
Final grade: 5.0 (out of max 5.0). My thesis explored the use of deep neural networks for audio signal denoising.

EXPERIENCE

Amazon AWS AI Tübingen, Applied Scientist Intern Sep 24 - Feb 25
Extending speculative decoding with early-exits. I analyzed EAGLE/EAGLE2 architectures and explored their integration with early-exit techniques. Contributed insights that improved the 405B LLM cloud-scale speculative decoding pipeline.

Sages, AI Engineer Apr 21 - Sep 22
OCR and NLP document processing pipelines to automatically adapt electronic documents to accessibility compliance for users.

Samsung R&D Warsaw, NLP Intern Jul 19 - Apr 21
Optimisation of NLP models for edge uses, reproducing MLMs such as BERT, wrapping the models in Java and Android libraries.

PUBLICATIONS

Universal Properties of Activation Sparsity in Modern Large Language Models UniReps, NeurIPS 2025
Filip Szatkowski, P. Będkowski, A. Devoto, J. Dubiński, P. Minervini, M. Piórczyński, S. Scardapane, B. Wójcik
A study revealing patterns of activation sparsity in modern LLMs, offering practical guidelines for model design and acceleration

Failure Prediction Is a Better Performance Proxy for Early-Exit Networks Than Calibration SPIGM, NeurIPS 2025
P. Kubaty, Filip Szatkowski, M. Jazbec, B. Wójcik
A new, failure prediction-based metric for assessing early-exit model quality.

Do LLMs Understand the Safety of Their Inputs? Training-Free Moderation via Latent Prototypes In review, 2025
M. Chrabaszcz, Filip Szatkowski, B. Wójcik, J. Dubiński, T. Trzciński, S. Cygert
Lightweight input moderation method for LLMs leveraging token latent representations.

Improving Continual Learning Performance and Efficiency with Auxiliary Classifiers ICML 2025
Filip Szatkowski, Y. Zheng, F. Yang, T. Trzciński, B. Twardowski, J. van de Weijer
Using adaptive computation with early-exit inference to improve continual learning performance while reducing inference cost.

Exploiting Activation Sparsity with Dense to Dynamic-k Mixture-of-Experts Conversion NeurIPS 2024
Filip Szatkowski, B. Wójcik, M. Piórczyński, S. Scardapane
Converting dense transformer model into granular MoE to accelerate inference through dynamic activation sparsity.

Sparser, Better, Deeper, Stronger: Improving Sparse Training with Exact Orthogonal Initialization ICML 2024
A. Nowak, Ł. Gniecki, Filip Szatkowski, J. Tabor
Novel sparse initialization using Givens rotations, which enables stable training of very-deep networks without normalization.

Adapt Your Teacher: Improving Knowledge Distillation for Exemplar-free Continual Learning WACV 2024
Filip Szatkowski, M. Pyla, M. Przewięźlikowski, S. Cygert, B. Twardowski, T. Trzciński
Improving continual learning techniques based on knowledge distillation.

Zero time waste in pre-trained early exit neural networks Neural Networks 2023
B. Wójcik, M. Przewięźlikowski, Filip Szatkowski, M. Wołczyk, K. Bałazy, B. Krzepkowski, I. Podolak, J. Tabor, M. Śmieja, T. Trzciński
Improving early exit classifiers through cascading and ensembling techniques, where I worked on NLP experiments.

SKILLS

- **Research.** Skilled in evaluating ideas, experimentation, and academic writing.
- **Expertise.** Inference efficiency, activation sparsity, speculative decoding, early exits, continual learning.
- **Programming.** Python, PyTorch and other relevant ML libraries, Bash, MLOps tools, Git.
- **Leadership.** Led several first-author research projects and successfully co-supervised students.
- **Communication.** Experience presenting research at conferences and mentoring.
- **Teamwork.** Collaborated in small research teams and led teams during ML in PL events.
- **Organisation.** Organised MLSS Kraków 2022, 2023, 2025; ML in PL Conference 2023, 2024; and ELLIS EDS 2025.
- **Languages.** Polish (native), English (fluent), Spanish (basic).