Financováno Evropskou unií
NextGenerationEU

NÁRODNÍ PLÁN OBNOVY

MINISTERSTVO PRŮMYSLU A OBCHODU

Využití nástrojů umělé inteligence v boji s dezinformacemi v médiích

Aplikovaný výzkum pro vývoj nástrojů

CEDMO 2.0 NPO

MPO 60273/24/21300/21000

Studie č. 2

Datum: 17. 12. 2025

Editoři: Jiří Matas, Jan Čech, Ondřej Bojar, Tomáš Kroupa, Tomáš Pajdla

Autoři výsledku: Tomáš Kroupa, Jakub Mareček, Jan Drchal, Branislav Bošanský, Karel Horák, Tomáš Pevný, Jan Čech, Vojtěch Franc, Giorgos Tolias, Ondřej Chum, Radek Mařík, Tomáš Pajdla, Jan Šedivý, Jan Hůla, Peter Polák, Dominik Macháček, Ondřej Bojar, Ondřej Dušek, Patrícia Schmidtová, Filip Rechtorík, Danil Semin, Ondřej Plátek, Miroslav Hrabal, Josef Jon

Využití nástrojů umělé inteligence v boji s dezinformacemi v médiích 1
Aplikovaný výzkum pro vývoj nástrojů 1
CEDMO 2.0 NPO 1
MPO 60273/24/21300/21000 1
Studie č. 2 1
SA01 Modely důvěryhodnosti v sítích 4
SA02 Odpovídání na otázky v otevřené doméně respektující kauzalitu 4
SA03 Extrakce faktů a názorů z veřejně dostupných zdrojů 5
SA04 Robustní vysvětlitelné učení pro detekci generativní AI 10
SA04.1 Optimalizace detektorů pro velmi nízké počty falešně pozitivních syntetických dat 11
SA04.2 Studie robustnosti existujících detektorů syntetického obsahu na běžné obrázkové modifikace 12
Citlivosti velkých obrazových modelů na obrazové úpravy 12
Adversární učení pro zvýšení robustnosti detektorů na obrazové úpravy 13
Studie posunu distribuce ve struktuře JPEG souborů 13
SA04.3 Ztrátová a bezeztrátová zero-shot detekce syntetických obrázků 14
Ztrátová zero-shot detekce syntetických obrázků 15
Bezeztrátová zero-shot detekce syntetických obrázků 16
SA05 Detekce generovaných videí s pomocí rychlých pohybů v obličeji 18
Detekce syntetické a deepfaky řeči 20
SA06 Detekce generovaných obrázků z vlastností obličejů 22
Výstupy 22
SA07 Detekce kopií a modifikací známých obrázků a videí 25
SA08 Detekce manipulace veřejného mínění v exteriéru 29
SA09 Nástroje pro moderování diskuzí s přítomností trolů 30
SA10 Nástroje pro analýzu vysvětlitelnosti a robustnosti velkých jazykových modelů 33
Guardrail model na bázi BERTu 35
Doladění menších jazykových modelů (LLM) pro provoz v místním prostředí 36
Generování kódu pomocí 10B modelu 36
SA11 Nástroje pro zvýšení produktivity zpravodajských médií 36
Pokrok v Retrieval-Augmented Generation (RAG) 37
Chatbot ADAM 38
Optimalizace kontextuální návaznosti 38
Vlastní embeddingové modely: 38
On-premise infrastruktura (NVIDIA Spark) 39
Guardrail for a Coding LLM 39
Safe Coding LLM 39
NSS 39
MHMP Klasifikátor 39
ADAM zvýraznění textu odpovědi 39
SA12 Nástroje pro analýzu velkých textových korpusů v médiích 40
SA13 Nástroj pro analýzu konzistence generovaných textů 41

SA14 Nástroje pro monitoring hlasových zpráv ve více jazycích 43
SA15 Nástroje pro detekci faktických chyb ve výstupech LLM 45
SA16 LLM pro podporu mediální gramotnosti v prostředí českého školství 46
Odborné publikace dedikované projektu CEDMO 2.0 NPO **55**
Publikované články 55

SA01 Modely důvěryhodnosti v sítích

Pracovali jsme na vylepšování algoritmů pro modelování reputace v sítích se vzájemnou komunikací mezi uživateli. Zpočátku jsme se snažili najít inspiraci ve stávající literatuře, jelikož v tomto oboru bylo prozkoumáno mnoho různých přístupů. Nakonec jsme se však vrátili k modifikaci námi vyvinutého přístupu *SHAPETrust* (viz Bandhana, A. Kroupa, T. and García, S. *Trust in Shapley: A Cooperative Quest for Global Trust in P2P Network.* Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, 2024. pp. 132-140). Cílem bylo zvýšení jeho odolnosti vůči škodlivým uživatelům. V této souvislosti jsme navrhli úpravu využívající Owenovu hodnotu namísto Shapleyho hodnoty. Owenova hodnota se vyznačuje prací s koaliční strukturou. My jsme zvolili takovou koaliční strukturu, která redukuje počet permutací použitých při výpočtu tak, že účastníci, kteří obdrží nulovou důvěru alespoň od jednoho dalšího účastníka, mají nižší šanci získat vysoké hodnoty globální důvěry. Současně jsme upravili i definici externí části koaliční hry, aby tento způsob výpočtu nemohl být zneužit ke škodlivým manipulacím.

Dalším bodem našeho výzkumu byla optimalizace výpočtu SHAPETrust v decentralizovaných sítích. Z původního článku jsme věděli, že interní hru lze vypočítat decentralizovaně. Problém však představovala externí hra, jejíž definice obsahuje minimum a je výpočetně náročnější. Tento problém jsme vyřešili nalezením uzavřené formy výpočtu, která zásadně snižuje celkovou složitost SHAPETrust a tím výrazně zvyšuje jeho škálovatelnost.

Vytvořili jsme balíček v programovacím jazyce Julia, který obsahuje implementace všech výše uvedených úprav a možností výpočtu. Balíček je dostupný na adrese https://github.com/ruttejan/ShapeTrust.jl. Informace o instalaci a použití balíčku je k nalezení v README souboru balíčku. Ke zprávě připojujeme rovněž náš článek o této problematice, který připravujeme k publikaci.

SA02 Odpovídání na otázky v otevřené doméně respektující kauzalitu

Pracovali jsme na třech “přípravných” úkolech pro učení s podmínkami kauzálních modelů:

- odhadech kauzalních modelů. Jednak na metodě ExDBN, která odhaduje kauzální modely ve smyslu dynamických Bayesovských sítí (https://arxiv.org/abs/2410.16100), jednak na jejím rozšíření o práci s matoucími proměnnými (confounders). Tuto práci vede dr. Aleš Wodecki a dr. Pavel Rytíř.

- teoretických otázkách algoritmů pro učení parametrů hluboké neuronové sítě s omezeními. To jde formalizovat jazykem tzv. krotké geometrie, kterou studujeme s dr. Allenem Gehretem (dříve ve Vídni a LA, https://www.mat.univie.ac.at/~agehret/)
- praktických implementacích, zejména v rámci knihovny train (https://github.com/humancompatible/train), která implementuje zatím první algoritmus (založený na stochastické aproximaci pro účelovou funkci i omezení). Tuto práci vede Andrii Kliachkin.

Předpokládáme, že i v Q1 2025 budeme pokračovat na těchto přípravných úkolech. Dr. Wodeckého nahradí dr. Jana Lepšová, která se zaměří na práci s Andrii Kliachkinem.

Pokračovali jsme v práci na třech "přípravných" úkolech pro učení s podmínkami kauzálních modelů:
● odhadech kauzalních modelů se zavádějícími proměnnými ("confoundingem"). Zejména jsme vyvinuli metodu ExMAG, která odhaduje kauzální modely se zavádějícími proměnnými (https://arxiv.org/abs/2503.08245). Článek jsme odeslali na konferenci NeurIPS, která přijímá jen malé procento zaslaných prací.
● teoretických otázkách algoritmů pro učení parametrů hluboké neuronové sítě s omezeními. To jde formalizovat jazykem tzv. krotké geometrie, kterou studujeme s dr. Allenem Gehretem a dr. Gilles Bareillem. Článek připravujeme pro odeslání do SIAM Journal on Optimization.
● praktických implementacích, zejména v rámci knihovny train (https://github.com/humancompatible/train), která implementuje zatím čtyři algoritmy (založené na stochastické aproximaci pro účelovou funkci i omezení). Tuto práci vede mgr. Andrii Kliachkin a dr. Jana Lepšová. Související článek jsme odeslali na konferenci NeurIPS, která přijímá jen malé procento zaslaných prací.
Předpokládáme, že i v Q3 2025 budeme pracovat na učení s omezením z kauzalních modelů, což povede dr. Jana Lepšová a Andrii Kliachkin.

SA03 Extrakce faktů a názorů z veřejně dostupných zdrojů

Pracovali jsme na dvou základních úkolech: na přípravě datové sady pro extrakci faktů a na nástroji PromptOpt (https://github.com/aic-factcheck/prompt_opt), který výrazně zrychluje přípravu promptů pro velké jazykové modely.

Vytvořili jsme vlastní datovou sadu orientovanou na extrakci pojmenovaných entit s dodatečnými informacemi (např. role pro osoby), událostí, jejich provázání s entitami a extrakci časového určení platnosti. Vzhledem k náročnosti byla tato datová sada tvořena iterativně – prvních několik samplů bylo vyrobeno ručně a na jejich základě byl metodami indukce promptu (viz níže) vytvořen prompt pro, v té době nejnovější "uvažující" (reasoning) modely od společnosti OpenAI (varianty modelů o1 a novějšího o3), které jsme použili k vygenerování anotací pro další vzorky. Tyto syntetické vzorky byly následně opraveny anotátorem a celý proces se opakoval. Díky využití modelů jsme byli schopni vytvořit kvalitní datovou sadu s výrazně nižšími nároky na práci lidských anotátorů.

Pro extrakci citací jsme ve výchozí fázi jako základní datovou sadu zvolili dataset SiR (Sources in iRozhlas) (https://ufal.mff.cuni.cz/anotace-citacnich-frazi-v-datech-irozhlas/sir-10) vytvořený na ÚFAL.

Tento dataset je rozsáhlý a velmi kvalitní, bohužel obsahuje pouze informace o osobě (zdroji) a frázi v textu určující typ citace (prohlásil, uvedl, apod.), neobsahuje však samotné citace a jejich význam.

Z tohoto důvodu jsme navrhli anotační metodologii a vytvořili první verzi datasetu CiteCS, který obsahuje informace o mluvčím, zdroji, ale i přesné znění citace, celý zdrojový text a normalizovanou verzi citace. Normalizovaná citace doplňuje kontext, tedy mimojiné nahrazuje zájmena. Jako příklad normalizovaná verze reálné citace *"Zde se jedná o zachování funkčního systému zaměstnaneckých benefitů."* může vypadat takto: *"U bodu číslo dva zákona o jednotném měsíčním hlášení zaměstnavatele se jedná o zachování funkčního systému zaměstnaneckých benefitů."* Námi vytvořený dataset má zatím menší rozsah 45 vzorků, ten je však pro metody optimalizace promptů i jejich přibližnou evaluaci dostatečný.

Naše úloha staví na opakovaném použití velkých jazykových modelů (LLMs). Vstupem tohoto typu modelů je prompt složený s instrukční části a vstupních dat (query), které jsou v našem případě tvořeny vstupním textem (novinářský text, zápis jednání poslanecké sněmovny, apod.). Pro úlohy, které řešíme je běžně třeba tvořit velké množství promptů. Ruční návrh a ladění promptu je velmi zdlouhavý proces s nejasným výsledkem a proto jsme se tuto část rozhodli automatizovat. Následující obrázek ilustruje úlohu extrakce:

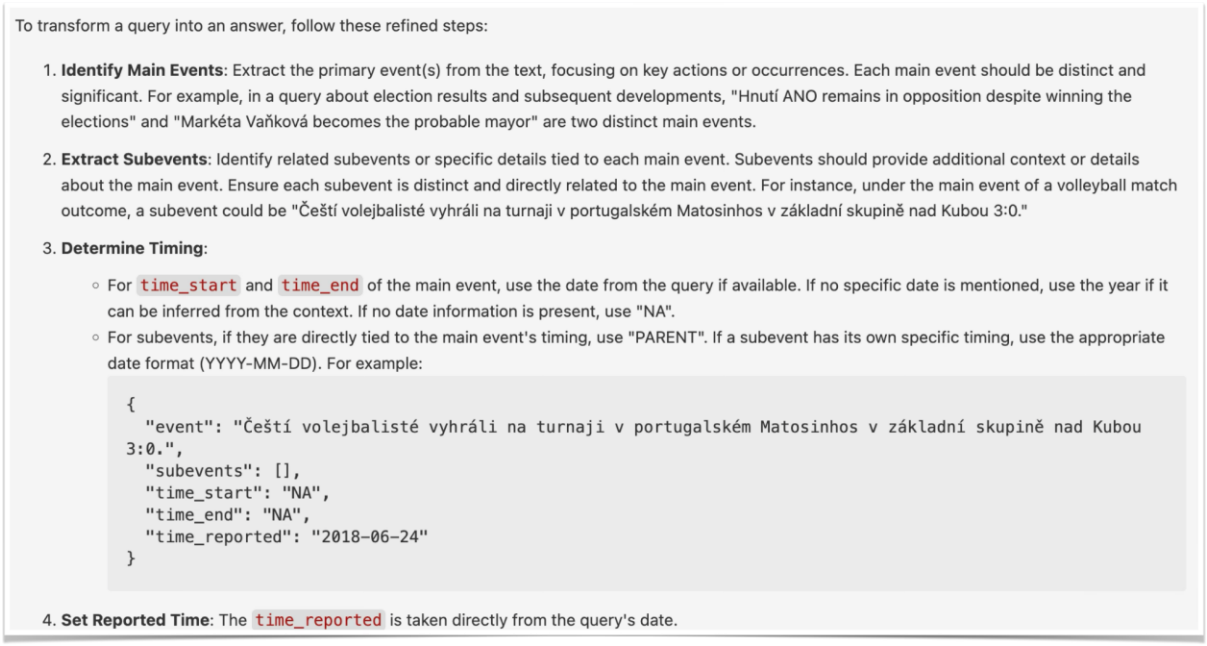
Nejprve jsme provedli analýzu dostupných řešení, nicméně vzhledem k jejich limitacím jsme se v rámci projektu soustředili na vývoj nástroje PromptOpt pro automatizované tvoření a ladění promptu na základě malých trénovacích dat (cca jednotky až nízké desítky vzorků, více vzorků je pak používáno pro testování). Pro účely vývoje a ověřování funkcionality jsme ručně vytvořili několik datových sad souvisejících s extrakcí znalostí z textů (extrakce tvrzení, informací o osobách a další) a také dataset CiteCS (viz výše).

Navrhli a implementovali jsme novou verzi PromptOpt (předchůdce nástroje byl členy týmu vyvíjen v rámci AIC, ČVUT):

- Stávající implementace PromptOpt nevyhovovala účelům tohoto projektu, kde se zaměřujeme specificky na extrakci znalostí z textů (mj. osoby a jejich role, organizace, lokace, události a jejich časová rozpětí, citace apod.). Důvodem je striktní orientace na strukturovaný výstup, pro kterou bylo třeba implementovat robustní metody omezování formátu generovaného výstupu (tzv. constrained generation).
- Dalším rozšířením PromptOpt, které bylo třeba řešit pro tento projekt, je podpora vícestupňových pipeline: experimenty na různých datech ukázaly, že není vhodné optimalizovat prompt pro jednu komplexní úlohu jako celek, místo toho je lepší úlohu rozdělit na sadu dílčích úloh, pro něž jsou prompty optimalizovány nezávisle.
- PromptOpt řeší úlohu black box optimalizace: v rámci projektu jsme porovnali jak optimalizační metody typu hill climber, tak meta-heuristiky (zejména evoluční algoritmy). Ukázalo se, že pro úlohy řešené v rámci tohoto projektu jsou metody lokálního prohledávání dostatečné – nejdůležitější je nalézt kvalitní výchozí řešení tzv. indukcí promptu.
- V rámci projektu jsme se zaměřili zejména na návrh kvalitních meta-promptů pro generování výchozích řešení (zmíněná indukce) a pro jejich iterativní zlepšování (modifikující operátory typu mutace, prohledávající okolí aktuálního řešení).
- Navrhli jsme rovněž meta-prompty pro hodnocení kvality generovaných výstupů (model-based metriky porovnávající predikci LLM s “gold” výstupem trénovacích/testovacích dat.) Numerické hodnocení kvality výstupů příslušného generovaného promptu je zásadním stavebním kamenem potřebným pro optimalizaci generovaných promptů. Protože mají výstupy mnoha úloh extrakce znalostí a citací strukturovanou formu (reprezentovanou typicky jako JSON), vyvinuli jsme rovněž metriku ObjectAligner, numericky porovnávající dvě struktury typu JSON. ObjectAligner je konfigurovatelná metrika, přičemž konfigurace je řešena jako rozšíření zaběhlého formátu JSONSchema.
- V rámci PromptOpt jsme úspěšně testovali přístupy typu Chain of Thought (CoT), kdy model před generováním samotné odpovědi produkuje tokeny úvah nad řešením – bylo ukázáno, že tento přístup vede na mnoha úlohách k výrazným zlepšením. Signifikantních zlepšení jsme dosáhli s použitím nejnovějších jazykových tzv. *reasoning* modelů, které jsou pro přístup podobný CoT přímo laděny. Jednalo se zejména o destilované verze modelů DeepSeek modely QwQ, Qwen3 a GPT-OSS.

Nejprve jsme provedli experimenty se zaměřením na datasety extrakce zdrojů. Následně jsme testovali generování promptů pro datasety SiR a CiteCS. Testovali jsme více modelů s různým rozsahem podpory českého jazyka zejména modely Llama 3.1-3.3, Cohere Command R+, Cohere Aya Expanse, MistralAI Ministral, Google Gemma3, Qwen, QwQ a také otevřené modely společnosti OpenAI. Porovnávali jsme tzv. reasoning (thinking) modely s běžnými LLM. Testovali jsme i komerční modely společnosti OpenAI.

Následující obrázek ukazuje úvodní část promptu pro extrakci událostí získaného nástrojem PromptOpt pouze na základě malé trénovací sady:



To transform a query into an answer, follow these refined steps:

1. **Identify Main Events**: Extract the primary event(s) from the text, focusing on key actions or occurrences. Each main event should be distinct and significant. For example, in a query about election results and subsequent developments, "Hnutí ANO remains in opposition despite winning the elections" and "Markéta Vaňková becomes the probable mayor" are two distinct main events.
2. **Extract Subevents**: Identify related subevents or specific details tied to each main event. Subevents should provide additional context or details about the main event. Ensure each subevent is distinct and directly related to the main event. For instance, under the main event of a volleyball match outcome, a subevent could be "Čeští volejbalisté vyhráli na turnaji v portugalském Matosinhos v základní skupině nad Kubou 3:0."
3. **Determine Timing**:
   - For `time_start` and `time_end` of the main event, use the date from the query if available. If no specific date is mentioned, use the year if it can be inferred from the context. If no date information is present, use "NA".
   - For subevents, if they are directly tied to the main event's timing, use "PARENT". If a subevent has its own specific timing, use the appropriate date format (YYYY-MM-DD). For example:

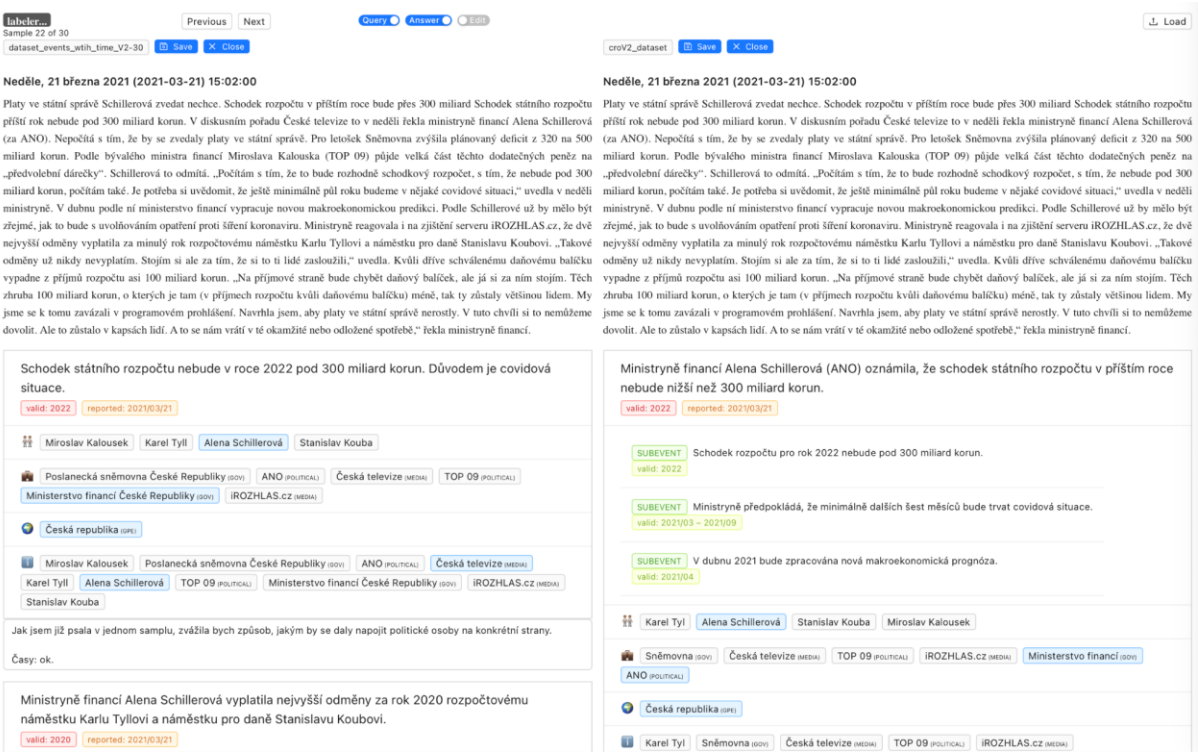
```
{
  "event": "Čeští volejbalisté vyhráli na turnaji v portugalském Matosinhos v základní skupině nad Kubou
3:0.",
  "subevents": [],
  "time_start": "NA",
  "time_end": "NA",
  "time_reported": "2018-06-24"
}
```

4. **Set Reported Time**: The `time_reported` is taken directly from the query's date.

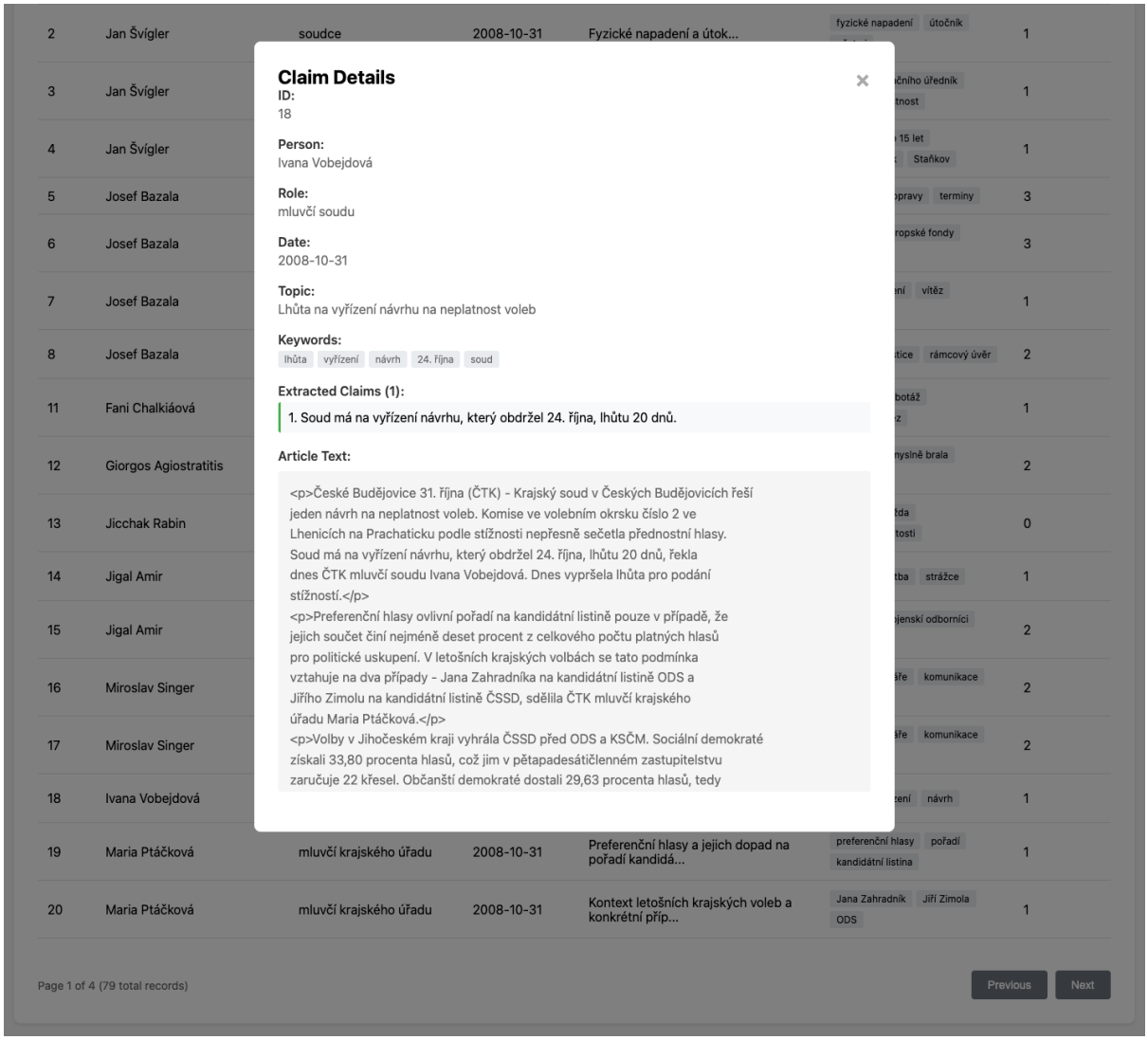
PromptOpt typicky používá dva modely: “optimizer” a “target”. Model “optimizer” pomocí meta-promptů hledá optimální prompt pro “target”. Pro optimizer volíme nejsilnější dostupný model s vysokou schopností provádět “reasoning” - v této úloze se nejlépe osvědčila plná verze modelu OpenAI GPT5. Nejlepší otevřený “optimizer”, který jsme schopni provozovat na vlastní infrastruktuře se v současné době ukázal být model OpenAI GPT-OSS-120B, velmi dobře fungují i modely Qwen3. Jako “target” model v současnosti preferujeme OpenAI GPT-OSS-20B a menší verze modelů Qwen3.

Podúlohy extrahující samotná tvrzení i informace o osobách umíme řešit velmi dobře, přičemž PromptOpt je schopen iterativně generovat instrukce (prompty) kontinuálně zvyšující se kvality. Problematičtější se jeví úlohy zahrnující precizní označování částí textů (text tagging/token classification). Zkvalitnění výstupů pro tento typ úloh se aktivně věnujeme: otestovali jsme řadu reprezentací kódování vstupu a výstupu a nyni implementujeme specializované procedury, pro vynucený formát generovaných textů (constrained generation).

Následující obrázek ukazuje nástroj pro náhled a anotaci dat vyvinutý v rámci projektu.



Dalším souvisejícím tématem, které jsme řešili byla tvorba metrik pro vyhodnocování halucinací v generovaných textech. Zlepšili jsme starší model AlignScoreCS a nově testovali řešení s použitím velkých jazykových modelů (LLM-as-a-judge). Velkými jazykovými modely jsme schopni dosahovat marginálně lepších výsledků za cenu výrazně větší výpočetní složitosti (encoder modely interně použité pro AlignScoreCS mají řádově o jeden až dva řády méně parametrů).



Pro praktické využití nalezených promptů jsme navrhli a implementovali jednoduchou aplikaci, která umožňuje import existujících dat (z důvodu interní spolupráce jsme se zaměřili na interní formát zpráv, který používá ČTK), extrakce osob, jejich role a výroků a témy, ke které se vztahují. Následně aplikace umožňuje uložení extrahovaných artefaktů do databáze. Aplikace taktéž poskytuje základní prostředí pro vyhledávání dle osob, data, nebo klíčového slova, přičemž po zobrazení detailu ukáže extrahované výroky i s původním článkem pro kontrolu.

SA04 Robustní vysvětlitelné učení pro detekci generativní AI

Vývoj modelů pro generování syntetických obrázků je jedním z nejrychleji rostoucím odvětví AI a ukazuje se, že s nástupem kvalitnějších generátorů je stále těžší spoléhat se na pro člověka těžko zřetelné artefakty a jiné forenzní stopy. Z tohoto důvodu využívá většina dnešních přístupů k detekci syntetických obrázků různé typy hlubokých modelů, které jsou používány pro samotné generování nebo sémantické zpracování obrazových dat. Tyto často dosahují velmi dobré přesnosti pokud již generovaná data viděli při učení, ale nejsou schopny kvalitně generalizovat na data z předem neviděných generátorů. Společně s tím je často skloňovaným tématem také snížená přesnost detektorů po aplikaci obrazových úprav jako

JPEG komprese nebo změna velikosti, ke kterým dochází například při sdílení obsahu na sociálních sítích. V této souvislosti jsme se zabývali robustností detekce syntetických dat ve třech různých rovinách. V první rovině jsme se zaměřili na otázku, zda lze robustnost detekce syntetických dat zajistit již na úrovni rozhodovací hranice, tedy optimalizací detektorů pro extrémně nízké míry falešně pozitivních rozhodnutí, a to i v situacích, kdy jsou k dispozici pouze omezená trénovací data a kdy klasické ztrátové funkce selhávají. Ve druhé rovině jsme si položili otázku, kterých signálů hluboké modely pro své rozhodování používají (šum nebo sémantický obsah) a jak zajistit, aby detektory poskytovaly spolehlivá rozhodnutí? Ve třetí rovině jsme se podívali na to, zda-li můžeme robustní detekci zajistit s dostatečně silným hlubokým modelem jenž nebyl při svém trénování vystaven přítomností artefaktů, které jsou typické pro synteticky generovaná data.

SA04.1 Optimalizace detektorů pro velmi nízké počty falešně pozitivních syntetických dat

V [SA04-1] jsme se zaměřili na vývoj a testování metod pro optimalizaci detektorů s velmi nízkou mírou falešných poplachů v úloze detekce skrytých nebo syntetických dat. Úspěšně jsme rozšířili přístupy TopPush a PatMat typu LEOT (Loss with Explicitly Optimized Threshold) [SA04-2] na hluboké modely, včetně EfficientNet [SA04-16], což je kritické pro jejich použití v daných úlohách, kde již lineární klasifikátory nedosahují požadované přesnosti. Obě navržené modifikace DeepTopPush a DelayedPatMat jsme otestovali, jak na lineárních tak hlubokých modelech a porovnali je s existujícími přístupy. Výsledky ukázaly (Fig. SA04-1), že naše metody dosahují o jeden až dva řády nižší míry falešných poplachů než dříve používané ztrátové funkce, přičemž si zachovávají srovnatelnou nebo vyšší přesnost detekce. Dále jsme prokázali, že detektory lze efektivně trénovat i s výrazně menším množstvím pozitivních vzorků, což výrazně snižuje výpočetní náklady.

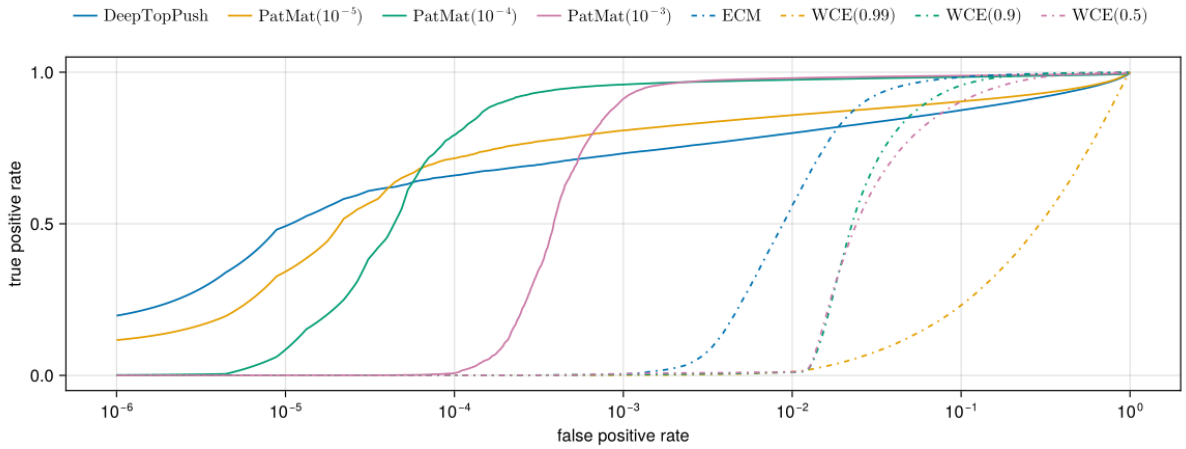


Fig. SA04-1. ROC křivky (s logaritmicky vyobrazenou mírou falešně pozitivních vzorků) lineární detektorů optimalizovaných na detekci skrytých dat v obrázcích, tréninková sada obsahovala 10% obrázků se skrytými daty (stego images), křivky byly odhadnuty na testovacích datech.

SA04.2 Studie robustnosti existujících detektorů syntetického obsahu na běžné obrázkové modifikace

Ukazuje se, že již zmíněné vysoké přesnosti dosahují detektory syntetického obsahu často kvůli zanedbání rozdílů mezi reálnými a syntetickými daty, které byly použity pro trénink.

Například článek [SA04-6] naznačuje, že detektory založené na hlubokých sítích fungují částečně jako detektory JPEG komprese nebo artefaktů specifických škálovacích faktorů. Obecně se proto můžeme zabývat tím jak různá historie vzniku daného obrazového signálu ovlivňuje detekci syntetického obsahu.

Citlivosti velkých obrazových modelů na obrazové úpravy

V článku [SA04-3] jsme se zaměřili na zkoumání detekovatelnosti jednotlivých kroků obrazového předzpracování pomocí velkých vision modelů, jako jsou CLIP, DINOv2 a (supervised) ViT [SA04-7][SA04-8][SA04-9], které jsou často základem pro detektory uměle generovaného obrazového materiálu. Pomocí kontrolovaných kroků předzpracování jsme vytvořili obrázky ze surových dat (tj. bez jakéhokoliv předzpracování) a analyzovali, jak různé operace – například demosaicing, denoising, škálování nebo JPEG komprese – ovlivňují stopy zanechané v datech. Vyhodnotili jsme 29 různých kroků předzpracování a prokázali, že tyto modely dokáží rozlišit mezi jednotlivými typy předzpracování s vysokou přesností, přičemž nejvíce detekovatelné jsou výrazné zásahy, jako je JPEG komprese a změna měřítka, zatímco jemné úpravy, například mírné doostření, jsou pro modely obtížnější. Dosavadní výsledky naznačují, že schopnost rozpoznání jednotlivých obrazových úprav zůstává u některých modelů vysoká i po aplikaci dalších obrazových úprav, viz Fig. SA04-2.

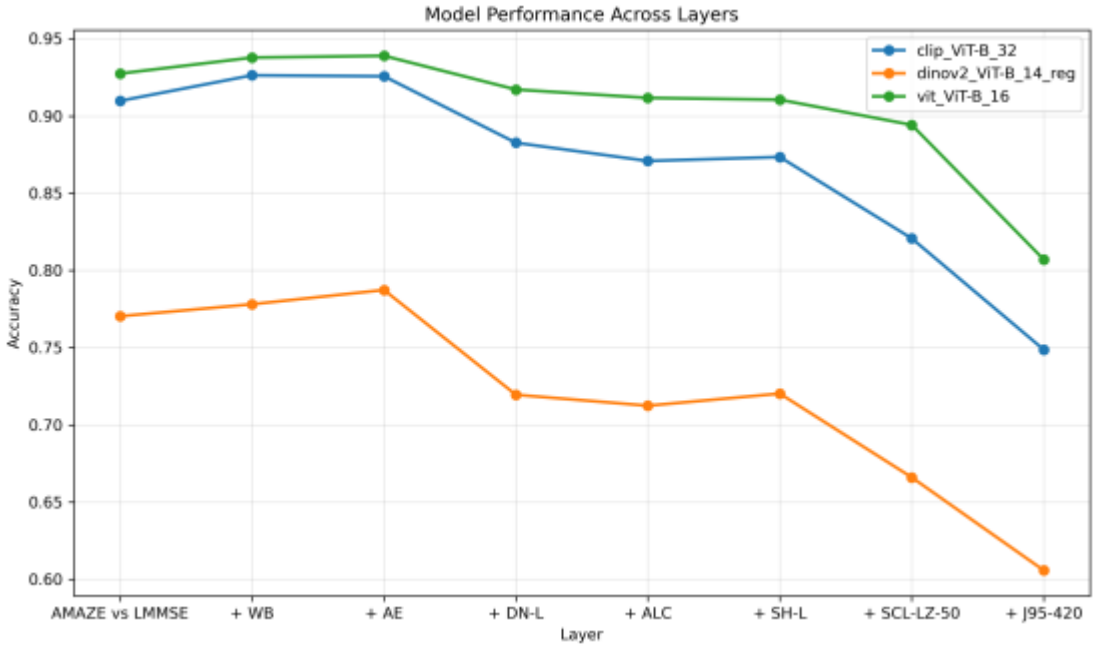


Fig. SA04-2. Přesnost detektorů mezi dvěma různými demosaicing algoritmy po aplikaci dalších obrazových úprav jako je například vyvážení bíle (WB), automatická expozice (AE) nebo JPEG komprese (J95-420).

Získané výsledky ukazují, že studované modely vnímají odlišnosti způsobené předzpracováním obrazu a mohou tak sloužit jako základ pro vývoj robustnějších detektorů syntetického obsahu.

Adversární učení pro zvýšení robustnosti detektorů na obrazové úpravy

Cílem bylo ověřit zda adversární učení napomáhá zvýšení robustnosti detektorů na obrazové úpravy a degradace, které jsou schopny zmást tradičně učené detektory. Zároveň jsme s pomocí adversárních útoků chtěli poodhalit, zda detektory založené na velkých vision

modelech zakládají své detekce na sémantickém obsahu obrázků nebo na šumu, jelikož při nízké síle útoku není možné měnit obsah, pouze šum.

V porovnání s předchozími studiemi [SA04-10], které zkoumali robustnost adversárně učených modelů na výrazné degradace obsahu, jsme se zaměřili na úpravy s velmi nízkou intenzitou, které nemění perceptuální vnímání. Alternativní verze datasetů jsme vytvořili s pomocí 12 různých obrazových úprav (5 úrovní intenzit) jako například různé kompresní nebo škálovací algoritmy. Zjistili jsme, že v úloze detekce uměle generovaných obrázků jsou rozdíly mezi modely trénovanými s různým poměrem adversárních vzorků zanedbatelné, ale zároveň nejhorších výsledků - přesnost na čistých a degradovaných datech - je dosaženo při trénování bez čistých dat.

Klasifikátory založené na výstupní vrstvě velkých vision modelů vykazují obecně vysokou robustnost na degradace i bez adversárního učení. Výjimkou jsou detektory založené na semi-supervised přístupech, jako např. DINOv2, kde adversární učení zvyšuje robustnost, avšak pouze zanedbatelně. Dosavadní výsledky z porovnání různě trénovaných velkých vision modelů v této úloze naznačuje, že klasifikátory založené na modelech CLIP jsou sice méně robustní na degradace než modely typu DINOv2, ale lépe generalizují na data z předem neviděných generátorů, což se shoduje s výsledky citlivostní studie z předchozí sekce.

Studie posunu distribuce ve struktuře JPEG souborů

Z předešlých studií víme, že JPEG komprese patří mezi hlavní kontributory falešných detekcí v úloze rozpoznávání generovaného obrazu, a proto je nanejvýš důležité porozumět i struktuře tohoto formátu. Naše hlavní motivace však pochází z úlohy detekce skrytých dat v obrázcích, kde v posledních letech je stále častější ukládat zprávy ve struktuře formátu na místo samotného obrazového signálu [SA04-4]. Detekce takových zpráv je složitá vzhledem k množství implementací, které jsou v souladu se standardem, ale zároveň produkují různé struktury. Ilustraci tohoto fenoménu jsme demonstrovali pomocí analýzy počtu unikátních hlaviček a markerů v čase, viz Fig. SA04-3, které zaznamenaly ve sledovaném období stálý nárůst.

Identifikovali jsme tři různé scénáře posunu distribuce. V první řadě to je uvedení nového režimu enkódování (sekvenční vs progresivní), ke kterému například došlo po vydání knihovny MozJPEG a následným rozšířením okolo roku 2014. V druhé řadě jsme se zabývali malými změnami, ke kterým dochází při uvedení nových verzí či větvení vývoje populárních knihoven. V poslední řadě jsme analyzovali rozdíly v distribuci při nahrávání obrazových dat na šest různých sociálních sítí - Facebook, Instagram, BlueSky, Pinterest, X a LinkedIn. Z našeho výzkumu [SA04-4] vyplývá, že každá z těchto platforem využívá vlastní konfiguraci enkódování s proprietárními kvantizačními tabulkami, specifickým nastavením metadat a aplikačních segmentů, které je možné použít k identifikaci zdroje daného obrázku. Z pohledu detekce generovaného obsahu nám tento výzkum umožňuje získat více informací o historii dat, díky čemu můžeme například zajistit rovnoměrné zastoupení jednotlivých platforem v trénovací sadě.

V úloze detekce skrytých zpráv v neobrazové složce formátu JPEG jsme porovnali námi navržené metody využívající hluboké modely pro hierarchická data (HMill) a hierarchickou stromovou vzdálenost (HTD) s existujícími metodami založených na statické extrakci příznaků

- HashTrick, MalJPEG [SA04-5]. Ve všech uvedených scénářích naše metody dosahují nejvyšší přesnosti a vykazují robustnost k posunu v distribuci dat.

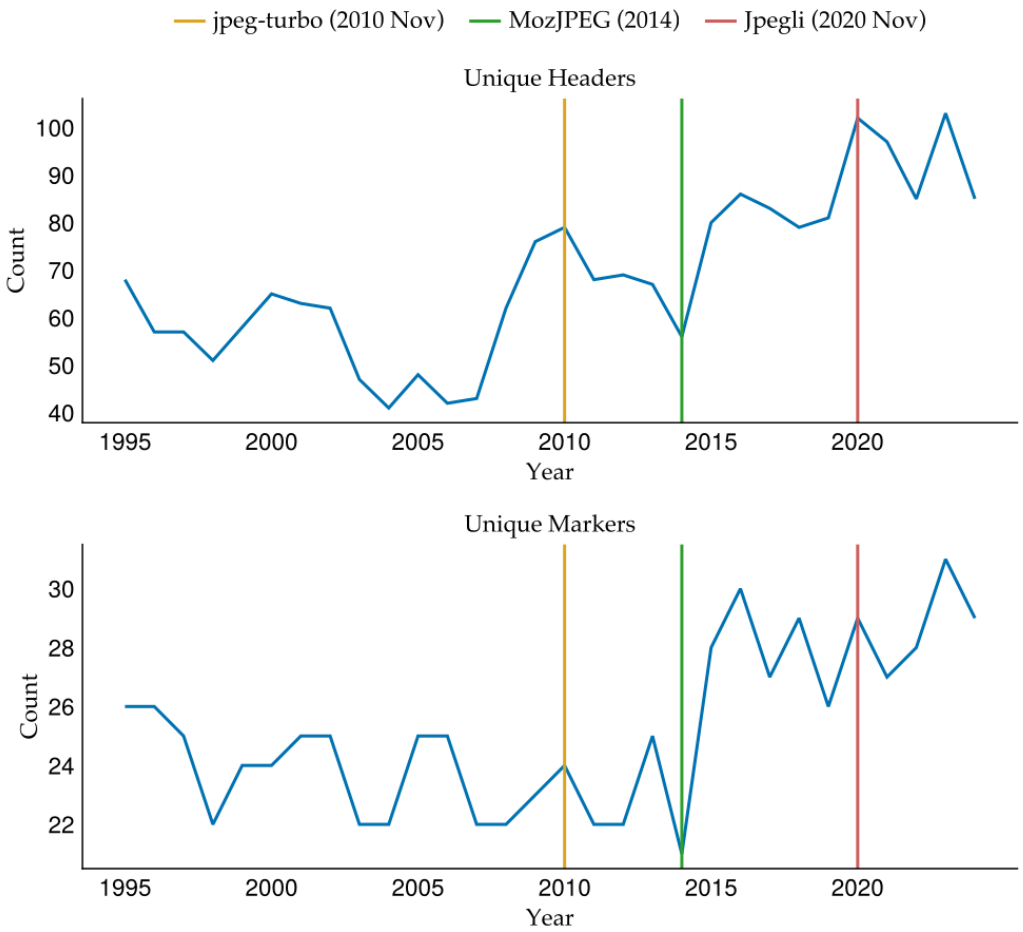


Fig. SA04-3. Evoluce formátu JPEG. Grafy ukazují nárůst unikátních počet hlaviček, markerů. Svislými čarami jsou znázorněny vydání důležitých knihoven pro enkódování formátu JPEG.

SA04.3 Ztrátová a bezeztrátová zero-shot detekce syntetických obrázků

Zero-shot detekce syntetických obrázků označuje přístup, při němž je detektor schopen identifikovat synteticky generovaná obrazová data bez explicitního využití syntetických vzorků během trénování. Tento přístup je klíčový zejména v kontextu rychle se vyvíjejících generativních modelů, kde se statistické vlastnosti syntetických dat neustále mění a znemožňují efektivní průběžné přeučování detektorů. Schopnost zero-shot detekce je proto zásadním předpokladem dlouhodobě robustní a škálovatelné forenzní analýzy obrazových dat. Zkoumání ztrátových i bezeztrátových přístupů k zero-shot detekci představuje důležitý krok k pochopení vztahu mezi kompresí obrazových dat, zachováním forenzních stop a detekovatelností syntetických obrázků v reálných scénářích.

Ztrátová zero-shot detekce syntetických obrázků

V první části jsme se zaměřili na detekci syntetických obrazových dat prostřednictvím jejich převodu do diskrétní latentní reprezentace a následného sekvenčního modelování. Výchozím bodem byl model Vector-Quantized Variational Auto-Encoder (VQ-VAE) [SA04-12], který umožňuje kvantizaci obrazových dat do diskrétního latentního prostoru. Tento proces má z principu ztrátový charakter a vede k částečné degradaci informace obsažené v původním obrazu, včetně jemných forenzních stop. Nicméně tato forma reprezentace výrazně redukuje rozměrnost vstupních obrazových dat, což se jeví jako klíčové pro výpočetně efektivní zpracování při neustále rostoucím rozlišení synteticky generovaných obrázků.

Na tomto základě jsme využili architekturu Vector-Quantized Generative Adversarial Network (VQ-GAN) [SA04-14], která kombinuje diskrétní latentní prostor s adversariálním učením a umožňuje získat sekvenci latentních tokenů reprezentujících obrazová data. Tyto sekvence jsme dále modelovali pomocí autoregresních jazykových modelů založených na transformerech, které jsou schopny efektivně zachytit dlouhodobé závislosti v diskrétních sekvencích. V tomto nastavení jsme adaptovali existující metody detekce syntetického textu (např. [SA04-15]) založené na analýze věrohodnostní funkce jazykového modelu a aplikovali je přímo na latentní sekvence obrazových dat. Experimenty byly inspirovány přístupy typu LlamaGen (VQ-GAN + LLaMA) [SA04-13] a jejich cílem bylo analyzovat vztah mezi mírou komprese forenzních informací v syntetických obrázcích a jejich detekovatelností na základě věrohodnosti generovaných sekvencí. Dosavadní výsledky ukazují, že takto navržené přístupy jsou schopné detekovat synteticky generovaná data, nicméně ne v tak dobré míře jako existující alternativní přístupy. V této části bychom v budoucnu rádi zkoumali rozdílné způsoby kvantizace v latentních prostoru, tak aby nedocházelo k příliš velkému poškození forenzní informace.

Současně s touto linií výzkumu jsme se zabývali možností exaktního vyjádření věrohodnostní funkce VQ-VAE modelu prostřednictvím destilace do traktabilních pravděpodobnostních modelů. Cílem bylo získat výpočetně efektivní a explicitně vyhodnotitelnou reprezentaci obrazových dat, což je klíčové zejména při analýze rozsáhlých kolekcí obrázků, například na sociálních sítích. Tento destilační přístup k traktabilnímu vyjádření věrohodnosti obrazových dat byl publikován v [SA04-11]. Významnou výhodou tohoto rámce je možnost formulace obecných pravděpodobnostních úkolů nad obrazovými daty, což otevírá cestu ke zvýšení vysvětlitelnosti detekčních metod a například umožnit označení pouze těch součástí reálných scén, které jsou generovány pomocí AI nástrojů (inpainting).

V rámci ztrátové zero-shot detekce jsme zkoumali také přístupy založené na architektuře CLIP, které pracují se silně komprimovanou sémantickou reprezentací obrazu a umožňují zero-shot analýzu bez explicitního trénování na syntetických datech, avšak s omezenou citlivostí na nízkoúrovňové forenzní stopy jak ukazují naše dosavadní experimenty.

Dále jsme se zabývali metodami rekonstrukce chyby. Tyto jsou typicky založeny na architekturách VAE, které projektují vstupní obrázek do latentního prostoru a následně jej rekonstruují zpět do obrazového formátu. Chyba se pak počítá mezi původním vstupním obrázkem nezatíženým ztrátou informace a jeho rekonstrukcí po ztrátě informace vzniklou průchodem generativním modelem, přičemž menší odchylky mohou indikovat, že se jedná o synteticky generovaný obsah. Jako základní model jsme si zvolili různé varianty modelů Stable Diffusion, které byly schopné generovat vysoce realistické obrázky. Pro zlepšení kvality detekce jsme je doplnili o BLIP textový enkodér [SA04-17], který poskytuje textový popis

vstupního obrázku. Testované detektory byly hodnoceny na obrázcích z hlubokých generativních modelů vyvinutých v různých historických etapách, tak aby se posoudila jejich schopnost generalizace a robustnosti při identifikaci syntetického vizuálního obsahu. Tyto metody byly schopny solidní detekce syntetických data avšak nedosahovali tak dobré výkonnosti jako existující metody podobného typu.

Bezeztrátová zero-shot detekce syntetických obrázků

V další části jsme se zaměřili na bezeztrátovou zero-shot detekci syntetických obrázků pomocí autoregresních modelů aplikovaných přímo v prostoru obrazových dat. Na rozdíl od ztrátových přístupů zde nedochází k žádné kvantizaci, a modely typu GPT nebo stavové modely (state-space models) explicitně modelují plnou pravděpodobnostní distribuci obrazů reprezentovaných jako dlouhé sekvence pixelů nebo jejich přesných diskrétních transformací. Tento přístup je bezeztrátový v tom smyslu, že zachovává veškerou informaci obsaženou v původním obrazu, včetně potenciálních forenzních stop.

Zvláštní pozornost jsme věnovali stavovým modelům, které jsou schopny efektivně pracovat s extrémně dlouhými sekvencemi a tím překonávají praktická omezení klasických transformerových architektur při modelování obrazů ve vysokém rozlišení. V tomto nastavení testujeme možnost zero-shot detekce syntetických obrázků čistě na základě autoregresní věrohodnosti celého obrazu, bez využití syntetických dat při trénování a bez rizika degradace informace způsobené ztrátovou kompresí. Tato část výzkumu se v současnosti nachází ve fázi vyhodnocování experimentálních výsledků.

Reference

[SA04-1] V. Mácha, L. Adam, V. Šmídl, A. Ker, and T. Pevný, “Optimizing detectors for very low false positive rates with a case study in steganography,” *IEEE Transactions on Information Forensics and Security*, in review.

[SA04-2] V. Mácha, L. Adam, and V. Šmídl, “DeepTopPush: Simple and scalable method for accuracy at the top,” *arXiv preprint* arXiv:2006.12293, Mar. 27, 2023. doi: 10.48550/arXiv.2006.12293.

[SA04-3] J. Franců, M. Papež, and T. Pevný, “Detection of image processing pipelines using pretrained vision models: A forensic analysis,” 2026, to be submitted.

[SA04-4] M. Zorek, T. Pevný, and R. Böhme, “Agility for steganalysis: Dealing with distribution drift in JPEG file structures,” in *Proc. IEEE European Symposium on Security and Privacy (EuroS&P)*, 2026, in review.

[SA04-5] A. Cohen, N. Nissim, and Y. Elovici, “Maljpeg: Machine learning based solution for the detection of malicious JPEG images,” *IEEE Access*, vol. 8, pp. 19,997–20,011, 2020.

[SA04-6] P. Grommelt, L. Weiss, F.-J. Pfreundt, and J. Keuper, “Fake or JPEG? Revealing common biases in generated image detection datasets,” *arXiv preprint* arXiv:2403.17608, Mar. 28, 2024.

[SA04-7] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” *arXiv preprint*arXiv:2103.00020, Feb. 26, 2021. doi: 10.48550/arXiv.2103.00020.

[SA04-8] M. Oquab *et al.*, “DINOv2: Learning robust visual features without supervision,” *arXiv preprint*arXiv:2304.07193, Feb. 2, 2024. doi: 10.48550/arXiv.2304.07193.

[SA04-9] A. Dosovitskiy *et al.*, “An image is worth 16×16 words: Transformers for image recognition at scale,” *arXiv preprint* arXiv:2010.11929, Jun. 3, 2021. doi: 10.48550/arXiv.2010.11929.

[SA04-10] K. Kireev, M. Andriushchenko, and N. Flammarion, “On the effectiveness of adversarial training against common corruptions,” *arXiv preprint* arXiv:2103.02325, Jan. 4, 2022. doi: 10.48550/arXiv.2103.02325.

[SA04-11] A. Hadžić, M. Papež, and T. Pevný, “Distillation of a tractable model from the VQ-VAE,” in *Proc. 8th Workshop on Tractable Probabilistic Modeling*, 2025.

[SA04-12] A. van den Oord and O. Vinyals, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[SA04-13] P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan, “Autoregressive model beats diffusion: LLaMA for scalable image generation,” *arXiv preprint* arXiv:2406.06525, 2024.

[SA04-14] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12,873–12,883, 2021.

[SA04-15] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in *Proc. International Conference on Machine Learning (ICML)*, pp. 24,950–24,962, 2023.

[SA04-16] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. International Conference on Machine Learning (ICML)*, pp. 6,105–6,114, 2019.

[SA04-17] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” Feb. 15, 2022, arXiv: arXiv:2201.12086. doi: 10.48550/arXiv.2201.12086.

SA05 Detekce generovaných videí s pomocí rychlých pohybů v obličeji

Základní myšlenkou výzkumu v této subaktivitě je pozorování, že když lidé mluví, jejich pohyb je pro ně charakteristický a možná dokonce unikátní pro každého jedince. Tyto pohybové

vzorce při mluvení jsou ovlivněny kombinací fyzických, psychologických a kulturních faktorů. Jde například o kývání hlavou, naklánění hlavy, pohyby rtů a obličejových svalů, mrkání očí, gesta rukou, apod.

Hypotéza je, že řečové pohyby není možné snadno dokonale zfalšovat. A tyto pohybové anomálie, které neodpovídají dané identitě, jsme schopni použít k detekci falešného videa nebo dokonce k časové lokalizaci, kde došlo k falšování uvnitř jinak autentického videa. Tento přístup má několik výhod. Klasické detektory deepfake obvykle pracují tak, že jsou natrénovány na velkých sadách, kde jsou příklady reálných a falešných videí, konstruovaných známými technikami. Současné metody založené na klasifikátoru na principu hlubokých sítí velmi přesně dokáží detekovat falešný obsah, pokud byly natrénovány na konkrétním generátoru a známé technice falšování. Problém je ale s generalizací [SA05-1]. Nové nebo neznámé techniky falšování jsou velmi obtížně detekovatelné. Souvisí to s tím, že klasifikátor se přeučí na konkrétní signálový otisk konkrétního generátoru. Naučí se detekovat anomálie na velmi nízké úrovni signálu. Mimo špatné generalizace je dalším velkým problémem nerobustnost vůči poškození signálu kvůli kompresi nebo snížení rozlišení, které otisk generátoru efektivně vymažou nebo výrazně oslabí. Přestože jsme významně vylepšili generalizaci na neviděnou deepfake techniku pomocí adaptace velkého fundamental modelu a dosáhli state-of-the-art výsledků na mnoha standardních testovacích sadách [SA05-2], přístup detekce přes příznaky vyšší úrovně by měl být nezávislý na použité technice generování falešného videa, málo citlivý vůči negativním faktorům kvality videa, teoreticky by měl fungovat až do té úrovně, dokud fungují extraktory pohybu. Navíc výsledek není "black-box", je interpretovatelný pro člověka, což u detekce na nízké signálové úrovni možné není.

Naším úkolem bylo navrhnout kompaktní deskriptor, jakousi signaturu, která z krátkého několika-sekundového videa extrahuje vektorovou reprezentaci. Vektory pro stejnou identitu budou co nejblíže, zatímco pro různé identity budou daleko.

Testovali jsme descriptor inspirovaný prací [SA05-3], který byl manuálně zkonstruován na základě korelací tzv. Facial Action Units (FAUs), což jsou vlastně aktivace mimických svalů v obličeji [SA05-4]. Deskriptor jsme vylepšili pomocí strojového učení. Vstupem je 20 FAUs a orientace hlavy mluvčího (3 úhly – roll, pitch, yaw) v čase extrahované pro každý snímek. To dává vstupní tensor (20+3) × T, kde T je počet snímků. Takový tensor reprezentuje krátký segment videa, např. 10 sekund. Descriptor, kterému říkáme "Talking Motion Signature" (TMS), učíme diskriminativně, pomocí projekce vstupního tensoru konvoluční neuronovou sítí. Máme velkou trénovací sadu vytvořenou ze sady VoxCeleb [SA05-3], kde je cca 600 identit, pro každou identitu je k dispozici 12 videí o délce 20 sekund obsahující souvislou řeč. Tedy naučíme klasifikátor, neuronovou síť, který rozpozná identity.

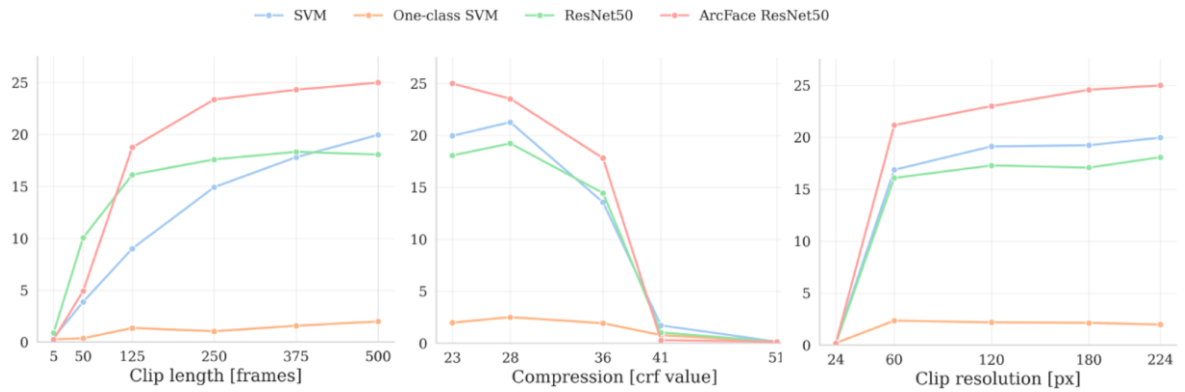


Fig. SA05-1. Přesnost rozpoznání identity v procentech v závislosti na délce vstupního videoklipu, kompresi a rozlišení videa. Výsledek je významně lepší než náhodná šance 1/600, a naše nejlepší metoda (ArcFace ResNet50) překonává původní metodu [SA05-3] (SVM).

Takto natrénovaný descriptor (TMS) jsme srovnali s původním manuálně zkonstruovaným [SA05-3], výsledky jsou zobrazeny ve Fig. SA05-1. Výsledky ukazují, že natrénovaný descriptor je diskriminabilnější, tzn. umožňuje rozpoznání identity s vyšší přesností, a je odolnější vůči zkracování sekvence, kompresi i změně rozlišení vstupního videa.

Fig. SA05-2. Objektivní hodnocení kvality imitátorů Baracka Obamy. Prototyp descriptoru Baracka Obamy **p** byl natrénován modelem. Query descriptor **q**(t) je odhadnutý z testovacích videí. Výsledná statistická kvalita je pak $1/N \Sigma \mathbf{p}^T\mathbf{q}(t)$.

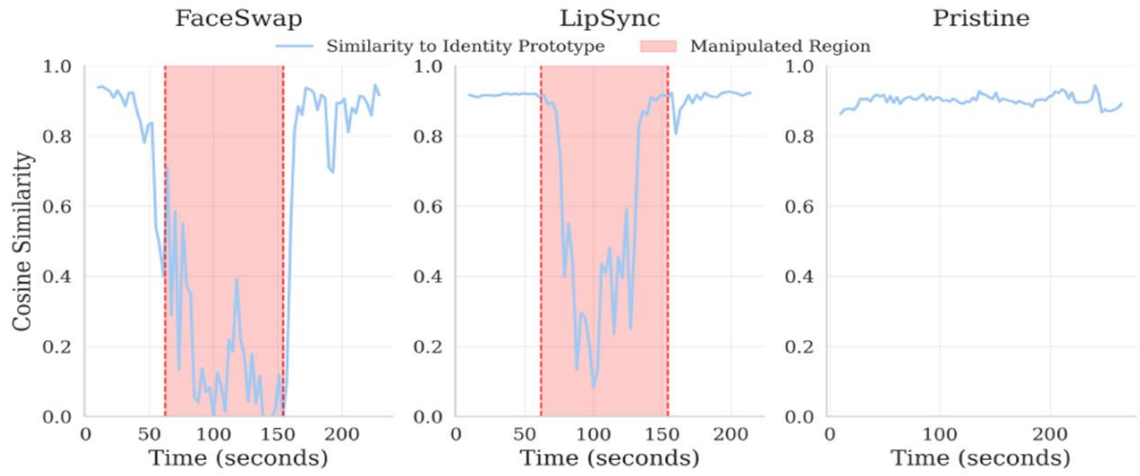


Fig. SA05-3. Ukázka detekce deepfakes. Cílová identita, president Zelenskyi, má naučený prototyp desriptor p. Grafy ukazují cosinovou podobnost mezi prototypem a query tesstovacím descriptorem. Vpravo je autentické video s presidenta Zelenského s přirozenou fluktuací podobnosti kolem 0.8. Vložením deepfakes s jiným řečníkem (president Macron), podobnost významně poklesne.

Navržený descriptor jsme dále použili ve dvou testovacích aplikacích: Objektivní hodnocení kvality imitace (Fig. SA05-2) celebrity a detekce deepfake pomocí face-swap a LipSync (Fig. SA05-3). Tato práce je podrobně popsána v našem článku [SA05-6], který je zařazen do přílohy této zprávy.

V průběhu řešení projektu, po diskusi s fact-checkery, se ukázalo, že mimo řečových pohybů je třeba zkoumat i samotnou řeč, protože falešné zvukové nahrávky a zvukové stopy jsou velice časté. Náš příspěvek v tomto směru popíšeme v následující sekci.

Detekce syntetické a deepfaky řeči

- **Kontext a cíle subaktivity**

V rámci projektu CEDMO 2.0 byl v roce 2025 zahájen samostatný výzkumný směr zaměřený na detekci syntetické a podvržené (deepfake) řeči. Motivací této aktivity je rychlý rozvoj generativních systémů řeči (TTS, voice conversion, multimodální deepfakes) a s tím spojená potřeba robustních, interpretovatelných a jazykově přenositelných detekčních metod, využitelných v oblasti bezpečnosti, médií a veřejné komunikace.

Subaktivita byla v roce 2025 realizována ve spolupráci týmu doc. Petr Pollák, Ing. Jiří Šmíd, Bc. Radek Kubica a prof. Roman Čmejla a od listopadu 2025 byla rozšířena o multimodální výzkum (audio–video) ve spolupráci s Ing. Cristianem Davidem Rios Urrego.

- **Akustická detekce podvržené řeči (DNN a datová infrastruktura)**

V oblasti detekce syntetické řeči na bázi hlubokých neuronových sítí tým navázal na předchozí výzkum využívající architektury ResNet s jednotřídní klasifikací pomocí OC Softmax. V modifikované verzi byl tento přístup ověřen na datech generovaných více systémy syntézy řeči (TTS) a voice conversion (VC), přičemž byla potvrzena jeho schopnost:

- efektivně detekovat podvrženou řeč vytvořenou různými generátory,
- generalizovat na neznámé systémy a nové akustické podmínky.

Klíčovým předpokladem dalšího rozvoje bylo vytvoření jazykově specifické české databáze podvržené řeči (CTUSpoof). V rámci projektu byla tato databáze systematicky rozšiřována jak z hlediska počtu mluvčích a promluv, tak z hlediska rozmanitosti použitých syntetických nástrojů. Data jsou shromažďována pro tři hlavní účely:

- trénování syntezátorů řeči v češtině (zdrojová data),
- testování detekčních systémů na reprezentativním evaluačním setu,
- trénování detektorů s využitím realistické augmentace (hluk, RIR).

V roce 2025 byly natrénovány základní verze českých syntezátorů ParlerTTS a SpeechT5 na rozsáhlém korpusu ParlaSpeechCZ (≈1000 hodin), které umožnily osvojení české výslovnosti a prosodie. Pro druhou fázi (fine-tuning) byla připravena kvalitnější data z databází SPEECON, GlobalPhone, CZKCC, CtuTest a CzLecDSP, určená ke zlepšení akustické kvality syntetických promluv.

- **Akustické markery a bayesovská interpretovatelná detekce**

Paralelně s DNN přístupem byl v rámci SA05 rozvíjen alternativní a interpretovatelný směr založený na Bayesovské detekci změn v řečovém signálu. Tento přístup vychází z hypotézy, že syntetická řeč vykazuje narušenou mikro-variabilitu a fyziologicky nepřirozené časové struktury, které lze kvantifikovat pomocí pravděpodobnostních modelů. Za tím účelem byla vyvinuta a systematicky ověřena rodina tří Bayesovských detektorů pro detekci různých akustických markerů:

- BSCD (Bayesian Step Changepoint Detector) – citlivý na náhlé změny amplitudy,
- BLCD (Bayesian Linear Changepoint Detector) – zachycující pozvolné driftové změny,
- RBD (AR Recursive Bayesian Changepoint Detector) – modelující rezonanční a autoregresní nestability.

Detektory byly testovány na referenčních databázích ASVspoof 2019 a ASVspoof 2021 v rámci standardizovaných evaluačních protokolů, kde dosahovaly stabilní separability mezi bona fide a spoof řečí (typicky AUC ≈ 0.84–0.88, u vybraných markerů > 0.95). Důležitým validačním krokem bylo jejich ověření na prodloužené fonaci pacientů s Parkinsonovou chorobou, kde se prokázalo, že stejné markery odlišují patologickou a zdravou řeč, což potvrzuje jejich fyziologickou interpretovatelnost a robustnost.

Výsledky byly shrnuty v článku [SA05-7] Change-Based Acoustic Markers for Robust and Interpretable Deepfake Speech Detection (v recenzním řízení, BSPC), který je hlavním publikačním výstupem této aktivity v roce 2025.

- **Multimodální detekce: synchronizace řeči a pohybu rtů**

Od listopadu 2025 byla aktivita SA05 rozšířena o multimodální výzkum audio-video deepfake detekce ve spolupráci s Ing. Cristianem Ríosem. Cílem této části je ověřit, zda vizuální informace o artikulačním pohybu (rtů) může doplnit akustickou složku a zvýšit robustnost detekce.

Výzkum je založen na analýze časové synchronizace mezi řečí a pohybem rtů, s využitím segmentového zpracování (1s okna) a neuronových architektur s mechanismy pozornosti. Pro počáteční experimenty byly vybrány veřejné anglické databáze (např. FakeAVCeleb), umožňující rychlou validaci metodiky. Paralelně se připravuje využití českých videozáznamů zdravých kontrol z medicínských databází, kde budou vytvářeny kontrolované audio manipulace (časový posun, nahrazení syntetickou řečí) pro testování jazykové přenositelnosti. Tato multimodální větev je koncipována jako komplementární rozšíření akustické detekce, s cílem vytvořit základ pro hybridní audio-video systémy v dalším období řešení projektu, zejména v kontextu detekce manipulovaného mediálního obsahu.

- **Shrnutí a další kroky**

V roce 2025 byla v rámci SA05 vybudována metodicky různorodá základna pro detekci syntetické a deepfake řeči. Byly paralelně rozvíjeny:

- DNN přístupy s důrazem na česká data,
- Bayesovské detektory akustických markerů,
- multimodální audio-video metody zaměřené na synchronizaci řeči a artikulace.

V následujícím období se aktivity zaměří na integraci těchto přístupů do hybridních detektorů, rozšíření českých databází a publikaci výsledků v mezinárodních časopisech.

Reference

[SA05-1] Nela Petrželková, and Jan Čech. Detection of Synthetic Face Images: Accuracy, Robustness, Generalization. In Proc. German Conference for Pattern Recognition. 2025.

[SA05-2] Andrii Yermakova, Jan Čech, Jiří Matas, Mario Fritz. Deepfake Detection that Generalizes Across Benchmarks. In Proc. WACV, 2026. To Appear.

[SA05-3] Bohacek, Matyas, and Hany Farid. "Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms." Proceedings of the National Academy of Sciences 119.48 (2022): e2216035119.

[SA05-4] Ekman, Paul, and Wallace V. Friesen. "Facial action coding system." Environmental Psychology & Nonverbal Behavior (1978).

[SA05-5] Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "Voxceleb2: Deep speaker recognition." arXiv preprint arXiv:1806.05622 (2018).

[SA05-6] Ivan Samarskyi, Jan Čech. Talking Motion Signatures for Identity Recognition. In Proc. VISAPP, 2026. To Appear.

[SA05-7] Roman Čmejla: *Change-Based Acoustic Markers for Robust and Interpretable Deepfake Speech Detection.* Biomedical Signal Processing and Control, Elsevier — under review.

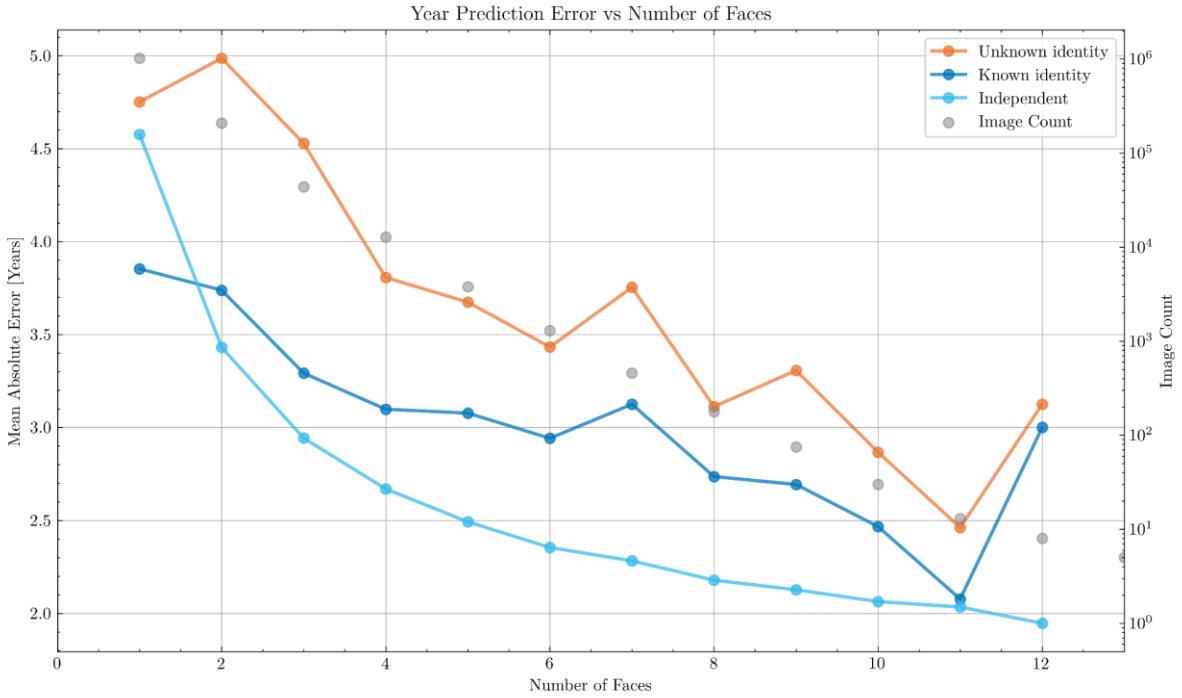
SA06 Detekce generovaných obrázků z vlastností obličejů

Naším cílem byl vývoj technologií pro ověřování pravosti multimediálního obsahu s využitím rozpoznávání obličejů a odhadu věku. Zaměřili jsme se zejména na detekci dezinformací týkajících se veřejně známých osobností, jejichž obrazová data bývají často manipulována. Vyvinuli jsme modely, databáze a softwarové nástroje umožňující ověřování identity a odhad věku osob z obrazových dat. Tyto metody byly integrovány do dvou nezávislých aplikací: pro datování fotografií a pro rychlé vyhledávání ve velkých databázích novinářských snímků.

Výstupy

1. **SOTA model pro odhad věku.** Vyvinuli jsme model pro odhad věku z obličejových fotografií založený na hlubokých neuronových sítích, trénovaný na rozsáhlých anotovaných datech. Model byl nezávisle evaluován v soutěži NIST (FRVT Age Estimation: https://pages.nist.gov/frvt/html/frvt_age_estimation.html) a dosahuje přesnosti srovnatelné s nejlepšími světovými řešeními.
2. **Rozsáhlá databáze obličejů.** Hlavní překážkou při vývoji našich modelů byla absence rozsáhlé databáze obličejových snímků s přesnou anotací identity a věku. Tento nedostatek jsme překonali vytvořením databáze s více než 1,5 milionu obličejů, jejichž anotace byly automaticky získány agregací veřejně dostupných zdrojů. Databáze odstranila klíčové omezení pro vývoj přesných modelů. Podrobnosti jsou uvedeny v článku [SA06-1].
3. **Metoda pro automatické datování fotografií.** Vyvinuli jsme metodu pro automatické datování fotografií, která využívá databázi osob se známým datem narození. Na

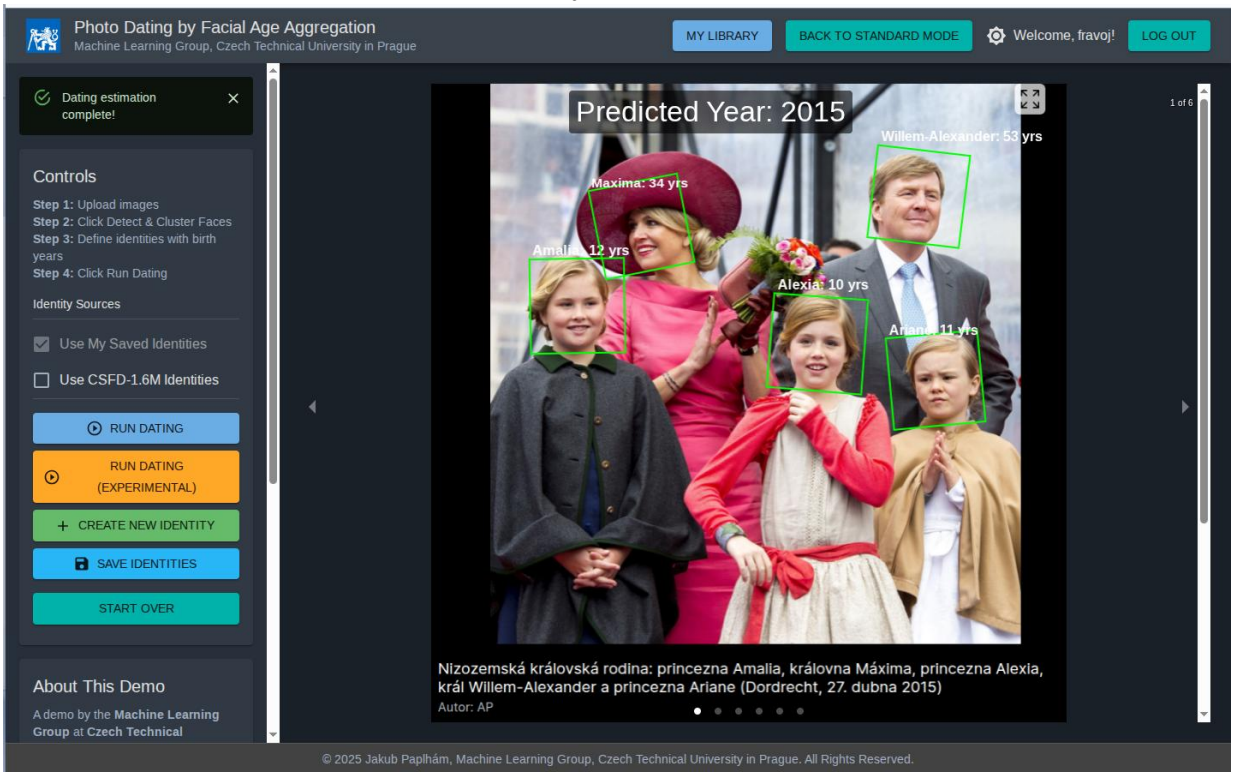
analyzovaném snímku jsou identifikovány osoby z databáze, je odhadnut jejich věk a na základě rozdílu mezi odhadovaným věkem a datem narození je statisticky určeno datum pořízení fotografie. Aktuální verze metody dosahuje přesnosti 2,5-5 let v závislosti na počtu detekovaných tváří. Podrobnosti jsou uvedeny v článku [SA06-1].



Přesnost datování jako funkce počtu identifikovaných tváří v obrázku.

4. **Webová aplikace pro datování fotografií.** Vyvinuli jsme webovou aplikaci pro automatické datování fotografií, která využívá metodu popsanou v článku [SA06-1]. Aplikace nabízí jednoduché uživatelské rozhraní umožňující nahrávání a automatické zpracování sekvencí obrázků. Využívá námi vytvořenou databázi známých osobností a zároveň podporuje automatické shlukování nových tváří a tvorbu uživatelských

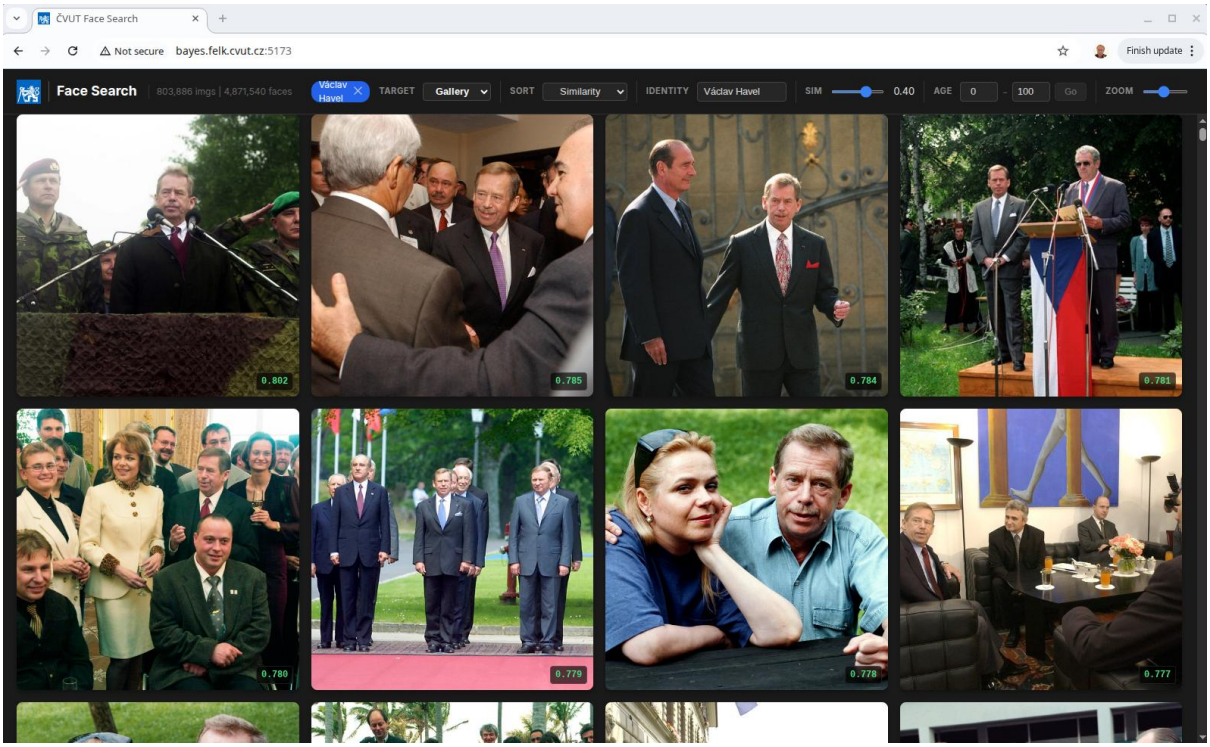
databází nových identit.



Screenshot z aplikace pro datování fotografií.

5. **Aplikace pro vyhledávání obrázků ve velkých databázích na základě identit a věku.** Vyvinuli jsme webovou aplikaci pro rychlé vyhledávání obrázků ve velkých databázích na základě identity a věku zobrazených osob. Aplikace byla otestována na reálné databázi 2 milionů novinářských fotografií poskytnuté ČTK. Nabízí jednoduché rozhraní, vyhledávání v reálném čase a následující funkce:
   - Vyhledávání obrázků podle neznámé identity na základě obličejové podobnosti.
   - Vyhledávání známých osob podle jména.
   - Třídění výsledků podle věku identit.

   Použitý model pro identifikaci je invariantní vůči věku osoby, a umožňuje tak nalezení snímků dané identity v různých životních obdobích.



Screenshot z aplikace pro vyhledávání obrázků podle identit a věku.

6. **Model pro odhad konfidence predikcí.** Zabývali jsme se odhadem konfidence predikcí získaných z modelů strojového učení, mezi něž patří i modely pro identifikaci osob a odhad věku z obličejových snímků. Navrhli jsme matematický rámec pro formulaci optimálního predikčního modelu, který umožňuje kvantifikovat nejistotu a odmítnout predikci v případech, kdy trénovací data neposkytují dostatek informací pro spolehlivé rozhodnutí. Výsledky jsou shrnuty v článku [SA06-2], připraveném k odeslání na konferenci.

Reference

[SA-06-1] J. Paplhám, V. Franc. Photo Dating by Facial Age Aggregation. In Proc. WACV, 2026. To Appear.
[SA-06-2] V. Franc, J.Paplhám. Epistemic Reject Option. ARXIV 2025. To be submitted.

SA07 Detekce kopií a modifikací známých obrázků a videí

Část 1: Detekce kopií obrazů jako nástroj pro ověřování faktů a propojování s ověřenými falzifikáty

Vyvinuli jsme modely pro detekci kopií obrazů, které určují, zda je daný obrázek kopií jiného existujícího obrázku. To slouží dvěma klíčovým účelům: **Fact-checking**: Pokud se v článku objeví upravený (photoshopovaný) obrázek, můžeme jej propojit s původním snímkem a zdrojovým článkem a ověřit tak jeho autenticitu.

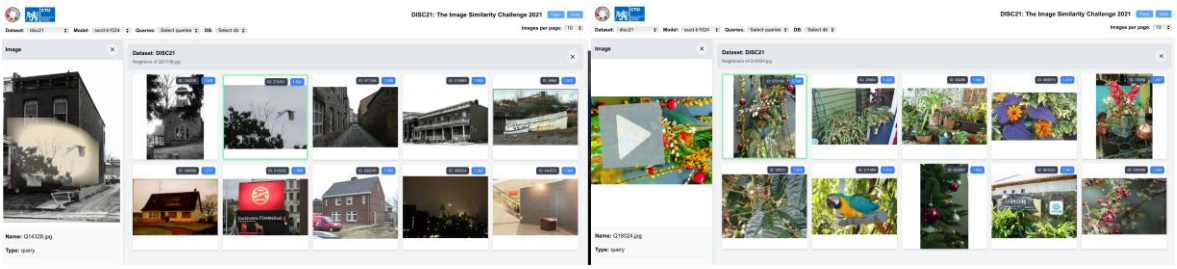
Fake detection: Pokud je nějaký obrázek v repozitáři označen jako falešný, mohou být jako falešné označeny i jeho kopie.

Nejmodernějším reprezentančním modelem pro detekci kopií obrazů je SSCD [1]. Natrénovali jsme model SSCD **s výrazně menšími prostředky** (4 GPU oproti 32 GPU použitým společností Meta/Facebook). Naše reimplementace dosáhla srovnatelné výkonnosti: globální průměrná přesnost činila 69,3 oproti publikovaným 71,0 na datasetu DISC.

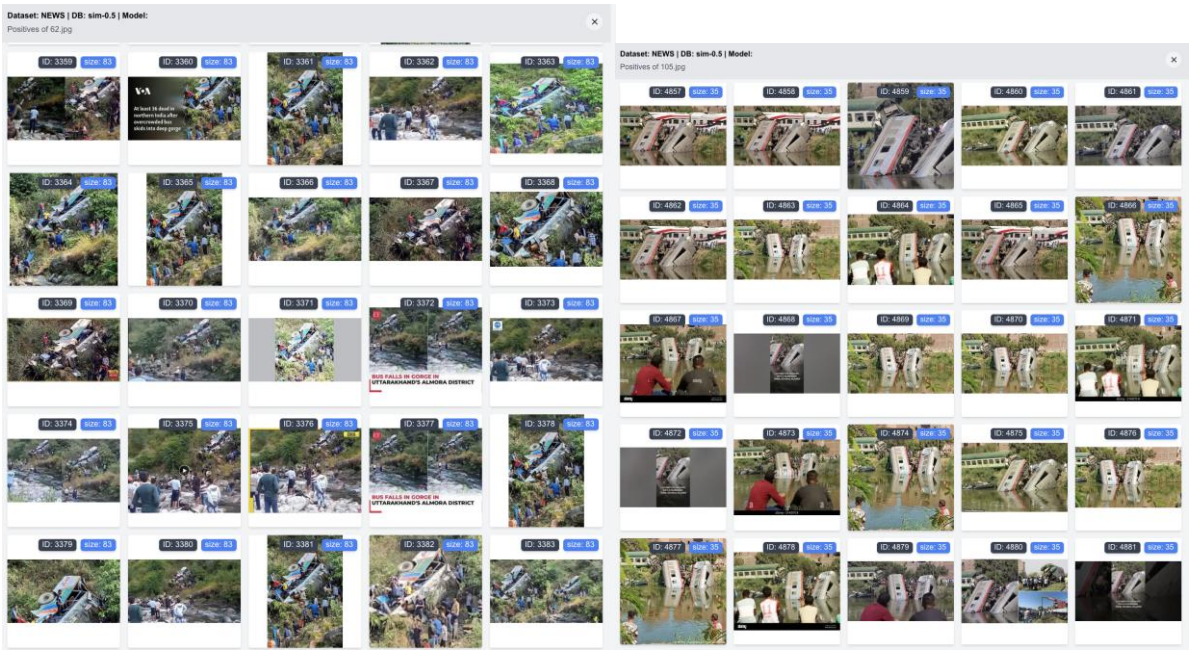
Pomocí SSCD jsme vytvořili dvě ukázkové aplikace:
Copy search – vyhledávání duplicit ve velkých kolekcích.
Copy clustering – seskupování kopií fotografií z rozsáhlých kolekcí stažených z novinových článků.

Pro clustering jsme využili podobnost SSCD k identifikaci kopií obrázků a následně spočítali komponenty souvislosti v grafu všech detekovaných kopií. Každá komponenta odpovídá klastru upravených verzí téhož původního snímku, a tím i téže události.

[1] Pizzi, E., Roy, S. D., Ravindra, S. N., Goyal, P., & Douze, M. (2022). A self-supervised descriptor for image copy detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14532-14542).



Snímek obrazovky z ukázky vyhledávání kopií na datasetu DISC21: Jsou zobrazeny dva příklady dotazů. Dotazový obrázek (vlevo) je upravená, photoshopovaná kopie jiného původního snímku. Původní obrázek je správně identifikován a nalezen i přes výrazné úpravy.



Snímek obrazovky z ukázky copy-cluster na obrazovém datasetu získaném z novinových článků: Zobrazeny jsou dva klastry, z nichž každý obsahuje photoshopované obrázky odvozené od stejného původního snímku.

Část 2: Vylepšené vyhledávání obrazů jako nástroj pro ověřování faktů (fact-checking)

Vylepšili jsme nástroje pro vyhledávání obrazů dvěma směry:
Vyhledávání na úrovni instancí napříč doménami: zlepšené vyhledávání konkrétních objektů, scén nebo míst v otevřeném prostředí (např. umělecká díla, produkty, předměty v domácnosti), tedy i mimo domény obsažené v trénovací sadě (např. památky, městská prostředí).
Komponované vyhledávání obrazů: prohledávání velkých kolekcí pomocí kombinace obrazového a textového dotazu. Text obvykle modifikuje určité aspekty dotazového obrázku, což umožňuje cílenější a jemněji odstíněné vyhledávání.

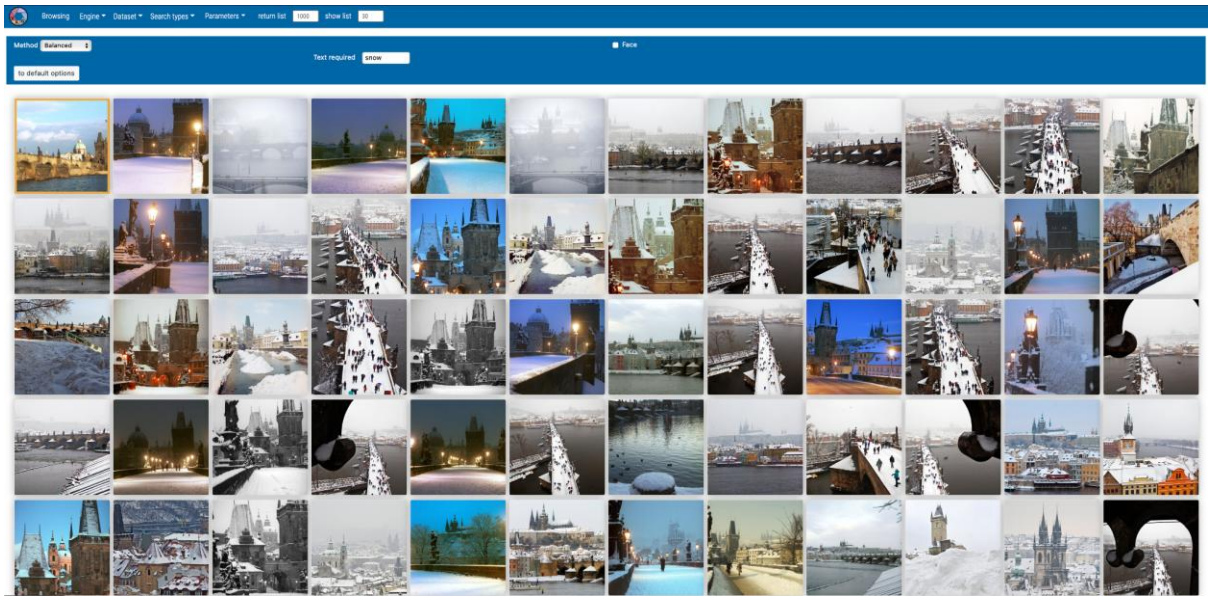
Tyto schopnosti jsou zásadní pro ověřování faktů, protože odhalování dezinformací často vyžaduje vyhledávání souvisejících obrázků napříč doménami nebo kombinaci obrazových a textových dotazů. **ČTK (Česká tisková kancelář) o tyto nástroje projevila zájem a v současnosti na jejich datech provozujeme reálnou demonstrační aplikaci**, která ukazuje, jak pokročilé vyhledávání posiluje procesy ověřování zpráv.

A. Komponované vyhledávání obrazů (nástroj pro ověřování faktů)
Navrhujeme metodu nevyžadující trénování, která využívá vizuálně-jazykové modely, jako je CLIP. Podobnosti obraz–obraz a text–obraz se počítají odděleně a následně se kombinují do vyváženého sjednoceného skóre. Míry podobnosti dále zpřesňujeme jednoduchými operacemi v prostoru reprezentací nebo na vstupních datech. Využitím sdíleného embeddingového prostoru projektujeme obrazy tak, aby lépe zachycovaly objekty a byly invariantní vůči variacím textového dotazu. Krok kontextualizace obohacuje dotazy o věrohodné alternativy, což zlepšuje interpretaci jemných textových úprav. Tento přístup výrazně překonává stávající metody v komponovaném vyhledávání obrazů na úrovni instancí (34 vs. 28 mAP).

Příklady komponovaného vyhledávání využívajícího dotazový obrázek (vlevo) a dotazový text (nahoře) k nalezení dotazovaného objektu v různých podmínkách.

Výše uvedená práce byla prezentována na konferenci NeurIPS 2025 (špičková konference v oblasti strojového učení) v následujícím článku: [Psomas, B., Retsinas, G., Efthymiadis, N., Filntisis, P., Avrithis, Y., Maragos, P., Chum, O. a Tolias, G., 2025, říjen. *Instance-Level*

Composed Image Retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.]



Snímek obrazovky z ukázky na 2 milionech obrázků ČTK: Dotazový obrázek (vlevo nahoře) zobrazuje Karlův most a textový dotaz ‚snow' vyhledá fotografie Karlova mostu pokrytého sněhem.

B. Vyhledávání obrazů na úrovni instancí napříč více doménami (nástroj pro ověřování faktů)

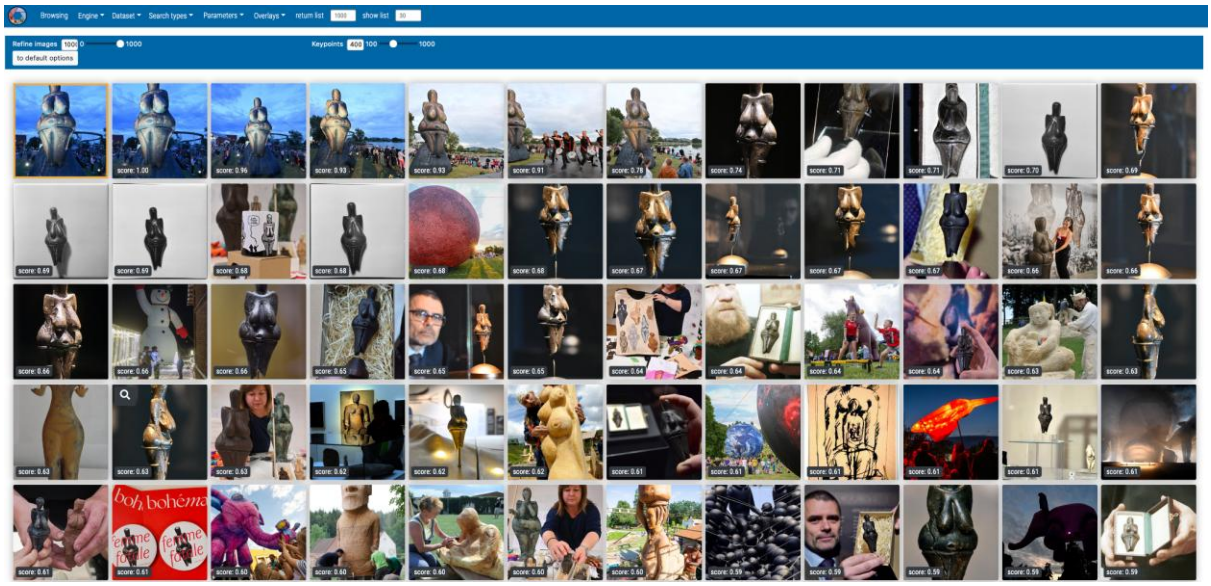
Zabývali jsme se vyhledáváním obrazů na úrovni instancí, jehož cílem je identifikovat všechny snímky konkrétního objektu (památky, uměleckého díla, produktu) ve velkých a různorodých kolekcích. Na rozdíl od doménově specifických systémů vyžadují reálné aplikace modely, které se dokážou zobecnit napříč doménami, protože trénovací a testovací instance jsou přirozeně disjunktní.

Představujeme model založený na podobnosti, který:
(i) spoléhá na korespondence lokálních deskriptorů namísto hrubých vizuálních rysů, čímž zavádí silné induktivní biasy pro interpretovatelnost a efektivitu;
(ii) zpřesňuje lokální podobnosti pomocí optimální přepravy s datově závislými vahami, která potlačuje neinformativní deskriptory;
(iii) agreguje silné korespondence pomocí naučitelného hlasovacího mechanismu do robustního skóre na úrovni celého obrázku.

Metoda se zobecňuje na neviděné domény bez potřeby rozsáhlých trénovacích dat. Při vyhodnocení na osmi různorodých datasetech (památky, umělecká díla, produkty, multidoménové kolekce) konzistentně překonává existující přístupy v out-of-domain nastaveních, a přitom zůstává výpočetně lehká. Je 10× rychlejší a přibližně o 5 % lepší (mAP) na neviděných doménách než dosud nejlepší metoda.

Tato práce byla odeslána na ICLR 2025 a je aktuálně v recenzním řízení. [Pavel Suma, Giorgos Kordopatis-Zilos, Yannis Kalantidis, Giorgos Tolias. *ELViS: Efficient*

Visual Similarity from Local Descriptors that Generalizes Across Domains. In Proc. ICLR, 2026. In Review.]



Snímek obrazovky z ukázky na 2 milionech obrázků ČTK: Dotazový obrázek (vlevo nahoře) zobrazuje sochu a mezi nalezenými výsledky se objevuje mnoho variant téže sochy i téměř identická menší verze.

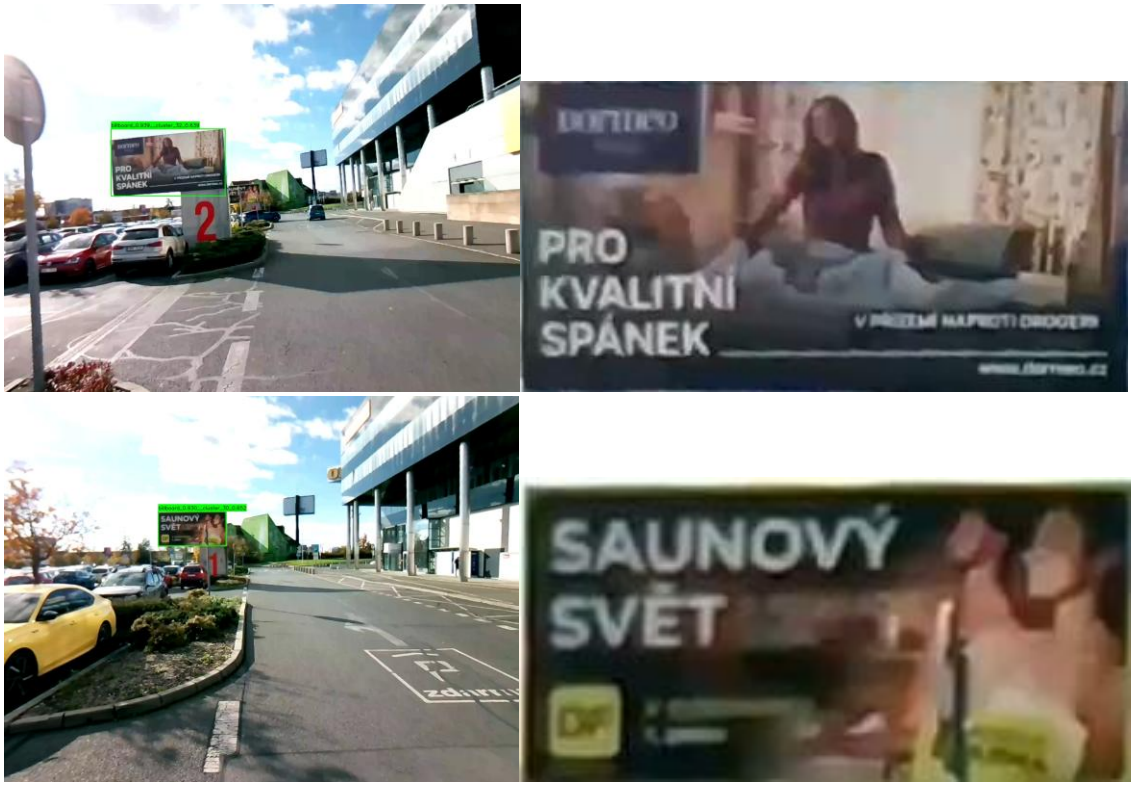
SA08 Detekce manipulace veřejného mínění v exteriéru

V rámci této úlohy jsme studovali metody vhodné pro efektivní lokalizaci nosičů reklam ve fotografiích a videu (například video z palubni kamery automobilu), následné extrakci a rektifikaci obsahu nadetekovaných nosičů; a vývoji metod robustního popisu mediální informace.

V prvních fázích jsme sbírali data pro trénování detekce billboardů a displejů (mobilních telefonů, laptopu, počítačových monitorů a televizních obrazovek). Jelikož obsah těchto mediálních nosičů může být libovolný, vyvýjeli jsme metody augmentace vhodné pro simulaci různého obsahu na stávajících trénovacích datech.

Nasbíraná data a vyvinuté augmentační metody byly použity k trénování detektoru založeného na yolo detektoru. Dále jsme provedli experimenty s klasifikaci detekovaného a extrahovaneho obsahu.

Finální metoda je schopna detekovat reklamní nosiče (billboardy, obrazovky a pod.) ve videu rychlostí 100-400 fps v závislosti na okolnostech (rozlišení videa, GPU, a pod.). Jednotlivá sdělení je možné klasifikovat do předem známých tříd, nebo geometricky rektifikovat a dále zpracovávat - například OCR, nebo detekce a rozpoznávání obličejů.



Příklady detekovaných nosičů - billboardů a rektifikovaného extrahovaného obsahu.

SA09 Nástroje pro moderování diskuzí s přítomností trolů

Diskusní příspěvky na sociálních sítí či doprovázející mediální zprávy umožňují lidem sdílet informaci, tříbit si a získávat nové informace či být exponován různým názorům. Šířená informace má tak velmi významný impakt na náš život, neboť významně ovlivňuje politické situace a volby. Bohužel, vytváří i prostředí, kdy se zájmové skupiny snaží prosadit své cíle šířením dezinformací, kdy dochází k systematickému a úmyslnému klamání. V řadě případů se původci uchylují k velmi agresivnímu napadání jednotlivých příspěvků s cílem rozbít jejich diskusi. Diskusní fóra se brání těmto potížistům, v řadě případů označovaných jako trollové, blokováním jejich kont a příspěvků. Vzhledem k rozsáhlému čím dále objemnějšímu proudu příspěvků však práce lidí-moderátorů je čím dále méně efektivní, v řadě případů vedoucí i na psychické potíže těchto lidí.

Cílem této úlohy je vytvořit funkční demonstraci konceptu nástroje, který usnadní moderátorovi detekovat takové příspěvky a jejich autory, kterým v dalších krocích omezí v publikačních aktivitách či končí zrušením konta k přístupu k diskusi. Činnost takových lidí se často projevuje jako tzv. trollení a v tomto směru je v dalším textu označujeme jako trolly, ačkoliv některé definice trollů vymezují jejich činnost částečně odlišně.

Práce na aktivite SA09 byla zahájena dle plánu v lednu 2025. Skupina se v prvních 4 měsících soustředila na rešeršní aktivitu v oblasti hodnocení stylu textu, hodnocení trolovosti pomocí multijazykových modelů postavených na architektuře BERT, experimentům v oblasti identifikaci témat v podmínkách datových sad s více než milionem dokumentů, pro které tradiční technologie již neškálují, a sledování vývoje výzkumu v oblasti temporálních grafových

neuronových sítí (GNN), kterými jsme chtěli aktivity přispívajících do diskusí hodnotit. Ačkoliv knihovny na zpracování statických grafů pomocí GNN poskytují velmi rozumnou prakticky aplikovatelnou podporu, v oblasti temporálních GNN je jejich vývoj mnohem pomalejší.

Nemalé úsilí bylo rovněž věnováno zajištění potřebných datových sad s diskuzními příspěvky. Podařilo se domluvit přístup přes API k datovým sadám s diskuzními příspěvky z Novinek.cz, které poskytla firma Newton Media, s.r.o. V lednu se doladily některé detaily poskytovaných datových položek. Pro období od 1.9.2024 do 1.4.2025 bylo poskytnuto 23950 původních zpráv a 1044732 diskusních příspěvků od 75140 diskutujících. Hloubkovou manuální analýzou jsme však zjistili, že řada diskusních příspěvků v datové sadě chybí v porovnání s WWW stránkami Novinky.cz. Autoři jsou identifikování pouze svým nejednoznačným jménem, ne unikátním identifikátorem profilu, což může vést k různému maskování autorů se stejným jménem. Nejsou zachovány ani návaznosti vzájemných reakcí diskutujících, ani počty lajků. Identifikované sekvence diskusí jsou krátké a neodpovídají těm prezentovaným na WWW stránkách. Je potřeba ihned poznamenat, že scrapovaní dynamických stránek třídy novinky.cz není vůbec jednoduchá záležitost a dostupné scrapovací nástroje jsou v tomto směru nedostatečně vybavené. V našem případě stav datové sady výrazně omezuje možnosti řešení budování modelu chování diskutujících pomocí temporálních grafových neuronových sítí.

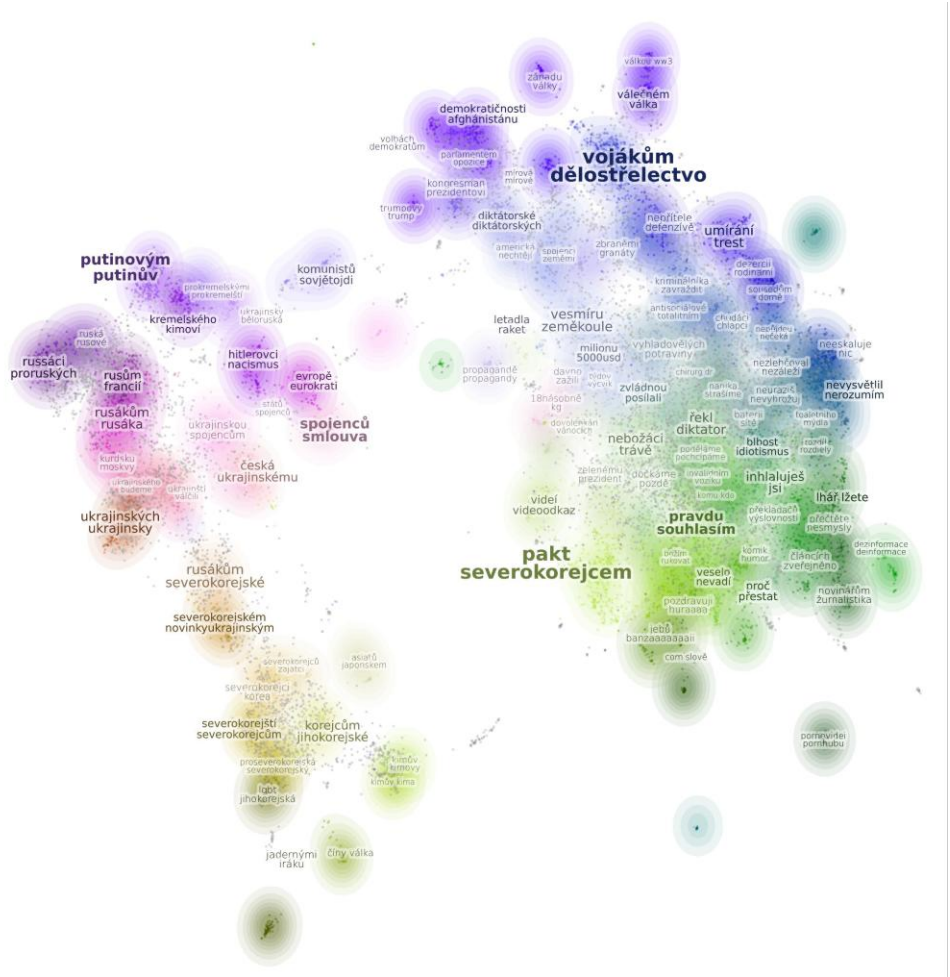
Vlastností specifickou pro českou kotlinu je celkový tón diskusních příspěvků. Příspěvky jsou v drtivé většině případů neutrální a negativní, vyjadřující pocity negativně naladěných či rozlobených diskutujících. Téměř pro každého diskutujícího s více příspěvky lze nalézt příspěvek, který lze považovat za projev trollení. S nadsázkou lze prohlásit, že pokud bychom každého diskutujícího prohlásili apriori za trola, pak dosáhneme přesnosti vyšší než 90%. Mělo proto smysl odstoupit od původního návrhu pracujícího s binární klasifikací troll/ne-troll a přistoupit ke konstrukci míry trollení s nějakou škálou. Tato míra může být i vektorem, který mapuje míru trollení v různých směrech.

Místo klasických technik postavených na hodnocení stylu, sentimentu, případně dalších specifických projevů trolení, jsme navrhli systém, který konstruuje skalární míru trollení pomocí regrese vnořených vektorů textu generovaných jazykovými modely postavených na architektuře BERT. Systém byl otestován na anotovaných datových sadách anglických textů, na kterém jsme dosáhli přesnějších výsledků v porovnání se systémy postavených na stylometrických příznacích. Nicméně pokusy o přenos metody na české příspěvky ukázal, že vlastnosti českého jazyka nemusí být dostatečně správně reprezentovány. Experimenty však naznačují, že by řešení mohla pomoci anotace alespoň malého datasetu s českými příspěvky. Vzhledem k celkovému negativnímu postoji českých příspěvků však vzniká otázka, jaký projev lze považovat pouze za klasický český negativismus a od jaké úrovně příspěvek hodnotit už jako projev trollení.

Kontaktovali jsme firmu TrollWall, která se údajně detekcí trollů měla zabývat. Firma se však zaměřuje pouze na detekci emotivních textů typu "hate speech". Rovněž řeší problém identifikace konkrétního jedince působícího na více platformách. Další jiný subjekt schopný poskytnout klasifikaci textů vystavených působení trolů jsme nezjistili.

Na datové sadě byl rovněž testovány možnosti použití témat, konkrétně témat vět. Jedná se o netriviální úlohu, neboť datová sada obsahuje 4261790 vět v diskusích. Na takové množství vektorů tradiční nástroje jako Top2Vec či Bertopic již nejsou připravené. Konkrétně, dílčí kroky

postavené na metodách UMAP a DBSCAN s kvadratickou náročnosti jsou v současných implementacích limitované na počty okolo 1 milionu vektorů. Nicméně vhodným vzorkováním lze témata získat. Řadu témat lze pak považovat za shluky obsahující věty používané typicky trolly, viz přiložený diagram témat vět. Experimentům věnovaných mírám zastoupení takových vět v příspěvcích se chceme věnovat v následujícím období projektu.



V posledních dvou měsících kalendářního roku 2025 jsme se rovněž věnovali možnosti použití velkých jazykových modelů (LLM) při rozboru diskusních příspěvků s ohledem na projevy chování trolů, případně použití propagandistických technik. V hodnocení propagandistických technik LLM poskytují dle publikovaných článků v současné době mnohem horší přesnost detekce. Nicméně, LLM se s příchodem ChatGPT, zvláště poslední verze 5.2 velmi zlepšují a použití specializovaných technik na bázi multi-agentních řešení výrazně zvyšují přesnost výsledků. Cílem současných prací je vytvořit anotovanou datovou sadu příspěvků s projevy trolení s pomocí LLM.

SA10 Nástroje pro analýzu vysvětlitelnosti a robustnosti velkých jazykových modelů

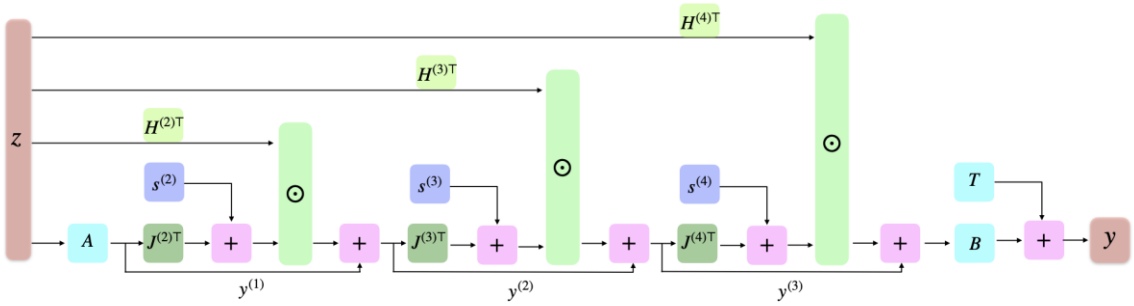
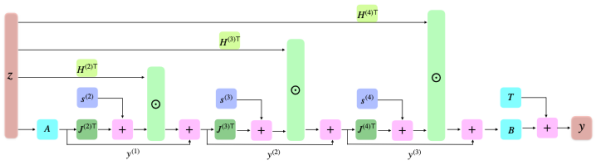


Figure 4.1 Schematic map of modified R-PolyNets for expansion of degree 3 with respect to the input z.

Výsledek (Chrysos et al.), publikovaný na CVPR 2023 demonstroval, že neuronové sítě s polynomiálními aktivačními funkcemi (příklad na obrázku výše) mohou při správné normalizaci signálu dosahovat srovnatelných výsledků jako sítě s klasickými aktovačními funkcemi v úlohách zpracování obrazu. Prostudovali jsme (Chrysos et al.) model, vytvořili explicitní polynomiální parametrizaci, popsali algebraické parametry prostoru funkcí, který model dokáže vytvořit a míru redundance modelu. Náš výsledek ukazuje, že model je často neschopen generovat libovolné polynomiální fukce ale zároveň je přeparametrizován. Nabízí se tak možnost vytvořit modely, kterí budou efektivnější a zároveň expresivnější. Zároveň jsme schopni pro daný model vyhodnotit jeho efektivitu a expresivitu a tak lépe porozumět možnostem a omezením daného modelu. Výsledek plánujeme publikovat na CVPR 2026.

Na počátku projektu jsme započali zkoumat popis a analýzu modelů založených na transformerech a polynomiálních nelinearitách. Vyšli jsme z prací [A] J. Kileel, M. Trager, J. Bruna. On the Expressive Power of Deep Polynomial Neural Networks**.** NeurIPS 2019 a [B] N W Henry, G L Marchetti, K Kohn: Geometry of Lightning Self-Attention: Identifiability and Dimension. CoRR abs/2408.17221(2024). Tyto práce ukázaly, jak lze použít teorii a metody algebraické geometrie k zodpovězení základních otázek ohledně kapacity, singularit a složitosti hlubokých neuronových sítí s polynomiálními nelinearitami a s nejjednodušší variantou architektur založených na transformerech. Tyto matematické modely umožňují dokazatelně vysvětlit vlastnosti sítí používaných ve velkých jazykových modelech.

Naše práce postupovaly ve třech směrech. Zaprvé jsme analyzovali architekturu hlubokých polynomiálních síti, která byla popsána v [C] G G Chrysos, B Wang, J Deng, V Cevher; Regularization of Polynomial Networks for Image Recognition. IEEE/CVF CVPR 2023, abychom přesně popsali model a kapacitu této sítě. Tato architektura je důležitá, protože dosahuje stené úspěšnosti jako architektura založená na RELU, ale lze ji mnohem snadněji matematicky analyzovat. Zadruhé pracujeme na rozšíření výsledků [B], abychom zahrnuli dva důležité bloky, které v [B] chyběly: (a) “skip connections” a (b) normallizaci “self-attention” bloku. Poslední směr, který jsme započali, se věnuje aproximaci velmi složité algebraické funkce, kterou lze využít pro detekci nereálných (fake) konfigurací v obrazech standardními architekturami jako jsou transformery a zároveň chceme využít schopností standardních architektur modelovat statistiku ténovacích dat, kde samotné algebraické metody typicky selhávají.



$$\overline{\varphi}_1(\boldsymbol{z}) = A\boldsymbol{z}$$
$$\overline{\varphi}_i(\boldsymbol{z}, \boldsymbol{y}^{(i-1)}) = (H^{(i)\top}\boldsymbol{z}) \odot (J^{(i)\top}\boldsymbol{y}^{(i-1)} + \boldsymbol{s}^{(i)}) + \boldsymbol{y}^{(i-1)}$$
$$\overline{\varphi}(\boldsymbol{z}) := \overline{\varphi}_n(\boldsymbol{z}, \overline{\varphi}_{n-1}(\boldsymbol{z}, \dots \overline{\varphi}_2(\boldsymbol{z}, \overline{\varphi}_1(\boldsymbol{z})))).$$
$$f_w(\boldsymbol{z}) := B\overline{\varphi}(\boldsymbol{z}) + \boldsymbol{T}.$$

Schéma třívrstvého modelu R-PolyNets a jeho algebraický popis, kterým lze studovat vlastnosti modelu.

Dále jsme pokračovali jsme v práci na analýze polynomiálních sítí R-PolyNets (Chrysos et al.) a dospěli k několika zajímavým výsledkům (Smolíkova 2005): 1) Popsali jsme explicitní vzorec pro rekurzivní konstrukci Jakobiánu mapy konstrukce koeficientů modelů R-PolyNets, která umožňuje explicitní výpočet dimenze prostoru funkcí, které zadý model dokáže modelovat. Tím jsme pro daný model schopni měřit jeho expresivitu a porovnávat modely mezi sebou. 2) Odvodili jsme přesný vzorec pro výpočet dimenze jedné vrstvy R-PolyNets a vzorec pro výpočet horní meze dimenze pro modely s n vrstvami. Na základě experimentů jsme předložili conjecture těsnější meze, viz následující obrázek.

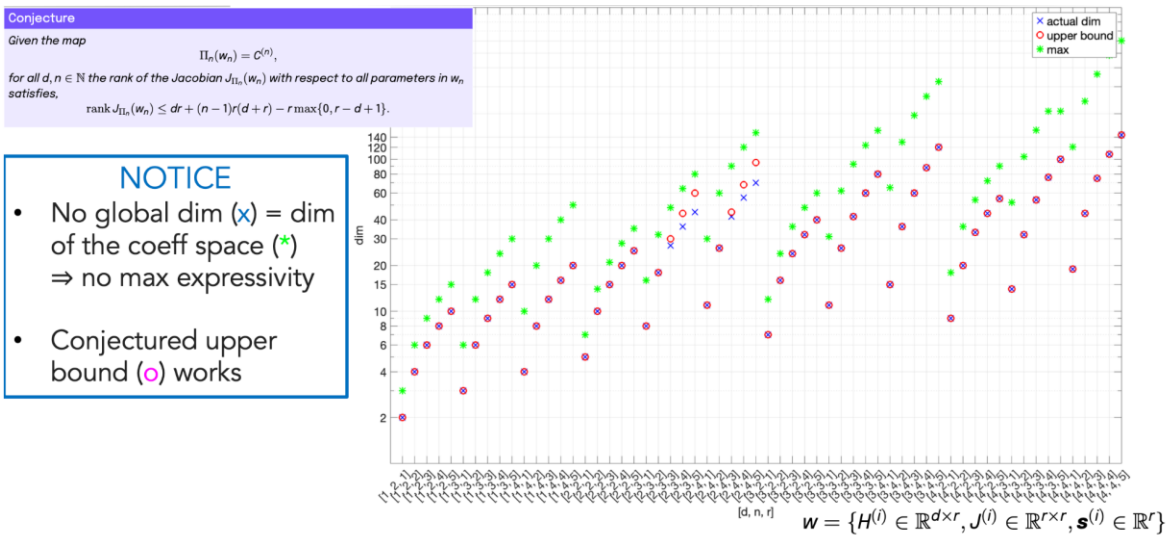


Figure: Plot comparing *'the upper bound'* from Conjecture, a total (*'max'*) number of polynomials of degree *n* in *d* variables and a computed (*'actual dim'*) dimension of the neuromanifold.

3) Ukázali jsme, že množiny modelů, které odpovádají stejnému zobrazení, jsou velké, a tedy že model je značně redundantní, což otevírá možnost navrhnout efektivnější modely se stejnou expresivitou. Momentálně pracujeme na potvrzení nebo vyvrácení navržené těsné horní meze na dimenzi modelu a na příspěvku na konferenci.

Ukazuje se, že řadu otázek ohledně polynomiálních ML modelů lze zodopvědět matematickými nástroji. Kláčovou otázkou ale zůstává, nakolik lze tyto závěry aplikovat na nepolynomiální sítě, V minulém období jsme se zapožali zabývat otázkou, jak dobře dokáře obecný ML model, např. vácevrstvý perceptron nabo transformer, aproximovat danou polynomiální model. Studujeme konkrétní důležitý problém v hledání korespondencí mezi obrazy, který může být použit k detekci falešných a halucinovaných obrazů. Práce (Fløystad 2005) odvodila složitou polynomiálná funkci, která dokáže detekovat konfigurace korespondencí, které nemohly vzniknout jako obrazy reálného světa. Naše první experimenty (Sungatullina 2025) naznačují, že pro praktická situace je možno tuto složitou polynopmiální funkci lze dobře aproximovat relativně jednoduchým vícevrstvým perceptronem. Další

otázkou, nyní je, jak popsat model realizovaný perceptronem jednoduchou polynomiální mapou.

(*Chrysos et al.) Chrysos, Grigorios G.; Wang, Bohan; Deng, Jiankang; Cevher, Volkan, 2023. Regularization of Polynomial Networks for Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023).*

(Smolíkova 2005) N. Smolíková (T. Pajdla supervisor). Algebraic Structure of R-PolyNets Neural Networks. MSc thesis. Charles University in Pague. 2005.

(G. Fløystad, J. Kileel, G. Ottaviani). The Chow Form of the Essential Variety in Computer Vision. Journal of Symbolic Computation Volume 86, May–June 2018, Pages 97-119.

(Sungatullina 2025). D. Sungatullina, T. Pajdla. Preliminary results on the use of Chow form of the Exxential Variety for Image Matching Verification. Work in progress 2025.

Guardrail model na bázi BERTu

Natrénovali jsme guardrail model typu BERT s 400 miliony parametry. Primární funkcí tohoto modelu je zamítání generování nevhodného obsahu, jako jsou například požadavky na kód pro DDoS útoky nebo SQL injection skripty.

Na testovacích sadách od Amazon AWS dosahujeme působivé přesnosti přibližně 97 %. Model byl trénován s využitím syntetických dat, která jsme připravili za pomoci modelu Anthropic Claude 3.7.

Ve spolupráci s Ministerstvem financí jsme rovněž trénovali doménově specifické modely typu BERT pro generování embeddingů. Trénovací sada obsahovala přibližně 40 000 stránek PDF dokumentů, a provedli jsme plné doladění (fine-tuning) modelu Alibaba BERT s 300 miliony parametrů.

Výsledný model překonal OpenAI model text-embedding-3-small a dosáhl výkonu velmi blízkého modelu text-embedding-3-large, který je dnes de facto standardem pro generování embeddingů v mnoha systémech.

Doladění menších jazykových modelů (LLM) pro provoz v místním prostředí

Vidíme velký potenciál v doladění menších jazykových modelů s přibližně 10 miliardami parametrů. Nedávno byly představeny kompaktní počítače vybavené nejnovějšími čipy NVIDIA, což otevírá možnost provozovat inference těchto modelů lokálně za přijatelnou cenu. To je obzvláště důležité pro organizace, které požadují ochranu dat a preferují uchovávat svá data za firewallem.

Generování kódu pomocí 10B modelu

V současnosti se zaměřujeme na doladění modelu s 10 miliardami parametrů pro generování Python kódu. Pro tento účel jsme:

Synteticky vytvořili datovou sadu obsahující 700 000 dvojic instrukce–kód a pomocí Supervised Fine-Tuning (SFT) jsme model sladili s těmito úlohami.

V další fázi jsme vytvořili 10 000 syntetických trojic: otázku, správnou odpověď a nesprávnou odpověď. Tyto trojice jsme využili pro Direct Preference Optimization (DPO), což vedlo k výraznému zlepšení přesnosti.

Nakonec jsme aplikovali upravenou verzi Proximal Preference Optimization (PPO) s ověřitelnou odměnou, založenou na několika tisících preferenčních trojicích, pro finální sladění modelu.

Další kroky: V současnosti odhadujeme výkonnost modelu na obecných benchmarkových sadách. Tato fáze vyžaduje další úpravy a výsledky budeme sdílet v příští aktualizaci.

SA11 Nástroje pro zvýšení produktivity zpravodajských médií

Naše skupina se specializuje na výzkum využití Large Language Models (LLM) pro administrativní aplikace zaměřené na zvyšování produktivity práce. Výzkum se soustředí na dva klíčové směry. Prvním a nejdůležitějším je optimalizace LLM, a to s ohledem na minimalizaci počtu potřebných parametrů při zachování požadované přesnosti v konkrétní doméně. Druhým směrem je vývoj uživatelsky přívětivého a efektivního rozhraní, které umožní praktické nasazení modelů v reálných pracovních procesech.

V současnosti jsme se soustředili na vývoj aplikace, která umožňuje prohledávání rozsáhlých datových souborů. Například organizace, které kontrolují, zda jejich výrobky splňují zákonné požadavky v různých geografických oblastech, nebo v aplikacích, kde se provádí podobné kontroly. Aplikace má také velký potenciál například v právních kancelářích, obchodních organizacích a podobně.

V oblasti výzkumu v této souvislosti pracujeme na vylepšování algoritmů. Tato činnost zahrnuje tvorbu testovacích posloupností, trénink LLM, které generují reprezentace textu atd. Současná aplikace využívá algoritmus ze třídy Retrieval Augmented Generation (RAG).

Aplikace pracuje v několika krocích:

- Nejprve všechny dokumenty rozdělíme na jednotlivé stránky a každou stránku reprezentujeme vektorem.
- Fáze hledání odpovědi začíná vygenerováním vektoru pro otázku, po čemž následuje vyhledání nejvíce semanticky podobné stránky.
- Tuto stránku poté využijeme k vygenerování odpovědi pomocí LLM.

Tento přístup umožňuje efektivní a přesné zpracování velkých objemů dat a poskytuje relevantní odpovědi na komplexní otázky.

Pokrok v Retrieval-Augmented Generation (RAG)

V poslední době se zaměřujeme na zlepšení vyhledávání v dokumentech pomocí techniky Retrieval-Augmented Generation (RAG). V rámci tohoto úsilí jsme testovali různé strategie pro přepis dotazů (query rewriting), avšak zaznamenali jsme pouze mírné zlepšení přesnosti.

Dále jsme experimentovali s metodou Hypotetických dokumentových embeddingů (HyDE), která rovněž přinesla přínosy v určitých specifických nebo obtížných scénářích vyhledávání. Obě metody – přepis dotazů i HyDE – byly užitečné v okrajových případech, ale celkový výkon významně nezvýšily.

Současně jsme zkoumali různé přístupy ke zvýrazňování relevantních částí textu v rámci top-N dokumentů vrácených RAG systémem. Na našich testovacích datech v češtině jsme dosáhli přesnosti blížící se 90 %, což ukazuje na silný potenciál pro praktické využití.

Na základě dosavadního výzkumu jsme rozšířili naše aktivity v oblasti menších jazykových modelů a agentních systémů. Trénovali jsme několik kompaktních LLM modelů, které tvoří základ nového agentního systému Alquist Insight.

Tento systém umožňuje vytvářet aplikace složené z menších (<10B), na konkrétní úkoly zaměřených LLM modelů. Takový modulární přístup snižuje nároky na výpočetní výkon a zároveň zvyšuje přesnost při řešení úzce vymezených problémů v podnikových a administrativních procesech.

Pro agenta provádějícího klasifikaci textů jsme natrénovali model s 308 miliony parametrů, který generuje embeddingy pro následnou downstream neuronovou síť provádějící vlastní klasifikaci.

Probíhá vývoj modelů pro sumarizaci právních textů, extrakci dat z uživatelských dotazů pro e-commerce aplikace a pro personalizované doporučovací systémy, založené na kombinaci embeddingových a retrieval technik.
V projektu ADAM jsme pokračovali v integraci do cloudové infrastruktury a v posilování bezpečnostních mechanismů. Pracujeme na highlighting relevantních řádků v PDF dokumentech (zvýraznění textu použitého pro generování odpovědí). Dále jsme vyvinuli aplikaci pro generování syntetických dat určenou pro trénink modelů.

Klíčové vlastnosti této aplikace:

- využívá více metod generování textu (různé typy modelů a generativních postupů) k dosažení vysoké datové diverzity,
- do syntetických dat aktivně vkládáme typické chybové vzory (např. opakování slov při diktování, pravopisné chyby při zadávání z klávesnice), aby modely byly robustní vůči reálným vstupním chybám,
- integrujeme synonymické slovníky a varianty frázování pro maximální pokrytí různých způsobů vyjadřování,
- k aplikaci patří sada nástrojů pro kvantifikaci diverzity dat pomocí několika metrik (např. lexikální rozmanitost, distribuce n-gramů, embeddingová vzdálenost), čímž dokážeme hodnotit a optimalizovat kvalitu tréninkových sad.

Pokračujeme v trénování modelu pro generování Python kódu, přičemž vycházíme z Phi-4 (Microsoft) foundation modelu a z upravených dat popsaných v předchozí aktualizaci. V této fázi máme pozitivní výsledky:

Dosahuje výrazného zvýšení odolnosti modelu vůči vytváření vulnerability-prone a malicious kódu díky kombinaci SFT, DPO a dalších technik finetuningu použité dříve, Pokračujeme v rozšiřování syntetických a preferenčně označených dat pro další zlepšení kvalitativních i bezpečnostních vlastností generovaného kódu.

Tyto aktivity společně představují důležitý posun k praktickému nasazení agentních systémů a menších LLM v kontextu české veřejné správy i komerčního sektoru — s důrazem na bezpečnost, lokální provoz (on-premise) a použitelnou datovou diverzitu pro reálné aplikace.

Chatbot ADAM

Vyvinuli chatbota, který dovoluje pracovníkům v auditním orgánu odpovídat na dotazy z pravidel pro provádění operačních programů. Chatbot je postaven na architektuře RAG. Skládá se ze dvou SaaS aplikací administrativní a interaktivního interface. Administrativní aplikace slouží k monitorování činnosti a konfiguraci dokumentů. V současnosti ADAM pracuje s 49 PDF souborů. Klientská aplikace zobrazuje odpověď na dotaz a současně i tet, x kterého byla odpověď vytvořena, tak aby uživatel mohl provést kontrolu. ADAM byl součástí projektu EDIH, který získal AI Award za rok 2025.

Optimalizace kontextuální návaznosti

Provedli jsme výzkum implementace scénářů, kdy se aktuální dotaz odkazuje na předchozí konverzaci. Po testování metod vkládání celé historie do kontextu (které zvyšovalo latenci) jsme jako optimální kompromis zvolili metodu reformulace dotazu pomocí LLM s použitím předchozího kontextu. Tento přístup jsme validovali na vygenerovaných testovacích datech a docilujeme přesností více jak 90 procent v závislosti na typu dat.

Vlastní embeddingové modely:

Pokročili jsme ve vývoji vlastních LLM kodérů typu BERT. Na datovém souboru 40 tisíc PDF s předpisy evropských operačních programů jsme natrénovali model o velikosti 600 milionů parametrů. Výsledky ukazují, že model překonává text-embedding-3-small a dosahuje kvality srovnatelné s verzí large od OpenAI.

On-premise infrastruktura (NVIDIA Spark)

Úspěšně jsme zprovoznili kompletní RAG systém na vlastním hardwaru NVIDIA Spark se 128GB unifikované paměti. Systém nyní dosahuje podobné odezvy jako při využití OpenAI API. Embeddings generujeme modelem viz. předchozi odstavec. Pro generování odpovědí aktuálně využíváme model Qwen3-30B MoE a testujeme s metodologií LEO. V další fázi se plánujeme zaměřit na finetuning tohoto modelu specificky na českých datech pro další zvýšení kvality výstupů

Guardrail for a Coding LLM

V rámci vývoje modelu pro kódování v Python jsme vyvinuli guardrail model, který zvyšuje bezpečnost a blokuje uživatelské požadavky na generování malicious kódu. Vycházeli jsme z

modelu ModernBERT large s 395M parametry. Model jsme trénovali na 220k vzorcích. Model blokuje více jak 85% požadavků na generování nevhodného kódu. Model zpracuje , propustí 99% požadavků na trénování bezpečného kódu.

Safe Coding LLM

Pro bezpečné trénování Python kódu jsme trénovali model pro generování kvaltího kódu, který neobsahuje skryté zranitelnosti (vulnerabilities). Vycházeli jsme z Phi-4-mini-instruct s 3.8B parametry. Trénink měl dvě fáze 1) SFT s naší vlastní trénovací sadou s 700k vzorky.
a 2) DPO s více jak 40k dvojic vzorků. Ve výsledku jem překonali SOTA LLMs o 30% (absolute) na benchmarcích pro cybersecurity.

NSS

Pro Nejvyšší správní soud (NSS) jsme vyvinuli LLM asistenta, který připravuje výtahy z textů jednotlivých soudních případů u soudu nižích instaní. Vyšli jsme z foundational modelu Llama 3.1 s 8B parametry. Pro trénink jsme použili algoritmus LoRA a trénovací posloupnost s 12k vzorky. Délka kontextu je 60k tokens umožňující pracovat i s rozsáhlými právními dokumenty. Na české právní doméně podle Elo testů dosajuje natrénovaný model lepších výsledků než GPT-5.

MHMP Klasifikátor

Pro MHMP jsme natrénovali klasifikátor, který klasifikuje podání občanů k jednotlivým agentům podle tematiky podání. Vycházeli jsme z modelu:
Alibaba-NLP/gte-multilingual-base MLP, classification heads, 305M parameters
Trénováno na 14,487vzorcích. Výsledný systém pracuje jako hierarchický klasifikátor (topic → block → thematic block) Na testovací sadě dosahujem přesnosti 96% a 94% F1 s Optuna hyperparameter optimalizací.

ADAM zvýraznění textu odpovědi

Pro auditní orgán MF jsem updatovali bota ADAM. Vyvinuli jsme algoritmus, který zvýrazní text z kterého byla vytvořena odpověď. Tato úpravu zvyšuje rychlost orientace v textu pro ověření správnosti odpovědi. Extractivní model poro otázky a odpovědi pro RAG ADAM vychází z modelu microsoft/mdeberta-v3-base s Extractive QA head a má 276M parameterů. Model jsme předtrénovali na SQuAD v2 (130k samples), dosáhli jsme přesnosti 82% Exact Match a 86% F1. Model byl fine-tuned na syntetických datech AOMF.

SA12 Nástroje pro analýzu velkých textových korpusů v médiích

V rámci prvního období jsme navrhli a implementovali **cloudové řešení každodenního automatického stahování volně dostupných článků a komentářů z několika vícejazyčných zdrojů**. K tomu jsme použili Amazon Web Services (AWS), cloudovou platformu od společnosti Amazon, která nabízí dobrý poměr “cena/výkon” a snadnou škálovatelnost s přibývajícími datovými zdroji. Systém pro každodenní automatické stahování dat je prvním krokem pro vybudování systému pro automatickou identifikaci dezinformací v médiích.

Zatím jsme schopni denně stahovat data ze dvou zdrojů:
Google News - služba společnosti Google, která agreguje zprávy a novinky z různých zdrojů po celém světě. Z tohoto zdroje jsme schopni dostat přibližně 500 - 600 MB textových dat pro pracovní den a cca 300 - 350 MB textových dat pro víkendový den. Zdroj obsahuje data v 35 jazycích.
Reddit - populární sociální zpravodajský a diskuzní web, který umožňuje uživatelům zveřejňovat obsah, jako jsou odkazy, obrázky nebo textové příspěvky, a následně je hodnotit pomocí hlasování. Z tohoto zdroje jsme schopni dostat cca 60 MB dat pro pracovní den a 40 MB dat pro víkendový den. Do budoucna plánujeme rozšíření tohoto datového zdroje, Reddit je velmi populární platforma, a tedy zde je potenciál získat z něj násobně více dat.

Dále pracujeme na rozšíření datasetu - máme **rozpracovaný scraper na Telegram**, po jeho dokončení plánujeme stahovat data s technickými diskuzemi - Stack Exchange, Hacker News. Po shromáždění dostatku dat začneme s experimenty s modely.

Tento projekt se zaměřuje na využití veřejně dostupného mediálního obsahu a uživatelsky generovaných dat s cílem vyvinout metodu pro automatickou detekci dezinformací (fake news) a identifikaci společenských trendů. Klíčovým zdrojem informací jsou textová data získávaná z médií a sociálních sítí, která slouží jako základ pro následnou analýzu prostřednictvím jazykových modelů.

V aktuálním období jsme se soustředili primárně na fázi získávání dat. To zahrnovalo jak automatizovaný sběr aktuálních textových dat, tak retrospektivní shromažďování historických dat. Za tímto účelem byly vyvinuty specializované nástroje pro web scraping, které nám umožnily pravidelně získávat obsah z českých i zahraničních zpravodajských serverů. Současně jsme se zaměřili na získávání dat ze sociálních sítí, konkrétně z platforem Reddit a Telegram, kde se nám podařilo extrahovat množství relevantního textového materiálu.

Získaná data jsou periodicky ukládána ve strukturovaném formátu JSON, což umožňuje jejich efektivní zpracování a porovnání v čase. Analýza se zaměřuje zejména na změny v obsahu a jazykových trendech napříč různými časovými obdobími. K tomu využíváme velké jazykové modely, které jsou buď dotrénovávány na konkrétní časové intervaly, nebo v některých případech trénovány zcela od začátku.

V průběhu tohoto období jsme experimentálně ověřili celý postup na několika open-source modelech, například LLaMA 3 a GPT-2. Výsledkem je sada modelů, z nichž každý je přizpůsoben určitému časovému úseku. Pro tyto modely následně vyhodnocujeme tzv. generační pravděpodobnosti pro sadu připravených vět. Z těchto pravděpodobností lze pozorovat, jak se mění „populárnost“ daných vět v závislosti na časovém kontextu.

Prvotní experimenty ukazují, že pro známé a mediálně sledované události vykazuje generační pravděpodobnost očekávané chování — tedy v období, kdy je daná událost aktuální, je pravděpodobnost jejího generování výrazně vyšší. Výkonnost jednotlivých modelů se však liší v závislosti na jejich architektuře a velikosti. Z tohoto důvodu jsme začali pracovat na vývoji objektivní metriky, která by umožnila systematicky porovnat modely mezi sebou. Tato metrika bude klíčová jak pro výběr nejvhodnějších modelů, tak i pro kvantitativní hodnocení úspěšnosti celé navržené metody.

SA13 Nástroj pro analýzu konzistence generovaných textů

V rámci této aktivity byly vyhotoveny dvě studie, které pracovně nazýváme jmény “Neural CSP” a “2D Early Exit”. Kromě původně plánované studie (Neural CSP), která má za cíl porozumění konzistence generovaných textů jazykovými modely jsme rozpracovali techniku, která umožní klasifikaci textů v podstatně kratší době, než kterou by vyžadovala klasifikace pomocí standardních přístupů (2D Early Exit). Na základě této techniky jsme vytvořili nástroj, který umožní efektivně detekovat hledaný jev ve velkém jazykovém korpusu.

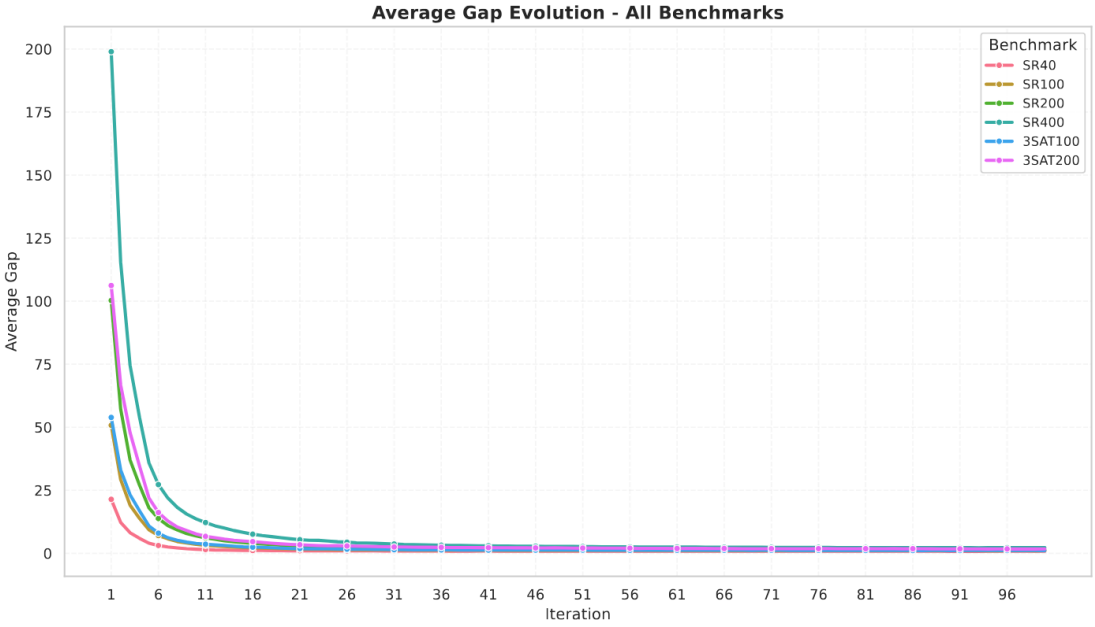
Neural CSP

V rámci studie **Neural CSP**, která se věnuje konzistenci textů, jsme vytvořili generátor syntetických dat umožňující testovat chování modelu přesně kontrolovatelným způsobem. Tato data obsahují sady faktů, které mohou či nemusí být konzistentní, což nám dovoluje studovat reakce modelu na konfliktní informace. Příkladem nekonzistence může být trojice tvrzení: *„Každý člověk je smrtelný“*, *„Sokrates je člověk“* a *„Sokrates je nesmrtelný“*.

Na těchto syntetických datech jsme otestovali několik modelů založených na architektuře **grafových neuronových sítí (GNN)** nebo **Transformerů**. Grafové sítě využíváme především pro jejich snazší interpretovatelnost. Experimenty ukázaly, že zatímco GNN lze škálovat na velmi rozsáhlé množiny faktů, v nichž spolehlivě detekují rozpory, Transformery často selhávají v případech, kdy počet faktů nutných k detekci nekonzistence překročí hranici deseti.

Součástí studie byla také analýza mechanismu, který modelům detekci nekonzistence umožňuje. Výsledky ukazují podobnost mezi empirickým chováním natrénovaných modelů a solvery pro semidefinitní programování (SDP), které optimalizují cílovou funkci pomocí gradientního sestupu. Tyto solvery lze mimo jiné využít právě k detekci logických rozporů a naše analýza potvrzuje, že neuronové sítě pro tuto úlohu využívají analogický princip.

Mechanismus se pokouší nalézt takovou interpretaci, aby byla všechna tvrzení splněna. Pokud se to nepodaří, je množina faktů vyhodnocena jako nekonzistentní. V korektně naučené síti probíhá hledání této interpretace postupně skrze jednotlivé vrstvy. Každou vrstvu lze v tomto kontextu vnímat jako jeden gradientní krok, který optimalizuje funkci maximalizující počet splněných tvrzení. Následující obrázek ilustruje, jak se počet nesplněných tvrzení (osa y) snižuje v průběhu jednotlivých vrstev (osa x). V tomto případě grafová neuronová síť během několika vrstev úspěšně nachází interpretaci, která splňuje maximální počet tvrzení.

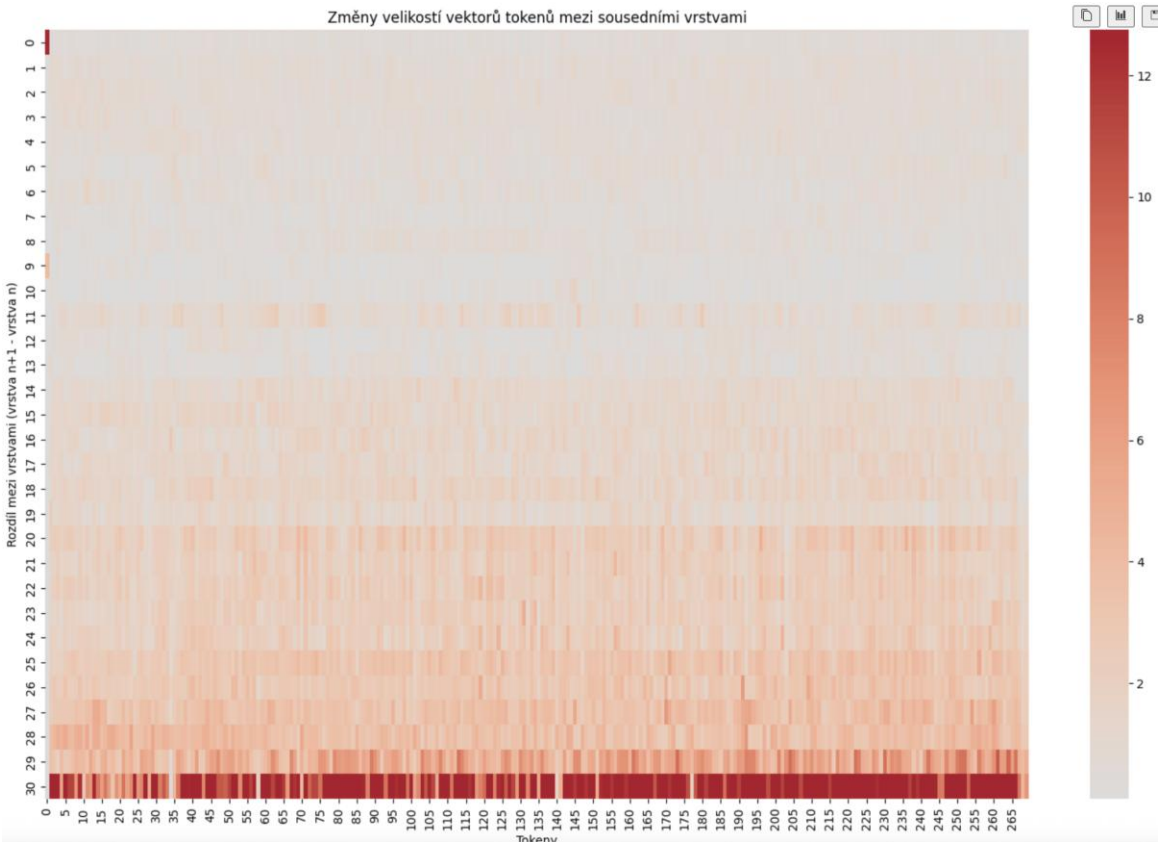


Tyto výsledky byly sepsány v publikaci “Neural Approaches to SAT Solving: Design Choices and Interpretability”, která byla přijata do žurnálu “International Journal of Approximate Reasoning” (IJAR) v prosinci 2025 a v publikaci “Geometric Reasoning in the Embedding Space”, která byla publikována v žurnálu “Machine Learning and Knowledge Extraction” (MAKE) v srpnu 2025.

2D Early Exit

V rámci výzkumu metod pro zrychlení klasifikace textu pomocí jazykových modelů jsme vyhodnotili několik přístupů, které jsou ortogonální k běžně užívané kvantizaci či destilaci. Jako nejperspektivnější se jeví technika **dynamické alokace výpočtu** pro jednotlivé segmenty textu. Tato metoda vychází z experimentů popsaných v předchozí studii (Neural CSP). Úspory je dosaženo tím, že model nezpracovává všechny věty vstupního textu v plné hloubce všech vrstev, ale adaptivně rozhoduje o potřebném výpočetním úsilí pro každou větu zvlášť.

Již v předchozí studii jsme demonstrovali, že neuronová síť hledá interpretaci vstupních tvrzení progresivně napříč vrstvami. U jednodušších tvrzení je interpretace stabilizována dříve, což znamená, že zbývající vrstvy již výsledek neovlivňují a lze je bez ztráty informace vynechat. Tento jev ilustruje následující obrázek, který zobrazuje míru změny reprezentace jednotlivých tokenů (osa x) v závislosti na hloubce vrstvy (osa y) u modelu Llama 3.2. Barevná škála zde znázorňuje velikost odchylky (normu rozdílu) vektorové reprezentace tokenu mezi dvěma po sobě jdoucími vrstvami.



Výsledná technika umožňuje v průměru **zkrátit čas inference o 90 %** při zanedbatelném dopadu na přesnost.

Výsledky této studie jsou popsány v článku s názvem “2D Early Exit”, který bude zaslán do žurnálu “Neural Computing” na začátku roku 2026.

SA14 Nástroje pro monitoring hlasových zpráv ve více jazycích

Práce na této aktivitě proběhla v prvním projektovém období podle plánu. V prvním období jsme se v KA2 SA14 zaměřili na podrobný průzkum aktuálních metod pro vyhodnocení latence u simultánního přepisu a překladu řeči, které je nezbytné pro spolehlivý výběr nejlepšího technického řešení. Zjistili jsme, že současná literatura nabízí několik metrik latence, avšak tyto metriky dosud nebyly empiricky porovnány a jejich hodnoty často vzájemně nekorespondují (viz například srovnání systémů na IWSLT 2023). Proto jsme se nejprve zaměřili na porovnání těchto metrik a identifikaci nejspolehlivější z nich. Výsledkem je studie “Better Late Than Never: Evaluation of Latency Metrics for Simultaneous Speech-to-Text Translation” (https://arxiv.org/abs/2509.17349), u níž v současné době usilujeme o publikaci na vhodné konferenci. Studie představuje první systematické vyhodnocení metrik latence pro simultánní překlad řeči (Simultaneous Speech-to-Text Translation; SimulST) z několika hledisek, jako jsou různé systémy, jazykové páry a fungování v režimu zpracování krátkých a

dlouhých promluv. Identifikovali jsme současná úskalí při vyhodnocování SimulST izolováním problémů v nejčastěji používaných metrikách, které jsou známy pod zkratkami AL, LAAL, DAL, AP, a ATD. Pro překonání těchto omezení navrhujeme novou metriku latence Yet Another Average Lagging (YAAL), která je lépe přizpůsobena režimu vyhodnocování latence při zpracování krátkých promluv. Naše analýza však také odhaluje inherentní nedostatky vyhodnocování latence systémů zpracovávajících krátké promluvy, což je další důvod k přechodu k vyhodnocování latence při zpracování dlouhých promluv jako ke spolehlivější alternativě. Dále jsme také ukázali, že resegmentace kandidátního překladu je nezbytná pro správné vyhodnocení systémů pracujících v režimu dlouhých promluv, a navrhli jsme vylepšený nástroj pro resegmentaci spojený s rozšířením YAAL pro hodnocení dlouhých promluv – LongYAAL. Podle výsledků YAAL a LongYAAL překonávají všechny ostatní metriky v obou režimech, čímž ustavují novou aktuálně nejlepší metriku pro Simultánní překlad.

Paralelně s hodnocením metrik latence se účastníme jako spoluorganizátoři workshopu IWSLT, který se zaměřuje na mluvenou řeč, a to jak uplynulém roce 2025, tak v budoucím ročníku 2026. V rámci naší role jsme navrhli soutěžní úlohu simultánního překladu českých mluvených projevů do angličtiny. Díky tomu získáváme srovnání našich systémů z SA14 a SA5 s ostatními vývojáři z celého světa. Aby bylo toto srovnání spolehlivé, potřebovali bychom vyhodnocení provést formou ruční anotace, a proto podáváme žádost o realokaci zdrojů původně určených pro MT Marathon na náklady související s ručním vyhodnocováním kvality výstupu. Pokud bude žádost úspěšná, mohli bychom srovnání rozšířit o lidské tlumočení jako další “systém” a provést detailnější hodnocení pomocí lidských anotátorů.

Rovněž jsme se věnovali implementaci testovacího prototypu systému na simultánní překlad řeči a jeho testování v reálných podmínkách, integraci automatického rozpoznání hlasu do tohoto prototypu, vyhodnocení ASR modelu whisper-large-v3 a studiu možností integrace libovolných překladových modelů do systému simultánního překladu řeči. Taky jsme se věnovali návrhu a vývoji další verze tohoto prototypu s přidanou možností odlíšení řečníků v reálném čase, která je sestavená z modulů na rozpoznání řeči (Whisper), odlíšení řečníků (Diart), a velkého jazykového modelu (DiarizationLM) na spájení jejich výstupů.

Dále jsme pracovali na studiu, implementaci a evaluaci simultánního přepisu řeči podle současného stavu poznání. Vycházeli jsme z práce Simul-Whisper, která implementuje model pro transkripci textu Whisper se simultánní policy AlignAtt, která podle současného stavu poznání dosahuje nejlepších výsledků. Implementaci Simul-Whisper jsme rozšířili a propojili s další implementací Whisper-Streaming, která poskytuje funkce na simultánní zpracování přepisu a překladu dlouhých promluv. Výsledkem je naše nová implementace SimulStreaming. SimulStreaming jsme vyhodnotili v soutěži IWSLT 2025 Simultaneous Speech Translation Shared Task. Řešení jsme popsali v článku s názvem “Simultaneous Translation with Offline Speech and LLM Models in CUNI Submission to IWSLT 2025”.

Od května letošního roku jsme se v souladu se zadáním zaměřili na vytvoření robustní datové sady pro automatický přepis řeči, která je nezbytnou prerekvizitou pro monitoring diskuzních pořadů. Hlavním výstupem je zveřejnění korpusu ParCzech4Speech, který byl osobně prezentovan na konferenci TSD 2025 (publikováno jako ParCzech4Speech: A New Speech Corpus Derived from Czech Parliamentary Data, https://doi.org/10.1007/978-3-032-02548-7_25). S rozsahem 2 695 hodin jde o největší otevřený zdroj pro české ASR, který na rozdíl od studiových nahrávek zachycuje autentické živé projevy. Z velkého objemu vstupních dat

jsme pomocí moderních nástrojů a filtrace extrahovali a zpracovali pouze úseky s vysokou shodou mezi zvukem a přepisem. Zpracování probíhalo pomocí nástroje WhisperX, který kromě přesného zarovnání umožňuje i pokročilé odlišení mluvčích (diarizaci). Pro co nejsnazší využití jsme dataset publikovali v repozitářích LINDAT a Hugging Face pod volnou licencí CC-BY.

Následná analýza dalších dostupných českých zdrojů ukázala, že pro účely monitoringu diskuzí nejsou stávající datasety vhodné. Často v nich chybí pro nás klíčové jevy, jako je překřikování více mluvčích (multi-talker). Proto jsme upustili od kompilace existujících dat a zaměřili se na tvorbu vlastních sad a vývoj specializovaného multi-talker ASR modelu. Vycházíme z aktuálních vědeckých poznatků, podle kterých je trénování a evaluace na synteticky upravených datech efektivním způsobem, jak model připravit na reálné a obtížné scénáře překryvů řeči. Do konce roku jsme proto připravili pipeline, která využívá kvalitní data z ParCzech4Speech pro generování těchto syntetických multi-talker trénovacích sad. V debatách České televize jsme detekovali segmenty s překřikováním a požádali o poskytnutí vzorku pro reálnou testovací sadu.

Do dubna 2025 plánujeme adaptaci existujících českých modelů pro multi-talker úlohy a jejich evaluaci na syntetických datech. V případě prodloužení projektu vznikne i manuálně anotovaná sada přímo z reálných debat ČT. Taková data v českém kontextu dosud chybí a pro ověření funkčnosti systémů v náročných akustických podmínkách jsou nezbytná.

SA15 Nástroje pro detekci faktických chyb ve výstupech LLM

ÚFAL v rámci této subaktivity vyvinul dva evaluační nástroje a provedl několik experimentů s detekcí chyb pomocí LLM a porovnání s lidskými anotátory.

Nástroj Factgenie, který umožňuje vyhodnocovat vygenerované texty pomocí anotací chybných slov nebo frází, jsme publikovali v první verzi už v září 2024. Následně byl v rámci projektu CEDMO2 NPO průběžně aktualizován a rozšiřován, zejména vzhledem k možnostem evaluace pomocí LLMs (integrace knihovny ollama pro lokální LLMs, podpora překrývajících se anotací, lepší vizualizace, integrace několika benchmarků, zjednodušení konfigurace a uživatelského rozhraní, podpora pro více anotací současně, výpočet mezianotátorské shody, šablony pro LLM prompty i lidskou anotaci). Kromě toho jsme zlepšili čitelnost a modularitu zdrojového kódu. Nástroj je plně funkční a k dispozici na https://github.com/ufal/factgenie pod otevřenou licencí MIT.

Vylepšení Factgenie jsme použili k experimentu, kde jsme na úloze anotace chyb v textu porovnali LLM a anotátory z crowdsourcingu s vlastními expertními anotacemi. Uvažovali jsme nejen chyby ve výstupu LLM na úloze generování textu z dat, ale i chyby LLM při úloze strojového překladu a také anotaci propagandistických technik v (lidmi psaném) textu, na kterou lze Factgenie a obdobné přístupy bez problémů aplikovat. Ukázali jsme, že mezianotátorské shody je pro tuto úlohu relativně náročné dosáhnout, ale LLM v některých

případech dosahují úrovně lidských anotátorů. Experiment jsme popsali v článku, který zatím nebyl přijat k publikaci, ale po úpravách na základě recenzí z konference COLM a časopisu TACL ho posíláme do recenze na relevantní workshop o evaluaci generovaných textů na konferenci EACL 2026. Preprint článku je k dispozici na https://llm-span-annotators.github.io/; data, použitelná jako benchmark k dalším experimentům, jsou dostupná na https://llm-span-annotators.github.io/ pod licencí Creative Commons.

Dále jsme vyvinuli metriku OpeNLGauge pro evaluaci NLG založenou čistě na ensemblu LLM běžících lokálně, tj. bez použití komerčních LLM jako je ChatGPT. Kromě evaluace faktičnosti/věrnosti vstupním datům je schopna (s jednoduchou změnou promptu) evaluovat i jiné aspekty kvality vygenerovaných textů (srozumitelnost, koherence aj.). Metrika snese srovnání s nejlepšími LLM metrikami používajícími komerční LLM a kromě toho díky anotaci jednotlivých chyb (podobně jako v experimentu výše) podává podrobnější feedback. Nástroj OpeNLGauge byl prezentován jako poster na specializovaném workshopu ACL GEM (Generation, Evaluation and Metrics) a jako řádný článek ve sborníku konference INLG (https://aclanthology.org/2025.inlg-main.19/). Nástroj je k dispozici pod licencí Creative Commons na https://github.com/ivankartac/OpeNLGauge .

Provedli jsme i několik experimentů s evaluací různých metrik (tradičních založených na překryvu slov, metrik s menšími předtrénovanými jazykovými modely, i LLM metrik) pro věrnost vstupním datům na úloze sumarizace textu – produkce krátkých shrnutí a nejdůležitějších bodů z textu (tzv. highlights). Z experimentů vyplývá, že použití LLM konkrétně na tuto úlohu nezaručuje lepší výsledky než mnohem jednodušší, starší evaluační metriky. Validace použití LLM lidskou evaluací na konkrétní úloze se tedy ukazuje jako důležitá. Pro tento typ evaluací jsme vytvořili benchmark, který bude v nejbližších měsících zveřejněn. Výsledky jsme prezentovali jako poster na workshopu ACL GEM, zároveň jsme odeslali článek k recenzi na konferenci LREC 2026.

V návaznosti na naši práci v oblasti anotace chyb jsme zahájili nové experimenty evaluující různé způsoby, jakými jazykové modely označovat slova a fráze v textu. V současnosti implementujeme a předběžně vyhodnocujeme několik přístupů: od kopírování textu s vloženými značkami až po sofistikovanější techniky, jako je extrakce chybných frází s disambiguací pomocí mechanismu attention nebo generování širšího kontextu pro zpřesnění anotace. Předpokládáme odeslání článku k recenzi v nejbližší době, pravděpodobně na konferenci ACL 2026.

SA16 LLM pro podporu mediální gramotnosti v prostředí českého školství

SA16 slouží jako technologická příprava a proof-of-concept pro možnost vývoje vlastních LLM vhodných pro úlohu KA3 Zvyšování digitální gramotnosti pomocí nástrojů umělé inteligence. Oproti obecně dostupným LLM, s nimiž KA3 průběžně pracuje, má vlastní LLM podstatnou

výhodu: jsou provozovány lokálně a citlivé informace (např. osobní data) tak nemusí opouštět příslušné instituce (např. jednotlivé školy nebo alespoň prostor českého Internetu). Práce na této aktivitě proběhla v prvním projektovém období podle plánu.

V dosavadních experimentech jsme se omezili na modely do velikosti 9B parametrů, protože jsme schopni je pohodlně trénovat metodami LoRA a QLoRA. Tyto modely jsou navíc dostatečně malé pro provozování na běžně dostupném hardware a ve kvantizované podobě by je bylo možné spouštět dokonce na koncových zařízeních uživatelů (například na uživatelově počítači nebo notebooku).

Pro evaluaci našich modelů jsme experimentovali s nedávno vydaným BenCzechMarkem (https://huggingface.co/blog/benczechmark, https://arxiv.org/abs/2412.17933), který nabízí automatické vyhodnocování open-source modelů v češtině pomocí kolekce autentických a strojově přeložených datasetů. Bohužel se během testování ukázalo, že BenCzechMark má pro naše využití řadu omezení. Za prvé nás v kontextu této subaktivity zajímá generování delšího textu, to ale BenCzechMark přímo netestuje, testuje převážně výběr testových odpovědí, krátké odpovědi a jazykové modelování. To je také ilustrováno i tím, že model Qwen2.5-72B, který je na BenCzechMarku momentálně druhý nejlepší, při praktickém nasazení v naší úloze velmi často generuje nesmyslný text a používá špatné nebo neexistující tvary slov. Za druhé BenCzechMark momentálně neumožňuje porovnání s komerčními modely jako GPT-4o a Claude 3.5 Sonnet, takže s nimi pak nemáme srovnání, které by ale bylo velmi vhodné, protože oba zmíněné modely již v mnoha úlohách velmi dobře s češtinou pracují. Za třetí je vyhodnocení celého BenCzechMarku poměrně časově i výpočetně náročné - vyhodnocení jednoho modelu nás stojí minimálně desítky GPU hodin na strojích s A100. To ale bohužel znamená, že se nám příliš nehodí pro vyhodnocování a porovnávání většího množství modelů v průběhu tréninku.

Kvůli výše zmíněných problémům s BenCzechMarkem se díváme i po alternativních metodách vyhodnocování. Nabízí se například metoda označovaná jako “LLM-as-a-judge”, která ale bohužel také trpí řadou problémů. Příkladem známého problému je skutečnost, že hodnotící LLM často preferuje svoje vlastní výstupy před výstupy ostatních systémů.

Jelikož obě tyto metody automatického vyhodnocování jsou nějakým způsobem problematické, hodila by se nám také možnost provést ruční hodnocení výstupů některých modelů. Proto bychom případné zbylé prostředky v kategorii ostatních osobních nákladů použili na **ruční hodnocení** a účastníky v experimentech v rámci KA2 SA16.

Další výzvou, se kterou se v rámci této subaktivity potýkáme, je nedostatek volně dostupných českých dat pro “instruction tuning”. Jednou z metod, která by nám potenciálně mohla s tímto problémem pomoci, je překlad dostupných anglických datasetů pomocí strojového překladu. Pro tyto účely jsme se rozhodli použít CUNI-MH, náš nový model pro anglicko-český strojový překlad, který se velmi dobře umístil v poslední soutěži ve strojovém překladu WMT24. Prozatím jsme pro experimenty vybrali datasety ultrachat_200k a ultrafeedback_binarized a ty jsme naším systémem strojově přeložili. Přeloženou verzi datasetu ultrafeedback_binarized jsme pak použili pro instruction tuning několika modelů (Gemma-2-9B, OLMO-7B) pomocí metody ORPO. Zatím se nám ale v tomto ohledu nepodařilo dosáhnout příliš zajímavých výsledků. Zatím nám není jasné, jestli je to z důvodu nevhodně zvolených hyperparametrů, kvůli případné nedostatečné kvalitě strojově přeloženého

datasetu, nebo jestli je problém ve zvolené evaluaci. Plánujeme proto v těchto experimentech pokračovat.

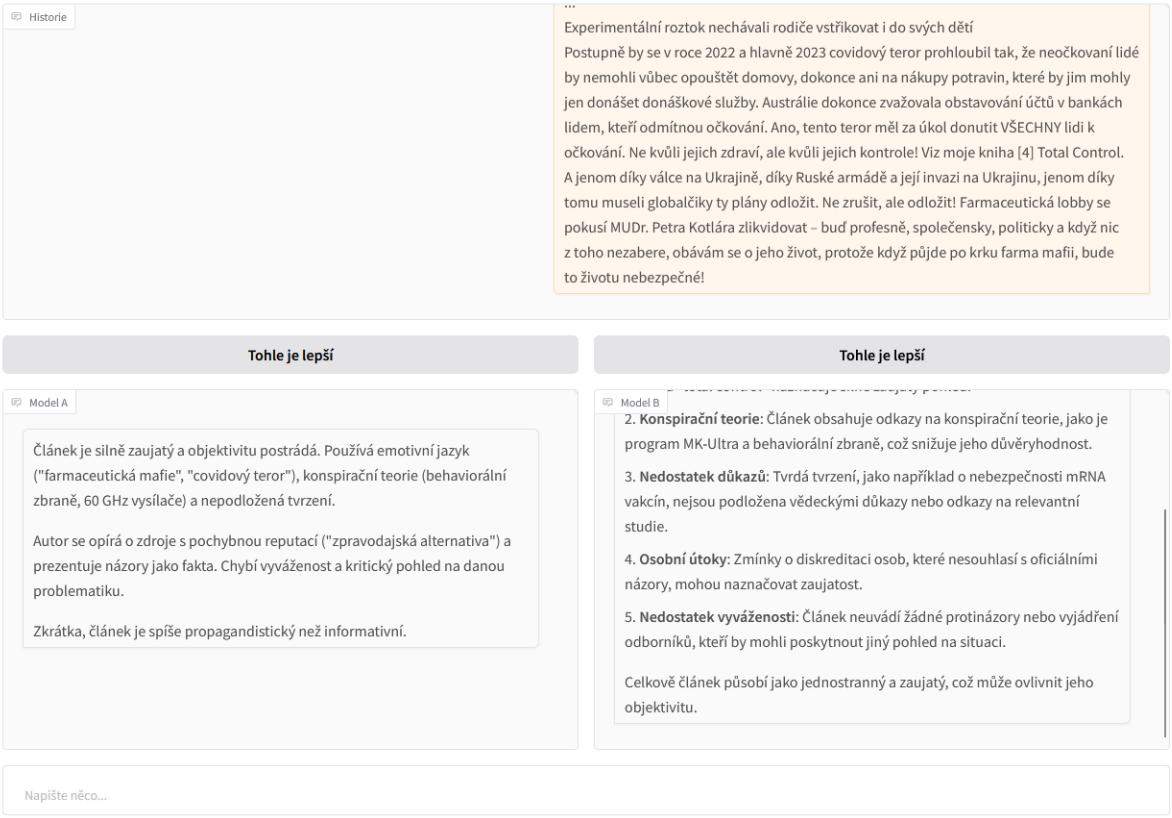
Na začátku monitorovacího období jsme pokračovali v experimentech s dotrénováním menších modelů na českých datech. Jelikož se nám ale zatím nepodařilo uspokojivě vyřešit automatické testování, rozhodli jsme se od těchto experimentů prozatím odklonit a pracovat na ostatních částech této subaktivity.

Oproti prvnímu sledovacímu období, kde jsme experimentovali převážně s malými modely do 10 miliard parametrů, jsme v tomto období začali experimentovat s většími dostupnými modely, konkrétně Llama 3.3 70B Instruct, DeepSeek R1 70B (Llama 3.3 70B Instruct dotrénovaná na výstupech velkého DeepSeek R1 modelu) a Gemma 3 27B. Tyto modely již do jisté míry zvládají češtinu, alespoň podle našeho subjektivního hodnocení. Umožňují nám také vyhodnotit, jak velké modely jsme schopni po technické stránce pro jednu školní třídu, tj. skupinu asi 30 lidí, nasadit na námi dostupném hardware. Použití těchto modelů nám pomůže odladit zbytek úkolů v rámci subaktivity. Je však třeba si uvědomit, že provozování těchto větších modelů stále potřebuje hardware výkonnější, než můžeme očekávat v prostředí českých základních a středních škol. Proto bychom se v budoucnu rádi vrátili k pokusům s menšími modely a vyhodnotili, zda jsou pro účely této subaktivity dostačující a zda bychom nějaký takový dostatečně malý model byli schopni spustit přímo na zařízeních, která se fyzicky nacházejí v dané škole.

V oblasti uživatelského rozhraní pro studenty jsme experimentovali se dvěma open-source projekty: LibreChat a Open WebUI. Oba projekty jsou webové aplikace vzhledově velmi podobné komerčním alternativám jako ChatGPT nebo Claude.ai. Open WebUI jsme interně úspěšně nasadili a testovali.

Pro účely výukových hodin jsme ale dospěli k závěru, že je tento nástroj příliš uživatelsky komplikovaný - umožňuje uživateli nastavovat i velmi technické parametry ovlivňující generování modelu. Nakonec jsme se tedy rozhodli implementovat vlastní, zjednodušené, webové rozhraní. Oproti výše zmíněným nástrojům také uživateli ukazujeme výstupy dvou různých modelů zároveň a během práce od uživatele očekáváme, že bude vybírat, kterou odpověď v dané chvíli preferuje. Tím získáme alespoň hrubou zpětnou vazbu, do jaké míry jsme námi nabízenými modely schopni nahradit komerční produkty jako například ChatGPT. Konkrétně tedy jsme mohli v rámci jedné výukové hodiny žákům ukazovat výstupy např. námi hostovaného modelu Gemma 3 a komerčního GPT-4o a zjistit tak, jestli studentům výstupy připadaly srovnatelné, nebo jestli jeden z modelů byl výrazně horší.

Zde je ukázka našeho webového rozhraní:



Námi připravené rozhraní jsme otestovali během května a června v celkem čtyřech pilotních hodinách pro studenty SŠ, v rámci hodin mediální výchovy nebo základů společenských věd. V rámci realizovaných lekcí se žáci zaměřili na identifikaci manipulativních technik v textu. Většina žáků již disponovala základními znalostmi v oblasti mediální gramotnosti a manipulativních postupů. První pilotní lekce byla poznamenána technickými obtížemi, které byly následně odstraněny. V časovém rozsahu jedné vyučovací hodiny se ukázalo jako náročné, aby žáci stihli přečíst celý výchozí text, formulovat prompty pro jazykové modely a zároveň připravit kvalitní výstupní prezentaci. V důsledku toho dokončil prezentaci v požadované kvalitě pouze omezený počet žáků.

Na základě těchto zjištění byla další výuka realizována v rozsahu dvou vyučovacích hodin. Tato lekce již probíhala bez technických problémů. Žáci formulovali srozumitelné a cílené prompty a systematicky porovnávali odpovědi obou použitých jazykových modelů. Celá druhá vyučovací hodina byla věnována prezentaci závěrů, hodnocení kvality výstupů modelů, reflexi vlastní práce a společné diskusi. Součástí diskuze byla rovněž témata bezpečného používání umělé inteligence a jazykových modelů ve vzdělávacím kontextu.

Pro podporu vyučujícího jsme připravili následující pracovní list:

Pracovní list pro učitele

Mediální gramotnost, manipulativní texty a AI

Anotace

Cílem aktivity je doplnit přechozí výuku kriticky přemýšlení o informacích na internetu. Žáci si vyzkouší rozpoznávání manipulativních technik v textech pomocí nástrojů umělé inteligence, konkrétně velkých jazykových modelů (LLM). Žáci budou analyzovat články, hodnotit jazyk a diskutovat o tom, jakým způsobem může být veřejné mínění ovlivňováno. Zároveň si uvědomí úskalí použití AI, především nutnost kriticky analyzovat její výstupy.

Cíle hodiny

- Naučit se používat AI/LLM společně s kritickým myšlením
- Za asistence AI rozpoznat manipulativní techniky v textech
- Porozumět výhodám a úskalím použití AI při analýze textu
- Zlepšit schopnost argumentace a prezentace výsledků ve skupině
- Otestovat potenciál lokálních LLM (modely mohou běžet přímo na škole, ochrana dat)

Evokace a úvod

Předpokládané předchozí znalosti:

- Úvod do tématu dezinformací: *Jak rozlišit názor od faktu? Proč jsou emoce v textu důležité? Pět základních otázek, budeme se soustředit na otázku "Jak?"* (např. podle Učíme o dezinformacích (JSNS))

Diskuze: *Kdo dnes četl něco na sociálních sítích nebo jinde na internetu? Je důležité vědět, jestli je to pravda? A jak pravdivost ověřujete?*

Použití AI pro analýzu textu

- Jaké AI nástroje znáte? (např. ChatGPT je jeden z druhů LLM)
- V čem nám mohou pomoci – a v čem škodit?
- Zkusíme si "něco jako" ChatGPT, které může běžet lokálně ve škole.

Problémy související s LLM:

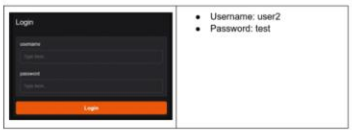
- Halucinace – model nezná koncept pravda vs. lež, generuje dobře znějící text, vymýšlí si
- Podlézavost – model se snaží zalíbit
- Ochrana osobních dat – co se o vás dozví?
- Aktuálnost - znají jen události do data trénování samotného modelu

Přístup k nástroji

Adresa nástroje: https://quest.ms.mff.cuni.cz/cedmollm/

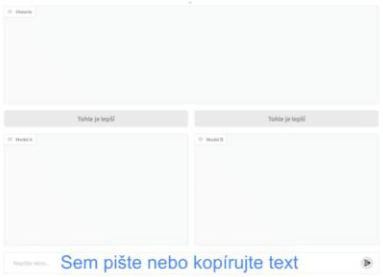
Krok 1: Přihlášení

- Username: cedmo
- Password: ufal



- Username: user2
- Password: test

Krok 2: Uživatelské rozhraní



Zadání úkolu

1. Jako vstup pro LLM použijte přiložené vybrané části těchto článků:
   - PravyProstor.net: Zmocněnec slovenské vlády vyrazil do války proti farmaceutické mafii
   - Aktualne.cz: Vakcíny mění DNA, oslovil jsem Obamu i Kennedyho, říká Ficův muž. Opírá se o Pekovou
   - Rozsah textu se může zdát větší, ale chceme mimo jiné ukázat, že LLM nám dokáží podobné úkoly zrychlit
2. V rozhraní LLM použijte jeden z následujících promptů (nebo libovolný jiný) a zkopírujte za něj vstupní text:

   > *Analyzuj následující text a identifikuj manipulativní nebo zavádějící jazykové prvky:*

   > *Procházej text větu po větě a pokud narazíš na nějaký manipulativní prvek, vypiš ho.*

   > *Jak konkrétně se snaží autor manipulovat čtenářem? Vypiš příklady manipulativního jazyka v následujícím odstavci:*

   > *Obsahuje text emotivní výrazy? Jaké emoce se snaží vyvolat?*

   > *Obsahuje text logické klamy? Např. argumenty strachem, ad hominem nebo whataboutismus?*

3. Úkolem studentů je ve skupinkách identifikovat co největší počet manipulativních technik: argumentační fauly, hraní na city, atd. Ve sdílené prezentaci (možnost připravit šablonu prezentace předem) zapíší své poznatky a následně je představí před třídou. Měli by si také dávat pozor na úskalí použití LLM, představená v úvodu hodiny.

Závěrečná prezentace a reflexe

Studenti představí výsledky svojí práce, tedy nalezené manipulační techniky v textu, ale i problémy, na které při použití LLM narazili, ve sdílené prezentaci. Následně:

- Skupiny sdílejí závěry, diskutují, porovnávají výstupy
- Učitel shrne a doplní – doplňující otázky: *Dalo se výstupům LLM věřit? Co vám připadalo podezřelé?*
- Případné soutěžní hodnocení:
  - Množství nalezených problematických úseků
  - Kvalitu prezentace a argumentace
  - Velké plus pokud identifikovali nějaký problém (zavádějící vyjádření nebo úplný nesmysl) ve výstupu LLM nebo LLM dokonce k podobnému vyjádření cíleně dovedli – jedním z cílů je zjistit, kdy se LLM "rozbije".

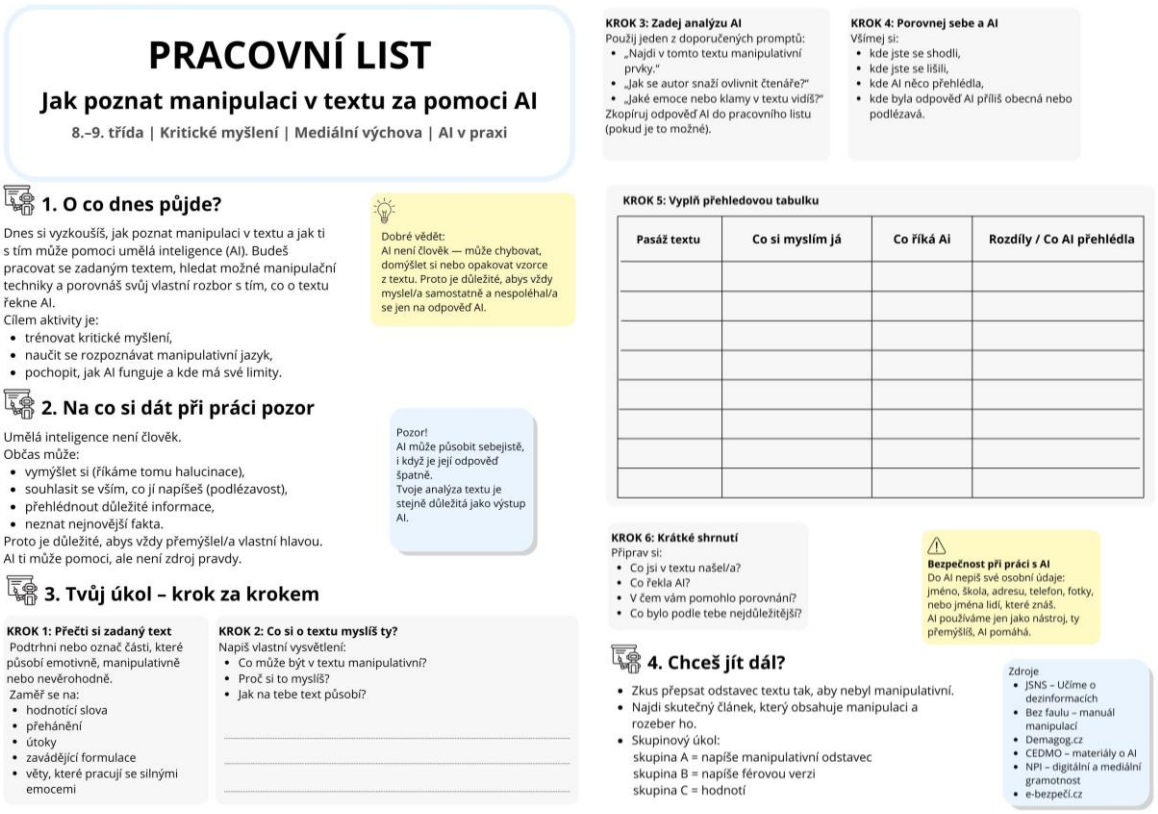
⚠️ **Upozorněte studenty, že výstupy AI nejsou objektivní pravda. Je třeba k nim přistupovat kriticky a zamýšlet se nad tím, jak by mohlo být použití těchto nástrojů zneužitelné (např. k podpoře konspiračních teorií). Ověrovat pravost informací na internetu a záměry jejich autorů je i s použitím AI náročné, ale je důležité nevzdávat se a nezačít si myslet, že "stejně lžou všichni."**

Doporučené zdroje

- Rizika AI
- Učíme o dezinformacích (JSNS)
- Mediální vzdělávání na SOŠ a učilištích: Výukový plán a příklady dobré praxe (JSNS)
- Seznam manipulativních technik
- Videomateriály:
  - Kovy - 5 základních otázek
  - Kovy - Otázka "Jak"

LLM nástroje mohou být užitečným pomocníkem v rozpoznávání manipulací a dezinformací. Je však nutné je používat kriticky, s ověřováním a mediální gramotností.

Další pilotní hodiny jsme připravili pro žáky ZŠ v průběhu listopadu. Pro potřeby mladších žáků jsme připravili zjednodušenou verzi pracovního listu.



PRACOVNÍ LIST

Jak poznat manipulaci v textu za pomoci AI

8.–9. třída | Kritické myšlení | Mediální výchova | AI v praxi

1. O co dnes půjde?

Dnes si vyzkoušíš, jak poznat manipulaci v textu a jak ti s tím může pomoci umělá inteligence (AI). Budeš pracovat se zadaným textem, hledat možné manipulační techniky a porovnáš svůj vlastní rozbor s tím, co o textu řekne AI.
Cílem aktivity je:

- trénovat kritické myšlení,
- naučit se rozpoznávat manipulativní jazyk,
- pochopit, jak AI funguje a kde má své limity.

Dobré vědět:
AI není člověk — může chybovat, domýšlet si nebo opakovat vzorce z textu. Proto je důležité, abys vždy myslel/a samostatně a nespoléhal/a se jen na odpověď AI.

2. Na co si dát při práci pozor

Umělá inteligence není člověk.
Občas může:

- vymýšlet si (říkáme tomu halucinace),
- souhlasit se vším, co jí napíšeš (podlézavost),
- přehlédnout důležité informace,
- neznat nejnovější fakta.

Proto je důležité, abys vždy přemýšlel/a vlastní hlavou. AI ti může pomoci, ale není zdroj pravdy.

Pozor!
AI může působit sebejistě, i když je její odpověď špatně.
Tvoje analýza textu je stejně důležitá jako výstup AI.

3. Tvůj úkol – krok za krokem

KROK 1: Přečti si zadaný text
Podtrhni nebo označ části, které působí emotivně, manipulativně nebo nevěrohodně.
Zaměř se na:

- hodnotící slova
- přehánění
- útoky
- zavádějící formulace
- věty, které pracují se silnými emocemi

KROK 2: Co si o textu myslíš ty?
Napiš vlastní vysvětlení:

- Co může být v textu manipulativní?
- Proč si to myslíš?
- Jak na tebe text působí?

KROK 3: Zadej analýzu AI
Použij jeden z doporučených promptů:

- „Najdi v tomto textu manipulativní prvky."
- „Jak se autor snaží ovlivnit čtenáře?"
- „Jaké emoce nebo klamy v textu vidíš?"

Zkopíruj odpověď AI do pracovního listu (pokud je to možné).

KROK 4: Porovnej sebe a AI
Všímej si:

- kde jste se shodli,
- kde jste se lišili,
- kde AI něco přehlédla,
- kde byla odpověď AI příliš obecná nebo podlézavá.

KROK 5: Vyplň přehledovou tabulku

Pasáž textu	Co si myslím já	Co říká AI	Rozdíly / Co AI přehlédla

KROK 6: Krátké shrnutí
Připrav si:

- Co jsi v textu našel/a?
- Co řekla AI?
- V čem vám pomohlo porovnání?
- Co bylo podle tebe nejdůležitější?

Bezpečnost při práci s AI
Do AI nepiš své osobní údaje: jméno, škola, adresu, telefon, fotky, nebo jména lidí, které znáš.
AI používáme jen jako nástroj, ty přemýšlíš, AI pomáhá.

4. Chceš jít dál?

- Zkus přepsat odstavec textu tak, aby nebyl manipulativní.
- Najdi skutečný článek, který obsahuje manipulaci a rozeber ho.
- Skupinový úkol:
  skupina A = napíše manipulativní odstavec
  skupina B = napíše férovou verzi
  skupina C = hodnotí

Zdroje

- JSNS – Učíme o dezinformacích
- Bez faulu – manuál manipulací
- Demagog.cz
- CEDMO – materiály o AI
- NPI – digitální a mediální gramotnost
- e-bezpečí.cz

Na základě realizovaných hodin na základní škole jsme zaznamenali rozdíly oproti výuce na střední škole. Žáci ZŠ prokazovali schopnost pojmenovat základní manipulativní techniky, nicméně měli obtíže s jejich samostatnou identifikací v konkrétních, autentických textech. Při práci s jazykovými modely potřebovali výrazně strukturovanější zadání, průběžné vedení a podporu při formulaci promptů i při interpretaci výstupů.

Ve srovnání se žáky střední školy byla práce žáků ZŠ více orientována na samotné porozumění textu a méně na jeho kritickou analýzu. Zatímco studenti SŠ dokázali své předchozí znalosti mediální gramotnosti poměrně plynule aplikovat v praxi a samostatně porovnávat výstupy různých modelů, u žáků ZŠ se ukázala potřeba výraznější role pedagoga jako facilitátora celého procesu.

Získané poznatky potvrzují, že využití jazykových modelů ve výuce mediální gramotnosti je smysluplné napříč věkovými skupinami, avšak vyžaduje výrazné přizpůsobení metodiky, tempa a míry podpory konkrétní cílové skupině. Zatímco na SŠ může AI nástroj fungovat jako prostředek pro rozvoj samostatného kritického myšlení, na ZŠ je jeho efektivní využití podmíněno pečlivě strukturovanou výukou a aktivním pedagogickým vedením.

Pro porovnání užitečnosti a spolehlivosti výsledného systému pro hodnocení odpovědí žáků a popagace dalšího vývoje ve vyhodnocovnání se do SA přidala ve 3. období podpora Sensemaking tasku na CLEF 2025 a podpora dalšího ročníku ve splupráci s OECD a PF CUNI. Příspěvek tasku Sensemaking na CLEF 2025 je detailně popsán ve 3. zprávě k realizaci projektu.

V tasku Sensemaking na CLEF 2026 připravuje náš tým dva tracky, zbytek dodá externí tým, co se projektu CEDMO 2.0 neučástí. Pro popularizaci jsme připravili veřejně dostupné popisy a materiály k vidění na Evaluator Sensemaking task at CLEF 2026.
V prvním tracku bude model muset pro daný krátký text, otázku a odpověd určit hodnocení otázky na škále 0-10. V druhém tracku bude navíc mít model přístup ke kódovacímu rubriku, který mu blíže popíše, jak odpovědi hodnotit. Rubrika se bude skládat ze tří částí. Část “full credit” popisuje, kdy je odpověd plně správně. Část “partial credit” popisujíce, kdy si odpověd zaslouží částečné přiznání bodů, i když není plně správně. Část “no credit”, která blíže popisuje, kdy nedat body žádné, a může být i prázdná. Samozřejmě, pokud žádnou z těchto částí nelze aplikovat, neměli by být přiznány žádné body.
Otázky a texty získáváme z několika zdrojů.
Seznam:

- otázky a texty z veřejně dostupných učebnic
- otázky vytvořené pro účely jiné studie k materiálům pro popularizaci vědy
- vlastní otázky a texty vytvořené z databáze demagog.cz, včetně ano-ne kontrolních otázek na porovnání schopností factcheckingu s pomocí ořezaného vysvětlovacího článku
- otázky a texty z PISA testů dodaných OECD

Hodnocení a odpovědi získáváme také z několika zdrojů. OECD dodá reálné odpovědí žáků na PISA testy a mi přidáme anotátory vytvořené a hodnocené odpovědi pro naše materiály.
Dodatečně máme v plánu vytvořit jako stříbrná data strojově vytvořené a hodnocené odpověďi pro dříve zmíněné otázky zkontrolované a upravené anotátory.
Dále máme v plánu vytvořit bronzová lidmi nekontrolovaná data překladem dříve zmíněných dat do několika jazyků modelu OpenEuroLLM, aby bylo možné potencionálně naše data použít pro výrobu benchmarku pro tento model.
Seznam těchto jazyků:

- Angličtina
- Čeština
- Portugalština
- Němčina
- Maďarština
- Finština
- Dánština
- Švédština
- Srbština
- Řečtina
- Irština
- Rumunština

Strojově vytvořený dataset vyrábíme iterativně a to tak, aby co nejlépe kvalita modelů vyvýjených na jeho devsetu předpovídala kvalitu na devsetu odpovědí reálných žáků. Znamená to pro každou iteraci devsetu vyvinout nové baseline modely, čímž navíc získáme možnost pochopit blíže i devset odpovědí reálných žáků. Chování modelů na datasetech zkoumáme pomocí jednoduchých unigramových a bigramových podobností, clustrování dimensionálně redukované matice správnosti, rozhodovacích stromů a lineární regrese.

Za účelem bezproblemovéhového propojení dat od různých anotátorů a zvýšení mezianotátorské shody navíc spolu s OECD spolupracujeme na vytvoření detailního

instruktářního dokumentu o tom, jak mají uměle vytvořené odpovědi od studentů a rubriky vypadat. Doufáme, že se tak více propojí svět vyučujících a svět vývojářů nástrojů.

Pro task jsme připravili také několik baseline založených na velkých, malých i triviálních modelech, abychom vyloučili to, že je dataset moc jednoduchý, a odfiltrovali příliš jednoduché položky.
Seznam baseline:

- několik metod, založených na jednoduchý modelech jako porovnání součtu a manuálně nastaveného prahu, KNN, rozhodovací stromy, atd. Budou používat vlastnosti jako TF-IDF, unigramové, bigramové a semantické podobnost odpovědi a otázky, odpovědi a kontextu, odpovědi a rubriky, délku odpovědi, slovní rozmanitost odpovědi atd.
- model DeBERTa-v3 finetunovaný na datasetu mnli a dofinetunovaný na mixu mnli a subsetu našeho malého dev datasetu
- ensamble několika modelů pod 30b parametrů, například gpt-oss-20b
- gpt-oss-120b

Pro všechny tyto baseliny máme už připravené první iterace a pracujeme na druhých.

Odborné publikace dedikované projektu CEDMO 2.0 NPO

Publikované články

1. Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe et al.: Findings of the IWSLT 2025 Evaluation Campaign. In: Proc. of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-272-5, pp. 412-481, 2025 https://doi.org/10.18653/v1/2025.iwslt-1.44
2. Petra Barančíková, Ondřej Bojar: Intrinsic vs. Extrinsic Evaluation of Czech Sentence Embeddings: Semantic Relevance Doesn't Help with MT Evaluation. In: Proc. of Machine Translation Summit XX: Volume 2, European Association for Machine Translation (EAMT), Geneva, Switzerland, ISBN 978-2-9701897-1-8, pp. 265-275, 2025 https://aclanthology.org/2025.mtsummit-1.20/
3. Gilles Bareilles, Allen Gehret, Johannes Aspman, Jana Lepšová, Jakub Mareček. Deep Learning as the Disciplined Construction of Tame Objects. ArXiv preprint arXiv:2509.18025. https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
4. Marko Čechovič, Natália Komorníková, Dominik Macháček, Ondřej Bojar: Corpus of Cross-Lingual Dialogues with Minutes and Detection of Misunderstandings. In: 28th International Conference on Text, Speech and Dialogue (Part II), Springer, Cham, Switzerland, ISBN 978-3-032-02551-7, pp. 301-312, 2025 https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
5. Armin Hadžić, Milan Papež, Tomáš Pevný. Distillation of a tractable model from the VQ-VAE. In Proc. 8th Workshop on Tractable Probabilistic Modeling, 2025. https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
6. Miroslav Hrabal, Josef Jon, Martin Popel, Nam Hoang Luu, Danil Semin, Ondřej Bojar: CUNI at WMT24 General Translation Task: LLMs, (Q)LoRA, CPO and Model Merging. In: Proc. of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-179-7, pp. 232-246, 2024 https://doi.org/10.18653/v1/2024.wmt-1.16
7. Miroslav Hrabal, Josef Jon, Martin Popel, Ondřej Bojar: CUNI at WMT25 General Translation Task. In: Proc. of the Tenth Conference on Machine Translation, Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-341-8, pp. 680-687, 2025 https://doi.org/10.18653/v1/2025.wmt-1.44
8. Josef Jon, Ondřej Bojar: Finetuning LLMs for EvaCun 2025 token prediction shared task. In: Proc. of the Second Workshop on Ancient Language Processing, Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-235-0, pp. 221-225, 2025 https://doi.org/10.18653/v1/2025.alp-1.29

9. Jussi Karlgren, Ekaterina Artemova, Ondřej Bojar, Marie Isabel Engels, Vladislav Mikhailov, Pavel Šindelář, Erik Velldal, Lilja Øvrelid: Overview of ELOQUENT 2025: Shared Tasks for Evaluating Generative Language Model Quality. In: Lecture Notes in Computer Science, Vol. 16089, Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proc. of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Springer, Cham, Switzerland, ISBN 978-3-032-04354-2, ISSN 1611-3349, pp. 224-241, 2025 https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
10. Ivan Kartáč, Mateusz Lango, Ondřej Dušek: OpeNLGauge: An Explainable Metric for NLG Evaluation with Open-Weights LLMs. In: Proc. of the 18th International Natural Language Generation Conference, Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-321-0, pp. 292-337, 2025 https://aclanthology.org/2025.inlg-main.19/
11. Ivan Kartáč, Mateusz Lango, Ondřej Dušek: OpeNLGauge: An Explainable Metric for NLG Evaluation with Open-Weights LLMs.Vienna, Austria, Jul 2025 (GEM poster)
12. Tom Kocmi, Eleftherios Avramidis, Rachel Bawden et al.: Findings of the WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet. In: Proc. of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-179-7, pp. 1-46, 2024 https://aclanthology.org/2024.wmt-1.1/
13. Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis et al.: Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In: Proc. of the Tenth Conference on Machine Translation, Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-341-8, pp. 355-413, 2025 https://doi.org/10.18653/v1/2025.wmt-1.22
14. Dominik Macháček, Peter Polák: Simultaneous Translation with Offline Speech and LLM Models in CUNI Submission to IWSLT 2025. In: Proc. of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-272-5, pp. 389-398, 2025 https://doi.org/10.18653/v1/2025.iwslt-1.41
15. Jiří Mírovský, Barbora Hladká: SouDeC 2.0 - Source Detection and Classification. Data/software, Charles University, Prague, Czech Republic, Dec 2025 https://quest.ms.mff.cuni.cz/soudec/
16. Kristýna Onderková: Insight discovery in structured data. Hanoi, Vietnam, Oct 2025 (YNLG poster)
17. Kristýna Onderková, Ondřej Plátek, Zdeněk Kasner, Ondřej Dušek: FreshTab: Sourcing Fresh Data for Table-to-Text Generation Evaluation. In: Proc. of the 18th International Natural Language Generation Conference, Association for Computational Linguistics, Kerrville, TX, USA, ISBN 979-8-89176-321-0, pp. 108-121, 2025 (Best short paper award) https://aclanthology.org/2025.inlg-main.7/
18. Kristýna Onderková, Ondřej Plátek, Zdeněk Kasner, Ondřej Dušek: Sourcing Fresh Resources for Table-to-Text Generation Evaluation. Vienna, Austria, Jul 2025 (GEM poster)

19. Kristýna Onderková, Ondřej Plátek, Zdeněk Kasner, Mateusz Lango, Ondřej Dušek: Generating Data Insights with LLMs by Querying Tables. Prague, Czech Republic, Apr 2025 (ML Prague poster)
20. Sara Papi, Peter Polák, Dominik Macháček, Ondřej Bojar: How "Real" is Your Real-Time Simultaneous Speech-to-Text Translation System?. In: Transactions of the Association for Computational Linguistics, Vol. 13, Association for Computational Linguistics, ISSN 2307-387X, pp. 281-313, 2025 https://doi.org/10.1162/tacl_a_00740
21. Nela Petrzelkova, Jan Cech. Detection of Synthetic Face Images: Accuracy, Robustness, Generalization. In Proc. DAGM GCPR, 2025. https://doi.org/10.1007/978-3-032-12840-9_29
22. Jan Pijálek, Karel Vlk, Ondřej Bojar: MEEDAV: A Synchronous Web Viewer for EEG, Eye-Tracking and Speech Data. Accepted for publication in: ACAIN 2025, 2025 https://drive.google.com/drive/folders/1CIA7petNtiIRloCcP1Z0v81OYO93Jyj9
23. Peter Polák, Ondřej Bojar: Long-Form End-to-End Speech Translation via Latent Alignment Segmentation. In: Proc. of the 2024 IEEE Spoken Language Technology Workshop (SLT), IEEE, Piscataway, USA, ISBN 9781479971305, pp. 1076-1082, 2024 https://doi.org/10.1109/SLT61566.2024.10832264
24. Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, Giorgos Tolias. Instance-Level Composed Image Retrieval. In: Proc. NeurIPS, 2025 https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
25. Pavel Rytir, Ales Wodecki, Georgios Korpas, Jakub Marecek. ExDBN: Exact learning of Dynamic Bayesian Networks. ArXiv preprint arXiv:2410.16100, 2024. https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
26. I. Samarskyi, J. Cech. Talking Motion Signatures for Identity Recognition. In Proc. VISAPP, 2026. To Appear. https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
27. Patrícia Schmidtová, Ondřej Dušek, Saad Mahamood: Faithfulness Metrics Do Not Generalize Well: A Case Study in Summarization.Vienna, Austria, Jul 2025 (GEM poster)
28. Pavel Šindelář, Ondřej Bojar: Overview of the Sensemaking Task at the ELOQUENT 2025 lab: LLMs as Teachers, Students and Evaluators. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS, Aachen, Germany, pp. 1330-1359, 2025 https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
29. Hening Wang, Leixin Zhang, Ondřej Bojar: Human and Machine: Language Processing in Translation Tasks. In: Proc. of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024), Association for Computational Linguistics, Online, pp. 243-250, 2024 https://aclanthology.org/2024.icnlsp-1.27/
30. A. Yermakov, J. Cech, J. Matas, M. Fritz. Deepfake Detection that Generalizes Across Benchmarks. In Proc. WACV, 2026. To Appear. https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
31. Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Andre Araujo, Ondrej Chum. UDON: Universal Dynamic Online distillatioN for generic image representations. In: Proc. NeurIPS, 2024

https://neurips.cc/virtual/2024/poster/94005
32. Nikolaos-Antonios Ypsilantis, Ondřej Chum. Co-Segmentation without any Pixel-level Supervision with Application to Large-Scale Sketch Classification. In: Proc. Asian Conference on Computer Vision (ACCV), 2024 https://doi.org/10.1007/978-981-96-0972-7_20
33. Nikolaos-Antonios Ypsilantis, Kaifeng Chen, André Araujo, Ondřej Chum. Infusing fine-grained visual knowledge to Vision-Language Models. Proc International Conference on Computer Vision (ICCV) -- Workshop on "What is Next in Multimodal Foundation Models", 2025. To Appear. https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA
34. Vladislav Stankov, Matyáš Kopp, Ondřej Bojar. ParCzech4Speech: A New Speech Corpus Derived from Czech Parliamentary Data. In: Ekštein, K., Konopík, M., Pražák, O., Pártl, F. (eds) Text, Speech, and Dialogue. TSD 2025. Lecture Notes in Computer Science(), vol 16029. Springer, Cham. Print ISBN 978-3-032-02547-0, Online ISBN 978-3-032-02548-7, 2026 https://doi.org/10.1007/978-3-032-02548-7_25
35. J. Paplhám, V. Franc. Photo Dating by Facial Age Aggregation. In Proc. WACV, 2026. To Appear. https://drive.google.com/drive/folders/1jcHisufqpxCn_fSFWz1JJqXTmKVNZ2zA