

# kinit

## V2.1 Správa o výsledkoch komunikácie, diseminácie a exploitácie

Názov projektu	Robustnosť indikátorov dezinformačného obsahu generovaného AI vo viacjazyčnom online priestore
Akronym	RobIndAI
Kód projektu	09I01-03-V04-00059
Začiatok projektu	01. 11. 2024
Trvanie projektu	20 mesiacov

# Obsah

1	Úvod.....	3
2	Plán komunikácie, diseminácie a exploitácie.....	4
2.1	Cieľové skupiny .....	4
2.2	Komunikačné kanály, nástroje a aktivity .....	4
2.3	Otvorený prístup a manažment výskumných dát .....	5
2.4	Využitelnosť a udržateľnosť.....	5
2.5	Merateľné ukazovatele .....	6
3	Výsledky komunikácie, diseminácie a exploitácie .....	7
3.1	Webová stránka projektu.....	7
3.2	Sociálne médiá.....	10
3.3	Vedecké publikácie a prezentácie .....	12
3.4	Repozitáre zdrojových kódov a datasetov .....	16
3.5	Odborné fóra a okrúhle stoly.....	19
3.6	Popularizačné udalosti .....	21
4	Záver.....	23

# 1 Úvod

Tento dokument predstavuje správu o výsledkoch komunikácie, diseminácie a exploítácie projektu RobIndAI. Súčasťou dokumentu je aj plán (Kap. 2), vytvorený v prvotných fázach projektu, ktorý mal zabezpečiť, že relevantné informácie o projekte a jeho výsledkoch budú dostupné cieľovým skupinám v správnom čase a dostupnej forme. Výsledky v Kap. 3 nadväzujú na tento plán s identifikáciou finálnych hodnôt merateľných ukazovateľov, spolu s opisom prípadných zmien oproti pôvodnému plánu.

## 2 Plán komunikácie, diseminácie a exploitácie

Na maximalizáciu dopadu výsledkov a výstupov projektu bol vytvorený plán komunikácie, diseminácie a exploitácie. Tento plán je prispôsobený potrebám jednotlivých cieľových skupín a má zabezpečiť efektívne šírenie informácií a dosahovanie požadovaných výsledkov projektu.

### 2.1 Cieľové skupiny

Komunikačné aktivity sú cieleňé na nasledovné skupiny:

- **Verejnosť** – zvyšovanie povedomia o prínosoch investícií do vedy a výskumu v rámci projektu RobIndAI širokej verejnosti na Slovensku.
- **Policajné authority, mediálni profesionáli, experti v oblasti dezinformácií, regulátori mediálneho priestoru** – komunikácia výsledkov projektu RobIndAI pre ich potenciálne širšie využitie nie len v rámci boja proti dezinformáciám.
- **Výskumníci v oblasti AI** – šírenie nových poznatkov, ktoré budú výstupom projektu RobIndAI v rámci vedeckej komunity.

### 2.2 Komunikačné kanály, nástroje a aktivity

Na efektívne šírenie informácií k uvedeným cieľovým skupinám sú použité:

- **Webová podstránka projektu** – bude obsahovať ucelený opis projektu RobIndAI od prvej fázy projektu a následne bude dopĺňaný o výsledky a výstupy projektu. Web slúži ako podklad pre iné komunikačné aktivity, ale navštevovaný bude aj ako primárny zdroj.
- **Sociálne médiá** – aktuality o projekte budeme šíriť cez účet KInIT (LinkedIn, Facebook, Twitter/X), ako aj cez účty partnerských projektov (predovšetkým VIGILANT), ale aj účty členov riešiteľského kolektívu.
- **Vedecké publikácie a prezentácie** – dosiahnuté výskumné výsledky diseminujeme prostredníctvom publikácií na najlepších fórach (A/A\* CORE konferencie, ako napr. ACL alebo EMNLP, Q1/Q2 časopisy). Šírenie výsledkov podporíme aktívnou účasťou na vedeckých podujatiach.

- **Repozitáre zdrojových kódov a datasetov** – softvérové artefakty, modely a dátové vzorky zverejníme prostredníctvom vhodných portálov: GitHub, HuggingFace, Zenodo.
- **Odborné fóra a okrúhle stoly** – aktívnou účasťou na existujúcich multidisciplinárnych fórach prešírime informácie o projekte a jeho výsledkoch medzi viacerými cieľovými skupinami (workshopy s policajnými autoritami v rámci VIGILANT, konferencie CEDMO, Dezinformácie a Demokracia).
- **Popularizačné udalosti** – prezentácie na udalostiach ako Noc Výskumníkov alebo Deň Európy, príp. popularizačných podcastoch alebo YouTube kanáloch, pomôžu dostať výsledky projektu do povedomia širšej verejnosti, čím zároveň prispejeme k popularizácii vedy a výskumu.

## 2.3 Otvorený prístup a manažment výskumných dát

Pri vedeckých publikáciách bude uprednostňovaná politika otvoreného prístupu, t. j. preferencia takých konferencií a časopisov, ktoré aj v rámci najvyššej kvality (A/A\* CORE konferencie, Q1/Q2 časopisy) umožňujú voľný prístup k publikovaným článkom. Pokiaľ to bude možné, výsledky budú taktiež dostupné na preprintových archívoch, ako napr. arXiv alebo TechRxiv. Projekt bude zverejňovať dáta v súlade s princípmi FAIR (angl. Findable, Accessible, Interoperable, Reusable). Dáta budú ukladané do otvorených repozitárov (napr. Zenodo), pokiaľ to neodporuje komerčným záujmom alebo ochrane duševného vlastníctva.

## 2.4 Využitelnosť a udržateľnosť

KInIT zabezpečí transfer technológií a komercializáciu výstupov projektu RobIndAI prostredníctvom aktívnej účasti v európskych projektoch (VIGILANT, vera.ai, CEDMO, DisAI, AI-CODE) a spolupráce so slovenskými firmami ako Gerulata a TrollWall, pričom výsledky budú zdieľané formou know-how a prezentované na vedeckých konferenciách a sociálnych sieťach. Dlhodobá udržateľnosť je zabezpečená zverejňovaním výskumných dát a modelov vo verejných repozitároch v súlade s princípmi otvorenej vedy, čím sa výstupy sprístupnia širokej výskumnej a odbornej komunite. Duševné vlastníctvo bude chránené autorskými právami a prípadnými patentmi, pričom KInIT bude pri komerčnom využití spolupracovať s partnermi z akademickej a komerčnej sféry, investormi a právnymi odborníkmi, a to vždy v rámci jasne definovaných etických hraníc zabraňujúcich zneužitiu know-how.

## 2.5 Merateľné ukazovatele

V Tab. 1 je prehľadne zachytená štruktúra nástrojov, aktivít, cieľových skupín a kľúčových ukazovateľov (KPI). V porovnaní s opisom projektu v rámci žiadosti sme KPI rozdelili na KPI výstupov, ktoré predstavujú priamo výstupy a výsledky projektu, a KPI dopadu, ktoré predstavujú odozvu cieľových skupín. KPI výstupov predstavujú **hlavné merateľné ukazovatele**, ktoré merajú množstvo práce vykonanej v rámci projektu. KPI dopadu sú **vedľajšie merateľné ukazovatele**, ktoré sú len nepriamo ovplyvniteľné množstvom práce v rámci projektu a závisia aj od externých vplyvov, ako napr. algoritmy sociálnych sietí.

Cieľové skupiny	Nástroj + aktivita	Plán KPI výstupov	Plán KPI dopadu
Policajné authority Mediálni profesionáli Výskumníci v oblasti AI Expertí v oblasti dezinformácií Regulátori mediálneho priestoru Verejnosť	Webová stránka projektu	# <b>webových stránok</b> $\geq 1$ ;	# návštev webového sídla $\geq 1\ 000$ ;
Policajné authority Mediálni profesionáli (Twitter/X, LinkedIn) Výskumníci v AI (Twitter/X, LinkedIn) Expertí v oblasti dezinformácií (Twitter/X, LinkedIn) Verejnosť (Facebook)	Sociálne médiá	# <b>postov</b> $\geq 10$ ;	# impresií $\geq 10\ 000$ ; # interakcií $\geq 200$ ;
Výskumníci v oblasti AI	Vedecké publikácie a prezentácie	# <b>vedeckých článkov</b> $\geq 3$ ; # <b>odborných prezentácií</b> $\geq 5$ ;	# citácií (2 roky po skončení projektu) $\geq 30$ ;
Výskumníci v oblasti AI Expertí v oblasti dezinformácií	Repozitáre zdrojových kódov a datasetov.	# <b>zverejnených modelov</b> $\geq 3$ ; # <b>zverejnených datasetov</b> $\geq 2$ ;	# stiahnutí artefaktov $\geq 10$ ;
Policajné authority Mediálni profesionáli Expertí v oblasti dezinformácií Regulátori mediálneho priestoru	Odborné fóra a okrúhle stoly	# <b>účastí na odborných fórach</b> $\geq 2$ ;	
Verejnosť	Popularizačné udalosti	# <b>účastí na popularizačných udalostiach</b> $\geq 2$ ;	

**Tab. 1** Prehľad plánovaných KPI.

### 3 Výsledky komunikácie, diseminácie a exploitácie

Táto kapitola obsahuje výsledky aktivít a výstupov, spolu s identifikáciou dosiahnutých minimálnych (niektoré KPI dopadu ešte rastú) hodnôt merateľných ukazovateľov, ktoré boli dostupné v čase prípravy tejto správy.

#### 3.1 Webová stránka projektu

V Tab. 2 sú prehľadne zobrazené plánované a dosiahnuté KPI projektu.

Cieľové skupiny	Plán KPI výstupov	Plán KPI dopadu	Dosiahnuté KPI výstupov	Dosiahnuté KPI dopadu
Policajné authority Mediálni profesionáli Výskumníci v oblasti AI Experti v oblasti dezinformácií Regulátori mediálneho priestoru Verejnosť	# <b>webových stránok</b> ≥ 1;	# návštev webového sídla ≥ 1 000;	# <b>webových stránok</b> = 1;	# návštev webového sídla ≥ 6 135;

**Tab. 2** Prehľad KPI pre webovú stránku.

Webová stránka projektu RobIndAI pozostáva z viacerých častí, prioritne z plného a krátkeho profilu projektu. Plný profil projektu RobIndAI ([link](#)) bol umiestnený v rámci webového sídla KInIT-u a obsahuje detailný opis projektu, riešiteľský tím, ako aj výstupy projektu. Boli vytvorené dve jazykové mutácie, v slovenčine a angličtine, periodicky aktualizované o nové výstupy. Tento profil je umiestnený v rámci projektového portfólia KInIT-u, pričom viaceré podstránky (napr. prehľad projektov KInIT-u celkovo, prehľad projektov špecifických tímov, prehľad vybraných projektov riešiteľov projektu RobIndAI) obsahujú krátky profil projektu v rámci webovej stránky (predstavenie projektu RobIndAI prostredníctvom stručného abstraktu). Počet návštev webového sídla v rámci dosiahnutých KPI dopadu zahŕňa počet návštev podstránok KInIT-u obsahujúcich krátky alebo plný profil projektu RobIndAI.

Na Obr. 1 je zobrazená časť obsahu plného profilu projektu RobIndAI, kde v spodnej časti je viditeľné financovanie projektu z Plánu obnovy a odolnosti SR spolu s príslušnými logami vyplývajúcimi z povinnej publicity projektu. Na Obr. 2 je zobrazený príklad obsahu krátkeho profilu projektu RobIndAI v rámci webovej stránky.

KInIT O nás Vyskum Vzdelávanie Careers Novinky Partneri Kontakt Zapojte sa EN

Home > Research > Web & User Data Processing > RobIndAI: Robustnosť indikátorov dezinformačného obsahu generovaného AI vo viacjazyčnom online priestore

# RobIndAI Project

**PROJECT**

**Duration:**  
11/2024 - 06/2026

**Funding agency:** Plán obnovy

**Project type:** Scientific project

**Principal Investigator:**  
Jakub Šimko

## RobIndAI: Robustnosť indikátorov dezinformačného obsahu generovaného AI vo viacjazyčnom online priestore

Projekt RobIndAI bojuje proti zneužívaniu AI na generovanie dezinformačných textov pomocou zvýšenia robustnosti metód detekcie strojovo-generovaného textu. Zameraním RobIndAI je viacjazyčný obsah (najmä jazyky stredo európskeho informačného priestoru) v doméne novinových článkov a obsahu sociálnych médií. RobIndAI je koncipovaný ako rozšírenie projektu VIGILANT (Horizontu Európa), ktorého je KInIT riešiteľom.

Cieľom projektu RobIndAI je výskum metód a modelov umelej inteligencie na zvýšenie robustnosti indikátorov dezinformačného obsahu (z webu a sociálnych médií) s orientáciou najmä na detekciu strojovo generovaného textu. Vzhľadom na schopnosť moderných jazykových modelov generovať vysokokvalitný text v rôznych jazykoch, ktorý je pre človeka nerozoznatelný, je obava zo zneužitia tejto technológie rastúca (napr. medzinárodné dezinformačné kampane). Spôhlivá detekcia strojovo generovaného textu a jeho rozlíšenie od originálneho textu písaného človekom je v tomto ohľade podstatným a veľmi žiadaným indikátorom.

V projekte RobIndAI budú použité metódy a modely zamerané na spracovanie textu a jeho klasifikáciu, fundamentálne multilingválne, prispôbené predovšetkým potrebám stredo európskeho informačného priestoru. V rámci projektu bude realizovaná porovnávacia štúdia efektívnosti existujúcich metód detekcie textu generovaného prostriedkami umelej inteligencie v stredo európskych jazykoch. Štúdia efektívnosti bude zameraná okrem identifikácie použiteľnosti jednotlivých detekčných metód v daných jazykoch aj na vyhodnotenie ich odolnosti voči existujúcim útokom a technikám zabránenia detekcie. **Oproti už prebiehajúcejmu Horizon Europe projektu VIGILANT, prinesie RobIndAI pokročilejšie metódy spracovania textu** (predovšetkým založené na najnovších veľkých jazykových modeloch), **regionálnu a obsahovo-doménovú špecifickosť metód** (spolu s novým datasetom zameraným na náš región), **dôraznejšie porovnanie rôznych možností detekcie** (osobitný model pre každý jazyk vs. spoločný model pre všetky jazyky), ako aj **robustnosť voči novým sofistikovanejším útokom**.

Projekt vychádza z predpokladu, že strojovo-generovaný text pomocou AI modelov má charakteristické vzory, ktoré je možné identifikovať pomocou analytických metód a umelej inteligencie. Z hľadiska orientácie na dezinformácie, projekt predpokladá, že strojovo-generovaný text je pozitívnym indikátorom masovo šírených dezinformácií v online priestore.

RobIndAI využíva moderné metódy strojového učenia, spracovania prirodzeného jazyka a analýzu dát na riešenie problému detekcie strojovo-generovaného textu v online médiách. Kľúčovým faktorom tiež bude získavanie kvalitných tréningových dát a rôznorodý dataset (obohatený o parafrázované texty) pre zabezpečenie účinnosti modelov v reálnom svete.

Financované EÚ NextGenerationEU prostredníctvom Plánu obnovy a odolnosti SR v rámci projektu č. 09I01-03-V04-00059.



Financované  
Európskou úniou  
NextGenerationEU



Manage consent

Obr. 1 Webová stránka s plným profilom projektu RobIndAI.

[O nás](#)
[Výskum](#)
[Vzdelávanie](#)
[Careers](#)
[Novinky](#)
[Partneri](#)
[Kontakt](#)
[Zapojte sa](#)
[EN](#)

Home » Výskum » Projects

# Projects

[NA TEJTO STRÁNKE](#)  
[Scientific projects](#)  
[Innovation projects](#)  
[Industry research projects](#)  
[Common good projects](#)  
[Tools and models](#)

## Scientific projects

**SensAI Project**

**SensAI: Presadzovanie etiky a ľudských práv v umelej inteligencii pri spracovaní jazykov s nízkymi zdrojmi**

01/2025 – 06/2026 →

Projekt SensAI sa venuje stavaniu mostov medzi AI, ľudskými právami a etikou, s osobitným zameraním na jazyky s nízkymi zdrojmi. SensAI stavia na európskom projekte ALFIE a rozširuje jeho etické...

**Exciting news**

EMA is part of the European Malware Analysis Project.

**EMA: Explainable Malware Analysis**

01/2025 – 06/2026 →

EMA is a Recovery and Resilience Plan project aimed at explainable malware analysis. KInIT collaborates with universities in Slovakia – Comenius University and Slovak University of Technology in Bratislava. With...

**NSlant project**

**NSlant: Ako médiá skresľujú informácie v digitálnej dobe**

12/2024 – 02/2025 →

V dnešnej digitálnej ére sa internet stal hlavným zdrojom správ pre masu. Hoci ponúka bezprecedentný prístup k obrovskému množstvu informácií, takýto digitálny nadbytok prináša aj výzvy pre transparentnosť a zvyšuje...

Manage consent

**RobIndAI Project**

**RobIndAI: Robustnosť indikátorov dezinformačného obsahu generovaného AI vo viacjazyčnom online priestore**

11/2024 – 06/2026 →

Projekt RobIndAI bojuje proti zneužívaniu AI na generovanie dezinformačných textov pomocou zvýšenia robustnosti metód detekcie strojovo-generovaného textu. Zameraním RobIndAI je viacjazyčný obsah (najmä jazyky stredoeurópskeho informačného priestoru) v doméne novinových...

**Gepero project**

**GEPERO: Generovanie personalizovaného obsahu vo výskume kvality informácií**

01/2024 – 06/2026 →

Projekt GEPERO sa zameriava na výskum a vývoj nových metód a modelov generovania personalizovaných textov v mnohých jazykoch určených pre výskum kvality informácií na webe a sociálnych médiách. Prioritne je...

**AI-Auditology: Social Media AI Algorithms Auditing**

08/2024 – 06/2025 →

The goal of the AI-Auditology project is to fundamentally change the oversight of social media AI algorithms, like recommender systems or search engines, and their tendencies to spread and promote...

Obr. 2 Webová stránka s krátkym profilom projektu RobIndAI.

## 3.2 Sociálne médiá

V Tab. 3 sú prehľadne zobrazené plánované a dosiahnuté KPI projektu.

Cieľové skupiny	Plán KPI výstupov	Plán KPI dopadu	Dosiahnuté KPI výstupov	Dosiahnuté KPI dopadu
Policajné authority Mediálni profesionáli (Twitter/X, LinkedIn) Výskumníci v AI (Twitter/X, LinkedIn) Experti v oblasti dezinformácií (Twitter/X, LinkedIn) Verejnosť (Facebook)	# postov ≥ 10;	# impresií ≥ 10 000; # interakcií ≥ 200;	# postov = 12;	# impresií ≥ 5 068; # interakcií ≥ 228;

**Tab. 3** Prehľad KPI pre sociálne médiá.

Počtom postov zverejnených na sociálnych sieťach sme prekročili plán, napriek tomu sme v čase písania tejto správy nedosiahli plánovaný počet impresií, ktorý sa ukázal ako veľmi ambiciózný. Počet interakcií približne zodpovedá plánovanému počtu. Je možné, že zverejňovanie príspevkov v neskorších fázach projektu, keď boli dosiahnuté zaujímavé výsledky, nezabezpečili plánovaný ohlas. Hoci plánovaný počet impresií nezodpovedá plánu, cílením prioritne na LinkedIn v anglickom jazyku sme zasiahli úzko špecializovanú odbornú verejnosť, čo má pre projekt RobIndAI vyššiu hodnotu ako vyšší počet náhodných zobrazení na Facebooku (napr. v prípade platenej reklamy).

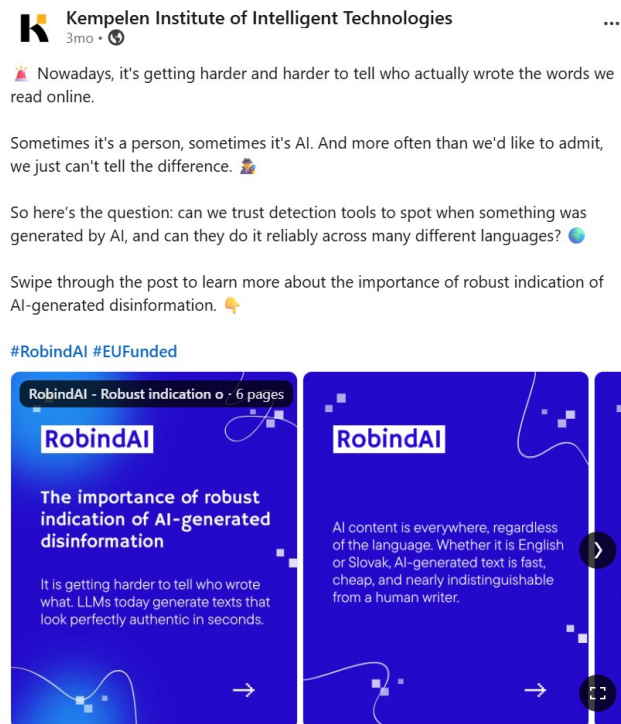
V Tab. 4 sú prehľadne zobrazené publikované príspevky na sociálnych sieťach LinkedIn, X, Facebook a Bluesky, spolu s dátumom uverejnenia a odkazmi na pôvodný príspevok.

Dátum	Jazyk	Krátky opis	Info	Link
10.6.2025	EN	We are proud to introduce our new RobIndAI project! 🤖	post – predstavenie projektu RobIndAI	<a href="#">LinkedIn</a>
10.6.2025	EN	We are proud to introduce our new RobIndAI project! 🤖	post – predstavenie projektu RobIndAI	<a href="#">X</a>
10.6.2025	EN	We are proud to introduce our new RobIndAI project! 🤖	post – predstavenie projektu RobIndAI	<a href="#">Facebook</a>
8.7.2025	EN	🏆 Big news from the RobIndAI project!	post – 1. miesto v zdieľanej úlohe	<a href="#">LinkedIn</a>
8.7.2025	EN	🏆 Big news from the RobIndAI project!	post – 1. miesto v zdieľanej úlohe	<a href="#">Facebook</a>
8.7.2025	EN	🏆 We've joined the Voight-Kampff Generative AI Detection 2025 shared task	post – 1. miesto v zdieľanej úlohe	<a href="#">X</a>

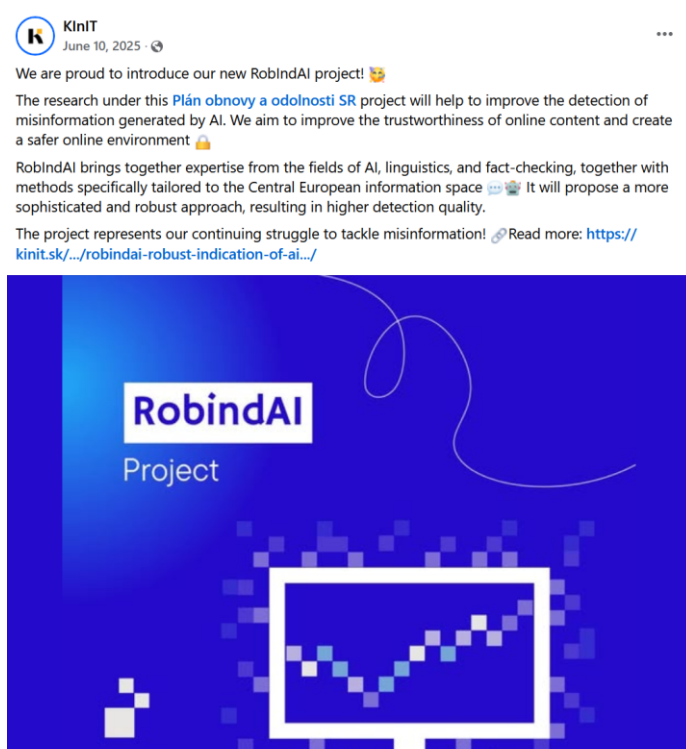
17.2.2026	EN	🔥 We like to think AI text detection is solved. Then the real world proves otherwise.	post – predstavenie benchmarku CEAID	<a href="#">LinkedIn</a>
17.2.2026	EN	Think AI text detection is solved?	post – predstavenie benchmarku CEAID	<a href="#">BlueSky</a>
18.2.2026	EN	Just being accepted to the Findings of ACL 2026 (CORE A* conference)	post – oznámenie Dominika Macka o akceptácii článku CEAID na ACL 2026	<a href="#">LinkedIn</a>
6.3.2026	EN	🔥 Nowadays, it's getting harder and harder to tell who actually wrote the words we read online.	post – robustná detekcia strojovo-generovaného textu	<a href="#">LinkedIn</a>
26.5.2026	EN	Now, the machine-generated text detection benchmark dataset available in an anonymized version	post – oznámenie Dominika Macka o zverejnení datasetu CEAID	<a href="#">LinkedIn</a>
29.6.2026	EN	RobIndAI results presentations at HPC User Day 2026 and Slovak NLP meeting	post - Dominik Macko o finalizácii projektu	<a href="#">LinkedIn</a>

**Tab. 4** Zoznam publikovaných príspevkov pre sociálne médiá.

Na Obr. 3 a Obr. 4 sú zobrazené príklady publikovaných príspevkov týkajúcich sa projektu RobIndAI.



**Obr. 3** Príklad príspevku na sociálnej sieti LinkedIn s animovanou grafikou projektu RobIndAI.



Obr. 4 Príklad príspevku na sociálnej sieti Facebook o predstavení projektu RobIndAI.

### 3.3 Vedecké publikácie a prezentácie

V Tab. 5 sú prehľadne zobrazené plánované a dosiahnuté KPI projektu.

Cieľové skupiny	Plán KPI výstupov	Plán KPI dopadu	Dosiahnuté KPI výstupov	Dosiahnuté KPI dopadu
Výskumníci v oblasti AI	# vedeckých článkov $\geq 3$ ; # odborných prezentácií $\geq 5$ ;	# citácií (2 roky po skončení projektu) $\geq 30$ ;	# vedeckých článkov $\geq 4$ ; # odborných prezentácií = 7;	# citácií (2 roky po skončení projektu) $\geq 16$ ;

Tab. 5 Prehľad KPI pre vedecké publikácie a prezentácie.

Počtom vedeckých článkov aj počtom odborných prezentácií sme prekročili plánovaný počet. Počet vedeckých článkov uvedený vyššie zahŕňa minimálnu hodnotu tohto ukazovateľa dosiahnutú aj pri rovnomernom rozpočítaní príspevku projektu RobIndAI k danej publikácii pri článkoch obsahujúcich poďakovanie viacerým projektom. Teda ak článok obsahuje poďakovanie 2 projektom, do tejto hodnoty je započítaný len podiel 0,5. Celkový počet vedeckých článkov súvisiacich s projektom RobIndAI a obsahujúcich príslušné predpísané poďakovanie je teda vyšší (zoznam v Tab. 6). Počet citácií výstupov projektu RobIndAI je

plánovaný odmerať 2 roky po skončení projektu (vzhľadom na prebiehajúce publikácie najnovších výsledkov projektu v záverečnej fáze), ale napriek tomu už v čase písania tejto správy ku koncu projektu bol počet citácií viac ako polovičný, takže naplnenie tohto plánu je vysoko pravdepodobné.

Fórum	Nadpis	Stav	Link
CLEF 2025	mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection	publikované	<a href="#">link</a>
magazín Computer vydávaný IEEE Computer Society	Beyond Speculation: Measuring the Growing Presence of Large Language Model-Generated Texts in Multilingual Disinformation	publikované	<a href="#">DOI</a>
kapitola knihy vydanej Springer Nature	False Alarm or Real Threat? Trends in GenAI-Mediated Disinformation	publikované	<a href="#">DOI</a>
EMNLP 2025	A Rigorous Evaluation of LLM Data Generation Strategies for Low-Resource Languages	publikované	<a href="#">DOI</a>
ACL 2026	Authorship Attribution in Multilingual Machine-Generated Texts	publikované	<a href="#">link</a>
Findings of ACL 2026	CEAID: Benchmark of Multilingual Machine-Generated Text Detection Methods for Central European Languages	publikované	<a href="#">link</a>
SemEval 2026	mcdok at SemEval-2026 Task 13: Finetuning LLMs for Detection of Machine-Generated Code	akceptované	<a href="#">preprint</a>
EMNLP 2026	Increasing the Robustness of the Fine-tuned Multilingual Machine-Generated Text Detectors	v recenznom konaní	<a href="#">preprint</a>
EMNLP 2026	Interpretable Predictability-Based AI Text Detection: A Replication Study	v recenznom konaní	<a href="#">preprint</a>

**Tab. 6** Zoznam zverejnených vedeckých článkov.

Na Obr. 5 je zobrazený príklad poďakovania projektu RobIndAI v rámci publikovaného vedeckého článku. Aj pri mierne upravenom doslovnom poďakovaní je v každom článku korektne uvedené číslo projektu. V uvedenom príklade bol RobIndAI jediným projektom financujúcim výskum (a teda plný príspevok do uvedeného počtu vedeckých článkov), v rámci „Computational resources“ sú uvedené projekty zabezpečujúce výpočtové kapacity (tieto nie sú započítané do rozpočítania príspevku projektu do počtu vedeckých článkov ani v prípade, že nie sú takto vizuálne oddelené, keďže priamo nefinancujú výskumníkov ani publikáciu).

## Acknowledgments

Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00059.

**Computational resources.** Part of the research results was obtained using the computational resources procured in the national project *National competence centre for high performance computing* (project code: 311070AKF2) funded by European Regional Development Fund, EU Structural Funds Informatization of Society, Operational Program Integrated Infrastructure.


### Obr. 5 Príklad poďakovania projektu RobIndAI za financovanie výskumu v rámci publikovaného článku.

Zoznam v Tab. 7 obsahuje prehľad odborných prezentácií výsledkov projektu RobIndAI, vrátane pozvanej prednášky vedeckej a odbornej komunite zameranej na boj proti dezinformáciám (časť o detekcii dezinformačných textov generovaných AI modelmi), prezentácie na top NLP konferencii EMNLP (CORE A\*) a prezentácií odbornej komunite v rámci CLEF, HPC User Day a Slovak NLP. Napriek viacerým vedeckým článkom publikovaným bez prezentácie (napr. článok v top magazíne Computer, kapitola v monografii vydanej vydavateľstvom Springer Nature), príp. s prezentáciou po ukončení projektu (napr. ACL 2026), bol počet plánovaných odborných prezentácií prekročený.

Fórum	Nadpis	Forma
<a href="#">HPC User Day 2025</a>	Evaluation of Large Language Model Capabilities: Generation and Classification of Text Data	poster
<a href="#">Truth is in the Eyes of the Machines - Symposium</a>	Positive and Negative Aspects of Large Language Models in Tackling Online Disinformation	pozvaná prednáška
<a href="#">CLEF 2025</a>	mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection	online prezentácia
EMNLP 2025	A Rigorous Evaluation of LLM Data Generation Strategies for Low-Resource Languages	poster
HPC User Day 2026	LLMs as Generators and Evaluators	poster
<a href="#">Slovak NLP</a>	Interpretable Predictability-Based AI Text Detection: A Replication Study	poster
<a href="#">Slovak NLP</a>	LLMs as Generators and Evaluators	poster

Tab. 7 Zoznam odborných prezentácií.

Obr. 6 a Obr. 7 zobrazujú príklady prezentácií výsledkov RobIndAI projektu na rôznych fórach.




in W @ + o o

---

Research Education Results News Events Collaboration Work@CWI About

---



Research semester programmes

## Truth is in the Eyes of the Machines - Symposium

This symposium is part of the Research Semester Programme on Misinformation Detection and Countering in the era of Large Language Models.


**Share this page** 
[in](#) [W](#) [@](#) [+](#) [o](#) [o](#)

**When** 9 May 2025 from 9 a.m. to 9 May 2025 6 p.m. CEST (GMT+0200)

**Where** Turing Hall, CWI, Science Park 125, Amsterdam, Netherlands.

**Add** [Add event to calendar](#)

How do misinformation and hate speech fuel and influence each other? How can sustainable and FAIR data (Findable, Accessible, Interoperable and Reusable) be developed to independently investigate misinformation and hate speech? How effective are generative AI models at detecting and mitigating information disorders? These research questions will guide the Research Semester Programme on Misinformation Detection and Countering in the era of Large Language Models. The Programme will be organized along three days, with workshops, keynotes and breakout sessions. A symposium will take place on 9 May at CWI in Amsterdam.



RESEARCH SEMESTER PROGRAMME  
Truth is in the eyes of the machines

Other parts of this Research Semester Programme are on 23 May (Groningen) and 27 October 2025 (Amsterdam). The organizers are [Davide Ceolin](#) (CWI, [Human-Centered Data Analytics](#)), [Anastasia Giachanou](#) (Utrecht University, Department of Methodology and Statistics) and [Tommaso Caselli](#) (University of Groningen, Jantina Tammes School).

**Abstracts, bio's & slides of the speakers** Hide >

---


**Alba Garcia Seco de Herrera** - UNED v

---

**Anna Rogers** - University of Copenhagen v

---

**Dominik Macko** - KinT ^



**Title:** *Positive and Negative Aspects of Large Language Models in Tackling Online Disinformation*

**Slides of the talk**

**Abstract:** As Large Language Models (LLMs) have recently revolutionized many areas, tackling online disinformation is not an exception. In this keynote, Macko will dive into positive as well as negative aspects of LLMs. First, the emergence of LLMs has heightened concerns about automatic generation and spread of disinformation. Macko will show how he identified and analyzed vulnerabilities of LLMs that can be potentially misused to generate false news articles. A specific focus will be given to the ability to personalize generated content towards specific target audiences. He will follow-up with answering a question whether and how such generated texts can be detected in highly multilingual settings, especially by utilizing LLMs.

**Bio:** **Dominik Macko** focuses on robust detection of multilingual machine-generated text, especially in the context of tackling online disinformation. In the past, he was also dealing with anomaly and intrusion detection in IP networks and energy efficiency and security in the Internet of Things environment. He has actively participated in 3 EU-funded research projects, primarily in the research role. Macko has authored or co-authored over 50 publications in scientific journals and conferences, and regularly reviews submissions for renowned international conferences and scientific journals. Formerly, he was a guarantor of the study programme Information Security at the Slovak University of Technology and served as a member of the Scientific Board.

---

**Iryna Gurevych** - TU Darmstadt v

---

**Rodrigo Agerri** - HITZ Center - University of the Basque Country UPV/EHU v

---

**Sophie Morosoli** - AI, Media and Democracy Lab v

---

**Stefano Mizzaro** - University of Udine v

---

**Stephan Lewandowsky** - University of Bristol v

**Programme 9 May**

Time	Subject
09:00-09:30	Registration
09:30-11:00	Session 1 - 2 speakers Chair Anastasia Giachanou <ul style="list-style-type: none"> <li>Rodrigo Agerri</li> <li>Anna Rogers</li> </ul>
11:00-11:30	Coffee break
11:30-13:00	Session 2 - 3 speakers Chair Davide Ceolin <ul style="list-style-type: none"> <li>Iryna Gurevych</li> <li>Stephan Lewandowsky</li> <li>Dominik Macko</li> </ul>

**Obr. 6** Pozvaná prednáška zahŕňajúca výsledky projektu RobIndAI o detekcii dezinformačných textov generovaných AI modelmi.



**Obr. 7** Prezentácia výsledkov projektu RobIndAI na HPC User Day 2026.

### 3.4 Repozitáre zdrojových kódov a datasetov

V Tab. 8 sú prehľadne zobrazené plánované a dosiahnuté KPI projektu.

Cielové skupiny	Plán KPI výstupov	Plán KPI dopadu	Dosiahnuté KPI výstupov	Dosiahnuté KPI dopadu
Výskumníci v oblasti AI Experti v oblasti dezinformácií	# <b>zverejnených modelov</b> ≥ 3; # <b>zverejnených datasetov</b> ≥ 2;	# stiahnutí artefaktov ≥ 10;	# <b>zverejnených modelov</b> = 3; # <b>zverejnených datasetov</b> = 2;	# stiahnutí artefaktov ≥ 30;

**Tab. 8** Prehľad KPI pre repozitáre zdrojových kódov a datasetov.

Počty zverejnených modelov a datasetov zodpovedajú plánu. Okrem týchto repozitárov boli v rámci projektu RobIndAI zverejnené aj repozitáre zdrojových kódov v počte 6, čím projekt prispel k dlhodobej udržateľnosti a využiteľnosti výsledkov. Podľa dosiahnutého KPI dopadu už v čase písania tejto správy môžeme vidieť, že plán bol prekročený. Do tohto počtu sú započítané dokázateľné počty stiahnutí zverejnených datasetov z platformy Zenodo (36 zobrazení, 6 stiahnutí), ako aj počet vyžiadaných prístupov na stiahnutie modelu z platformy HuggingFace (24 akceptovaných prístupov na stiahnutie). K uvedenému dopadu prispel najmä výstup vo forme modelu mdok, ktorý sa umiestnil na prvom mieste v rámci zdieľanej úlohy konferencie CLEF 2025.

V Tab. 9 je zobrazený prehľad zverejnených repozitárov obsahujúcich výstupy projektu RobIndAI.

Typ	Názov	Link
zdrojový kód	CEAID: Benchmark of Multilingual Machine-Generated Text Detection Methods for Central European Languages	<a href="#">GitHub</a>
zdrojový kód	mcdok @ SemEval-2026 Task 13	<a href="#">GitHub</a>
zdrojový kód	mdok	<a href="#">GitHub</a>
zdrojový kód	mdok Detector of Machine-Generated Texts	<a href="#">GitHub</a>
zdrojový kód	Authorship Attribution in Multilingual Machine-Generated Texts	<a href="#">GitHub</a>
zdrojový kód	Increasing the Robustness of the Fine-tuned Multilingual Machine-Generated Text Detectors	<a href="#">GitHub</a>
model	DominikMacko/gemma-2-9b-it-multidomain-robust-mgt-detector	<a href="#">HuggingFace</a>
model	DominikMacko/mdok	<a href="#">HuggingFace</a>
model	DominikMacko/mdok-multiclass	<a href="#">HuggingFace</a>
dataset	CEAID	<a href="#">Zenodo</a>
dataset	CEAID Adversarial Subset	<a href="#">Zenodo</a>

**Tab. 9** Zoznam publikovaných repozitárov.

Na Obr. 8 a Obr. 9 sú zobrazené príklady zverejnených repozitárov modelu (resp. adaptéra modelu, ktorý je možné aplikovať na dostupný predtrénovaný model) a datasetu. V oboch typoch repozitárov sme dbali na uvedenie poďakovania projektu RobIndAI (viditeľné v spodnej časti obrázkov), ako aj na podmienený prístup a tzv. disclaimer, ktoré zabezpečujú akceptáciu využitia zverejnených výsledkov len na nekomerčné výskumné účely (ako vyplynulo z odporúčania etického vyhodnocovacieho procesu vzhľadom na citlivý obsah, využívajúc tiež anonymizáciu v prípade datasetov).

Published May 26, 2026 | Version v1

Dataset Restricted

## CEAID

Macko, Dominik

CEAID is a dataset (described in a paper) for machine-generated text detection benchmark for 7 Central European languages (Croatian, Czech, German, Hungarian, Polish, Slovak, and Slovenian) in two domains (news and social media). It contains 188,098 texts, of which about 23k are human-written and about 165k are generated by 8 multilingual large language models. The dataset has been anonymized to minimize amount of sensitive data by hiding email addresses, usernames, and phone numbers.

If you use this dataset in any publication, project, tool or in any other form, please, cite the paper.

### Disclaimer

Due to data source, the dataset may contain harmful, disinformation, or offensive content. MultiSocial dataset description states that based on a multilingual toxicity detector, about 8% of the text samples are probably toxic (from 5% in WhatsApp to 10% in Twitter). Although we have used data sources of older date (lower probability to include machine-generated texts), the labeling (of human-written text) might not be 100% accurate. The anonymization procedure might not successfully hide all the sensitive/personal content; thus, use the data cautiously (if feeling affected by such content, report the found issues in this regard to dpo[at]unit.sk). The intended use is for non-commercial research purpose only.

### Data Source

The dataset is a subset of a combination of data from MULTITuDev3 (news articles) and MultiSocial (social-media texts). The data contain at least 200 test samples per each domain and class (human vs. machine) for each of the selected Central European languages (Croatian, Czech, German, Hungarian, Polish, Slovak, and Slovenian). The machine texts are generated by 8 LLMs, 6 of which are the same across the two domains (Aya-101, GPT-3.5-Turbo-0125, Mistral-7B-Instruct-v0.2, OPT-IML-Max-30B, v5-Eagle-7B-HF, and Vicuna-13B), one is only in news domain (Llama-2-70B-chat-HF), and one is only in social-media domain (Gemini).

The dataset has the following fields:

- "text" - a text sample,
- "label" - 0 for human-written text, 1 for machine-generated text,
- "multi\_label" - a string representing a large language model that generated the text or the string "human" representing a human-written text,
- "split" - a string identifying train or test split of the dataset for the purpose of training and evaluation respectively,
- "language" - the ISO 639-1 language code identifying the detected language of the given text,
- "length" - word count of the given text,
- "source" - a string identifying the source dataset / platform of the given text,
- "domain" - "news" for news articles from MULTITuDE, "social\_media" for social-media texts from MultiSocial.

Basic statistics:

Language	News Train	News Test	Social Media Train	Social Media Test	All Train	All Test
cs (Czech)	7734	2328	11041	6073	18775	8401
de (German)	7764	2322	21038	9497	28802	11819
hr (Croatian)	7819	2348	14475	5993	22294	8341
hu (Hungarian)	7791	2350	14492	5957	22283	8307
pl (Polish)	7818	2336	16687	6971	24505	9307
sk (Slovak)	7664	2317	0	2026	7664	4343
sl (Slovenian)	7845	2354	0	3058	7845	5412
<b>Total</b>	<b>54435</b>	<b>16355</b>	<b>77733</b>	<b>39575</b>	<b>132168</b>	<b>55930</b>

Files

**Restricted**

The record is publicly accessible, but files are restricted. <a href="https://zenodo.org/account/settings/login?next=https://zenodo.org/records/20391070">Log in</a> to check if you have access.

**Request access**

If you would like to request access to these files, please fill out the form below.

You need to satisfy these conditions in order for this request to be accepted:

In order to share the dataset with you, please agree to the following terms:

1. You will use dataset strictly only for non-commercial research purposes. The request for access to the dataset must be sent from the official and existing e-mail address of the relevant university, faculty or other scientific or research institution (for verification purposes).
2. You will not re-share the dataset with anyone else not included in this request.
3. You will appropriately cite the paper mentioned in the dataset description in any publication, project, tool using this dataset.
4. You understand how the dataset was created and that the "human" label may not be 100% correct.
5. You acknowledge that you are fully responsible for the use of the dataset (data) and for any infringement of rights of third parties (in particular copyright) that may arise from its use beyond the intended purposes. The authors are not responsible for your actions.

**You are currently not logged in.** Do you have an account? [Log in here](#)

Your email address \*

Your full name \*

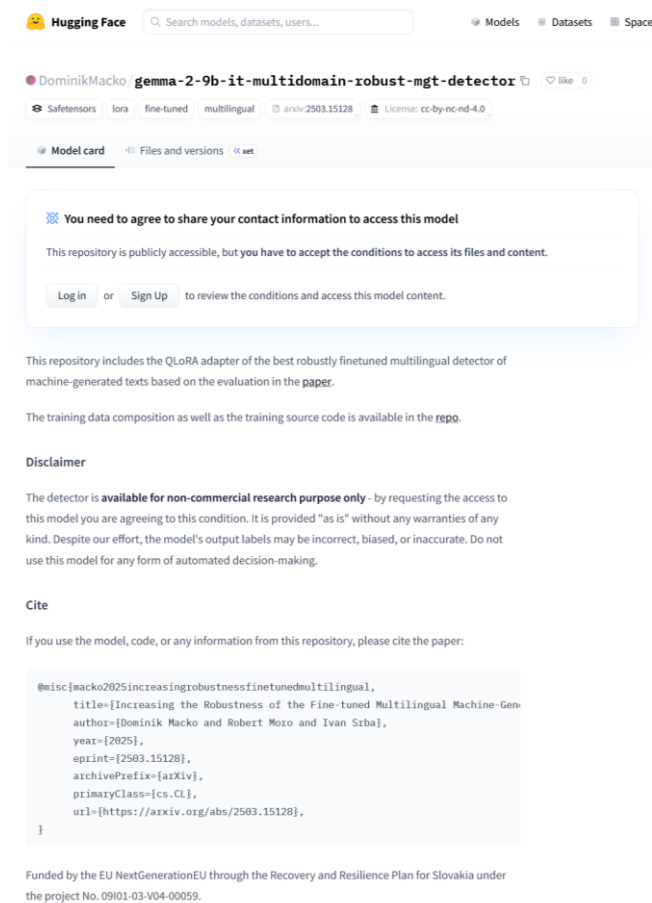
**Request message**

I agree that my full name and email address will be shared with the owners of the record

### Additional details

Funding **Government of Slovakia**  
RobInDAI 0991-01-V04-0009

## Obr. 8 Príklad zverejneného repozitára datasetu na Zenodo.



Obr. 9 Príklad zverejneného repozitára modelu na HuggingFace.

### 3.5 Odborné fóra a okrúhle stoly

V Tab. 10 sú prehľadne zobrazené plánované a dosiahnuté KPI projektu.

Cieľové skupiny	Plán KPI výstupov	Plán KPI dopadu	Dosiahnuté KPI výstupov	Dosiahnuté KPI dopadu
Policajné authority, Mediálni profesionáli Experti v oblasti dezinformácií Regulátori mediálneho priestoru	# účastí na odborných fórach ≥ 2;	-	# účastí na odborných fórach = 5;	-

Tab. 10 Prehľad KPI pre odborné fóra a okrúhle stoly.

Počet plánovaných účastí na odborných fórach bol dvojnásobne presiahnutý, najmä vďaka synergiám s projektami Horizon Europe VIGILANT (diskusia s partnermi v rámci konzorcia) a vera.ai (webinár European Broadcasting Union, stretnutie so zástupcami DG CONNECT

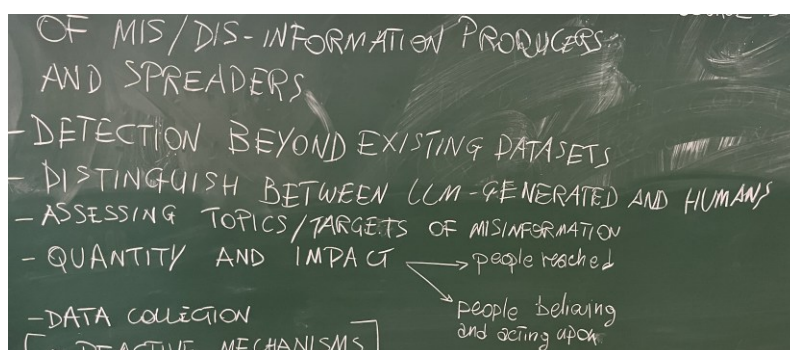
Európskej komisie), ktoré vzbudili záujem odbornej verejnosti najmä v oblasti boja proti dezinformáciám, a tým prispeli aj k exploitačným aktivitám projektu. Robustná viacjazyčná detekcia strojovo-generovaného textu bola v rámci diskusií zastúpená ako jeden z indikátorov kredibility online obsahu.

V Tab. 11 je zobrazený prehľad účastí na odborných fórach a okrúhlych stoloch so stručným opisom danej účasti.

Opis udalosti	Forma
Účasť na online stretnutí HE projektu vera.ai so zástupcami DG CONNECT Európskej komisie, kde boli komunikované aj RobIndAI výsledky robustnej viacjazyčnej detekcie strojovo-generovaného textu.	online stretnutie
Diskusia o možnostiach použitia a spoľahlivosti detekcie strojovo-generovaného textu v slovenčine pre potreby Investigatívneho centra Jána Kuciaka na analýzu neautentických účtov a koordinovaného chovania.	online konzultácia
Účasť na expertnom workshope pre stavbu potenciálneho projektového konzorcia zameraného na boj z dezinformáciami deň pred <a href="#">Truth is in the Eyes of the Machines - Symposium</a> , účastníkmi boli aj mediálni profesionáli z oblasti overovania faktov („factchecking“).	workshop
Účasť na <a href="#">webinári</a> HE projektu vera.ai pre mediálnych profesionálov s opisom výsledkov projektu RobIndAI v oblasti detekcie AI-generovaných dezinformácií.	webinár
Stretnutie KlnIT-u s priemyselným partnerom Gerulata zahrňujúce aj výsledky projektu RobIndAI	prezentácia

**Tab. 11** Zoznam účastí na odborných fórach a okrúhlych stoloch.

Na Obr. 10 a Obr. 11 sú zobrazené príklady z účastí na odborných fórach s využitím výsledkov projektu RobIndAI.



**Obr. 10** Príklad „brainstormovacej“ časti workshopu (predchádzajúceho konferencii Truth is in the Eyes of the Machines – Symposium), inšpirovanej priamo výsledkami projektu RobIndAI v oblasti detekcie LLM-generovaného obsahu.

## Multilingual machine-generated text (MGT) detection

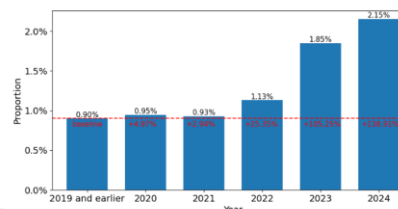
### What we do now

- Deal with English primary focus of NLP
- Benchmarking existing detection methods on multiple languages
  - MULTITuDE (11 languages), MultiSocial (22 languages)
- Increase robustness of detection methods
  - against authorship obfuscation (e.g., paraphrasing, adversarial attacks)
  - transferability to multiple languages
  - transferability to multiple domains (news vs. social)

Train	Test: en	Test: non-en	Difference
en	0.9292	0.6903	↓ 25.7%

### What we plan

- Estimate the prevalence of MGTs in MultiClaim
  - a dataset of multilingual fact-checked claims of social-media texts
  - providing empirical evidence of LLMs misuse for disinformation
  - extension to other datasets (e.g., elections)



12

Relevant papers: [MULTITuDE](#), [Authorship obfuscation](#), [MultiSocial](#), [KInIT at SemEval-2024 Task 8](#), [IMGTB](#), [Robust fine-tuning](#)

KInIT

Obr. 11 Príklady z účasti na prezentácii výskumu KInIT so zobrazenou časťou o plánoch a výsledkoch projektu RobIndAI v oblasti robustnej detekcie generovaného obsahu.

## 3.6 Popularizačné udalosti

V Tab. 12 sú prehľadne zobrazené plánované a dosiahnuté KPI projektu.

Cieľové skupiny	Plán KPI výstupov	Plán KPI dopadu	Dosiahnuté KPI výstupov	Dosiahnuté KPI dopadu
Verejnosť	# účastí na popularizačných udalostiach ≥ 2;	-	# účastí na popularizačných udalostiach = 2;	-

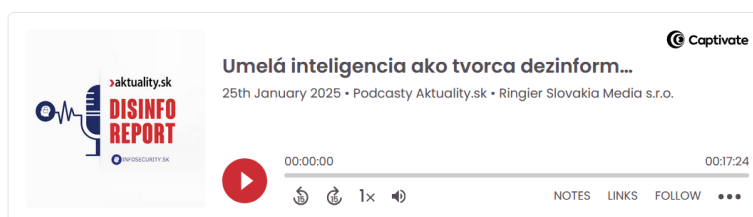
Tab. 12 Prehľad KPI pre popularizačné udalosti.

Počet účastí na popularizačných udalostiach bol naplnený podľa plánu. V rámci projektu RobIndAI sme využili možnosť popularizácie prostredníctvom podcastu zameraného špeciálne na dezinformácie, dostupného širokej verejnosti prostredníctvom viacerých kanálov (napr. iTunes, Spotify, web, Facebook). Okrem toho sme prezentovali výsledky výskumu projektu RobIndAI vo forme interaktívneho dema detektora strojovo-generovaného textu v rámci účasti na popularizačnej udalosti Európska noc vedy 2025.

V Tab. 13 je zobrazený prehľad účastí na popularizačných udalostiach so stručným opisom danej účasti. Obr. 12 a Obr. 13 zobrazujú informáciu o účasti na popularizačných udalostiach, kde boli okrem iného referované aj výsledky projektu RobIndAI v oblasti detekcie strojovo-generovaného textu.

Opis udalosti	Forma
Zverejnenie výsledkov projektu RobIndAI v oblasti detekcii dezinformačných textov generovaných veľkými jazykovými modelmi v rámci epizódy <a href="#">podcastu Disinfo Report</a> zverejneného na webe <a href="#">aktuality.sk</a> , ako aj <a href="#">infosecurity.sk</a>	podcast
Zverejnenie výsledkov projektu RobIndAI vo forme interaktívneho dema detektora viacjazyčného strojovo-generovaného textu v rámci udalosti <a href="#">Európska noc vedy 2025</a> .	demo

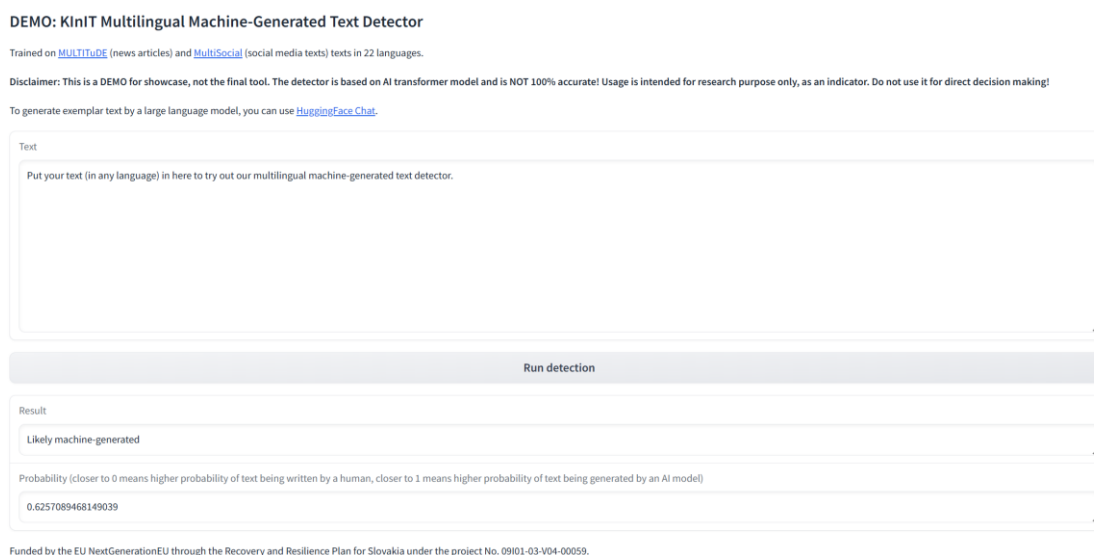
**Tab. 13** Zoznam účasti na popularizačných udalostiach.



Ako tieto ochranné mechanizmy fungujú? Aké sú spoľahlivé a ako ťažko sa dajú obísť? Ktoré jazykové modely sú bezpečné a ktoré majú čo dobiehať? Aké sú vôbec schopnosti umelej inteligencie tvoriť nové dezinformácie na požiadanie a čo ak chceme, aby ich ušila na mieru konkrétnej cieľovej skupiny ľudí?

Aj o tom v novej epizóde podcastu Disinfo Report projektu Infosecurity.sk hovorí Tonka Zsigmondová s výskumníkom Kempelenovho inštitútu inteligentných technológií Dominikom Mackom.

**Obr. 12** Príklad účasti na popularizačnej udalosti prostredníctvom podcastu Disinfo Report aj o výsledkoch projektu RobIndAI v oblasti detekcie dezinformačných textov vygenerovaných AI modelmi.



**Obr. 13** Príklad účasti na popularizačnej udalosti Európska noc vedy 2025 s výsledkami projektu RobIndAI vo forme dema detektora strojovo-generovaného textu.

## 4 Záver

Projekt RobIndAI úspešne naplnil, a vo väčšine oblastí aj výrazne prekročil, všetky plánované hlavné merateľné ukazovatele (KPI výstupov) komunikácie, diseminácie a exploitácie. Niektoré vedľajšie merateľné ukazovatele (KPI dopadu) bude potrebné vyhodnotiť 2 roky po ukončení projektu (napr. citácie), ale už aktuálne hodnoty naznačujú vysokú pravdepodobnosť ich naplnenia. Celkovo projekt RobIndAI dosiahol svoje komunikačné a diseminačné ciele a vytvoril základ pre dlhodobé využitie výsledkov výskumu odbornou komunitou, médiami i priemyselnými partnermi.