Herbert Woisetschläger

Boltzmannstrasse 3, 85748 Garching, Germany | herbert.woisetschlaeger@tum.de

Summary

Recent PhD graduate specializing in a wide range of machine learning applications, including Large Language Models (LLMs), generative AI, and agentic systems, in centralized and distributed settings. Research contributions have been published at major machine learning and systems conferences. Main research outcomes have been integrated into the production-level pre-training library for IBM's open-source LLMs, demonstrating real-world impact and technical excellence. Proven industry track record through prior full-time management consulting experience, focusing on cost efficiency programs and IT transformation projects for enterprise clients.

Work Experience

Research Associate (Full-time position) — Feb 2022 – now
Technical University of Munich — *Munich, Germany*

- Developed an LLM inference routing algorithm for service level guarantees and avg. 2× operating cost savings (published at NeurIPS 2025)
- Created a federated learning (FL) benchmark uncovering major performance challenges for applications at the network edge (published at ACM Middleware)
- Co-authored and published papers at leading systems and ML venues (incl. NeurIPS, ICLR, IJCAI, ACM Middleware)
- Won a competitive EUR 250K+ research grant in the first year of my PhD (< 5% acceptance rate, ID: DIK0446/01, granted by the Bavarian State Ministry of Economic Affairs, Regional Development, and Energy)

PhD Research Intern, LLM Data & Tools — May 2025 – Aug 2025
T.J. Watson Research Center, IBM Research — *Yorktown Heights, US*

- Developed a post-training system for improving the reliability of LLMs in multi-turn tool calling scenarios (incl. data capturing, synthetic data generation in a multi-agent setup, verification, and model fine-tuning)

PhD Research Intern, LLM Data & Tools — May 2024 – Aug 2024
T.J. Watson Research Center, IBM Research — *Yorktown Heights, US*

- Co-developed and integrated an efficient pre-training algorithm into the open source IBM Foundation Model Stack (published at ICLR 2025)
- Contributed a data quality assessment technique (and highly parallel code) to IBM's LLM pre-training dataset GneissWeb (paper under submission)

Management Consultant, CIO/COO Advisory — Mar 2020 – Jan 2022
Capgemini Invent — *Munich, Germany*

- Worked on corporate reorganization projects that lifted a total of EUR 3M+ in net efficiency gains for clients
- Contributed to business development initiatives that generated EUR 2M+ in revenue
- First 10 months of this role as working student in the CEO Office for the Central Europe Executive Committee
- Industrial Focus: Consumer Goods, High Tech Manufacturing, Financial Services

Intern, Strategy & Innovation Consulting — Jul 2019 - Jan 2020
Detecon Inc., Subsidiary of Deutsche Telekom / T-Mobile — *San Francisco, US*

Various Internships, Part-Time Jobs and Paid Project Assignments during Undergraduate Studies — 2015 - 2017
Ceconomy AG, Deutsche Bahn AG, Catholic Diocese Eichstätt — *Germany*

Education

PhD, Computer Science — Feb 2022 – Nov 2025
Technical University of Munich — *Munich, Germany*

- Dissertation: Efficient Federated Learning Systems in Resource-Constrained Environments under Emerging AI Regulation (Advisor: Prof. Hans-Arno Jacobsen)

MSc, Information Systems — Oct 2017 – Nov 2020
Technical University of Munich — *Munich, Germany*

- Focus: Software Engineering, Distributed Systems
- Thesis: Efficient Event Processing in the Context of Non-intrusive Load Monitoring (Advisor: Prof. Hans-Arno Jacobsen)

BSc, Business Administration — Oct 2014 – Sep 2017
Catholic University Eichstätt-Ingolstadt — *Ingolstadt, Germany*

- Focus: Information Management & Microeconomics
- Thesis: Globalization & Child Labor – What is the Link? (Advisor: Prof. Alexander Danzer)

Studies Abroad

Middleware Systems Research Group — May 2025, May 2024, Oct 2023
University of Toronto, Prof. Hans-Arno Jacobsen — *Toronto, Canada*

Data-centered and Secure Computing Research Group — Jul 2022 – Sep 2022
University of Stavanger, Prof. Chunming Rong — *Stavanger, Norway*

Undergraduate Semester Abroad (BComm.) — Jul 2016 – Dec 2016
Stellenbosch University — *Stellenbosch, South Africa*

Skills

- *Programming Languages*: Python (proficient), SQL (proficient), JavaScript (proficient), Bash (proficient)
- *Deep Learning Frameworks & Tools*: PyTorch (incl. DDP, FSDP), Sci-Kit Learn, Jax, Slurm, IBM LSF, Flower FL, Docker, Kubernetes, OpenShift, NVIDIA Nsight profilers for roofline modeling
- *Languages*: English (fluent), French (basic), German (native)

Research Projects & Grants

[G01] **F^{2}IML-BAU – Federated Incremental Deep Learning for Real-Time Video Analytics in the Construction Industry.** — 2023 – 2026, *252k EUR (290k USD)*
Competitive grant provided by the Bavarian Ministry of Economic Affairs, Regional Development and Energy based on a research proposal. Grant No. DIK0446/01. Self-acquired and project lead with budget responsibility.

Technical Program Committee Member/Reviewer

Annual AAAI Conference on Artificial Intelligence (AAAI)

2024, 2025

Conference on Neural Information Processing Systems (NeurIPS)

2024, 2025

International Conference on Artificial Intelligence and Statistics (AISTATS)

2024

International Conference on Learning Representations (ICLR)

2025, 2026

International Conference on Machine Learning (ICML)

2025

Workshop on Federated Learning in the Age of Foundation Models (FL-FM@NeurIPS)

2023

IEEE Transactions on Computational Social Systems

2024

IEEE Transactions on Mobile Computing

2024, 2025

IEEE Transactions on Parallel and Distributed Systems

2024

Invited Talks

A view on language model pre-training and inference: Data curation, sample reweighting, and energy-optimal inference for higher user satisfaction (in English) — Apr 2025, *Remote, Oslo, Norway*
Computer Science Seminar at University of Oslo (invited by Prof. Sabitha Maharjan)

A view on language model pre-training and inference: Data curation, sample reweighting, and energy-optimal inference for higher user satisfaction (in English) — Mar 2025, *Hong Kong, SAR*
Computer Science Seminar at Hong Kong University of Science and Technology (invited by Prof. Hans-Arno Jacobsen)

Federated Learning Prioritäten im Kontext des EU AI Act - Impulse für einen interdisziplinären Diskurs (in German) — Apr 2024, *Bayreuth, Germany*
Bayreuth Law Tech Talks at University of Bayreuth (invited by Prof. Ruth Janal)

Federated Learning Priorities under the EU AI Act (in English) — Apr 2024, *Munich, Germany*
Data Science Reading Group, Carl Zeiss AG (invited by Dr. Thomas Stecher)

Federated Learning Priorities under the EU AI Act (in English) — Mar 2024, *Remote, Cambridge, UK*
Flower Labs & University of Cambridge (invited by Prof. Nicholas Lane)

Föderiertes Lernen: Eine Reise Von Der Theorie Bis Zur Umsetzung (in German) — Mar 2023, *Karlsruhe, Germany*
M3 Conference, Heise Publishing (invited by Rainald Menge-Sonnentag)

Scientific Community Engagement & Outreach

ICML 2025 Workshop on Collaborative and Federated Agentic Workflows — Jul 2025, *Vancouver, Canada*
Workshop co-organizer (https://cfagentic.github.io)

Publications

Please find the latest overview of my research on Google Scholar. This list includes workshop papers.
Key: * = equal contribution.

[P10] **H. Woisetschläger**, R. Zhang, et al. "Dynamically Learned Test-Time Model Routing in Language Model Zoos with Service Level Guarantees". In: *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*. Dec. 2025.

[P09] D. Sow, **H. Woisetschläger**, et al. "Dynamic Loss-Based Sample Reweighting for Improved Large Language Model Pretraining". In: *The Thirteenth International Conference on Learning Representations (ICLR'25)*. Apr. 2025.

[P08] H. Gohari, S. Kadhe, Y. Shah, C. Adam, A. Adebayo, P. Adusumilli, F. Ahmed, N. Baracaldo Angel, S. Borse, Y. Chang, X. Dang, N. Desai, R. Eres, R. Iwamoto, A. Karve, Y. Koyfman, W. Lee, C. Liu, B. Lublinsky, T. Ohko, P. Pesce, M. Touma, S. Wang, S. Witherspoon, **H. Woisetschläger**, D. Wood, L. Wu, I. Yoshida, S. Zawad, P. Zerfos, Y. Zhou, and B. Bhattacharjee. "GneissWeb: Preparing High Quality Data for LLMs at Scale". Under Submission. Feb. 2025.

[P07] **H. Woisetschläger**, A. Erben, et al. "FLEdge: Benchmarking Federated Learning Applications in Edge Computing Systems". In: *Proceedings of the 25th International Middleware Conference (MIDDLEWARE'24)*. Dec. 2024, pp. 88–102.

[P06] R. Zhang, **H. Woisetschläger**, Shiqiang Wang, and HA. Jacobsen. "MESS+: Energy-Optimal Inferencing in Language Model Zoos with Service Level Guarantees". In: *NeurIPS 2024 Workshop on Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning (AFM@NeurIPS'24)*. Dec. 2024.

[P05] **H. Woisetschläger**, A. Erben, Shiqiang Wang, R. Mayer, and HA. Jacobsen. "A Survey on Efficient Federated Learning Methods for Foundation Model Training". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI'24)*. Aug. 2024, pp. 8317–8325.

[P04] **H. Woisetschläger**, A. Erben, et al. "Federated Learning Priorities Under the European Union Artificial Intelligence Act". In: *2nd Workshop on Generative AI and Law (GenLaw'24) in conjunction with the Forty-First International Conference on Machine Learning (ICML'24)*. July 2024.

[P03] H. Woisetschläger*, Simon Mertel*, Christoph Krönke, R. Mayer, and HA. Jacobsen. "Federated Learning and AI Regulation in the European Union: Who is Responsible? - An Interdisciplinary Analysis". In: *2nd Workshop on Generative AI and Law (GenLaw'24) in conjunction with the Forty-First International Conference on Machine Learning (ICML'24)*. July 2024.

[P02] **H. Woisetschläger**, A. Erben, et al. "Federated Fine-Tuning of LLMs on the Very Edge: The Good, the Bad, the Ugly". In: *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning (DEEM'24)*. June 2024, pp. 39–50.

[P01] Jiahui Geng, Zongxiong Chen, Yuandou Wang, **H. Woisetschläger**, Sonja Schimmler, R. Mayer, Zhiming Zhao, and Chunming Rong. "A survey on dataset distillation: approaches, applications and future directions". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI'23)*. Aug. 2023, pp. 6610–6618.