Príloha E:
V3.2 Metódy, časť 2

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**

Metódy, časť 2

Tento dokument opisuje výstup V3.2 projektu DisAI AMPLIFIED zameraný na metodologické prínosy projektu. Výstup V3.2 sa odovzdával v dvoch termínoch – prvá verzia dokumentu bola odovzdaná v 15. mesiaci riešenia projektu (september 2025). Druhá verzia – t. j. tento dokument – sa odovzdáva pri ukončení projektu (v prvej časti správy sme omylom uviedli, že druhá časť bude odovzdaná v 21. mesiaci riešenia projektu – to bol však podľa projektového plánu termín odovzdania pre druhú časť dátových množín, ktoré k tomu termínu aj naozaj boli zverejnené prostredníctvom webovej stránky projektu; metódy teda samozrejme odovzdávame v súlade s pôvodným, schváleným harmonogramom).

Keďže prvá časť správy k V3.2 už pokrývala základnú metódu LLM-asistovaného vyhľadávania existujúcich fact-checkov a pôvodnú verziu retrieval-based probingu pre vision-language modely, tento dokument tieto metódy len zasadí do kontextu a doplní najmä metódy, ktoré vznikli v nasledujúcej fáze riešenia. Výsledky dosiahnuté v rámci experimentov sú opísané v rámci prílohy A.

Rekapitulácia metód z V3.2, časť 1

LLM-asistované multilingválne vyhľadávanie existujúcich fact-checkov

Prvá metodická línia sa týka úlohy previously fact-checked claim detection (PFCD), teda identifikácie už existujúcich fact-checkov relevantných pre zadané tvrdenie/príspevok. Navrhnutý postup je dvojstupňový. V prvom kroku sa pomocou textového embedovacieho modelu vyhľadá množina kandidátskych fact-checkov. V druhom kroku generatívny jazykový model posudzuje, či je každý kandidát voči vstupnému príspevku skutočne relevantný.

- **Vstup:** Príspevok alebo tvrdenie v jednom z podporovaných jazykov a databáza existujúcich fact-checkov.
- **Medzikrok:** Vyhľadanie top-N kandidátov pomocou vektorovej podobnosti.
- **Výstup:** Filtrovaný alebo preusporiadaný zoznam relevantných fact-checkov, prípadne binárne rozhodnutie o relevancii dvojice príspevok – fact-check.
- **Kľúčová metodická otázka:** Ako formulovať inštrukcie pre LLM a ako zohľadniť jazyk, veľkosť modelu a prípadný preklad do angličtiny.

Evaluácia modelov analýzou top k nájdených výsledkov

Druhá pôvodná metodická línia vychádza z kritiky štandardných probing-ových prístupov založených na technike image-text matching-u. Pri tomto klasickom type probingu model o vzájomnej relevancii medzi vopred pripravenými pozitívnymi a negatívnymi dvojicami obrázok – text. Takéto dvojice však nemusia obsahovať prípady, ktoré sú pre konkrétny model najviac mätúce. Navrhnutá alternatíva preto namiesto ručne definovaných párov analyzuje top-k nájdené výsledky, t.j. páry, ktoré samotný model vo vyhľadávacom (retrieval) scenári považuje za najbližšie.

- **Vstup:** Obrázky, textové opisy a multimodálny model schopný mapovať obe modality do spoločného priestoru.

- **Postup:** Pre dopytový (query) obrázok alebo dopytový text sa vyhľadá top-k najbližších položiek z opačnej modality (t. j. hľadajú sa text zodpovedajúce obrázkom alebo obrázky zodpovedajúce textom).
- **Výstup:** Množina správnych aj chybných párov, ktoré umožňujú analyzovať reálne sklony modelu k zámene konceptov.
- **Kľúčová metodická otázka:** Ako korektne evaluovať výsledky v retrieval setup-e, kde je dátová množina spravidla riedko anotovaná.

Metódy doplnené v druhej fáze riešenia

Diagnostika jazykového skreslenia a skreslenia vo vyhľadávaní

Nadväzujúca metodika reaguje na pozorovanie, že výkonnosť PFCD systémov sa (niekedy výrazne) líši naprieč jazykmi a témami. Cieľom metodiky je nemerať len priemernú úspešnosť, ale identifikovať, či systém systematicky nezvýhodňuje určité jazyky, typy tvrdení, často vyhľadávané fact-checky a pod.

Jazykové skreslenie

Pod jazykovým skreslením máme na mysli situáciu, kedy model dosahuje podstatne lepšie výsledky pre jazyky s veľkým množstvom zdrojov (angl. high-resource; napr. angličtina), než pre jazyky s menším množstvom zdrojov (tréningový dát). Metodika skúma, či rozdiely súvisia s jazykom vstupného tvrdenia, jazykom fact-checku, jazykom inštrukcie alebo kombináciou týchto faktorov.

- Inštrukcie pre model sa pripravujú vo viacerých jazykoch a realizuje sa porovnávanie s anglickými inštrukciami.
- Výsledky sa vyhodnocujú oddelene pre jednotlivé jazyky, jazykové rodiny a typ skriptu (písma).
- Metodika umožňuje odlíšiť slabšie porozumenie vstupu od zhoršenia výsledkov v dôsledku nevhodne zvoleného jazyka inštrukcie.

Skreslenie vo vyhľadávaní

Pod skreslením vo vyhľadávaní máme na mysli tendenciu retrieval systému opakovane vyhľadávať určité fact-checky bez ohľadu na ich skutočnú relevanciu. Mohlo by ísť napr. o generické tvrdenia, často sa vyskytujúce naratívy alebo tematicky dominantné položky, ktoré sa v embedovacom priestore nachádzajú blízko mnohých vstupov.

- Analýza frekvencie výskytu fact-checkov medzi top-k výsledkami odhaľuje nadmerne často navracané položky.
- Porovnanie rozdelení kosínusovej podobnosti pre správne a nesprávne výsledky ukazuje, či samotná podobnosť postačuje na odlíšenie relevantných kandidátov.
- Tematické modelovanie resp. tematická klasifikácia pomáha určiť, či systém nezvýhodňuje dominantné témy, napríklad zdravotné alebo politické naratívy.
- LLM re-ranking možno následne použiť ako čiastočnú korekciu týchto typov skreslenia.

Metodika hodnotenia web-search asistentov: dôveryhodnosť zdrojov a podloženosť odpovedí

Ďalší metodický komponent sa venuje problému dialógových asistentov s integrovaným webovým vyhľadávaním. Takéto systémy môžu byť pre používateľov prístupnejšie, ale prinášajú nové riziko: odpoveď môže obsahovať citácie, a predsa sa opierať o nedôveryhodné alebo nedostatočne relevantné zdroje.

Metodika preto hodnotí dva navzájom odlišné aspekty: kvalitu citovaných zdrojov a mieru, do akej sú tvrdenia v odpovedi skutočne podložené citovanými zdrojmi. Dôležité je, že samotná prítomnosť citácií sa nepovažuje za postačujúci dôkaz spoľahlivosti odpovede.

Rámcový postup:

- Výber tvrdení z tém náchylných na dezinformácie.
- Formulácia otázok z dvoch odlišných používateľských rolí: overovateľ (fact-checker) a používateľ, ktorý tvrdeniu verí (keďže modely môžu mať tendenciu pre tieto dva typy užívateľov odpovedať odlišne).
- Získanie odpovedí od asistentov s webovým vyhľadávaním.
- Extrahovanie citovaných zdrojov a hodnotenie ich dôveryhodnosti.
- Rozklad odpovede na atomické tvrdenia.
- Overenie, či sú jednotlivé atomické tvrdenia podložené citovanými zdrojmi.
- Výpočet metrík pre dôveryhodnosť zdrojov a podloženosť odpovedí.

Rozšírená retrieval-based probing metodika RAP

Retrieval-Based Adversarial Probing (RAP) rozvíja pôvodnú top-k evaluáciu VLM. Jadrom metódy je, že model sa netestuje iba na statických kontrastívnych pároch, ale na príkladoch, ktoré sám vyhľadá ako najpodobnejšie. Tým sa získavajú chyby, ktoré sú pre daný model prirodzene najpravdepodobnejšie.

Rozšírenie oproti prvej časti V3.2 spočíva najmä v systematickej post-retrieval anotácii a v automatizovanom hodnotení pomocou reasoningového multimodálneho modelu. V prípade SVO-Probes sa vyhodnocuje nielen správnosť páru obrázok – popis, ale aj typ prípadného nesúladu: subjekt, sloveso alebo objekt.

- Model vygeneruje retrieval výsledky pre text-to-image alebo image-to-text scenár.
- Pre každú query položku sa vyberie top k výsledkov.
- Ľudskí anotátori posudzujú, či text realisticky opisuje obrázok.
- Pri nesúlade sa určuje typ chyby: subjekt, sloveso alebo objekt.
- Reasoningový VLM môže na základe rovnakej schémy vykonať automatickú anotáciu a jeho výstupy je možné porovnať voči ľudským anotáciám.

Efektívna adaptácia vision-language modelov pre multimodálny fact-check retrieval

Ďalšia metodická línia sa nezameriava čisto na evaluáciu multimodálnych modelov, ale na spôsoby, ako ich efektívne adaptovať na úlohu multimodálneho vyhľadávania existujúcich fact-checkov. Východiskom je otázka, či je možné všeobecný generatívny vision-language

model prispôsobiť na retrieval bez návrhu špecializovanej embedovacej architektúry alebo samostatnej retrieval hlavy.

Navrhnutý prístup je založený na kontrastívnom fine-tuning s využitím LoRA adaptéra. Reprezentácie sa extrahujú priamo zo skrytých reprezentácií generatívneho VLM. Tým sa testuje hypotéza, že veľké multimodálne modely už obsahujú užitočné retrieval-orientované reprezentácie, ktoré stačí len efektívne sprístupniť pre retrieval.

- Vytvorenie tréningových dvojíc alebo viacprvkových dávok z multimodálneho fact-check retrieval benchmarku.
- Spracovanie textu, OCR textu a obrázkov v spoločnom VLM.
- Extrakcia reprezentácií zo skrytých stavov modelu bez pridania špecializovanej retrieval hlavy.
- Kontrastívny tréning s LoRA adaptáciou, zameraný na priblíženie relevantných párov a oddialenie nerelevantných párov.
- Evaluácia pomocou retrieval metrík, napríklad HIT@1, HIT@5 a HIT@10.
- Analýza chýb pomocou reasoningového modelu s cieľom rozlíšiť chyby embedovania od inherentnej sémantickej podobnosti rôznych naratívov.

Dialógový fact-checkingový asistent ako integrovaná metodická architektúra

Posledný metodický komponent integruje viacero uvedených metód do konverzačného fact-checkingového asistenta. Na rozdiel od samostatnej klasifikácie relevancie alebo jednorazového vyhľadávania ide o interaktívny systém, v ktorom používateľ môže klásť doplňujúce otázky, žiadať vysvetlenia a postupne spresňovať overovanie tvrdenia a pod.

Rámcový postup:

- Retrieval: Systém vyhľadá kandidátske existujúce fact-checky a relevantné zdroje.
- Filtrácia: LLM vyberie relevantné fact-checky a môže pridať vysvetlenie, prečo sú relevantné.
- Sumarizácia: systém pripraví stručný prehľad dôkazov z fact-checkingových článkov.
- Predikcia pravdivosti: model agreguje dôkazy a navrhne predbežné označenie tvrdenia.
- Dialógová interakcia: používateľ môže klásť podotázky a systém udržiava kontext naprieč správami.
- Webové vyhľadávanie s kontrolou zdrojov: systém filtruje alebo preusporadúva zdroje podľa dôveryhodnosti.

Metodicky dôležitá je aj schéma overenia takéhoto systému. Využíva dva komplementárne režimy: škálovateľnú simuláciu pomocou LLM a menšie, ale kvalitatívne hodnotnejšie overenie s profesionálnymi fact-checkermi. Simulácia je vhodná na rýchle iterovanie počas vývoja, zatiaľ čo expertné hodnotenie zostáva potrebné pred nasadením v citlivých scenároch.

Súhrnná mapa metód

Metóda	**Hlavný problém**	**Vstup**	**Výstup**

LLM-asistované PFCD	vyhľadať už existujúce fact-checky relevantné pre nové tvrdenie	príspevok, databáza fact-checkov, kandidáti z embedovacieho retrievalu	filtrovaný zoznam relevantných fact-checkov, klasifikácia relevancie, …
Diagnostika jazykového skreslenia	zistiť, či systém funguje nerovnomerne naprieč jazykmi	viacjazyčné vstupy a viacjazyčné inštrukcie	výsledky členené podľa jazyka a odporúčania pre prompting
Diagnostika retrieval skreslenia	zistiť, či retrieval nadmerne preferuje určité fact-checky alebo témy	top-k výsledky, podobnosti, frekvencie a témy	identifikácia dominantných kandidátov a tematických nerovnováh
Hodnotenie web-search asistentov	posúdiť, či odpovede stoja na dôveryhodných a relevantných zdrojoch	odpovede asistentov, citácie a atomické tvrdenia	metriky dôveryhodnosti zdrojov a podloženosti odpovedí
RAP	testovať VLM na modelovo špecifických mätúcich prípadoch	obrázky, texty, top-k retrieval výsledky	post-retrieval anotácie, typy chýb, robustnejšia evaluácia VLM
Efektívna adaptácia VLM	zlepšiť multimodálny fact-check retrieval bez špecializovanej architektúry	multimodálne tréningové páry, generatívny VLM	retrieval použiteľné reprezentácie po kontrastívnom LoRA fine-tuningu
Dialógový fact-checking asistent	umožniť interaktívne overovanie tvrdení	tvrdenie, fact-checky, webové zdroje, dialógový kontext	konverzačná podpora overovania, vysvetlenia a sumarizácia dôkazov