Thomas Vakili

Department of Computer and Systems Sciences
Stockholm University
+46 8 16 16 59
thomas.vakili@dsv.su.se

Blekingegatan 16 lgh 1608
118 56 Stockholm, Sweden
+46 704 38 34 26
https://vakili.science

Academic Background

PhD in Computer Science — Stockholm University 2021–2026

Currently studying towards a PhD degree in computer science at the Department of Computer and Systems Sciences. The position is focused on privacy-preserving machine learning for natural language processing and my main supervisors are Professor Hercules Dalianis and Professor Aron Henriksson.

Visiting PhD Student — Pontificia Universidad Católica de Chile (UC) 2025–2025

A one-month research visit to UC. The purpose of the trip was to initiate a project about Spanish clinical NLP together with Assistant Professor Jocelyn Dunstan and her group.

Visiting PhD Student — CMM Center for Mathematical Modeling 2023–2023

Was awarded a scholarship to visit CMM in Chile for three and a half months. I worked on privacy-preserving machine learning under the guidance of Assistant Professor Jocelyn Dunstan.

MSc in Computer Science and Engineering — KTH Royal Institute of Technology 2013–2020

Studied the five-year engineering program which grants an engineering diploma (Swedish: *civilingenjörsexamen*) and an MSc in computer science. My specialization was natural language processing and machine learning.

Persian Philology — Uppsala University 2016–2017

Between the bachelor and master parts of the engineering program I studied Persian for a year at the Department of Linguistics and Philology. This improved my skills in my second mother tongue and also deepened my knowledge of linguistics.

Selected Industry Experience

Consultant — Netlight Q3 2019–Q1 2021

Netlight is a prestigious IT consultancy firm based in Stockholm and with offices in multiple European cities.

Data Engineer — Spotify (via Netlight) Q4 2020–Q1 2021

Worked as a data engineer in a multinational team at Spotify. The focus of this assignment was helping Spotify develop their data pipeline to improve their curation of core datasets.

Data Scientist — PostNord (via Netlight) Q3 2019–Q4 2020

PostNord is a major logistics company based in the Nordics. I worked as a data scientist and backend developer to build machine learning-based route planning and time series prediction.

Backend Developer — Truecaller (part time) Q2 2018–Q4 2018

Truecaller is a global company providing its users with caller ID and enables mobile payments. I worked with the micro-services and had the opportunity to learn working with Kafka, Cassandra, Play and many other technologies.

Teaching Experience

Thesis Supervision Spring 2022–

I have extensive experience as a supervisor at the master and bachelor level. Since 2022, I have supervised nine master theses and two bachelor theses.

Internet Search Techniques and Business Intelligence Fall 2022–

I am a lecturer and lab examiner in this master-level course. My lecture is about text classification for search engines, and I also examine and develop the labs in the course.

Natural Language Processing Spring 2022–

I am a teacher in the natural language processing course which is given to master students in my department. My responsibilities include holding multiple lectures as well as designing and examining lab assignments and group projects.

Language Technology Fall 2022

I tought a bachelor-level course about language technology. My duties included examining lab assignments as well as giving a lecture about lexical resources.

Digital Business Strategies and Change Management Fall 2021

I was a teaching assistant in a course on digital management in which I was responsible for grading essays and presentations.

Information Retrieval Spring 2019

During my studies at KTH, I was a teaching assistant in a course on information retrieval. I was responsible for grading lab assignments and helping students understand the material.

I also have extensive experience as a supervisor at the master and bachelor level. Since 2021, I have supervised 11 master students and four bachelor students.

Dissemination

Interview: The Secret Behind ChatGPT – How AI Works Sepember 2025

Interviewed for an popular-science article in the Swedish magazine *PC för Alla*. The goal of the article was to explain large language models to a layperson, without resorting to anthropomorphic metaphors. The original title in Swedish is "Hemligheten bakom Chat GPT – så fungerar AI".

Panelist: CALD-pseudo workshop at EACL 2024 March 2024

Participated as a panelist at the CALD-pseudo workshop. The panel discussion centered on pseudonymized data, which is one of my areas of expertise.

Invited speaker: Bravida Developer Summit March 2024

Title: AI och språkteknologi (AI and NLP). Presented the state of the art in NLP and the risks of large language models to a lay audience.

Keynote speaker: RISE Health Data Workshop October 2022

Title: Möjligheter och utmaningar inom Klinisk NLP (Possibilities and Challenges in Clinical NLP). Keynote talk about promising applications for clinical NLP and the challenges associated with handling private data.

Invited speaker: University of Michigan NLP4Health September 2022

Title: Automatic De-Identification at Stockholm University. Presented the latest research about automatically de-identified training data to the NLP4Health group at UMich.

Conference & Workshop Papers

T. Vakili, A. Henriksson, and H. Dalianis. 2025. Data-Constrained Synthesis of Training Data for De-Identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27414–27427.

T. Vakili, M. Hansson, and A. Henriksson. 2025. SweClinEval: A Benchmark for Swedish Clinical Natural Language Processing. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pp. 767–775.

L. Kiefer, J. O. Alabi, **T. Vakili**, H. Dalianis, and D. Klakow. 2025. Instruction-Tuning LLaMA for Synthetic Medical Note Generation in Swedish and English. In *Proceeedings of the 15th International Conference on Recent Advances in Natural Language Processing*, p. 557.

T. Vakili, A. Henriksson, and H. Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical BERT models: Privacy preservation with maintained data utility. *BMC Medical Informatics and Decision Making* 24 (1), p. 162.

C. Aracena, L. Miranda, **T. Vakili**, F. Villena, T. Quiroga, F. Núñez-Torres, V. Rocco, and J. Dunstan. 2024. A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 111–121.

T. Vakili, T. Hullmann, A. Henriksson, and H. Dalianis. 2024. When Is a Name Sensitive? Eponyms in Clinical Text and Implications for De-Identification. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pp. 76–80.

A. Lamproudis, T. O. Svenning, T. Torsvik, T. Chomutare, A. Budrionis, P. D. Ngo, **T. Vakili**, and H. Dalianis. 2023. Using a Large Open Clinical Corpus for Improved ICD-10 Diagnosis Coding. In *AMIA Annual Symposium Proceedings*, p. 465.

T. Vakili, and H. Dalianis. 2023. Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data. In *The 24rd Nordic Conference on Computational Linguistics*.

A. Dolk, H. Davidsen, H. Dalianis, and **T. Vakili**. 2022. Evaluation of LIME and SHAP in Explaining Automatic ICD-10 Classifications of Swedish Gastrointestinal Discharge Summaries. In *Scandinavian Conference on Health Informatics*, pp. 166–173.

O. Bridal, **T. Vakili**, and M. Santini. 2022. Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats. In *LREC 2022 Joint Workshop Language Resources and Evaluation Conference 20-25 June 2022*, p. 49.

O. Jerdhaf, M. Santini, P. Lundberg, T. Bjerner, Y. Al-Abasse, A. Jönsson, and **T. Vakili**. 2022. Evaluating Pre-Trained Language Models for Focused Terminology Extraction from Swedish Medical Records. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pp. 30–32.

T. Vakili, A. Lamproudis, A. Henriksson, and H. Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4245–4252.

T. Vakili, and H. Dalianis. 2022. Utility Preservation of Clinical Text After De-Identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 383–388.

T. Vakili, and H. Dalianis. 2021. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*.

Journal Articles

T. Vakili, A. Henriksson, and H. Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical BERT models: Privacy preservation with maintained data utility. *BMC Medical Informatics and Decision Making* 24 (1), p. 162.

J. Dunstan, **T. Vakili**, L. Miranda, F. Villena, C. Aracena, T. Quiroga, P. Vera, S. Viteri Valenzuela, and V. Rocco. 2024. A pseudonymized corpus of occupational health narratives for clinical entity recognition in Spanish. *BMC Medical Informatics and Decision Making* 24 (1), p. 204.

Theses

T. Vakili. 2025. Preserving the Privacy of Language Models. (PhD thesis, Stockholm University).

T. Vakili. 2023. Attacking and Defending the Privacy of Clinical Language Models. (Licentiate thesis, Stockholm University).

T. Vakili. 2020. A Method for the Assisted Translation of QA Datasets Using Multilingual Sentence Embeddings. (Master thesis, KTH – Royal Institute of Technology).

Language Skills

Swedish – native

English – fluent

Persian – advanced

Spanish – intermediate