

Research and Activities Report

Martin Bača

Supervisor: prof. RNDr. Gabriel Semanišin, PhD.

15.6.2026

Research Direction

Graph-based analysis of training data for efficient domain adaptation of language models

Research Direction

Adaptation of language models to a specific domain, with an emphasis on legal text

Main idea:

Study how the structure of training data influences the fine-tuning process and whether graph-based methods can help us select, weight, or interpret the most relevant training samples.

Research Direction

Representation of training dataset as a **graph**:

- **Nodes:** Individual documents
- **Edges:** Similarity of documents

Similarity of documents is a combination of semantic similarity, shared keywords, legal references and the domain specific relation

Research Problem

Hypothesis: Not all training data has the same impact

Research Problem

- LLMs are trained on vast amount of data
- We often do not know which documents are the most useful
- Some documents are typical representatives of the target domain.
- Some documents are very general.
- Some documents lie between multiple topics or domains.

Research Problem

The research problem is to understand how different types of documents help or hurt transfer learning.

Proposed Approach

Proposed Approach

To build a weighted graph from the training data:

- **Nodes:** documents, paragraphs, or text segments
- **Edges:** similarity or relationship between text units
- **Edge weights:** combination of cosine similarity, shared keywords, legal references, metadata, or other domain-specific signals

Proposed Approach

Graph algorithms can then be used to identify:

- Topic communities,
- Central documents,
- Boundary documents,
- Bridge documents between domains,
- Representative and non-representative training samples.

Role of Contrast

To understand what is specific to legal text, we need contrastive data.

Possible contrasts:

- Legal vs. non-legal (general) text,
- Different law domains (e.g. criminal law vs. civil law),
- Domain-specific documents vs. bridge documents.

Connecting the Graph with the Model

The graph helps us identify different roles of documents:

- Representative documents inside one topic,
- Central documents in a domain,
- Bridge documents connecting multiple topics,
- Boundary documents between legal and non-legal text.

Connecting the Graph with the Model

The research then asks whether these graph roles correspond to model behavior:

- Graph-near documents → close in the model's emb. space?
- Bridge documents → mixed or unstable representations?
- Do legal documents activate different model components than non-legal documents?
- Do different document types cause different LoRA/adaptation weight changes during fine-tuning?

Motivation

Fine-tuning large language models for a specific domain can be highly resource-intensive:

- Large domain-specific datasets,
- Significant GPU memory,
- Long training time,
- High computational and energy cost,
- Repeated experimentation with different training settings.

However, not all training data contributes equally to the adaptation process.

Related Problem: Where Is Domain Knowledge Stored?

A related research problem is whether specific knowledge is distributed across the whole model or concentrated in identifiable components.

This raises several questions:

- Do legal texts activate different parts of the model than non-legal texts?
- Are some layers, neurons, attention heads, or adaptation weights more important for legal-domain behavior?
- Can we identify model components that are strongly associated with a specific topic or domain?

Expected outcomes

Expected outcomes of the research:

- A graph-based methodology for analyzing training data,
- Identification of representative and bridge documents,
- Experiments on different data selection strategies,
- Analysis of how data structure affects transfer learning,
- Comparison between graph-based similarity and model representations,
- Recommendations for more efficient domain adaptation.



Research Activities

Project ADFIR

Project ADFIR

- Research in security incidents analysis
- Applying clustering methods to preprocessed data
- Building dashboards for results visualization
- **Extent:** 40 hours per month

Project ALTAML

Project ALTAML

- Research in 3-partite graph-based similarity of Court Decisions (CD)
- Representing CD as aggregation of embeddings of Law References and Keywords



Publications

Publications

CISIS 2026

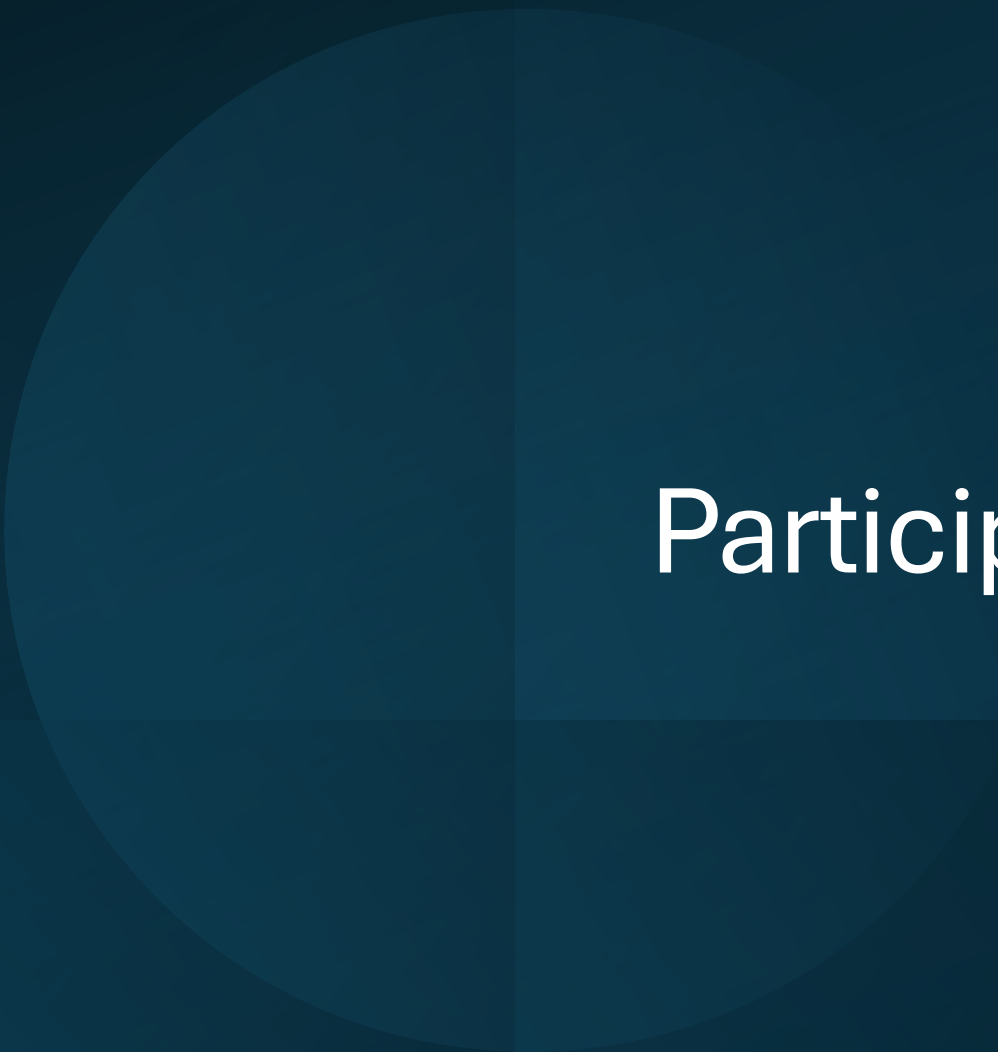
- **Paper title:** Identification of relevant digital forensic traces using time window aggregation and hierarchical clustering
- **Co-authors:** Sophia Petra Krišáková, Pavol Sokol
- **Date:** June 18th – 19th, 2026
- **Location:** Marbella, Spain

Teaching Activities

Teaching Activities

PAZ1b: Programming, Algorithms and Complexity

- 1st year Bachelor students
- **Type of activity:** Seminars/Exercises
- **Extent:** 4 hours per week



Participation in Events

Conferences

DFRWS 2026

- **Date:** March 24th – 27th, 2026
- **Location:** Linköping University, Linköping, Sweden
- **Role:** Attendee (no presentation)
- **Detail:** The Digital Forensics Research Conference in Europe

Workshops

V4AI Workshop in Zlín

- **Date:** April 13th – 14th, 2026
- **Location:** TBU, Zlín, Czech Republic
- **Project:** V4AI: Improving AI Use in Higher Education and Research Across the Visegrad Region
- **Role:** Attendee (no presentation)

Organization of Events

Promotional Events

Faculty DOD

- **Date:** May 2nd, 2026
- **Location:** Jesenná 5, Košice

Scientific Events

ŠVK – Informatics workshop

- **Date:** April 22nd, 2026
- **Location:** Jesenná 5, Košice
- **Role in organization:** Main coordinator in cooperation with Juraj Šebej

Other Events

Night of Mathematicians

- **Date:** February 10th, 2026
- **Location:** Jesenná 5, Košice
- **Role:** Helped organize a school event for students and staff and supervised a quiz station for participants.

FLL – First LEGO League

- **Date:** January 28th, 2026
- **Location:** Jesenná 5, Košice
- **Role:** Judge in category Challenge

Other Activities

Oponent's review of master's thesis

- **Theme:** Semantic and Structural Segmentation of Judicial Decisions
- **Student:** Bc. Attila Papán
- **Supervisor:** RNDr. Peter Gurský, PhD., univerzitný docent

Supervision of bachelor's thesis

- **Theme:** Design and Implementation of an AI Agent for Automated Product Discovery and Recommendation in an E-commerce Environment
- **Student:** Nazar Popovych

- **Theme:** Design and Implementation of an AI Agent for Automated Purchasing in an E-commerce Environment
- **Student:** Damián Kaszycki
- **Consultant:** RNDr. Šimon Horvát, PhD.

Thank you for your attention