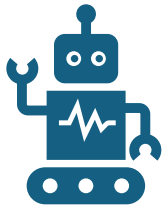
Research and Activities Report

Martin Bača

Research

What is Explainable AI (XAI)?



Explainable AI (XAI) studies methods that make AI systems understandable to humans

Focuses on explaining:

why a model made a decision

which inputs influenced the output

Relevant mainly for complex models:

neural networks

deep learning

large language models (LLMs)

Why XAI Matters (Especially for LLMs)

Enables

- debugging model behavior
- error analysis and bias detection
- trust and accountability

Critical in

- safety-sensitive applications
- regulated domains
- human-in-the-loop systems

For LLMs

- outputs are fluent but not guaranteed to be correct
- explanations help distinguish plausible text from reliable reasoning

My Research: Student-Oriented Explainable LLM Chatbot

Goal: build a chatbot for first-year students to help them

- orient themselves at the university
- get reliable answers to common questions

Architecture

- Retrieval-Augmented Generation (RAG)
- static knowledge base
- real-time search for frequently changing information

Research focus

- explainability of RAG-based LLMs
- understanding why a specific answer was generated
- detecting and communicating uncertainty to users

Problems that have been solved

What has been solved in XAI

LOCAL EXPLANATIONS FOR NEURAL NETWORK PREDICTIONS

ATTRIBUTION OF INPUTS TO OUTPUTS

MODEL-AGNOSTIC EXPLANATION METHODS

FAITHFUL EXPLANATIONS FOR SINGLE PREDICTIONS

Input Attribution & Local Explanations

“Which parts of the input contributed most to this input?

- *Largely solved*

Solutions

- Gradient-based attribution
- Perturbation-based methods
- Token-level relevance scores for LLMs

Model-Agnostic Explanation Methods

“How do we explain models we don’t control internally?”

- *Largely solved*

Solutions

- Local surrogate models
- Sampling-based approximations
- Post-hoc explanation frameworks

Faithfulness vs. Interpretability Trade-off

“Can explanations be both simple and faithful?”

- Largely understood problem

Status

- Trade-off is now **well-characterized**
- No longer treated as a mystery

LLMs Will Always Hallucinate, and We Need to Live With This

Auhors: Sourav Banerjee, Ayushi Agarwal, Saloni Singla

Brief Relevant Paper Presentation

Where Hallucinations Come From

Why Hallucinations Are Inevitable

Learning

Incomplete training data

Open-ended world

Understanding

Retrieval is undecidable

Intent is ambiguous

Generation

No global control

Self-reference & unpredictability

Hallucinations emerge at *every stage* of the LLM pipeline.

Problem Group 1: Knowledge & Understanding

Limits of Knowledge and Understanding

Training data is incomplete

No model sees the whole world

Retrieval is undecidable

No guarantee the model can find a known fact

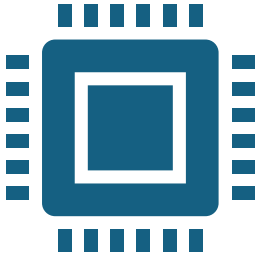
Intent classification is undecidable

Queries are ambiguous by nature

Even before generation, the model may already be wrong

Problem Group 2: Generation Is Uncontrollable

Limits of Knowledge and Understanding



Generation as computation

Halting is undecidable

The model can’t predict its own output

Expressive power

Any token sequence is possible

Includes contradictions & self-reference

Fluency does not imply correctness or consistency.

One-Slide Summary

Key Message of the Paper

Hallucinations are structural

They follow from:

Incompleteness

Undecidability

self-reference

Mitigation ≠ elimination

Teaching Activities

Winter Semester 2025

- **PAZ1a:** Programming, Algorithms and Complexity
  - 1st year Bachelor students
  - **Type of activity**: Seminars/Exercises
  - **Extent**: 4 hours per week

Winter Semester 2025

PAZ1a: Programming, Algorithms and Complexity

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

Participation in Events

Conferences

ITAT 2025

- **Date**: September 26th – 30th, 2025
- **Location**: Telgárt, Slovakia
- **Role**: Speaker
- **Paper title**: Design of Transformer-Based QA Systems for Logistic Data
- **Co-authors**: RNDr. Šimon Horvát, PhD.

Study Visits

V4AI Study Visit in Krakow

- **Date**: October 28th – 30th, 2025
- **Location**: AGH, Krakow, Poland
- **Project**: V4AI: Improving AI Use in Higher Education and Research Across the Visegrad Region
- **Role**: Attendee (no presentation)

Organization of Events

Promotional Events

Day of Open Doors (DoD)

- **Date**: November 7th, 2025
- **Location**: Moyzesova 9

Events/Conferences

AI2SEP Study Visit in Košice

- **Date**: November 4th – 6th, 2025
- **Location**: Jesenná 7, Košice
- **Role in organization**: Technical support (only participated on September 6th, due to illness)

ASEF Innolab 6 workshop

- **Date**: December 2nd - 5th, 2025
- **Location**: Šrobárova 2, Košice
- **Role in organization**: Technical Assistant and other small roles

Events/Conferences

Informatkovica

- **Date**: December 18th, 2025
- **Location**: Medická 4, Košice
- **Target audience**: Employees and students of ÚINF

Aj Ty v IT

- **Date**: November 29th, 2025
- **Location**: Jesenná 7 and Park Angelinum 9
- **Role in organization**: University coordinator - responsible for venue preparation, classroom setup, and coordination with reception/security staff

Thank you for your attention