Financováno Evropskou unií
NextGenerationEU

NÁRODNÍ PLÁN OBNOVY

MINISTERSTVO PRŮMYSLU A OBCHODU

Využití nástrojů umělé inteligence podporující transformaci médií

Umělá inteligence podporující transformaci médií

CEDMO 2.0 NPO

MPO 60273/24/21300/21000

Studie č. 4

Datum: 17. 12. 2025

Editoři: Gabriela Vostřezová, Petr Orálek, Petr Kozl

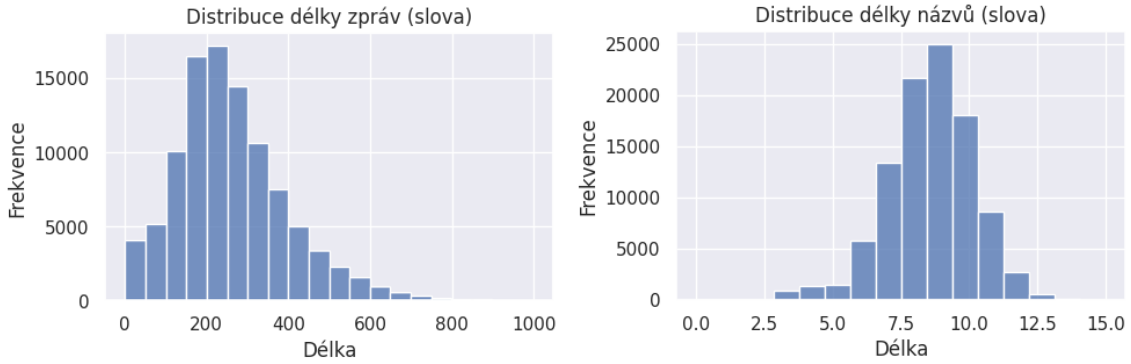
Autoři výsledku: Pavel Pecina, David Javorský, Ondřej Dušek, Kristýna Onderková, Filip Rechtorík, Ondřej Plátek, Viliam Lisý, Josef Jon, Miroslav Hrabal, Ondřej Bojar, Jakub Paplhám, Vojtěch Franc, Giorgos Tolias

SA1 Nástroj pro sémantické vyhledávání a vyhodnocování informací v rámci uzavřeného ekosystému ověřených informací pro redaktory a editory

Cílem této subaktivity je navrhnout a ověřit metodiky, které umožní vytvářet různé formy automatických shrnutí. Projekt vzniká ve spolupráci mezi Univerzitou Karlovou a ČTK a opírá se zejména o moderní techniky využívající velké jazykové modely (LLM). Součástí práce je rovněž porovnání těchto technik s tradičními neuronovými modely a návrh promptovacích strategií vhodných pro české zpravodajské prostředí.

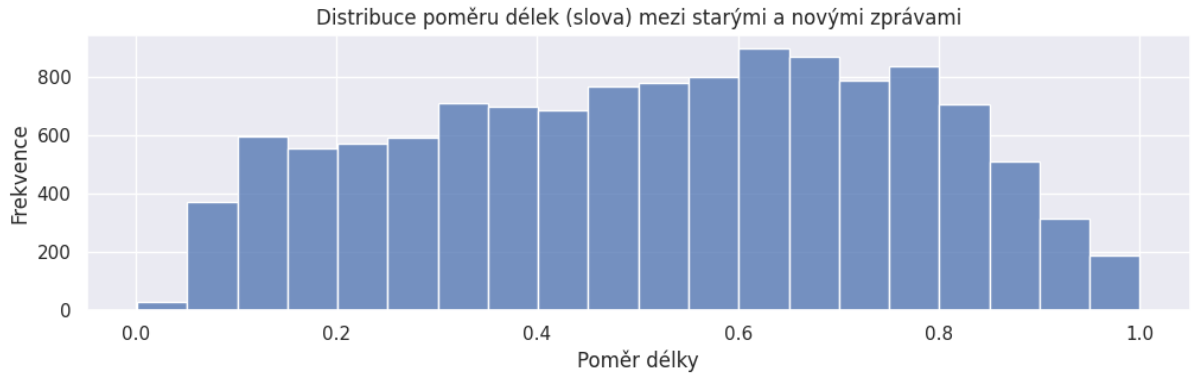
Datový korpus a modely

Na základě smlouvy s ČTK byl výzkumnému týmu zpřístupněn rozsáhlý archiv zpravodajských textů obsahující přibližně 600 000 článků publikovaných v letech 2006–2010. Tento korpus představuje primární datovou základnu projektu a umožňuje provádět jak kvantitativní analýzy, tak i testování sumarizačních přístupů v realistickém redakčním kontextu.



Obrázek č. 1: Distribuce délek zprav a jejich názvů.

Texty byly analyzovány z hlediska jejich délky a struktury. Například Obrázek č. 1 zobrazuje distribuci délek zpráv a jejich názvů. Můžeme vidět, že navětší množsvtí zprav má délku kolem 200 slov a délky názvů se pohybují mezi 8 až 10 slov.



Obrázek č. 2: Distribuce poměru délek zprav a jejich názvů.

Pozornost byla taktéž věnována vztahu mezi plnými verzemi zpráv a jejich aktualizovanými či redakčně upravenými variantami, hlavně jejich rozdílů v rozsahu. Obrázek č. 2 zachycuje

distribuci poměru délek mezi starými, původnými zprávami a jejich aktualizovanými, novými verzemi. Můžeme vidět, že distribuce skoro roznoměrne pokrývá celý rozsah poměrů v intervalu 0 až 1. Tato statistika poukazuje na příznivé vlastnosti datasetu, kde se dokáže najít skoro libovolný poměr, což se speciálně může hodit v případe poskytování příkladů pro velké jazykové modely (LLM). Obsahová stránka shrnutí byla také studována a dva příklady staré a nové zprávy jsou zobrazeny v Obrázkách č. 3 a č. 4. Jedná se o tu samou zprávu, která byla aktualizována dvakrát.

```
<p>Volduchy (Rokycansko) 1. května (ČTK) - Tragicky skončily oslavy příchodu
máje ve Volduchách na Rokycansku, kde spadla dnes k ránu na
dvaadvacetiletého muže májka. Mladík byl na místě mrtev a záchranářům se ho
už nepodařilo oživit. ČTK to dnes řekla mluvčí plzeňské záchranky Lenka
Ptáčková.</p>
<p>Podle policie se neštěstí stalo patrně při pokusu skupiny mladých lidí
májku pokácet. Policie případ vyšetřuje.</p>
```

```
<p>Volduchy (Rokycansko) 1. května (ČTK) - Tragicky skončily oslavy příchodu
máje ve Volduchách na Rokycansku, kde spadla dnes ráno na dvacetiletého
muže májka. Mladík na místě podlehl vážným zraněním a záchranářům se ho už
nepodařilo oživit. ČTK to dnes řekla mluvčí plzeňské záchranky Lenka
Ptáčková.</p>
<p>Podle policejní mluvčí Jolany Číhové se neštěstí stalo při pokusu skupiny
asi deseti mladých lidí z nedaleké obce Litohlavy májku pokácet. Případ
vyšetřuje na místě výjezdová skupina policie a u všech aktérů tragického
kácení zjišťuje, zda pili alkohol. "Zatím není jasné, kdo tragédii
způsobil. Míru zavinění určí až vyšetřování," uvedla mluvčí.</p>
```

Obrázek č. 3: Příklad staré zprávy (nahoru) a aktualizované zprávy (dolu).

```
<p>Volduchy (Rokycansko) 1. května (ČTK) - Tragicky skončily oslavy příchodu
máje ve Volduchách na Rokycansku, kde spadla dnes ráno na dvacetiletého
muže májka. Mladík na místě podlehl vážným zraněním a záchranářům se ho už
nepodařilo oživit. ČTK to dnes řekla mluvčí plzeňské záchranky Lenka
Ptáčková.</p>
<p>Podle policejní mluvčí Jolany Číhové se neštěstí stalo při pokusu skupiny
asi deseti mladých lidí z nedaleké obce Litohlavy májku pokácet. Případ
vyšetřuje na místě výjezdová skupina policie a u všech aktérů tragického
kácení zjišťuje, zda pili alkohol. "Zatím není jasné, kdo tragédii
způsobil. Míru zavinění určí až vyšetřování," uvedla mluvčí.</p>
```

```
<p>Volduchy (Rokycansko) 1. května (ČTK) - Tragicky skončily oslavy příchodu
máje ve Volduchách na Rokycansku, kde dnes ráno usmrtila dvacetiletého muže
májka. Mládež ze sousední vesnice ji pokácela, kmen se odrazil od
automobilu a mladíka zasáhl. Na místě podlehl vážným zraněním, záchranářům
se ho už nepodařilo oživit. Vyplývá to z informací záchranné služby,
policie a starostky obce.</p>
<p>Podle policejní mluvčí Jolany Číhové se neštěstí stalo před 07:00, když
skupina asi deseti mladých lidí z nedaleké obce Litohlavy májku pokácela.</p>
<p>"Takové ráno jsme nečekali. Tu mládež jsem neznala. Přišla ze sousední obce
k ránu, když už jsme přestali májku hlídat. Nikdo nepočítal s tím, že by ji
ještě někdo kácel. Na nedaleké lavičce pospávali poslední tři strážci.
Stožár májky, který skupinka kácela sekerkou, spadl na zaparkovaný
automobil, odrazil se od něj a zasáhl toho mladíka. Záchranáři ho už
nemohli oživit," uvedla starostka obce Miroslava Kodlová.</p>
<p>Případ vyšetřuje na místě výjezdová skupina policie a u všech aktérů
tragického kácení zjišťuje, zda pili alkohol. "Zatím není jasné, kdo
vlastně tragédii způsobil. Míru zavinění určí až vyšetřování," uvedla
policejní mluvčí.</p>
<p>Tradice stavění májky ve Volduchách už má dlouhou tradici. Stejně tak se
čas od času pokouší mládež z okolních obcí pestře ozdobený stožár postavený
na návsi ve stojanu zapuštěném do země pokácet. Nikdy to ovšem neskončilo
tragicky, řekla starostka. Neštěstí jí otřáslo. Nikoho ze skupinky, která
májku kácela, nezná. Na místo tragédie se hned zašla podívat.</p>
```

Obrázek č. 4: Příklad staré zprávy (nahoru) a aktualizované zprávy (dolu).

Další zajímavou charakteristikou datasetu byla délka “řetízků” aktualizovaných zpráv, t.j. počet aktualizací jedné zprávy. Výsledky jsou v Tabulce č. 1 a zachycují, že nejvíc zpráv má jenom jednou aktualizaci, avšak, překvapivě, jsou tam i takové zprávy, které mají 5 (2k zpráv), 8 (150 zpráv) nebo až 15 aktualizací (15 zpráv).

Počet aktualizací	1	2	3	4	5	6	7	8	9	10	11	12	14	15
Počet zpráv	43 075	24 152	12 183	5 196	2 060	942	399	168	45	10	22	12	14	15

Tabulka č. 1: Počer případů, kdy jedna zpráva má daný počet aktualizací.

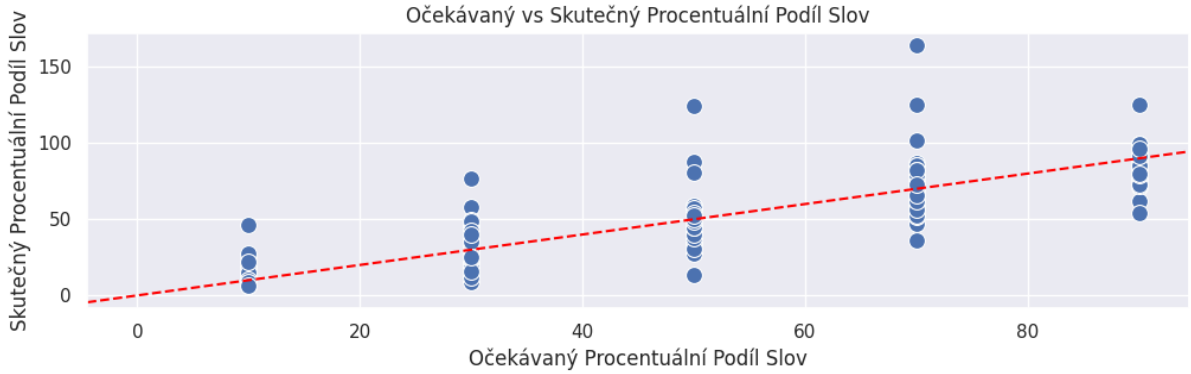
Dále byl studován poměr délek mezi vstupními texty a jejich souhrny či titulky, což představuje důležitou charakteristiku pro pozdější experimenty s řízením délky generovaných výstupů.

Experimenty

Projekt zkoumal využití LLM, konkrétně Llama 3.3, a případně dalších otevřených či interních variant. Tyto modely jsou schopny generovat souhrny různých délek i forem a poskytují flexibilní prostředí pro výzkum řízení délky, struktury a informativnosti výstupu. Pro referenční srovnání byly zařazeny také starší neuronové systémy, například model Transformer adaptovaný pro sumarizaci a model MLASK zaměřený na multimodální zpracování českých zpravodajských textů. Tyto systémy představují výkonnostní baseline a umožňují objektivně hodnotit přínos moderních LLM.

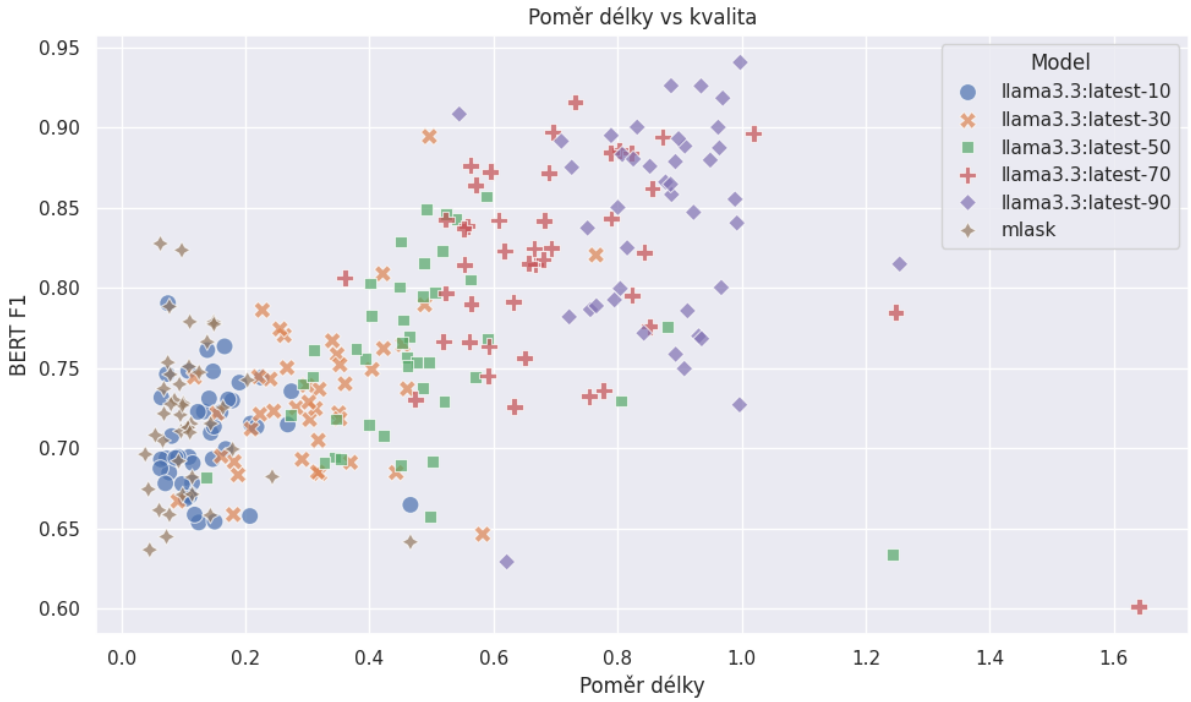
Součástí výzkumu bylo testování strategií pro výběr demonstračních příkladů, které jsou součástí promptu v rámci in-context learningu. Experimenty se zaměřily na dvě hlavní techniky výběru: volbu příkladů podobné délky a výběr příkladů na základě sémantické podobnosti. Ukázalo se, že vhodně zvolený kontext může výrazně zvýšit kvalitu generovaného shrnutí, a dosažené výsledky překonaly metody publikované v oblasti vícejazyčné sumarizace. Veřejně dostupné modely i promptovací postupy poskytovaly v českém prostředí jen omezeně použitelné výstupy.

Experimenty s LLM se zaměřily na nalezení takových promptovacích technik, které umožní spolehlivě řídit délku generovaného shrnutí. Nejprve byla testována jednoduchá instrukční schémata typu „Vygeneruj krátké shrnutí“, „Vygeneruj středně dlouhé shrnutí“ a „Vygeneruj dlouhé shrnutí“. Modely byly schopné rozlišit alespoň relativní délky, avšak nedařilo se jim konzistentně zachovávat absolutní počet slov.



Obrázek č. 5: Očekávaný vs skutečný poměr počtu slov v souhrnech a v původní zprávě. Červená přerušovaná čára zachycuje optimálni zkrácení.

Následně byla zvolena přesnější strategie založená na procentuálním vyjádření cílové délky, např. „Shrň text tak, aby výstup měl X % délky původního textu.“ Výsledky tohoto přístupu jsou zobrazeny v Obrázku č. 5. Tento přístup vykázal výrazně vyšší přesnost, zejména v okolí hodnoty 10 %. Zároveň se potvrdilo, že velmi krátké texty nelze dále smysluplně zkracovat a že modely mají tendenci shrnutí mírně podhodnocovat, tedy produkovat o něco kratší výsledky, než bylo zadáno.

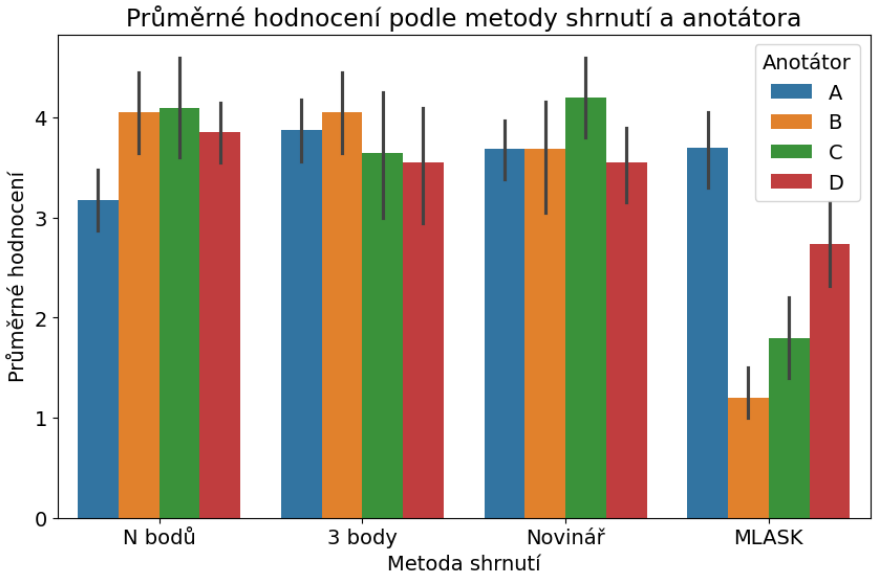


Obrázek č. 6: Kvalita souhrnů pro jednotlivé zkracující poměry.

Také jsme studovali kvalitu vygenerovaných souhrnů, co je zobrazeno na Obrázku č. 6. Můžeme pozorovat velkou variabilitu. Kromě délky byla zkoumána také forma shrnutí. Testovány byly instrukce vedoucí k tvorbě shrnutí ve třech bodech (3 body), shrnutí v několika bodech (N bodů) a shrnutí koncipovaná jako práce profesionálního editora (Novinář). Součástí experimentů bylo i porovnání s interním modelem MLASK vyvinutým kolegy, který je schpný generovat převážně krátká shrnutí.

Hodnocení

Výsledky se dále hodnotily jak automaticky, tak na základě zpětné vazby editorů ČTK. Pro účely ruční evaluace bylo vytvořeno 400 shrnutí k 100 vybraným zprávám, přičemž pro každý text byly vygenerovány čtyři různé varianty (3 body, N bodů, Novinář a MLASK). Tato sada byla poskytnuta editorům ČTK, kteří shrnutí posuzovali z hlediska informační úplnosti, stylistické kvality, srozumitelnosti a relevance pro redakční praxi. Výsledky tohoto hodnocení jsou zobrazeny na Obrázku č. 1.



Obrázek č. 1: Hodnocení jednotlivých typů shrnutí.

Hodnocení ukazuje rozdíly mezi jednotlivými metodami shrnutí i mezi anotátory. Metody založené na odrážkách („N bodů“ a „3 body“) dosahují stabilně vyšších průměrných známek napříč všemi čtyřmi anotátory, přičemž nejvýše hodnotí typicky anotátor C a anotátor B. Metoda „Novinář“ si vede podobně dobře, její výsledky jsou však o něco variabilnější. Naopak metoda MLASK vykazuje napříč anotátory výrazně nižší skóre a zároveň největší mezi-anotátorskou variabilitu, zejména u anotátora B. Celkově se tak jako nejrobustnější jeví odrážkové metody, které kombinují nejvyšší průměrné hodnocení s relativně nízkou variabilitou.

Metoda	Bez výhrad	Moc krátké / stručné	Moc podrobné / nadbytečné	Nejasné	Neúplné / chybí info	Špatná struktura	Špatná čeština
3 body	10	0	0	12	9	6	6
Novinář	9	0	7	9	11	5	7
MLASK	5	4	0	7	32	1	4
N bodů	9	0	18	6	13	8	4

Tabulka č. 2: Kvalitativní hodnocení zpráv.

Tabulka č. 2 zhrňuje kvalitativní hodnocení zpráv. V této evaluaci se annotátoři napříč metodami výrazně shodují v několika vzorcích: oceňují stručnost, přesnost a úplnost u 3-bodového shrnutí, zatímco MLASK považují za příliš stručné a často jen reprodukující první větu zprávy. Velmi často kritizují chybějící klíčové informace (kdo, kdy, kde, proč), faktické nepřesnosti a neobratnou nebo chybnou češtinu, která se objevuje zejména u přístupu

„Novinář“ a MLASK. U shrnutí v několika bodech opakovaně zmiňují nadbytečnou podrobnost a nepřehlednost. Společná je také silná citlivost na logiku textu, návaznost informací a celkovou kompaktnost shrnutí. Celkově se tak jako nejrobustnější metoda jeví shrnutí ve třech bodech, které kombinuje stručnost, přesnost i srozumitelnou strukturu.

V případě prodloužení projektu se další práce bude opírat o výsledky evaluace provedené ČTK. Předpokládá se implementace finální metodiky sumarizace, případně doplněná o pokročilejší promptovací strategie nebo techniky instruktážního učení.

SA2 Nástroje pro sémantické vyhledávání, práci s obsahem a jeho personalizaci pro veřejnost

Aktivita SA2 byla během prvního monitorovacího období připojena k aktivitě SA1. Tato změna byla poskytovatelem grantu schválena.

SA3 Vyvinutí nástrojů pro rozpoznávání obličejů a identifikaci osob, vyhledávání podobných fotografií a sémantické vyhledávání

Přehled a Cíle Aktivity

Cílem SA3 byl vývoj pokročilých nástrojů umělé inteligence pro automatickou analýzu a organizaci vizuálního obsahu. Hlavním záměrem bylo vytvořit sadu technologií, které umožní mediální agentuře efektivně využívat rozsáhlé digitální archivy i aktuální zpravodajství bez nutnosti spoléhat se výhradně na manuální textové popisky, takzvaná metadata.

Potenciál pro využití spočívá především ve třech oblastech:

1. **Zefektivnění práce editorů:** Automatické návrhy identit a klíčových slov zrychlují zařazení fotografií do fotobanky.
2. **Zhodnocení archivu:** Možnost dohledat historické fotografie osobností z doby, kdy ještě nebyly známé (a tudíž nejsou v metadatech jmenovány), nebo vyhledávat obecné koncepty, umožňují fotobance plně využívat veškerá archivní data.
3. **Kontrola kvality:** Automatická detekce chyb v popiscích na základě vizuální analýzy (např. nesoulad odhadovaného data pořízení a věku osoby).

Dosažené výsledky: Výzkumný tým vyvinul a nasadil do testovacího provozu funkční prototypy pro vyhledávání podle vizuální podobnosti, pro rozpoznávání tváří, odhad věku a multimodální vyhledávání (text-obraz). Všechny tyto nástroje prokazují vysokou úspěšnost.

Použitá Data

Na základě smlouvy s ČTK bylo výzkumnému týmu ČVUT zpřístupněno přibližně 2 000 000 fotografií z fotobanky, na kterých bylo detekováno celkem 11 000 000 tváří. Tato data byla použita výhradně k testování funkcí vyvinutých nástrojů.

Klíčovým aspektem práce s daty bylo rozhodnutí nevyužívat pro vyhledávání existující metadata (popisky). Vyvinuté nástroje fungují čistě na základě vizuální informace a jsou schopny obohatit databázi o informace, které v textových popiscích editorů chybí. Fotografie byly zpracovány ve dvou režimech:

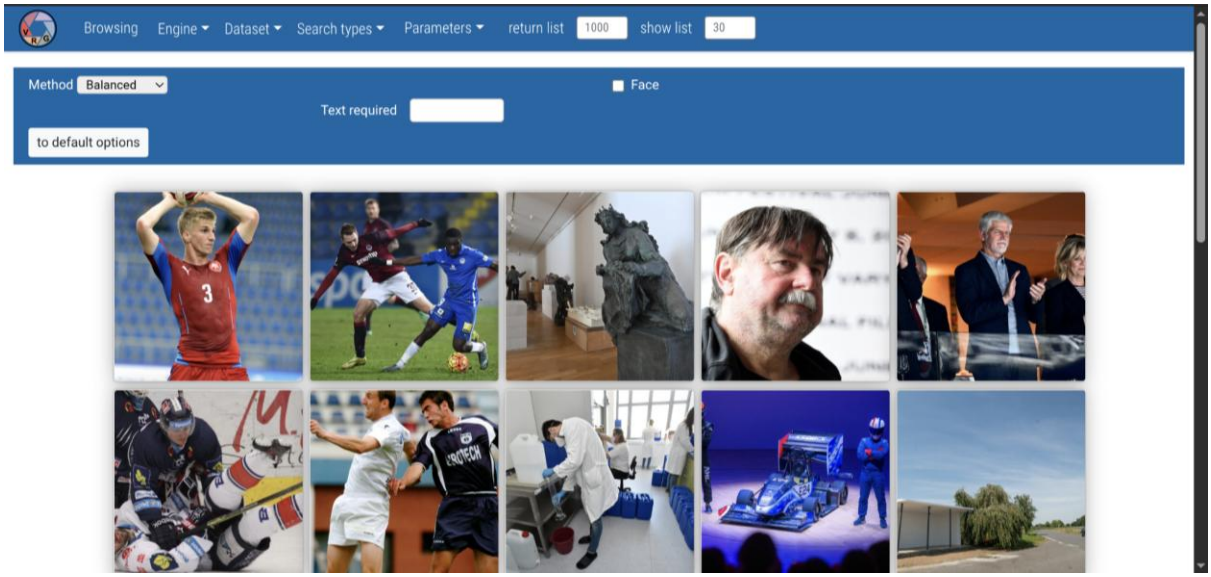
- Snížené rozlišení: Pro výpočet globálních příznaků pro vyhledávání podobnosti.
- Plné rozlišení: Pro detailní analýzu obličejů a odhad věku.

Vyvinuté Nástroje

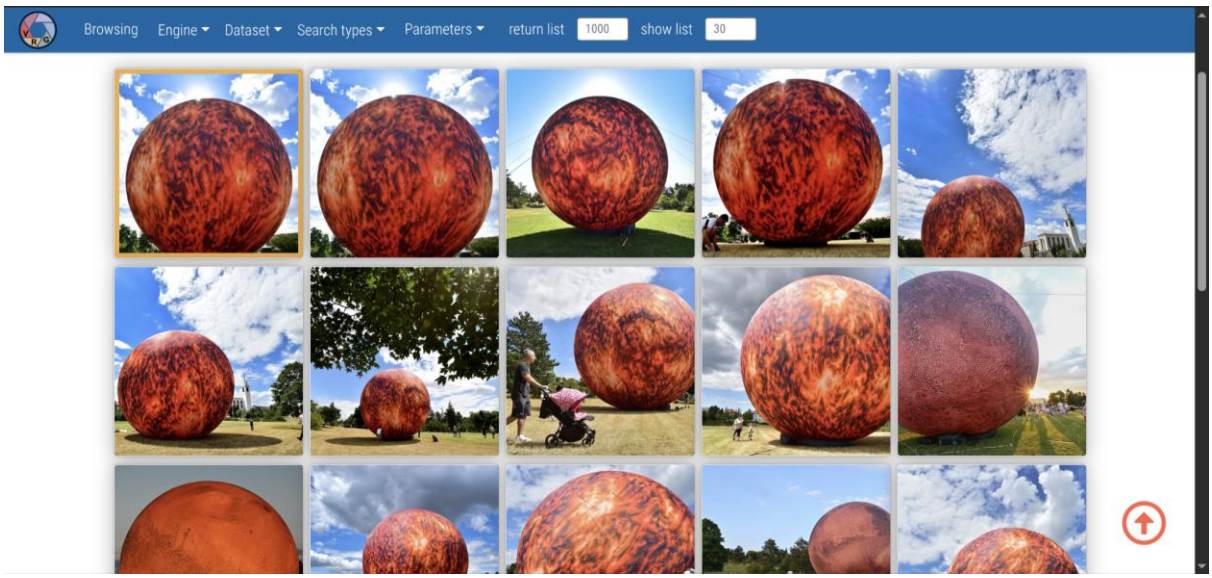
Všechny vyvinuté softwarové nástroje byly integrovány do demonstračních rozhraní, ke kterým má ČTK plný přístup. Řešení je v současné době ve fázi pokročilého testování a běží na reálných datech z archivu ČTK. Dále bylo pro ČTK vytvořeno API, které umožňuje přístup k funkcím pro zpracování obličejů a umožňuje editorům ČTK testování na interních datech.

Nástroj pro vyhledávání na základě podobnosti a sémantické vyhledávání

Na základě smlouvy s ČTK bylo týmu VRG ČVUT zpřístupněno přibližně 2 000 000 fotografií ve sníženém rozlišení. Tyto fotografie byly použity jako databáze pro vyhledávání.

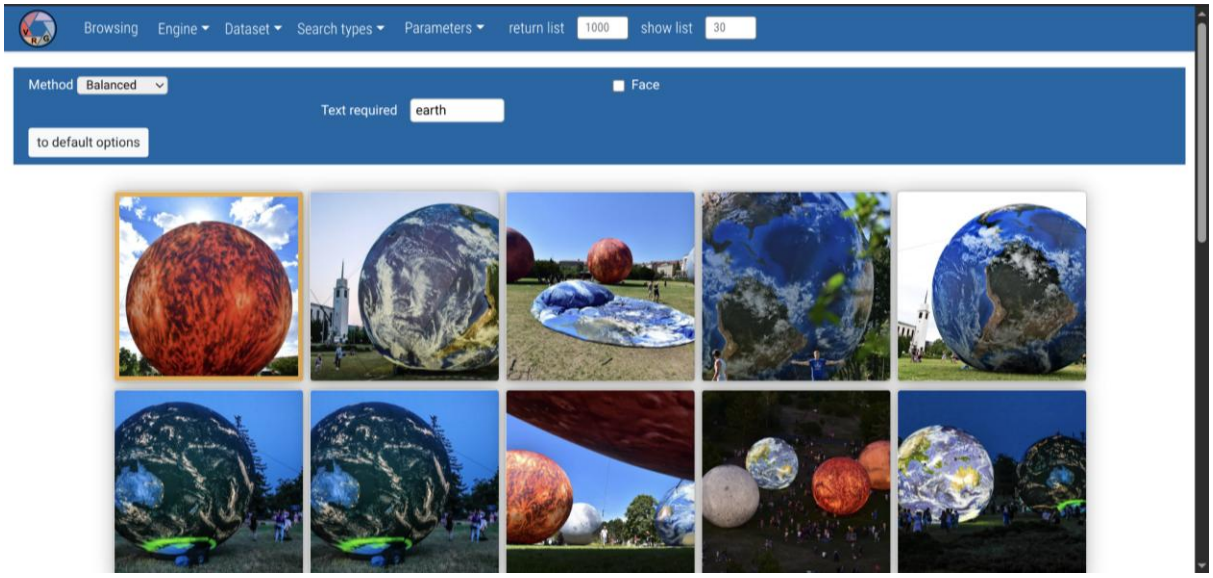


Obrázek č. X: Uživatelské rozhraní nástroje pro vyhledávání na základě podobnosti. Nástroj umožňuje vyhledávat na základě podobnosti k fotografii, na základě textu, nebo v kombinaci obojího.



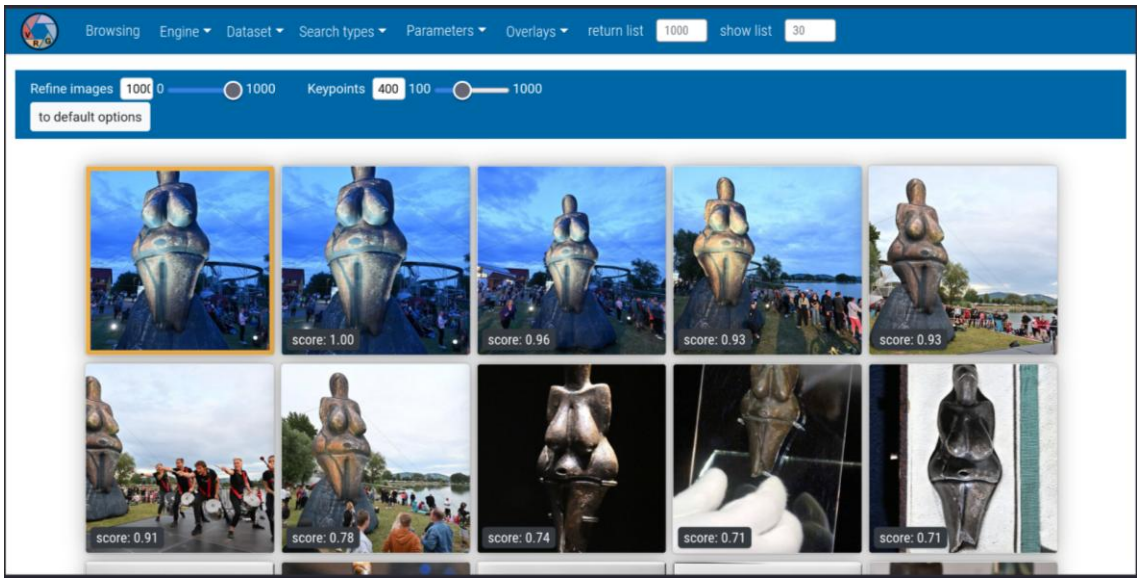
Obrázek č. X: Ukázka výsledku vyhledávání na základě podobnosti. Fotografie použitá k vyhledání je v levém horním rohu, zvýrazněna žlutě.

Vyhledávání funguje na základě podobnosti mezi dvěma obrázky, případně mezi obrázkem a textem. Podobnost je měřena v reprezentaci hluboké neuronové sítě. Jelikož reprezentaci lze spočítat pro libovolný obrázek nebo text, není pro vyhledávání potřeba mít fotografie popsány metadaty. Je proto možné prohledávat i archivní fotografie bez popisu.

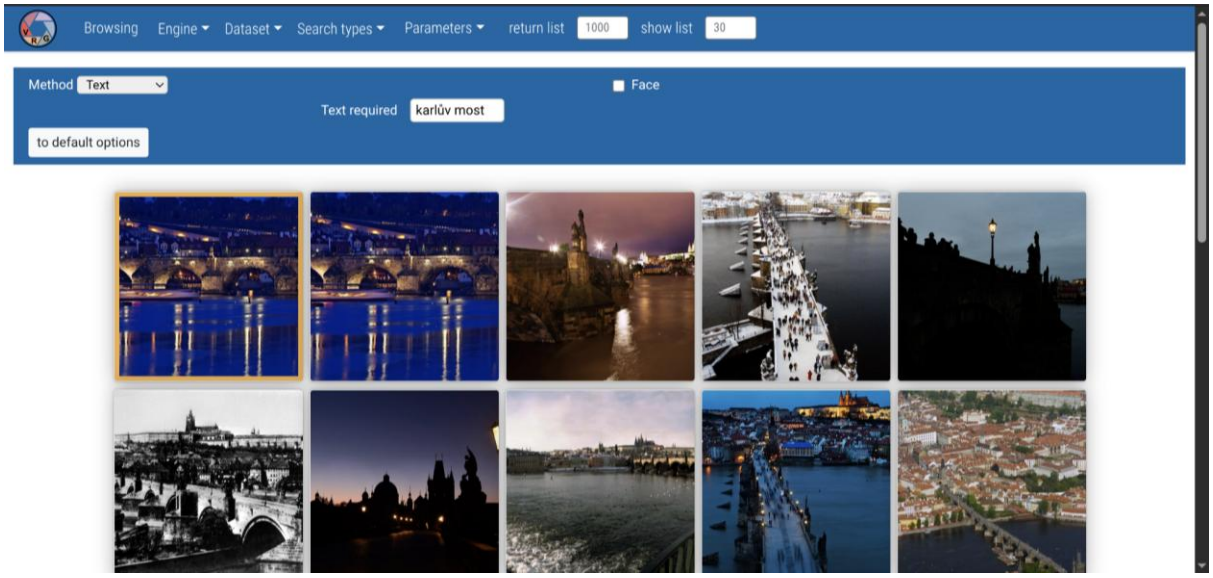


Obrázek č. X: Ukázka výsledku vyhledávání na základě podobnosti v kombinaci s textem. Fotografie použitá k vyhledání je v levém horním rohu, zvýrazněna žlutě. Vyhledávaný text "earth" způsobí, že nalezené obrázky pocházejí ze stejné expozice, ale zobrazují Zemi.

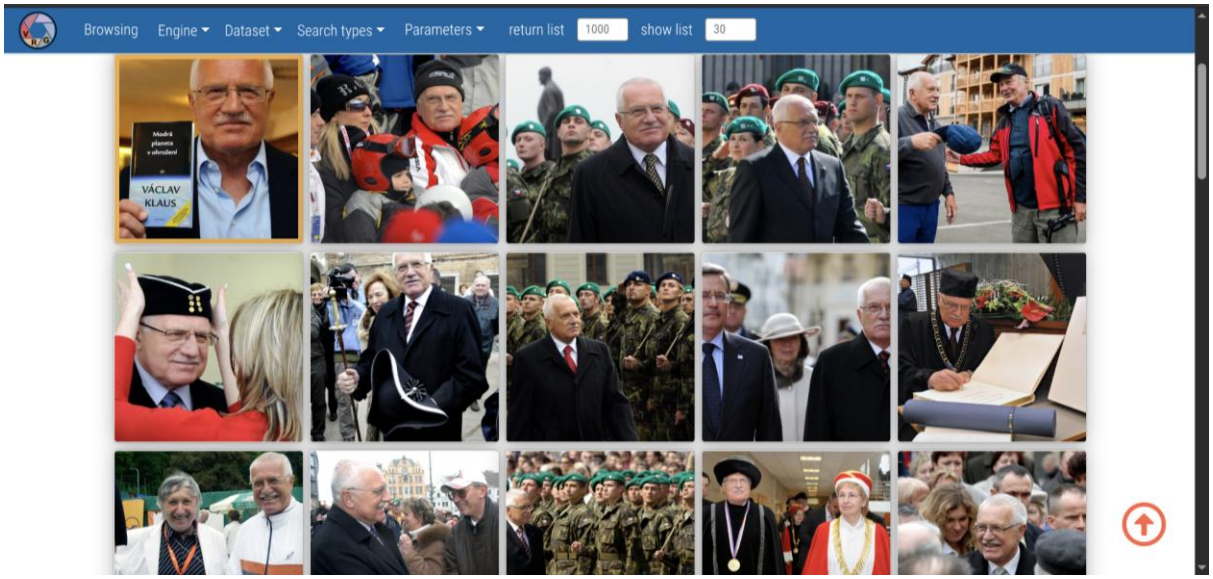
Nástroj umožňuje kombinovat vyhledávání pomocí obrázku a pomocí textu. Je proto možné vyhledávat konkrétní objekty specifikované obrázkem v kontextu specifikovaném textem.



Obrázek č. X: Ukázka výsledku vyhledávání na základě podobnosti. Fotografie použitá k vyhledání je v levém horním rohu, zvýrazněna žlutě.

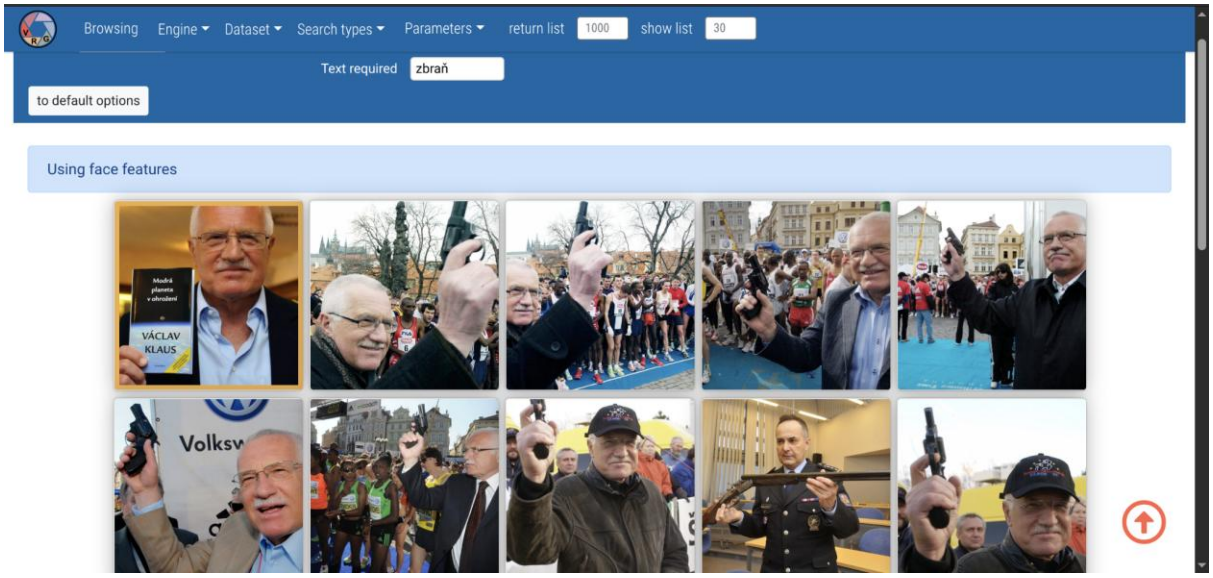


Obrázek č. X: Ukázka výsledku vyhledávání na základě textu "Karlův most". K vyhledávání nejsou použita žádná metadata nebo manuálně vytvořené tagy, používá se podobnost mezi hledaným textem a fotografií na základě výstupu neuronové sítě.



Obrázek č. X: Nástroj podporuje funkci vyhledávání na základě obličeje, převzatou z nástroje pro vyhledávání na základě identity a věku. Je pak možné vyhledávat konkrétní osoby v kombinaci s textovým popisem. Zde zobrazen Václav Klaus vyhledán v kombinaci s textem "čepice".

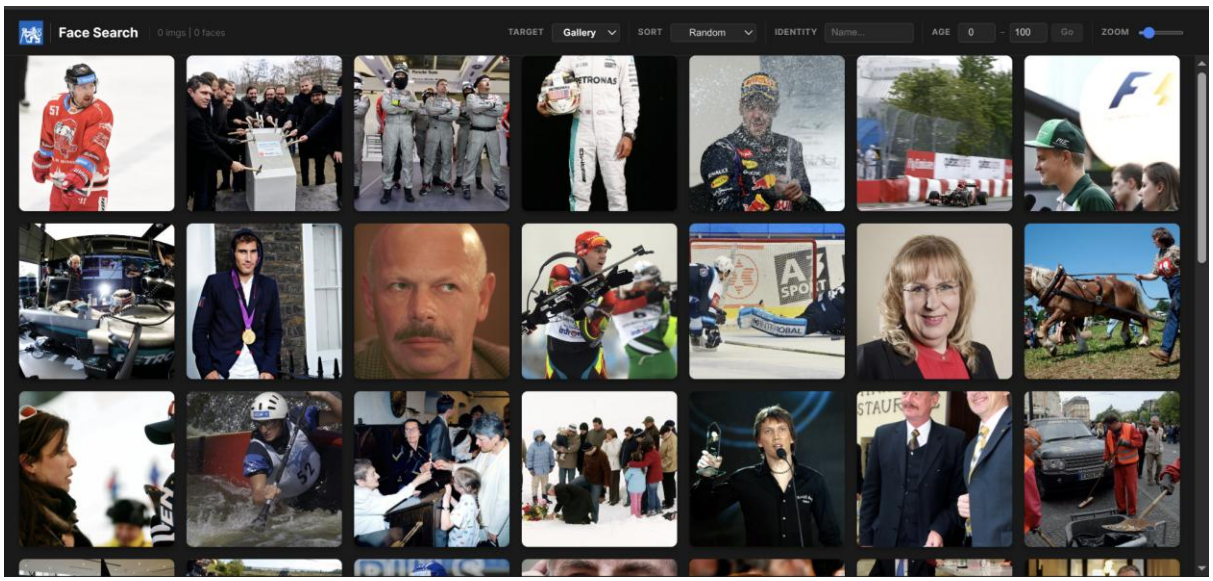
Nástroj dále umožňuje vyhledávat pomocí obličeje. V kombinaci s vyhledáváním pomocí textu je proto možné vyhledávat konkrétní osoby v libovolném kontextu specifikovaném textem. Veškeré tyto funkce opět nepoužívají metadata nebo manuálně vytvořené tagy.



Obrázek č. X: Nástroj podporuje funkci vyhledávání na základě obličeje, převzatou z nástroje pro vyhledávání na základě identity a věku. Je pak možné vyhledávat konkrétní osoby v kombinaci s textovým popisem. Zde zobrazen Václav Klaus vyhledán v kombinaci s textem "zbraň".

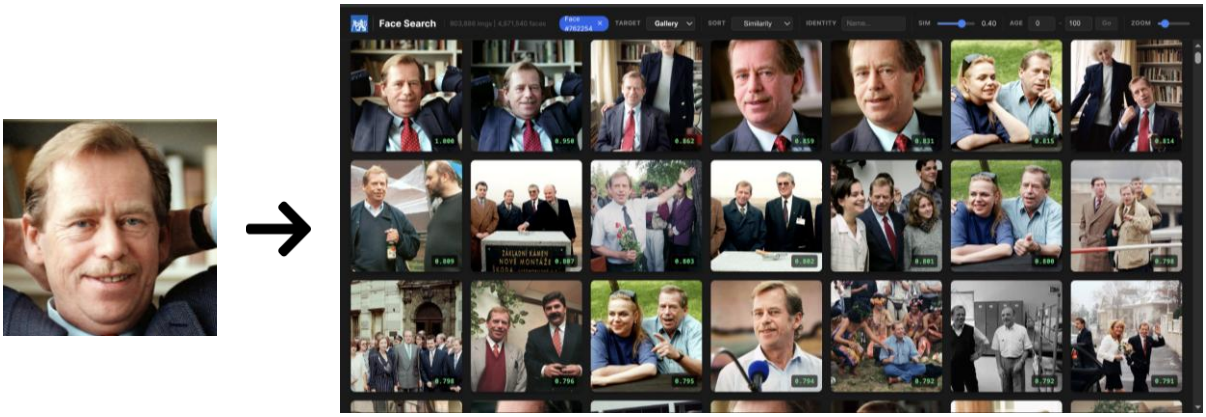
Nástroj pro vyhledávání na základě identity a věku

Na základě smlouvy s ČTK bylo výzkumnému týmu MLG ČVUT zpřístupněno přibližně 2 000 000 fotografií v plném rozlišení, na kterých bylo nalezeno 11 000 000 tváří. Tyto nalezené tváře tvořily zájmovou databázi pro vyhledávání.

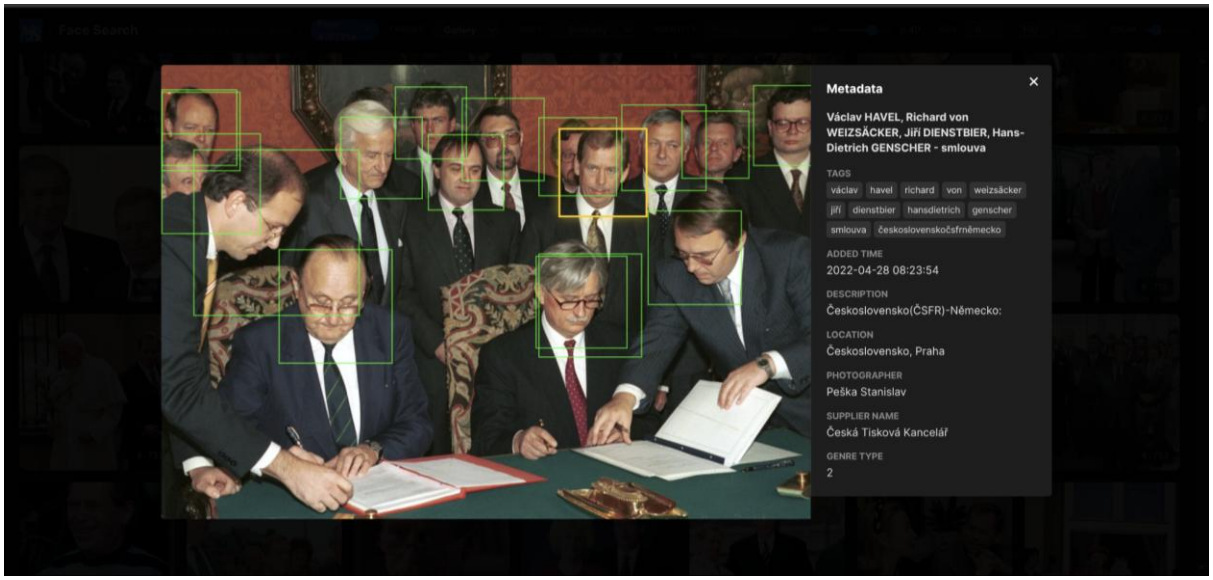


Obrázek č. X: Uživatelské rozhraní nástroje pro vyhledávání ve velkých databázích fotografií na základě identity a věku.

Nástroj umožňuje zvolit libovolnou tvář v databázi a vyhledávat další fotografie s vybranou identitou. Uživateli se fotografie zobrazují seřazené dle důvěryhodnosti nálezů.

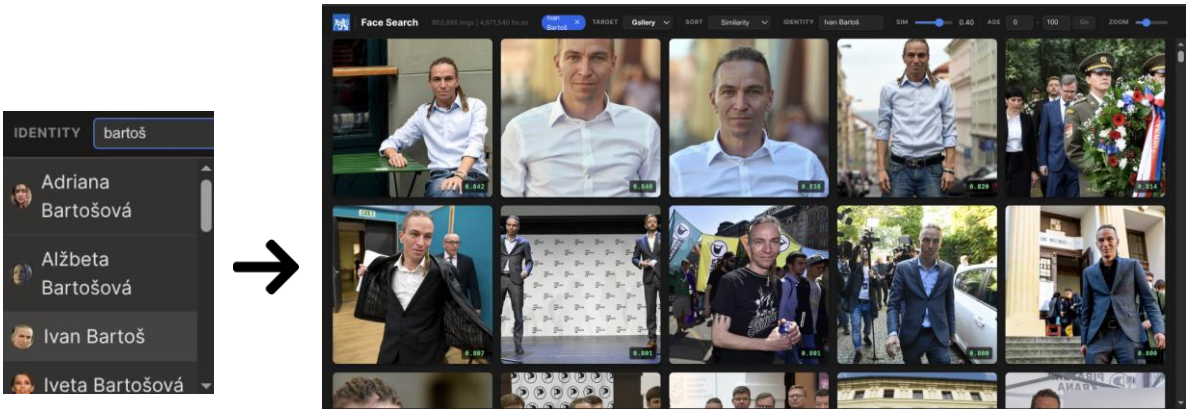


Obrázek č. X: Vyhledávání na základě identity obličeje; zde Václava Havla.

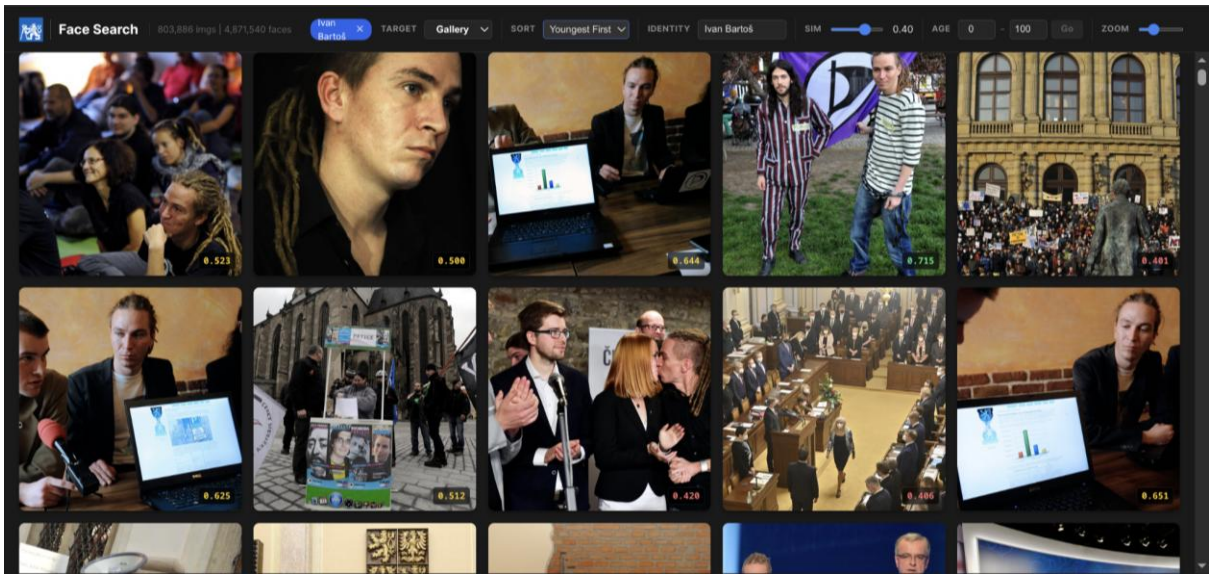


Obrázek č. X: Ukázka vyhledávání na základě identity obličeje; zde Václava Havla. Nalezený obličej je zvýrazněn žlutě.

Nástroj dále umožňuje vyhledávat na základě předem definovaných identit. Mediální agentura může udržovat vlastní seznam známých osob a pomocí tohoto seznamu vyhledávat fotografie, kde se identita vyskytuje. Jelikož je vyhledávání založeno čistě na vizuální podobnosti, je možné vyhledávat i na fotografiích, na kterých nebyly osoby označeny editorem.

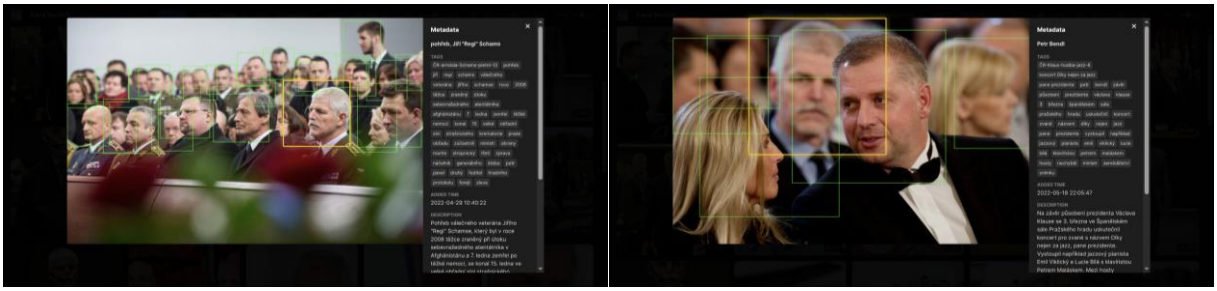


Obrázek č. X: Vyhledávání na základě jména; zde Ivana Bartoše.



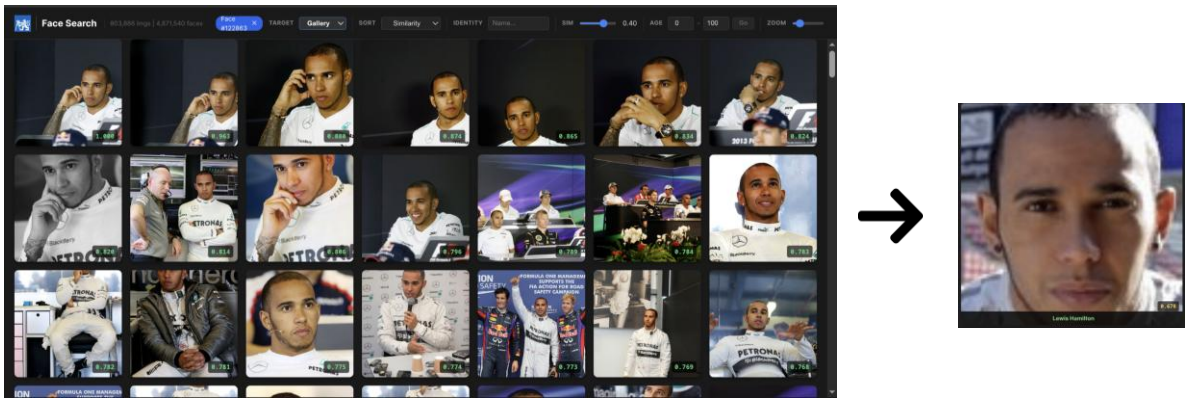
Obrázek č. X: Řazení výsledků na základě věku. Obrázky jsou seřazeny na základě vzhledu od nejmladší instance Ivana Bartoše po nejstarší.

Nástroj dále umožňuje výsledky řadit na základě odhadnutého věku vyhledávané osoby. Na základě odhadnutého věku nástroj také dokáže odhadovat datum pořízení fotografie. Je tedy schopen automaticky kontrolovat přiložená metadata a upozornit mediální agenturu na možné chyby v popisech fotografií. Díky odhadu věku je možné zakrýt obličej nezletilým.



Obrázek č. X: Příklad využití nástroje pro popis archivních fotografií.
Zde vyobrazeny nalezené fotografie s Petrem Pavlem před jeho vstupem do politiky.

Nástroj je vhodný zejména na organizaci archivních fotografií. Umožňuje zpětně dohledat fotografie se zájmovou identitou, aniž by tato identita musela být popsána. Je proto možné vyhledávat fotografie se známými osobnostmi z doby, kdy ještě známy nebyly.



Obrázek č. X: Příklad využití nástroje pro rozpoznání identity.
Zde vyobrazeno správné rozpoznání pilota Formule 1, Lewise Hamiltona.

Nástroj lze dále použít pro zefektivnění práce editora při popisování fotografií. Nástroj dokáže rozpoznat tváře známých osobností. Pro popis fotografie tak může editor obdržet strojově vytvořený návrh, který pouze zkontroluje a případně opraví.

Technické Řešení

Jádrem technického řešení jsou metody hlubokého učení a vektorové reprezentace dat.

1. **Sémantické vyhledávání a podobnost:** Využíváme modely které mapují text i obraz do společného vektorového prostoru. To umožňuje porovnávat sémantickou blízkost textového dotazu a obsahu fotografie.
2. **Rozpoznávání tváří:** Proces zahrnuje detekci obličeje, zarovnání a extrakci příznaků pomocí hlubokých konvolučních sítí. Každá tvář je reprezentována kompaktním vektorem, který je invariantní vůči osvětlení či póze.
3. **Odhad věku:** K predikci věku je využita samostatná neuronová síť trénovaná na odhad vizuálního věku z výřezu obličejů.

Výpočetní náročnost: Zatímco trénování modelů a prvotní indexace archivní databáze (zpracování 2 mil. fotografií) je výpočetně náročné a probíhá na strojích s grafickými kartami (GPU), samotné vyhledávání je optimalizováno pro rychlou odezvu v řádu milisekund, což umožňuje interaktivní práci s nástrojem v reálném čase.

SA4 Automatizace správy obsahu sociální sítě

Práce na této subjektivitě pobíhali na zároveň na pracovištích AIC FEL ČVUT a UFAL MFF UK. Shrnujeme je v tomto pořadí.

Správa účtů na sociálních sítích obnáší řadu repetitivních a časově náročných úkolů, jako je monitorování komentářů, reagování na ně za účelem zvýšení zapojení (engagement), přesdílení obsahu z jiných zdrojů nebo úprava formy původního obsahu tak, aby vyhovoval standardům a maximalizoval dosah na různých platformách.

V rámci této dílčí aktivity jsme se zaměřili na využití AI agentů založených na velkých jazykových modelech (LLM) k (částečné) automatizaci některých z těchto úkolů. Abychom umožnili testování a trénování AI agentů v bezpečném prostředí, prozkoumali jsme existující simulátory zapojení na sociálních sítích a identifikovali projekt Y social jako vhodné prostředí. V tomto rámci jsme vytvořili jednoduché prototypy agentů a zároveň jsme zapojili aplikační partnery projektu do diskuzí o konkrétních případech užití.

Diskuze ukázaly zájem především o automatizovaný monitoring a upozorňování (alerting), a také o automatizovanou tvorbu návrhů příspěvků cílených na specifické platformy. Celkový postoj byl však skeptický kvůli nízké spolehlivosti modelů.

Proto jsme se nakonec zaměřili na následující úkoly:

1. Automatizovaná tvorba poutavých návrhů příspěvků na základě existujícího obsahu, například vytvoření blogového příspěvku na základě vědeckého článku nebo příspěvku na sociální sítě na základě formální tiskové zprávy.
2. Zvyšování spolehlivosti textu generovaného agenty LLM vynucením držení se vstupu (kontextu) při standardním generování tokenů.
3. Zvyšování spolehlivosti textu generovaného agenty LLM rozšířením architektury LLM o explicitní primitivum pro kopírování z kontextu.
4. Vytvoření a pečlivé vyhodnocení spolehlivosti LLM agenta pro automatizaci ověřování faktů (fact-checking) prováděného serverem demagog.cz.
5. Generování přirozených popisných textů z datových tabulek, jako jsou např. sportovní výsledky.
6. Využití LLM agentů pro analýzu a popis časových řad, jako je např. vývoj cen akcií.

Automatizovaná tvorba poutavých příspěvků

Pro vytvoření a validaci systému pro automatickou optimalizaci zapojení u příspěvků na sociálních sítích potřebujeme datovou sadu obsahu s ukazateli míry zapojení, které vyvolal. Pro úlohu vytvoření poutavého blogového příspěvku z vědeckého článku jsme shromáždili datovou sadu článků a odpovídajících blogových postů, doplněnou o počet komentářů a „lajků“ pro měření zapojení. Pro optimalizaci příspěvků pro různé platformy jsme shromáždili datovou sadu příspěvků ze sítí X, Instagram a LinkedIn, včetně komentářů, lajků a počtu sledujících pro normalizaci dat.

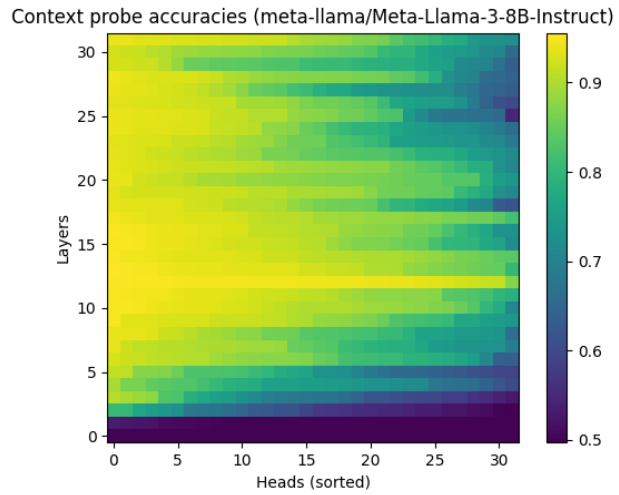
Na základě těchto datových sad jsme nejprve vyhodnotili schopnost LLM predikovat míru zapojení příspěvků. V době experimentů (začátek roku 2025) byla schopnost špičkových

(frontier) LLM modelů v této predikční úloze velmi omezená. Proto jsme se rozhodli pro tento úkol natrénovat menší model založený na architektuře BERT, což se jeví jako mnohem slibnější cesta.

Dalším krokem byl samotný „agentní cyklus“ (agentic loop), který navrhl obsah, vyhodnotil jej, vygeneroval návrhy na zlepšení a uložil je do pracovní paměti před vygenerováním dalšího návrhu. Tento proces konzistentně vylepšoval příspěvky na základě měřené metriky.

Zvyšování spolehlivosti LLM vynucením dodržování kontextu

Zkoumali jsme, jak může agent LLM rozhodnout, zda odpovědi, které generuje, pocházejí ze spolehlivých dat prezentovaných v jeho kontextu, nebo z nespolehlivých informací uložených ve vahách modelu. Navíc jsme vyvinuli metody, jak tento kompromis ovlivnit a podstatně zvýšit spoléhání se na informace v kontextu, kdykoli je to přínosné. Používáme kombinaci dvou metod. Jednou z nich je podvzorkování (subsampling) kontextu pro sledování změn v pravděpodobnosti produkování výstupu. Druhou je trénování lineárních sond (linear probes) k predikci toho, zda produkovaný výstup pochází z kontextu, a následná intervence do modelu prostřednictvím sondy za účelem posílení této vlastnosti. Naše experimenty ukazují, že tyto metody jsou efektivní, a do konce projektu plánujeme zveřejnit repozitář a odborný článek na toto téma.



Obrázek ukazuje přesnost predikce, zda daná odpověď pochází z vah nebo z kontextu, a to na základě různých hlav v různých vrstvách modelu Llama 3 3B.

Zvyšování spolehlivosti LLM pomocí explicitní schopnosti kopírování

Dalším způsobem, jak zvýšit spolehlivost LLM, je zlepšit jejich schopnost kopírovat informace přítomné v jejich kontextu na výstup bez jakýchkoli úprav. To je důležité zejména u „reasoning“ modelů (modelů se schopností uvažování), které neprodukují odpověď okamžitě, ale vytvářejí delší myšlenkové řetězce (chains of thought), ve kterých opakovaně přeformulovávají stejnou informaci v mírně odlišných kontextech. Pokud model v tomto procesu udělá chybu při kopírování, je pravděpodobné, že se chyba bude kumulovat a podstatně ovlivní konečný výsledek. Ověřili jsme, že LLM mohou mít s kopírováním v jednoduchých úlohách uvažování problémy.

Následně jsme se ve spolupráci s kolegy z VUT v Brně zaměřili na změnu architektury LLM tak, aby bylo kopírování podstatně spolehlivější, a to přidáním atomické operace kopírování úseku (spanu) z kontextu přímo na výstup. Počáteční experimenty ukazují, že model lze vyladit (fine-tune) tak, aby tuto dodatečnou operaci využíval. Celková spolehlivost se však v tuto chvíli nezlepšila. Model je schopen používat operaci kopírování a vést smysluplné konverzace, ale aktuální verze je v benchmarkových testech horší kvůli občasným chybám v identifikaci úseku ke kopírování a nižší přesnosti při generování atomických tokenů. Plánujeme na modelu dále pracovat a vyzkoušet alternativní architektury a režimy ladění (fine-tuning).

Automatizace ověřování faktů pro demagog.cz

Během projektových schůzek jsme identifikovali novou příležitost pro agenty LLM při pomoci serveru demagog.cz s tvorbou fact-checků (ověřování faktů). Je to vynikající úloha pro zlepšování spolehlivosti agentů, a to vzhledem k velmi vysokým nárokům na úplnost a kvalitu zdrojů, přísným stylistickým pravidlům pro finální podobu fact-checku a jasnému procesu zpětné vazby, kdy zkušení editoři poskytují juniorním analytikům zpětnou vazbu k návrhům jejich práce.

Naše počáteční experimenty s agenty pro hloubkový průzkum (deep research agents) od OpenAI a Google Gemini byly slibné, ale rozhodně ne dostačující. Proto jsme naše řešení založili na open-source frameworku Open Deep Research od LangChain. Během vývoje našeho vlastního agenta však byly proprietární deep research agenty překonány špičkovými reasoning modely. Od příchodu GPT-5 byly tyto modely schopny vytvářet fact-checky užitečné pro demagog.cz, a proto jsme se zaměřili na pomoc s architekturou začlenění těchto agentů do jejich interních systémů.

Na jejich základě jsme vytvořili šablony promptů a statický neadaptivní pracovní proces agenta (workflow) pro generování fact-checků. Toto workflow průběžně využíváme ke generování plně automatizovaných ověření pro výroky, které lidští analytici právě verifikují. Každý vygenerovaný fact-check je následně hodnocen lidskými analytiky na základě přísného kontrolního seznamu kritérií. Tento proces vytváří databázi využitelnou pro vyhodnocování schopností současných LLM v této úloze, kterou lze rovněž využít pro fine-tuning nebo další automatizované zlepšování workflow. Předběžné výsledky dosavadních anotací jsou shrnuty v následující tabulce.

Celkové hodnocení AI generovaného fact checku annotátorem	Počet
Vše bez problémů, s hrdostí zveřejnitelné.	0
Zveřejnitelné beze změn, ale bylo by vhodné provést úpravy.	9
Zveřejnitelné po zásadnějších změnách.	7
Nezveřejnitelné, ale ušetřilo to lidskému fact-checkerovi nějaký čas.	1
Bylo by lepší, kdyby to neexistovalo. Čtením fact-checker ztratil víc času, než mu to ušetřilo.	0

Generování popisných textů z datových tabulek

Na pracovišti ÚFAL jsme se v rámci této subaktivity věnovali vývoji nástroje SQuirreL pro generování popisných textů – zajímavostí – z datových tabulek, jako jsou např. sportovní výsledky. Navrhli jsme přístup využívající LLM v kombinaci s SQL dotazy nad tabulkou: systém nejprve pomocí LLM vygeneruje základní myšlenku dané zajímavosti. V dalším kroku toto LLM převede na odpovídající SQL dotaz (dotaz v tabulce slouží k vyhledání faktických dokladů dané zajímavosti). Se základní myšlenkou a výsledky SQL můžeme pak vygenerovat výslednou zajímavost včetně jasného dokladu. Systém je pomocí reflektivních LLM promptů schopen odhalit vlastní chybu a vygenerovat opravený text. Máme funkční prototyp systému SQuirreL a provedli jsme detailní lidskou evaluaci přes crowdsourcing, ale kvalita výstupu na benchmarcích (LogicNLG a náš vlastní benchmark FreshTab viz níže) zatím nepřekonává přímé promptování LLM, zejména z důvodu vyšší citlivosti nástroje SQuirreL na konzistenci dat. Data v benchmarcích jsou založena na ručně vytvořených tabulkách z Wikipedie, kde často dochází k nekonzistencím (např. data různých typů ve stejném sloupci). Aktuálně se proto pokoušíme posílit robustnost systému.

Pro danou úlohu generování zajímavostí z tabulek jsme vytvořili novou metodu nazvanou FreshTab pro automatické sestavení nových benchmarků na základě Wikipedie a Wikidat, abychom mohli evaluovat systémy s novými tabulkami, kterým použité LLM nemohly být vystaveny během předtrénování. To zajišťuje objektivnější evaluaci LLM a předchází zkreslení výsledků vlivem kontaminace trénovacích dat. Výsledky naší práce byly publikovány na konferenci INLG (International Conference on Natural Language Generation, CORE-B, https://aclanthology.org/2025.inlg-main.7/), kde příspěvek získal ocenění Best Short Paper. Tytéž výsledky byly prezentovány i jako poster na prestižním workshopu GEM (Generation, Evaluation, and Metrics) pořádaném v rámci konference ACL. FreshTab je dostupný na Githubu pod otevřenou licencí MIT (https://github.com/Kristyna-Navitas/FreshTab).

Využití LLM agentů pro analýzu a popis časových řad

Dále jsme se věnovali experimentům s využitím LLM pro analýzu a popis časových řad (např. časový vývoj cen akcií), který zahrnuje agentické přístupy, kde LLM nejprve opakovaně generuje a pouští kód v jazyce Python a přímo tak pracuje s relevantními daty, což má za účel zvýšení přesnosti konečného výstupu. Vývoj si vyžádal některá rozšíření nástroje Factgenie (viz KA2 SA15). Evaluovali jsme výstupy několika různých LLM v různých nastaveních, kdy srovnáváme přímé promptování s použitím různých variant generování kódu. Ukazuje se, že na benchmarku TimeSeriesExam mají varianty s použitím kódu cca o několik procent vyšší úspěšnost. Aktuálně evaluaci rozšiřujeme na další benchmarky a připravujeme sepsání sepsání experimentů do konferenčního příspěvku, pravděpodobně na konferenci ICML 2026.

SA5 Nástroj pro spolehlivý překlad důvěryhodných textů do cizích jazyků

Cílem této subaktivity je zvýšit **spolehlivost strojového překladu** důvěryhodných textů (zejm. zpravodajství ČTK) do cizích jazyků a umožnit jejich **bezpečné publikování i v zahraničí.** Systém musí umět minimalizovat rizika překladových chyb, upozornit na problematická místa a současně být prakticky nasaditelný v redakčních workflow, včetně práce s formátovanými dokumenty a webovým obsahem.

Modely a experimenty: LLM pro překlad i odhad spolehlivosti

Jedním z prvních kroků bylo vyhodnocení kvality překladu prostřednictvím velkých jazykových modelů. Provedli jsme široké srovnání dostupných modelů a pro jazyky a typy textů uvažované v projektu jsm na základě výsledků zvolili 3 modely, se kterými pracujeme dále: **Granite 3, EuroLLM a Aya**. U modelu Granite 3 jsme ověřili, že i když výchozí varianta neposkytuje dostatečně kvalitní překlady, je možné ji **výrazně zlepšit dotrénováním** postupem ověřeným v našich předchozích pracích. Modely **EuroLLM** a **Aya Expanse** se v našich experimentech ukázaly jako velmi silné již bez dalšího tréninku a v předběžných testech se zdá, že u překladu prostého textu je lze ještě dále zlepšit.

Odhad spolehlivosti a mezinárodní evaluace (WMT)

V článku *CUNI at WMT25 General Translation Task* jsme dále prezentovali vývoj a testování nových překladových systémů pro jazykové páry **en↔cs, cs↔de, cs↔uk a en↔sr**, založených na finetuning postupech velkých jazykových modelů (např. **EuroLLM-9B-Instruct**) metodami **LoRA** a **Contrastive Preference Optimization**. Tyto systémy slouží jako podklad pro další výzkum spolehlivého a doménově adaptovaného překladu.

Překlad formátovaných textů a datové sady pro evaluaci

Praktické nasazení strojového překladu v redakčních a lokalizačních workflow naráží na specifický problém: kromě samotného významu musí systém zachovat také **strukturu dokumentu a inline značky** (např. HTML tagy, XLIFF markery, odkazy, zvýraznění, placeholdery). Chyby ve značkách jsou často závažné – mohou rozbít webovou stránku, způsobit chyby ve vykreslení,nebo znemožnit automatické zpracování v navazujících krocích publikace. Pro tyto účely jsme vytvořili dataset zahrnující **webové stránky, PDF příručky a další dokumenty** v jazycích používaných menšinami v ČR (**čeština, ukrajinština, angličtina, vietnamština, ruština a další**) a připravili jeho popis do článku zaslaného na konferenci LREC (v době psaní textu v recenzním řízení).

V tomto článku o datasetu **CzechDocs** je tento problém řešen na autentických formátovaných dokumentech (HTML, DOCX, PDF), které byly převedeny do reprezentací vhodných pro evaluaci (zejména XLIFF). Dataset je záměrně „tagově hustý“: většina segmentů obsahuje alespoň jeden tag, což umožňuje systematicky měřit nejen kvalitu překladu textu, ale i schopnost tagy **správně kopírovat, zachovat a umístit**.

Dva hlavní přístupy k překladu s tagy

1) Detag-and-project (detagování a zpětné promítnutí tagů)
Klasický přístup známý z lokalizace spočívá v tom, že se modelu předá pouze „holý“ text bez tagů. Tagy se po překladu **znovu vloží** do cílového textu pomocí slovního zarovnání a heuristik pro projekci značek. Výhodou je, že překladač řeší čistě jazykový problém a lze použít jakýkoliv překladový engine, i takový, který není k překladu značek přizpůsobený. Nevýhodou je složitější pipeline (zarovnání, heuristiky) a riziko chyb při projekci (zejm. když dojde k přeformulování, přeskládání slovosledu apod.).

2) Přímý překlad se tagy ve vstupu (direct tagged input)
Alternativně se tagy modelu ukážou přímo ve vstupu (typicky v MOS formátu, nebo nahrazeny placeholdery, které mají být zkopírovány). Model má za úkol přeložit text a současně tagy **zachovat a umístit na odpovídající místa**. U LLM se ukazuje, že tento přístup může fungovat překvapivě dobře i bez speciálního tréninku, ale je citlivý na instrukce a na typ dokumentu.

V práci porovnáváme variantu, kdy je model pouze vyzván „přelož“, versus variantu, kde prompt **výslovně požaduje zachování markup značek**. Výsledek je prakticky důležitý: u některých modelů (v uvedené studii zejména u *gpt-4.1-nano*) vede explicitní instrukce k dramatickému zlepšení metriky, která hodnotí překlad *včetně tagů* (tagged BLEU), zatímco metriky na „odtagovaném“ textu (untagged BLEU) zůstávají podobné. Jinými slovy: dobrý prompt umí výrazně snížit riziko, že model tagy přenese nesprávně.

Model	Input	D&P	Prompt	Tagged BLEU	Untagged BLEU
	Plaintext	No	Normal	0.9	55.7
	MOS	No	Normal	25.4	54.6
gpt-4.1-nano	MOS	No	Explicit	87.8	55.4
	MOS	No	Explicit+Context	79.5	54.6
	MOS	Yes	Normal	89.0	50.4
	Plaintext	No	Normal	0.8	51.8
	MOS	No	Normal	75.7	52.4
aya-expanse-8B	MOS	No	Explicit	83.6	51.4
	MOS	No	Explicit+Context		
	MOS	Yes	Normal	89.6	49.5

Co z výsledků plyne pro nasazení

Z porovnání přístupů vyplývá několik praktických závěrů:

- **LLM umí tagy přenášet přímo**, ale spolehlivost výrazně roste, když se zachování tagů explicitně vynutí promptem.
- **Detag-and-project** může dosahovat velmi dobrých výsledků. Cena je vyšší složitost pipeline a citlivost na kvalitu alignmentu/heuristik.
- Mezi přístupy existuje **trade-off**: některé konfigurace zlepší „tagovou správnost“, ale mohou mírně zhoršit skóre na čistém textu (protože projekce tagů i segmentační omezení můžou omezovat přirozené přeformulování).

Tento rámec přesně zapadá do cíle SA5: aby redaktor mohl publikovat jazykové mutace bez znalosti cílového jazyka, musí systém hlídat nejen správnost významu, ale i **strukturální integritu** – a práce na CzechDocs ukazuje, jak takovou integritu realisticky měřit a jaké strategie (detagging+projekce vs. přímý tagged input s promptem) dávají v praxi nejlepší kompromis.

Rychlá revize překladu pomocí eyetrackingu

V rámci SA5 jsme prověřovali možnost **velmi rychlé revize** strojových překladů pomocí sledování očních pohybů (eyetrackingu). K sledování očních pohybů byl využit eyetracker Eyelink 1000 plus, který s frekvencí 1000 Hz snímá odhadovanou pozici pohledu uživatele v rámci obrazovky.

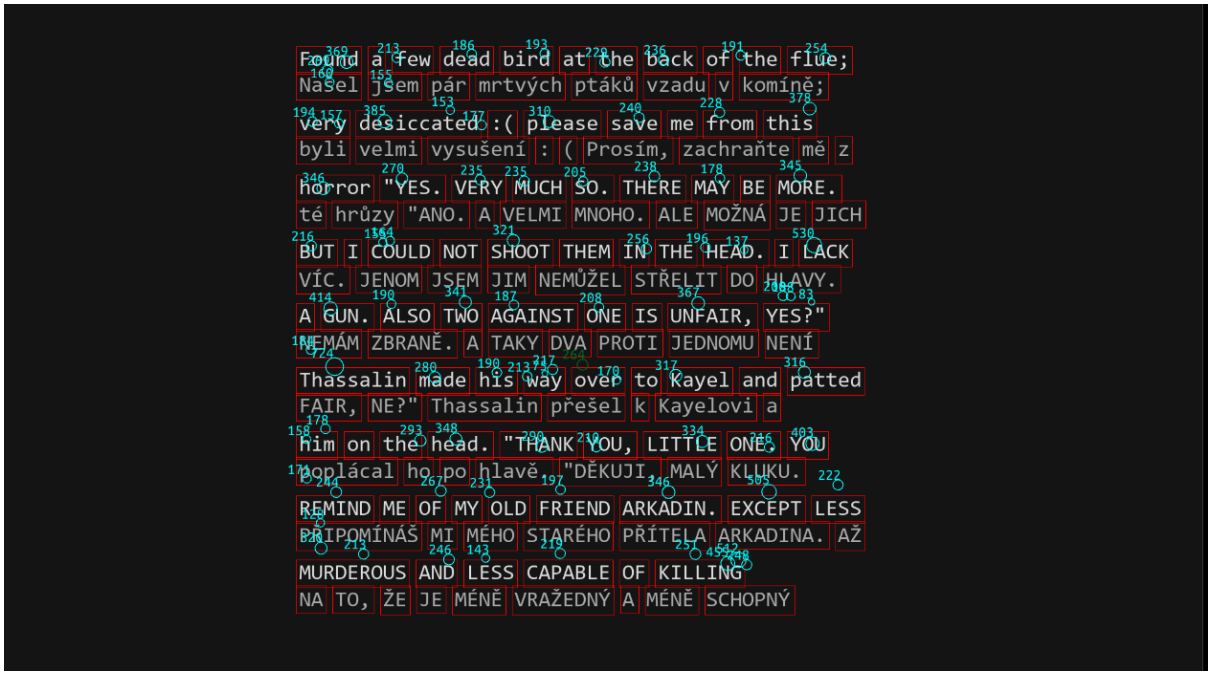
Motivace

Tradiční přístup k revizi ve strojovém překladu, kterým rozumíme ruční opravování nebo alespoň vyznačování chyb, má dva nedostatky: časovou náročnost a potřebu odborných anotátorů. V důsledku toho je dnešní revize překladu velmi drahá a zdlouhavá. Pokud by ovšem bylo možné chyby v překladu odhalit čistě na základě čtení, revize by se výrazně zefektivnily. Představou je například to, že se člověk u chyby zastaví nebo se vrátí na začátek věty, protože ho chyba zmate. Možné jsou ale i další potenciální vzory chování.

Cílem našich experimentů tedy bylo ověřit, zda je možné chyby odhalit pomocí eyetrackingu při čtení překladů, zda budou naměřená data dostatečně vypovídající a zda se v datech vyskytnou nějaké vzory, které v dostatečném množství případů vypovídají o chybě. Dále jsme zkoumali, do jaké míry ovlivňuje přesnost našeho přístupu čtenářova expertiza v oblasti lingvistiky a překladu.

Dosavadní postup

Po seznámení se s technickými aspekty měřícího přístroje Eyelink 100 plus i jeho dokumentací a iniciálním zapojení jsme pro experimenty zvolili knihovnu Pylink, která nabízí API nutné pro komunikaci s eyetrackingovým přístrojem a promítání experimentů uživateli.
Následovala implementace promítání zdrojového textu a jeho strojového překladu tak, aby se střídaly řádky zdrojového a přeloženého textu pro možnost jejich snadného porovnání čtenářem. Promítání má několik parametrů, se kterými lze experimentovat, abychom vyvážili efektivitu a přesnost nahrávání, např. velikost fontu, rozestupy mezi řádky či padding po stranách obrazovky. Nahraná data se ukládají do EDF (EylinkDataFormat), ve kterém jsou uloženy jednotlivé fixace očních pohybů v čase, obrazovka zobrazená v daný moment a další data. Dále ukládáme souřadnice oblastí, které na obrazovce zaujímají jednotlivá slova. Ty lze spárovat se souřadnicemi fixací očních pohybů z měření. Právě z tohoto párování se pokoušíme vypozorovat vzory, které by napovídaly, že se jedná o chybu v překladu.
Pro efektivní porovnávání překladu a originálu bylo důležité, aby se delší souvětí nerozešly napříč řádky a odpovídající slova v obou jazycích zůstaly pokud možno pod sebou. To je obecně netriviální problém, zejména kvůli volnému slovosledu v českém jazyce. Problém jsme vyřešili pomocí nástroje Awesome Align (https://arxiv.org/abs/2101.08231), díky kterému jsme schopni napárovat jednotlivá slova a zalomit jednotlivé řádky tak, abychom předešli desynchronizaci překladu a originálu.



Na obrázku lze vidět rozložení textu pro experiment a areas of interest jednotlivých slov, které reprezentují červené obdélníky. Dále jsou vidět jednotlivé fixace čtenáře, které jsou vyznačený modrými značkami, u značek je pak trvání fixace v milisekundách.

Data

Jako data pro tento experiment využíváme anotovaná data strojového překladu z angličtiny do češtiny z překladové soutěže WMT 2024 (https://www2.statmt.org/wmt24/), která zahrnují originální text, jeho strojový překlad a chyby vyznačené manuální revizí překladu. Chyby z anotace WMT 2024 považujeme za správně a úplně identifikované (ground truth) a experiment odvíjíme od nich. Analyzujeme oční pohyby s cílem odhalit vzorce, které se vyskytují při čtení těchto chyb.

Dosavadní výsledky

V tuto chvíli máme připravený experiment, zajištěný eyetracker a pracujeme na zajištění respondentů, abychom mohli nasbírat dostatek dat k analýze. Sběr dat jako takový by měl proběhnout během ledna/února 2026 a samotná datová analýza pak během března.

Budoucí práce

Po nasbírání dat se budeme věnovat analýze dat a hledání vzorů, které by napovídaly, že čtenář narazil na chybu v překladu. Výsledky analýzy plánujeme publikovat v konferečním sborníku (např. konference o automatickém překladu WMT 2026). Kód bude uveřejněn pod otevřenou licencí.