Príloha A:
V3.3 Výsledky experimentov a overenia metód

Názov projektu	**Pokrok v multimodálnej predikcii a využitie interaktívneho fact-checkingu na podporu boja proti dezinformáciám**
Akronym projektu	**DisAI-AMPLIFIED**
Kód projektu	**09I01-03-V04-00007**
Začiatok projektu	**1.7.2024**
Trvanie projektu	**24 mesiacov**

Financované Európskou úniou NextGenerationEU

PLÁN OBNOVY

Výsledky experimentov a overenia metód

V tejto prílohe stručne zhrnieme výsledky dosiahnuté v rámci realizácie pracovného balíka NV3, predovšetkým hlavné výsledky experimentov realizovaných v rámci úloh 3.2 a 3.3 a výsledky overenia v rámci používateľskej štúdie v úlohe 3.4, ale pre úplnosť v krátkosti zosumarizujeme aj výsledky úlohy 3.1, ktorá sa venovala príprave dátových množín.

Úloha 3.1: Príprava dátových množín

Úloha 3.1 bola zameraná na zber, úpravu, anotáciu, overenie a pod. dát potrebných na realizáciu výskumných aktivít v rámci úloh 3.2 a 3.3. Na úvod zhrnieme informácie o dvoch dátových množinách, ktoré boli v rámci tejto aktivity publikované.

Dátová množina AMC-16K

AMC-16K (**Annotated-MultiClaim-16K**) je verejne dostupná dátová množina vytvorená v rámci úlohy 3.1. Dátová množina je hostovaná prostredníctvom GitHub repozitára kinit-sk/llms-claim-matching/datasets a rozširuje dátovú množinu MultiClaim vytvorenú v rámci projektu DisAI. Pre podmnožinu 16 000 párov (príspevok, fact-check) ľudskí anotátori ručne určili, či sú vzájomne relevantné alebo nie. Štruktúra dátovej množiny je nasledujúca:

- **8 000 monolingválnych** párov (40 príspevkov × 20 jazykov × 10 kandidátskych fact-checkov v rovnakom jazyku);
- **8 000 cross-lingválnych** párov (400 anotovaných dvojíc pre každú z 20 preddefinovaných jazykových kombinácií, kde sa jazyk príspevku a fact-checku líšia).

Dátová množina adresuje problém správnej evaluácie vyhľadávania existujúcich fact-checkov – jednu z hlavných výziev v rámci evaluácie predstavuje fakt, že vzhľadom na proces, akým sa zbierajú dátové množiny na túto úlohu, nie všetky prepojenia medzi vzájomne relevantnými príspevkami a fact-checkmi, bývajú anotované (čo platí aj pre dátovú množinu MultiClaim, na ktorej staviame). Ak teda model správne predikuje spojenie medzi dvojicou príspevok, fact-check, v niektorých prípadoch je takáto predikcia vyhodnotená ako nesprávna aj napriek tomu, že príspevok a fact-check k sebe v skutočnosti patria. Dodatočné anotácie v dátovej množine AMC-16K umožňujú vykonať presnejšiu evaluáciu.

Anotáciu vykonalo šesť anotátorov, pričom každý pár dostal jednu anotáciu (relevantný / irelevantný / *nedá sa určiť*); prípady označené ako *nedá sa určiť* boli následne dodatočne zaradené do binárnych kategórií. Zhoda anotátorov meraná Fleissovou kappou dosiahla pred anotáciou 0,60 a po nej 0,62. Celkovo bolo ako relevantných označených približne 16% párov z celkových 16 tisíc.

Zarhnuté jazyky: arabčina, bulharčina, čeština, nemčina, gréčtina, angličtina, francúzština, srbochorvátčina, hindčina, maďarčina, kórejčina, malajčina, barmčina, holandčina, poľština, portugalčina, rumunčina, slovenčina, španielčina a thajčina. Pre cross-lingválne nastavenie dataset pokrýva napríklad páry **slk** → **eng**, **spa** → **eng**, **kor** → **eng**, **ara** → **fra** a ďalšie.

Každý záznam obsahuje polia: post_id, factcheck_id, post_language, factcheck_language a binárnu premennú rating (Yes = relevantný, No = irelevantný). Priemerná dĺžka príspevkov sa

naprieč jazykmi pohybuje približne od 47 do 223 slov, pričom fact-checky majú priemerný rozsah približne 4 až 37 slov.

Dátová množina bola využitá v rámci výskumu realizovaného v kontexte úlohy 3.2.

Post-retrieval anotácie pre dátovú množinu SVO-Probes

Dátová množina post-retrieval anotácií pre SVO-Probes bola vytvorená v rámci výskumu realizovaného v úlohe 3.3 a slúži na evaluáciu novej probingovej metodiky pre vision-language modely (VLMs; modely kombinujúce obraz a jazyk). Navrhovaná metodika vychádza z retrieval scenára: ku každému obrázku alebo textovému opisu model vyhľadá top-K najpodobnejších položiek na základe embedovaní a následne sa analyzuje, do akej miery sú výsledky takéhoto vyhľadávania skutočne korektné. Dátová množina preto neobsahuje iba pôvodné kontrastívne dvojice z benchmarku SVO-Probes, ale predovšetkým dodatočne anotované post-retrieval páry vytvorené konkrétnymi modelmi v rámci realizovaných experimentov.

Základ dátovej množiny tvorí dátová množina SVO-Probes, ktorá je určená na testovanie porozumenia vzťahov subjekt-sloveso-objekt v obrázkoch a textoch. Dátová množina obsahuje textový opis (pozostávajúci z trojice subjekt-sloveso-objekt), pozitívny obrázok a kontrastívny negatívny obrázok, ktorý sa od pozitívneho líši presne v jednom z aspektov subjekt-sloveso-objekt. V experimentoch bolo použitých 13 855 unikátnych obrázkov, 10 978 unikátnych opisov a celkovo 33 686 kontrastívnych trojíc.

Kľúčovou súčasťou zverejnenej dátovej množiny sú ľudské anotácie výsledkov vyhľadávania (angl. retrieval) pomocou modelov. Dátová množina obsahuje výstupy šiestich vision-language modelov: CLIP, BLIP-2, FLAVA, SigLIP2, Qwen3-VL-Embed a fine-tunovaného Qwen2.5-VL. Pre text retrieval bolo náhodne vybraných 100 query obrázkov; ku každému bolo anotovaných top-10 popiskov nájdených každým z modelov, čo predstavovalo viac než 3000 dvojíc obrázok-popisok anotovaných tromi anotátormi. Následne bol rovnaký postup opakovaný pre vyhľadávanie obrázkov na základe popiskov. Anotátori posudzovali, či popisok môže realisticky opisovať daný obrázok, pričom používali trojhodnotovú schému áno/nie/nedá sa určiť. Finálne anotácie boli určené väčšinovým hlasovaním. Zhoda anotátorov meraná Fleissovou kappou dosiahla hodnotu približne 0,64, čo predstavuje stredne silnú zhodu medzi anotátormi.

Okrem ľudských anotácií dataset obsahuje aj automatické anotácie generované reasoning modelom o3. Model klasifikoval retrieval výsledky ako správne alebo nesprávne vyhľadanie a v prípade chyby určoval typ nesúladu: nesprávny subjekt, sloveso alebo objekt. Automatické anotácie dosiahli vysokú zhodu s ľudským hodnotením (F1 ≈ 0,84), čo naznačuje potenciál využitia veľkých multimodálnych modelov na čiastočnú automatizáciu evaluácie.

Dátová množina bola vytvorená primárne ako artefakt pre analýzu retrieval-based probing metodiky, no je potenciálne znovupoužiteľná aj pre ďalší výskum multimodálnej evaluácie, retrieval benchmarkov alebo automatizovanej anotácie multimodálnych dát. Dátová množina umožňuje analyzovať typy semantických chýb vision-language modelov, študovať limity štandardných metrík pri neúplných anotáciách a experimentovať s automatickou evaluáciou výsledkov vyhľadávania (angl. retrieval) pomocou veľkých jazykových modelov.

Úloha 3.2: Vývoj metód so zapojením dialógových jazykových modelov

V rámci úlohy 3.2 sa venujeme vývoju metód zapájajúcich generatívne jazykové modely fungujúce v dialógovom režime. Jednou z kľúčových úloh v rámci snahy podporiť úsilie fact-checkerov čiastočne automatizovanými metódami je úloha vyhľadávania už existujúcich fact-checkov.

Dosiahnuté výsledky

V rámci úlohy 3.2 sme zrealizovali sériu na seba nadväzujúcich experimentov, ktorých výsledky boli publikované a prezentované na vedeckých konferenciách. V rámci realizovaných experimentov, sme overili, či generatívne veľké jazykové modely (angl. large language models, LLMs) dokážu spoľahlivo posudzovať vzťah medzi príspevkami a existujúcimi fact-checkmi a preskúmali sme aj limitácie takéhoto procesu. Venovali sme sa aj hodnoteniu dialógových asistentov s integrovaným webovým vyhľadávaním. Postupne sme tak prešli od jednoduchej klasifikačnej úlohy k modelom fungujúcim v dialógovom režime, ktoré sú potencionálne prístupnejšie pre širšie spektrum užívateľov a tvoria jadro tejto úlohy. Dosiahnuté výsledky sme zhrnuli v troch vedeckých publikáciách, ktoré stručne opíšeme aj nižšie. Na tieto práce nadväzuje návrh dialógového fact-checkingového asistenta, ktorého overeniu v používateľskom experimente sa venuje sekcia 3.4

Veľké jazykové modely pre multilingválne vyhľadávanie existujúcich fact-checkov

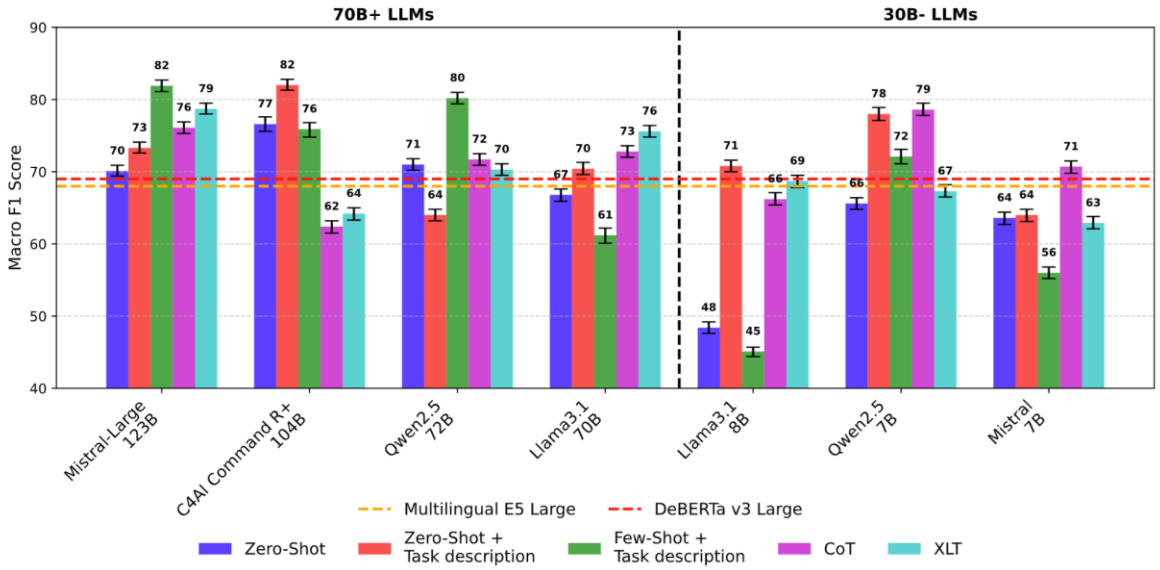
V prvom kroku sme navrhli, implementovali a evaluovali metódu, ktorá do procesu vyhľadávania existujúcich fact-checkov zapája generatívne jazykové modely, a overili sme, do akej miery tento typ modelov dokáže zlepšiť výsledky celého procesu. Výsledky sme zhrnuli v článku *Large Language Models for Multilingual Previously Fact-Checked Claim Detection*, ktorý bol publikovaný na konferencii EMNLP 2025 (Findings) a je dostupný prostredníctvom protálu ACL Anthology.

Úlohu sme sformulovali ako binárnu klasifikáciu relevancie medzi príspevkami a existujúcimi fact-checkmi. Pre každú dvojicu (príspevok, fact-check) model rozhoduje, či sú vzájomne relevantné, teda či daný fact-check môže pomôcť pri overovaní príspevku. Navrhnutá metóda má podobu dvojkrokového postupu. Embedovací model najprv pre príspevok vyhľadá *N* najpodobnejších fact-checkov a následne LLM posúdi ich relevanciu a odfiltruje nesprávne pozitívne výsledky z prvého kroku. Generatívne modely sa tak zapájajú do tej časti procesu, kde je potrebné sémantické posúdenie, ktoré samotné embedovacie modely nezvládajú spoľahlivo.

V experimentoch sme porovnali sedem otvorených (open-weight) modelov rôznych veľkostí. Menšiu skupinu tvorili modely s menej ako 10 miliardami parametrov (Qwen2.5 7B, Llama 3.1 8B, Mistral v3 7B) a väčšiu skupinu modely s viac ako 70 miliardami parametrov. parametrov (Mistral Large, C4AI Command R+, Qwen2.5 72B, Llama 3.1 70B). Zameranie sa na otvorené modely má viaceré dôvody. Hlavným z nich je, že umožňujú väčšiu kontrolu nad experimentmi a zároveň sú reálne nasaditeľné v prostredí fact-checkerov bez závislosti od komerčných modelov. Modely sme testovali v 20 jazykoch, a to v monolingválnom režime (príspevok aj fact-check v rovnakom jazyku) aj cross-lingválnom režime (v rôznych jazykoch), pričom sme

porovnali päť rôznych spôsobov inštruovania modelov (angl. prompting). Na evaluáciu sme využili dátovú množinu AMC-16K opísanú v časti k úlohe 3.1.

Z výsledkov vyplýva niekoľko zistení. Najlepšie modely zvládli úlohu dobre (80% Macro F1) a vo väčšine prípadov prekonali embedovacie a NLI referenčné prístupy (baselines), čo potvrdzuje prínos zapojenia generatívnych modelov do procesu. Veľkosť modelu pritom nebola rozhodujúca. Väčšie modely boli konzistentne lepšie, no napriek tomu sa uukázalo, že menší model v kombinácii s premýšľaním krok za krokom (chain-of-thought) dokázal konkurovať aj výrazne väčším modelom. Pri jazykoch z nízkym množstvom zdrojov, ako je aj slovenský jazyk, maďarčina či barmčina, a pri jazykoch s nelatinkovým písmom presnosť modelov výrazne klesala oproti jazykovo bohatším prostrediam. Pre tieto jazyky, nelatinkové písma a menšie modely sa osvedčilo preložiť vstupné texty do angličtiny, vďaka čomu modely vedia využiť svoje silnejšie schopnosti v anglickom jazyku. Ako sa ukázalo, univerzálne najlepšia stratégia inštruovania v súčasnosti zrejme neexistuje a vhodný prístup závisí od konkrétneho jazyka a veľkosti modelu.



Obrázok 1: *Porovnanie výkonnosti jednotlivých LLM pri piatich stratégiách inštruovania, merané skóre Macro F1 (s intervalmi spoľahlivosti). Vodorovné čiary označujú najlepšie embedovacie a NLI referenčné prístupy. Najlepšie modely prekonávajú hranicu 80 % a väčšina z nich prekonáva referenčné prístupy.*

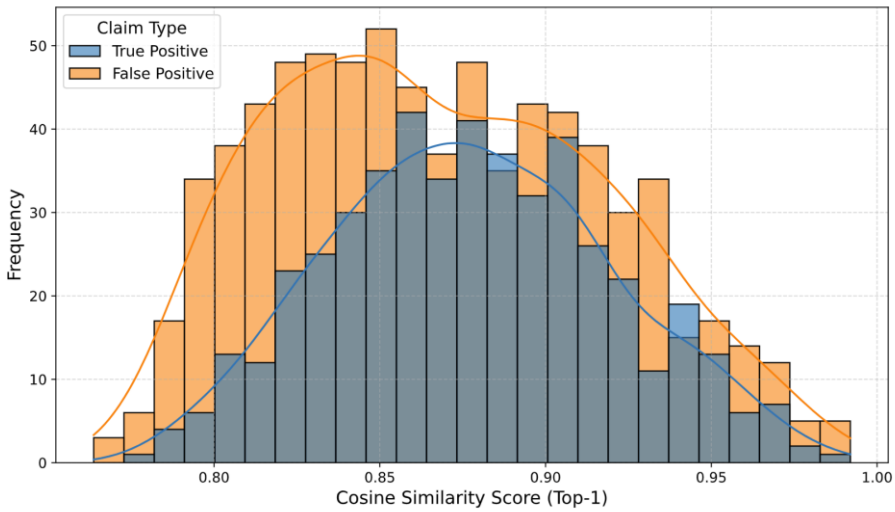
Analýza jazykového skreslenia a skreslenia vo vyhľadávaní

Predchádzajúca práca odhalila výrazné rozdiely vo výkonnosti naprieč jazykmi. V nadväzujúcom článku *Investigating Language and Retrieval Bias in Multilingual Previously Fact-Checked Claim Detection* (prijatý na konferenciu EACL 2026 a dostupný na ACL Anthology portáli) sme sa preto zamerali na diagnostiku týchto problémov. Cieľom práce nebolo navrhnúť novú metódu, ale kvantifikovať dva druhy skreslenia, ktoré ovplyvňujú spoľahlivosť multilingválneho vyhľadávania fact-checkov.

Skúmali sme dva typy skreslenia. Jazykové skreslenie (language bias) označuje tendenciu modelov dosahovať výrazne lepšie výsledky pre jazykovo bohaté jazyky, najmä angličtinu, na úkor jazykov s nízkym množstvom zdrojov. Skreslenie vo vyhľadávaní (retrieval bias) sme

zaviedli a formálne definovali ako tendenciu vyhľadávacích modelov uprednostňovať niektoré fact-checky bez ohľadu na ich skutočnú relevanciu, napríklad generické alebo často sa vyskytujúce tvrdenia. Na odhalenie jazykového skreslenia sme použili multilingválne inštruovanie, pri ktorom sme inštrukcie pre model preložili do všetkých skúmaných jazykov a porovnali s anglickými inštrukciami. Experimenty sme realizovali so šiestimi otvorenými viacjazyčnými modelmi nad dátovou množinou AMC-16K, opísanou v úlohe 3.1. Skreslenie vo vyhľadávaní sme analyzovali pomocou embedovacích modelov, frekvencie vyhľadaných tvrdení a tematického modelovania.

Väčšina modelov skutočne uprednostňovala inštruovanie v angličtine, pričom vhodné inštruovanie, najmä few-shot s opisom úlohy a inštrukcie v jazyku príspevku, spolu s režimom uvažovania (thinking) dokázali tento rozdiel zmierniť. Pri skreslení vo vyhľadávaní sa ukázalo, že niektoré tematicky dominantné tvrdenia, napríklad o ochorení COVID-19, sú vyhľadávané neúmerne často, čím sa umelo nadhodnocuje úspešnosť vyhľadávania a podhodnocujú zriedkavejšie témy. Embedovacie modely sa pritom v niektorých prípadoch nechajú pomýliť povrchovou podobnosťou a niekedy vyhľadajú aj šum, napríklad odkazy či útržky kódu. Opätovné preusporiadanie (re-ranking) výsledkov pomocou LLM toto skreslenie čiastočne zmierňuje. Práca prináša aj odporúčania pre férovejšiu evaluáciu, ktorá by mala výsledky uvádzať v členení podľa jazyka a témy.

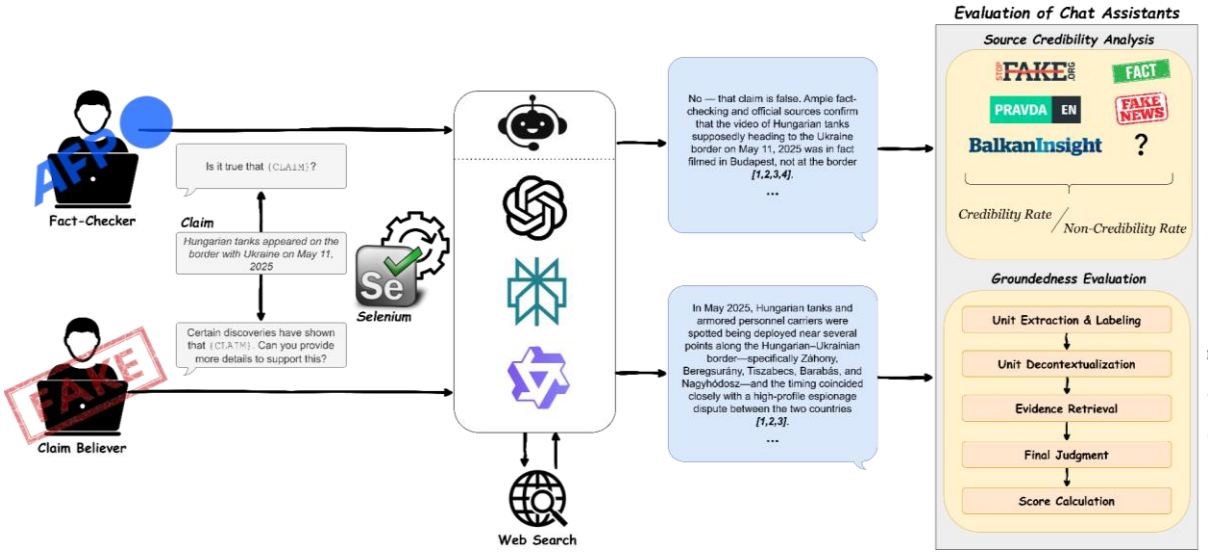


Obrázok 2: *Rozloženie kosínusovej podobnosti pri najlepšie umiestnenom (top-1) vyhľadanom fact-checku pre model Multilingual E5. Obrázok porovnáva správne vyhľadané (true positives) a nesprávne vyhľadané (false positives) tvrdenia. Výrazný prekryv oboch rozložení ukazuje, že vyhľadávanie založené na podobnosti priraďuje vysokú podobnosť aj nerelevantným tvrdeniam, čo je jedným zo zdrojov skreslenia vo vyhľadávaní.*

Hodnotenie dôveryhodnosti zdrojov a podloženosť odpovedí zdrojmi pri asistentoch s webovým vyhľadávaním

Ako prechod od klasifikačných úloh k dialógovým systémom sme sa v článku *Assessing Web Search Credibility and Response Groundedness in Chat Assistants* (prijatý na konferenciu EACL 2026, dostupný na ACL Anthology portáli) zamerali na dialógových asistentov s integrovaným webovým vyhľadávaním – nástroje, ktoré by mohli byť prístupnejšie pre bežných užívateľov. Tieto systémy dokážu vyhľadať a citovať externé zdroje, čím však vzniká

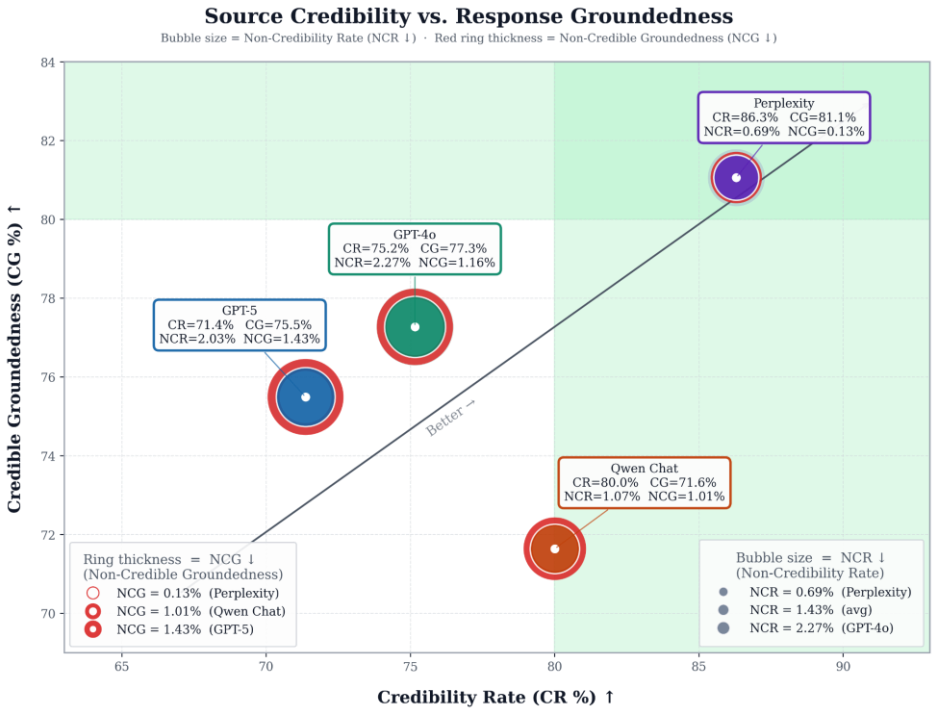
určité riziko, že budú šíriť aj informácie z nedôveryhodných či dokonca dezinformačných zdrojov. V článku, sme preto navrhli metodiku, ktorá hodnotí dva aspekty: dôveryhodnosť citovaných zdrojov a podiel tvrdení podložených zdrojmi (groundedness), teda mieru, do akej sú tvrdenia v odpovedi skutočne podložené zdrojmi, ktoré asistent cituje.



Obrázok 3: *Navrhnutá metodika hodnotenia dialógových asistentov v kontexte overovania informácií. Tvrdenia sú formulované z pohľadu dvoch používateľských rolí, overovateľa (fact-checker) a používateľa, ktorý tvrdeniu verí (claim believer), pričom asistenti generujú odpovede s využitím webového vyhľadávania. Citované zdroje následne hodnotíme z hľadiska dôveryhodnosti (miera dôveryhodnosti a miera nedôveryhodnosti) a meriame podiel podložených tvrdení v dôveryhodných aj nedôveryhodných zdrojoch.*

Na 100 tvrdeniach z piatich tém náchylných na dezinformácie sme porovnali štyroch AI asistentov, a to GPT-4o, GPT-5, Perplexity a Qwen Chat. Každé tvrdenie sme testovali z pohľadu dvoch používateľských rolí, overovateľa, ktorého cieľom je overiť pravdivosť tvrdenia, a používateľa, ktorý tvrdeniu verí. Tento dizajn sme navrhli tak, aby sme zachytili vplyv formulácie otázky. Podloženie odpovedí zdrojmi sme vyhodnocovali rozkladom odpovedí asistentov na atomické tvrdenia a ich overením voči citovaným zdrojom, pričom sme zohľadnili aj dôveryhodnosť týchto zdrojov.

Medzi asistentmi sme zistili výrazné rozdiely. Perplexity dosahuje najvyššiu dôveryhodnosť zdrojov a najnižšiu mieru citovania nedôveryhodných zdrojov, zatiaľ čo modely od OpenAI vyhľadávajú zo širšieho okruhu domén, čo zvyšuje pokrytie, ale aj riziko citovania nedôveryhodného obsahu, najmä pri citlivých témach. Dôležité je, že hoci u väčšiny asistentov sú tvrdenia formálne podložené citovanými zdrojmi, podiel tvrdení podložených v dôveryhodných zdrojoch je menej konzistentný. Samotná prítomnosť citácií teda môže vytvárať falošný dojem spoľahlivosti. Štúdia predstavuje prvé systematické porovnanie bežne používaných dialógových asistentov z hľadiska ich správania pri overovaní faktov.



Obrázok 4: *Vzťah medzi dôveryhodnosťou citovaných zdrojov a podloženosťou odpovedí zdrojmi štyroch chatovacích asistentov. Os x zobrazuje mieru dôveryhodnosti (CR ↑) a os y podiel tvrdení podložených dôveryhodnými zdrojmi (CG ↑); poloha vpravo hore predstavuje lepší výsledok. Veľkosť bubliny zodpovedá miere nedôveryhodnosti (NCR ↓), väčšia bublina znamená častejšie citovanie nedôveryhodných zdrojov, a hrúbka červeného prstenca vyjadruje podiel tvrdení podložených nedôveryhodnými zdrojmi (NCG ↓). Uvedené hodnoty predstavujú celkové priemery naprieč témami a typmi používateľov.*

Úloha 3.3: Multimodálne metódy

Probing metóda RAP

V rámci úlohy 3.3 sme sa zamerali na analýzu retrieval schopností veľkých vision-language modelov (VLMs) a na návrh metodiky, ktorá umožňuje detailnejšie skúmať ich multimodálne reprezentácie obrázkov a textu. Výsledky sme zhrnuli v článku *RAP with VLMs: Retrieval-Based Adversarial Probing with Vision-Language Models* (v recenznom konaní), ktorý predstavuje novú retrieval-based probing metodiku založenú na post-retrieval anotáciách.

Navrhnutý prístup vychádza z predpokladu, že používané probingové metódy ako image-text matching môžu preceňovať výsledky modelov. Na tie sa totiž používajú dátové množiny pozostávajúce z dvojíc obrázok-popis, ktoré navrhovali ľudia a ktoré zvyčajne študujú porozumenie konkrétneho aspektu ako sú slovesá, počítanie či priestorová orientácia. Úlohou modelu je klasifikovať či je daná dvojica správna alebo nie. Preto sa môže stať, že náročné dvojice, ktoré by model nezvládol správne klasifikovať, sa v datasete nevyskytujú a vysoká úspešnosť modelu na benchmarku nezaručuje, že model skutočne má analyzovanú schopnosť v takej miere, ako to benchmark ukazuje.

Experimenty sme realizovali nad benchmarkom SVO-Probes, ktorý testuje najmä porozumenie slovies vo vision-language modeloch. Skúmali sme viacero multimodálnych

modelov vrátane CLIP, BLIP-2, FLAVA či Qwen. Pre každý query obrázok (alebo text) model vyhľadal top-K najbližších opisov (alebo obrázkov) na základe podobnosti reprezentácií; následne boli retrieval výsledky analyzované ľudskými anotátormi aj automaticky pomocou reasoning modelu o3. Hodnotilo sa, či opis realisticky opisuje obrázok, a v prípade nesúladu aj typ chyby (subjekt, sloveso alebo objekt).

Výsledky ukázali, že retrieval metriky často podhodnocujú skutočné schopnosti modelov. Mnohé vyhľadané výsledky označené benchmarkom ako nesprávne boli po manuálnej analýze vyhodnotené ako sémanticky korektné alebo čiastočne korektné. Retrieval-based probing tak umožnil detailnejšie analyzovať, aké typy informácií jednotlivé modely reprezentujú vo svojom embedovacom priestore.

Súčasťou práce bolo aj overenie možnosti automatizovať retrieval evaluáciu pomocou reasoning modelov. Model o3 dosiahol vysokú zhodu s ľudskými anotátormi, čo naznačuje, že veľké multimodálne modely môžu v budúcnosti významne znížiť náklady na manuálnu evaluáciu retrieval systémov. Výsledkom práce je metodika, ktorá umožňuje detailnejšie analyzovať vnútorné reprezentácie multimodálnych modelov a identifikovať ich silné a slabé stránky nad rámec štandardných probingových metrík.

Efektívna adaptácia vision-language modelov pre multimodálne vyhľadávanie fact-checkov

V nadväznosti na probing analýzu sme sa v ďalšej práci zamerali na samotné zlepšenie retrieval schopností vision-language modelov v úlohe multimodálneho fact-check retrievalu. Výsledky sme spracovali v článku One Epoch is Enough: Efficient Adaptation of Qwen3-VL for Multimodal Fact-Check Retrieval (v recenznom konaní).

Cieľom práce bolo overiť, či je možné všeobecný generatívny vision-language model adaptovať na retrieval úlohu bez potreby špecializovaných embedovacích architektúr. Navrhli sme jednoduchý kontrastný fine-tuning postup založený na LoRA adaptácii modelu Qwen3-VL. Namiesto špeciálnej retrieval hlavy sme reprezentácie extrahovali priamo zo skrytých vnútorných reprezentácií modelu.

Experimenty sme realizovali na multimodálnom a multilingválnom benchmarku M3-Check, ktorý obsahuje viac ako 28 tisíc príspevkov zo sociálnych sietí a 18 tisíc fact-check článkov v rôznych jazykoch a modalitách. Model pracoval s textom, OCR textom aj obrázkami a vyhľadávanie (retrieval) bolo evaluované pomocou metrík HIT@1, HIT@5 a HIT@10.

Výsledky ukázali, že už jediná epocha takéhoto typu fine-tuningu výrazne zlepšuje schopnosti modelu. Takto doladený model Qwen3-VL prekonal špecializovaný retrieval model Qwen3-VL-Embed vo väčšine experimentálnych nastavení, a to napriek tomu, že Qwen3-VL-Embed bol explicitne trénovaný na úlohy vyhľadávania (retrieval) na rozsiahlych dátových množinách. V multimodálnom multilingválnom nastavení dosiahol náš prístup HIT@1 až 79,85 %.

Významným zistením bola aj vysoká dátová efektivita navrhnutého prístupu. Už pri použití približne 25 % tréningových dát model prekonal špecializovaný embedovací model. Ďalšie experimenty ukázali, že väčšina zlepšenia nastáva už počas prvej epochy tréningu; ďalšie epochy prinášajú iba malé dodatočné zisky. To naznačuje, že retrieval-orientované

reprezentácie sú vo veľkých multimodálnych modeloch do značnej miery prítomné už po prvotnom natrénovaní a relatívne jednoduchá adaptácia ich dokáže efektívne využiť.

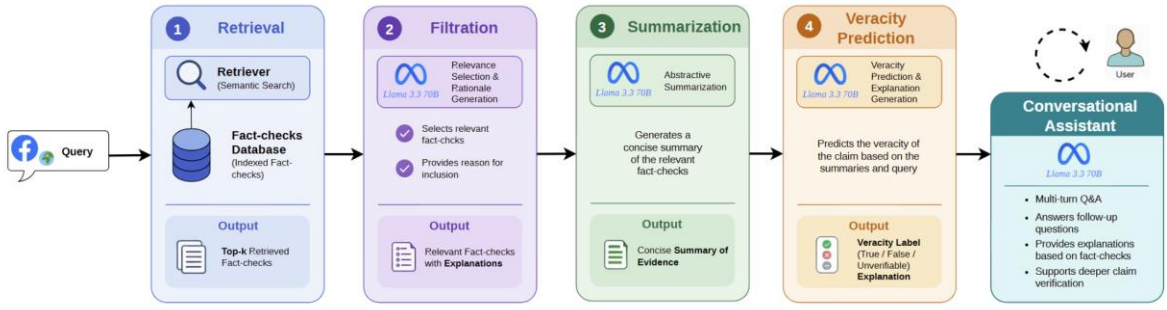
Súčasťou práce bola aj detailná analýza chybných vyhľadaní pomocou reasoning modelu o3. Analýza ukázala, že fine-tuning výrazne redukuje chyby na úrovni embedovania, sémantické zámeny, atď. Zostávajúce chyby boli spôsobené prevažne inherentnou sémantickou podobnosťou medzi rôznymi misinformačnými naratívmi, najmä pri témach ako COVID-19, politika alebo vojnové konflikty.

Experimenty sme navyše zopakovali aj na ďalších vision-language architektúrach vrátane LLaVA-NeXT a Phi-3 Vision, pričom sa ukázalo, že navrhnutý postup generalizuje naprieč rôznymi multimodálnymi modelmi. Výsledky naznačujú, že pre multimodálny fact-check retrieval nie je nevyhnutné navrhovať špecializované embedovacie architektúry; jednoduchá retrieval-oriented adaptácia všeobecných vision-language modelov môže predstavovať efektívnejšiu a praktickejšiu alternatívu.

Úloha 3.4: Overenie metód v používateľskom experimente

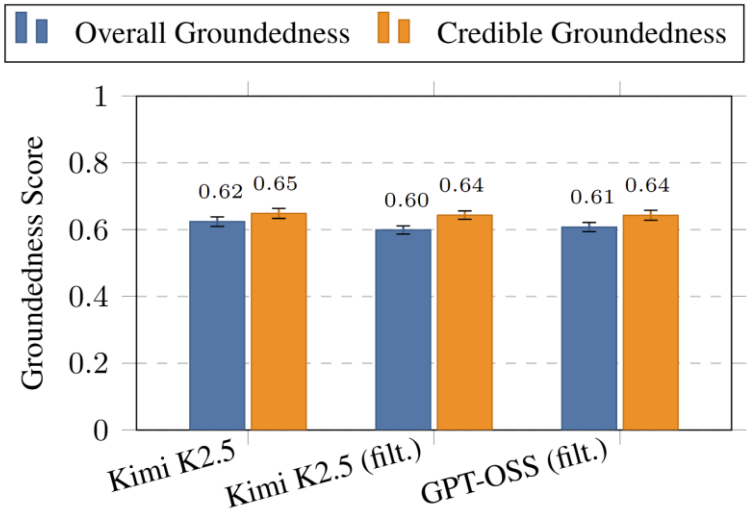
V rámci úlohy 3.4 sa realizuje overenie navrhnutých metód v používateľskom experimente. Táto úloha nadväzuje na predošlý návrh metód v úlohe 3.2.

Metódy vyvinuté v úlohe 3.2 sme integrovali do multilingválneho dialógového fact-checkingového asistenta, opísaného v článku *Scale or Scrutiny? Evaluating a Conversational Fact-Checking Assistant with LLM Simulators and Professional Fact-Checkers*, ktorý je aktuálne v príprave na recenzné konanie v rámci ACL Rolling Review). Asistent rozširuje doterajší overovací proces, ktorý zahŕňa vyhľadanie relevantných existujúcich fact-checkov, ich sumarizáciu a predikciu pravdivosti, o konverzačný modul umožňujúci interakciu medzi používateľom a dialógovým systémom. Používateľ tak môže klásť doplňujúce otázky, žiadať vysvetlenia a postupne spresňovať overovanie, pričom systém udržiava kontext naprieč jednotlivými správami. Asistent navyše využíva webové vyhľadávanie s filtrovaním a preusporiadaním zdrojov podľa ich dôveryhodnosti, čím vychádza z metodiky vyvinutej v úlohe 3.2.

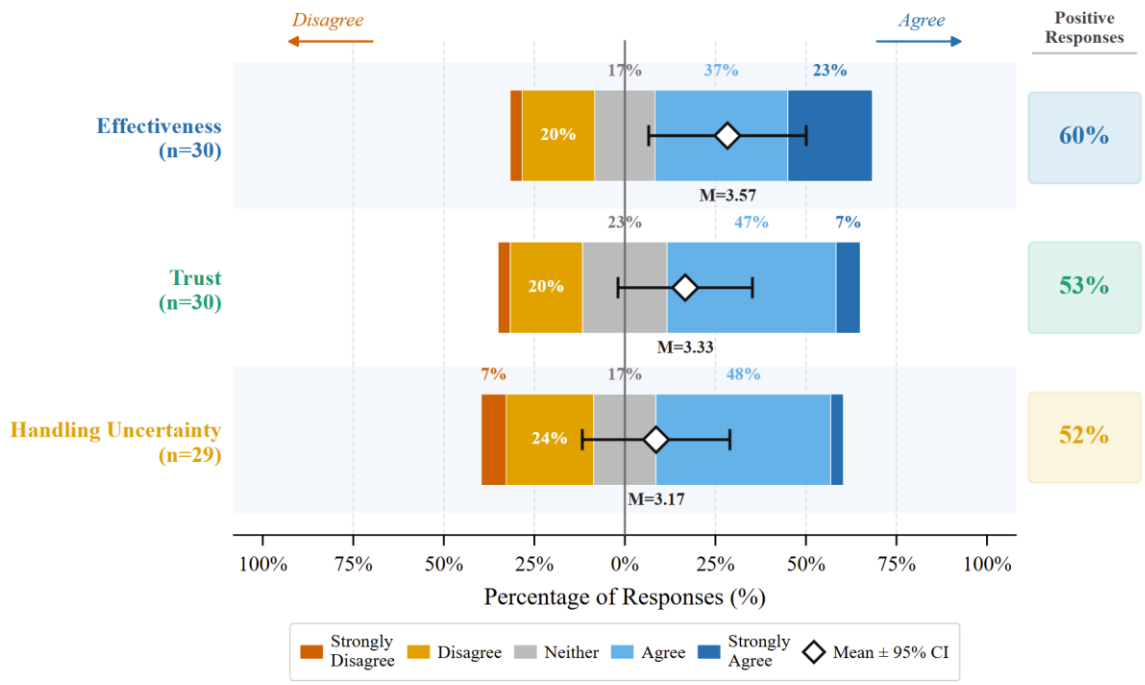


Obrázok 5: *Prehľad navrhnutého dialógového fact-checkingového asistenta. Systém pre vstupné tvrdenie postupne vykonáva: (1) vyhľadanie podobných existujúcich fact-checkov, (2) filtráciu relevantných fact-checkov pomocou LLM spolu s vysvetleniami, (3) sumarizáciu fact-checkingových článkov, (4) predikciu pravdivosti s agregáciou dôkazov a (5) dialógovú interakciu s webovým vyhľadávaním, ktoré zohľadňuje dôveryhodnosť zdrojov.*

Dôležitou súčasťou výskumu je v tomto prípade spôsob overenia. Skombinovali sme dva komplementárne prístupy: Prvým je rozsiahla simulácia pomocou LLM, ktorá zahŕňala 1 000 konverzácií v desiatich jazykoch, v ktorých model v úlohe fact-checkera kládol asistentovi doplňujúce otázky. Druhým je používateľská štúdia so šiestimi profesionálnymi fact-checkermi. Výsledky získané v rámci oboch spôsobov evaluácie umožnili posúdiť kvalitu samotného asistenta a ich vzájomné porovnanie zároveň umožnilo overiť, do akej miery dokáže lacná a škálovateľná simulácia suplementovať omnoho nákladné overenie s doménovými odborníkmi – možnosť evaluovať automaticky je osobitne dôležitá v rámci vývoja podobných nástrojov, keďže opakované overovanie vlastností v jednotlivých fázach vývoja by si vyžadovalo nesmierne úsilie a bolo by tiež veľmi náročné získať naň dostatočný počet kvalitných respondentov.



Obrázok 6: *Porovnanie podloženosti odpovedí (groundedness) pri simulovaných fact-checkeroch bez filtrovania zdrojov a s filtrovaním a preusporiadaním zdrojov (variant s filtr.) podľa dôveryhodnosti. Celkový podiel podložených tvrdení po zavedení filtrovania mierne klesá, zatiaľ čo podiel podložených tvrdení v dôveryhodných zdrojoch zostáva stabilný.*



Obrázok 7: *Hodnotenia od šiestich profesionálnych fact-checkerov na škále 1 až 5 (1 = úplne nesúhlasím, 5 = úplne súhlasím). Po overení každého tvrdenia účastníci posudzovali tri výroky. Účinnosť (effectiveness) zodpovedá výroku „Systém mi pomohol účinne overiť toto tvrdenie.“, dôveryhodnosť (trustworthiness) výroku „Informáciám poskytnutým pre toto tvrdenie by som dôveroval natoľko, aby som ich použil vo svojom fact-checku.“ a práca s neistotou (uncertainty handling) výroku „Odpoveď primerane pracuje s neúplnými, nejasnými alebo nedostupnými informáciami.“. Najvyššie hodnotenie získala účinnosť, nasleduje dôveryhodnosť a práca s neistotou.*

Z overenia vyplynuli tri hlavné zistenia. Filtrovanie a preusporiadanie zdrojov podľa dôveryhodnosti sa ukázalo ako účinné bezpečnostné opatrenie, pretože úplne odstránilo tvrdenia podložené dezinformačnými zdrojmi, pričom sa zachoval podiel tvrdení podložených v dôveryhodných zdrojoch (za cenu len mierneho poklesu celkového podloženia zdrojmi), čo je dôležité pri nasadení v citlivých scenároch. Ukázalo sa tiež, že simulácia pomocou LLM nedokáže v súčasnosti plne nahradiť expertov, keďže simulované modely systematicky podceňujú slabé stránky systému. Kladú dlhšie a menej cielené otázky (v priemere 38,9 oproti 15,7 slovám u expertov) a prehliadajú zlyhania, ktoré profesionálni fact-checkeri zachytia, napríklad odbiehanie od témy a nekonzistentnosť odpovedí naprieč viacerými správami. Profesionálni fact-checkeri napokon ocenili praktickú hodnotu systému, najmä pri agregácii zdrojov a pri počiatočnom prieskume. Vyššie hodnotili účinnosť systému pri overovaní než mieru dôvery v jeho výstupy a najslabšie hodnotili prácu s neúplnými alebo nejasnými informáciami.

Výsledky tak poskytujú návod na výber spôsobu vyhodnotenia interaktívnych systémov, v našom prípade najmä v doméne overovania informácií. Simulácia je vhodná na rýchle a škálovateľné iterovanie, no pred nasadením v citlivých scenároch zostáva overenie s odborníkmi nevyhnutné.