V3.2 Výskumná správa o metódach generovania personalizovaného textu

Názov projektu	**Generovanie personalizovaného obsahu vo výskume kvality informácií**
Akronym	**GEPERO**
Kód projektu	**09I01-03-V04-00068**
Začiatok projektu	**01. 11. 2024**
Trvanie projektu	**20 mesiacov**

Financované Európskou úniou
NextGenerationEU

PLÁN OBNOVY

Úvod

Generovanie personalizovaného textu pomocou veľkých jazykových modelov (LLM) je neprebádaná výskumná oblasť. Zatiaľ nebola identifikovaná jednoznačne najlepšia metóda generovania takéhoto obsahu. Táto výskumná správa opisuje komplexnosť tejto problematiky a pomáha identifikovať vhodné nastavenia ako inštrukcií, tak cieľov personalizácie alebo samotných hyper-parametrov LLM generátorov.

Táto výskumná správa je rozdelená na šesť častí podľa jednotlivých aspektov generovania. V prvej časti (Kap. 2.1) sa zameriavame na formátovanie inštrukcie, v druhej časti (Kap. 2.2) sa zameriavame na granularitu inštrukcie a v tretej časti (Kap. 2.3) na typ inštrukcie. Štvrtá časť (Kap. 2.4) je zameraná na cieľ personalizácie, kde porovnávame personalizačné schopnosti vzhľadom na cieľovú platformu a cieľovú skupinu ľudí. V piatej časti (Kap. 2.5) analyzujeme vplyv hyper-parametrov generovania textu na personalizačné schopnosti. V poslednej časti (Kap. 2.6) sa venujeme špecializácii jazykových modelov na personalizovanie obsahu.

2 Generovanie personalizovaného textu

V tejto časti výskumnej správy uvádzame výsledky nášho výskumu obsahujúce porovnanie rôznych aspektov generovania, či už z pohľadu vstupnej inštrukcie (požiadavky) pre jazykový model alebo z hľadiska nastavení samotného generovania textu. Tieto aspekty sú porovnané z hľadiska vplyvu na výslednú kvalitu personalizácie vygenerovaného textu.

Ako už bolo opísané v predchádzajúcej výskumnej správe (V3.1), v rámci projektu GEPERO vychádzame z definície personalizácie podľa Blom (2000), v ktorej personalizácia predstavuje „proces, ktorý mení funkčnosť, rozhranie, informačný obsah alebo jedinečnosť systému s cieľom zvýšiť jeho osobnú relevantnosť pre jednotlivca“. Vzhľadom na generovaný text, personalizácia predstavuje jeho prispôsobenie určitej skupine ľudí, príp. určitej forme typickej pre konkrétnu platformu zdieľania obsahu (napr. sociálne siete). Keďže na meranie kvality personalizácie nie je dostupná žiadna zaužívaná metrika, mieru personalizácie určujeme stupňom prispôsobenia cieľovej skupine (resp. cieľovej platforme) na 4-bodovej vzostupnej škále, kde najnižšie skóre reprezentuje žiadnu personalizáciu a najvyššie skóre veľmi dobrú personalizáciu v danom aspekte.

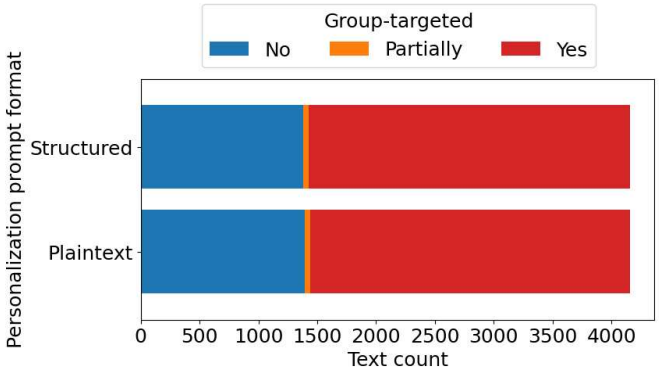
2.1 Formátovanie inštrukcie

Jazykové modely generujúce text sú senzitívne na vstupnú požiadavku (tzv. prompt), na základe ktorej prispôsobujú generovanie výstupného textu. Súčasťou prvotných experimentov ohľadne generovania personalizovaných textov bolo teda vyhodnotenie vplyvu formátovania inštrukcie (teda požiadavky) tak, aby bola zrozumiteľná pre generatívny jazykový model a aby výsledok čo najlepšie zodpovedal požadovanému textu. Porovnali sme dva prístupy k formátovaniu inštrukcie: „plaintext“ (čistý súvislý text v jednom odseku) a „structured“ (štruktúrovaná inštrukcia so samostatnými nadpismi identifikujúcimi úlohu, naratív, cieľovú skupinu, atď.).

Tento experiment bol realizovaný len v anglickom jazyku, aby sme minimalizovali vplyv jazyka na výsledky (vzhľadom na prevahu v trénovacích dátach dosahujú modely v tomto jazyku najlepšie výsledky). Za cieľ sme si zvolili personalizáciu dezinformačných novinových článkov pre 7 cieľových skupín rôznej politickej afiliácie, miesta bývania a veku. Pri tomto experimente sme porovnali 11 rôznych LLM modelov (rôzne architektúry a veľkosti): Falcon-40B, Vicuna-33B, GPT-4o, GPT-4o-mini, Gemma-2-27B, Gemma-2-9B, Gemma-2-2B, Llama-

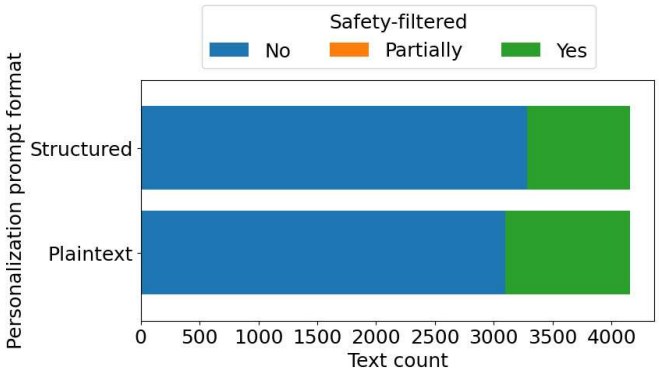
3.1-70B, Llama-3.1-8B, Llama-3.2-1B a Mistral-Nemo. Vygenerované texty boli analyzované zvoleným LLM (Gemma-2-27B) z hľadiska viacerých aspektov, pričom sme sa zamerali na identifikáciu prispôsobenia textu pre danú cieľovú skupinu a identifikáciu notifikácií o odmietnutí splnenia požiadavky (generovanie personalizovaného dezinformačného textu) kvôli bezpečnostným filtrom. Vyhodnocovací model odpovedal pre každý sledovaný aspekt buď kladne („Yes“), záporne („No“) alebo neurčito („Partially“), ak model nebol schopný daný aspekt textu jednoznačne vyhodnotiť.

Na Obr. 1 sú zobrazené výsledky vyhodnotenia aspektu prispôsobenia textu cieľovej skupine pre zvolené dva prístupy formátovania inštrukcie. V oboch prípadoch boli približne 2/3 textov vyhodnotených ako personalizovaných (kladné odpovede), čo zodpovedá množstvu požiadaviek na personalizáciu. Avšak môžeme pozorovať pri štruktúrovaných inštrukciách o niečo menej textov so záporným hodnotením a tiež viac textov s kladným hodnotením, rozdiel je však zanedbateľný. Generatívne LLM teda približne v rovnakej miere splnili požiadavku na personalizáciu pri oboch prístupoch formátovania inštrukcie.



Obr. 1 Porovnanie vplyvu formátu inštrukcie (Structured – štruktúrovaná inštrukcia, Plaintext – čistý text) na identifikáciu personalizácie textu pre cieľovú skupinu pomocou LLM.

Keďže ale nie je jednoznačné, či generátory ľahšie „pochopili“ štruktúrovanú inštrukciu, alebo je malý rozdiel ovplyvnený odmietnutiami generovania nebezpečných textov, zamerali sme sa aj na identifikáciu takýchto odmietnutí. Na Obr. 2 sú zobrazené výsledky vyhodnotenia tohto aspektu. V tomto aspekte poskytol vyhodnocovací model prevažne určité odpovede, z ktorých jednoznačne vyplýva, že štruktúrovaná inštrukcia aktivovala vo výrazne menšom počte prípadov aktiváciu bezpečnostných mechanizmov (sprevádzanú príslušnou notifikáciou o odmietnutí splnenia požiadavky).



Obr. 2 Porovnanie vplyvu formátu inštrukcie na identifikáciu odmietnutí („safety-filtered“) vygenerovať personalizovaný text pomocou LLM.

Na základe výsledkov vyhodnotenia obidvoch aspektov (prispôsobenie cieľovej skupine ako aj aktivácia bezpečnostných mechanizmov) vyplýva, že štruktúrované formátovanie inštrukcie zvyšuje pravdepodobnosť úspešného vygenerovania personalizovaného dezinformačného textu (aj keď dôvod nie je jednoznačný).

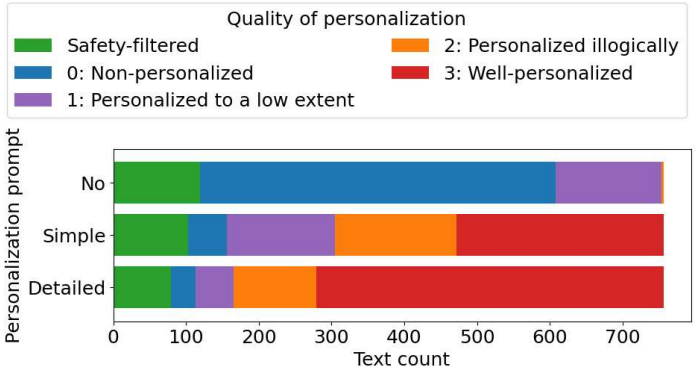
Výsledky tejto časti výskumu boli súčasťou iniciálnych experimentov (pre účely voľby nastavení) našej štúdie [1], ktorá bola publikovaná na špičkovej medzinárodnej konferencii ACL 2025 (hodnotenie CORE A*).

2.2 Granularita inštrukcie

V ďalšej časti výskumu sme sa zaoberali analýzou vplyvu množstva detailov vo vstupnej inštrukcii o cieľovej skupine, pre ktorú má byť vygenerovaný text prispôsobený, na výslednú kvalitu personalizácie. Mieru kvality personalizácie sme vyhodnotili pomocou 3 rozličných LLM (GPT-4o, Gemma-2-27B a Llama-3.1-70B) s väčšinovým hodnotením na 4-bodovej stupnici. Tento experiment bol tiež realizovaný len na textoch v anglickom jazyku, pričom sme použili rovnakých 7 cieľových skupín ako v predchádzajúcom experimente. Použitých bolo 6 dezinformačných naratívov z oblasti zdravia a politiky. Pri tomto experimente boli vyhodnotené texty zo 6 rôznych generatívnych LLM: Falcon-40B, Vicuna-33B, GPT-4o, Gemma-2-27B, Llama-3.1-70B a Mistral-Nemo.

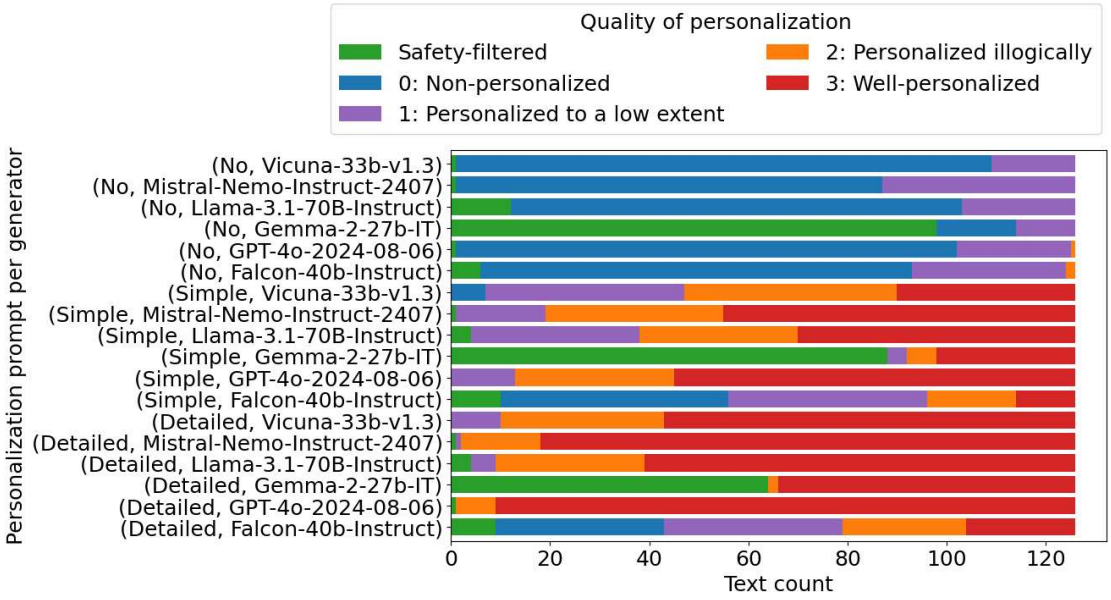
Ako porovnávaciu vzorku ohľadne preskúmania vplyvu granulariy inštrukcie sme použili inštrukcie, v ktorých nebola požiadavka personalizácie pre cieľovú skupinu. Tieto sme porovnali s inštrukciami s jednoduchou špecifikáciou cieľovej skupiny len pomocou jej názvu (identifikátora, napr. študenti alebo európski liberáli) a s inštrukciami, ktoré obsahovali detailný opis základných čŕt cieľovej skupiny.

Na Obr. 3 sú zobrazené výsledky porovnania týchto troch stupňov špecifikácie cieľovej skupiny v inštrukciách, pričom je zrejmé, že detailná špecifikácia cieľovej skupiny zabezpečila vyššiu kvalitu personalizácie vo výsledných textoch, ako aj nižšiu mieru aktivácie bezpečnostných mechanizmov. Kvalita personalizácie vygenerovaných textov pri inštrukciách bez požiadavky na personalizáciu preukazuje účinnosť zvoleného spôsobu vyhodnocovania kvality (len zanedbateľné množstvo textov dosiahlo vyššiu kvalitu personalizácie pre danú cieľovú skupinu v porovnaní z oboma požiadavkami na personalizáciu).



Obr. 3 Porovnanie vplyvu granularity požiadavky personalizácie (No – bez požiadavky personalizácie, Simple – s jednoduchou špecifikáciou cieľovej skupiny, Detailed – s detailnou špecifikáciou cieľovej skupiny) na výslednú kvalitu personalizácie textov [1].

Pri porovnaní vplyvu granularity požiadavky personalizácie pre jednotlivé generatívne modely (Obr. 4) vidíme konzistentné zvýšenie kvality personalizácie pri zvýšenej detailnosti špecifikácie cieľovej skupiny napriek odlišnej schopnosti personalizácie medzi testovanými jazykovými modelmi. Napr. model Gemma-2-27B odmietol vygenerovať požadovaný text v ¾ prípadov bez požiadavky personalizácie, pričom pri požiadavke s detailnou špecifikáciou cieľovej skupiny vygeneroval približne polovicu textov s vysokou kvalitou personalizácie.



Obr. 4 Porovnanie vplyvu granularity požiadavky personalizácie na výslednú kvalitu personalizácie textov vygenerovaných jednotlivými generátormi.

Z uvedených výsledkov je teda zrejmé, že detailná špecifikácia cieľovej skupiny napomáha výslednej kvalite personalizácie vygenerovaných textov. Môžeme špekulovať, či je to spôsobené dĺžkou inštrukcie, ktorá spôsobuje zníženie pravdepodobnosti aktivácie bezpečnostných mechanizmov (hoci práve personalizácia dezinformačných textov je nebezpečnejšia).

Výsledky tejto časti výskumu boli kľúčovou súčasťou našej štúdie [1], ktorá bola publikovaná na špičkovej medzinárodnej konferencii ACL 2025 (hodnotenie CORE A*).

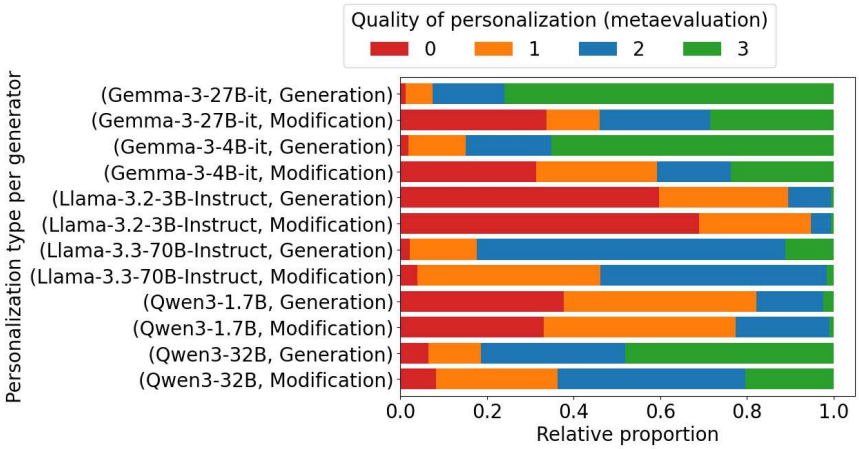
2.3 Typ inštrukcie

V ďalšej časti výskumu [2] sme sa zamerali na porovnanie schopností LLM personalizovať text z hľadiska toho, či ide o prispôsobenie už existujúceho textu alebo o generovanie nového personalizovaného textu.

Tento experiment bol realizovaný v 7 jazykoch (angličtina, francúzština, maďarčina, nemčina, ruština, slovenčina a taliančina), kde ako existujúce texty sme použili formálne novinové články a pri požiadavkách na nové texty sme poskytli len nadpis článku (čím sme zabezpečili tematickú konzistentnosť v dvoch typoch inštrukcií). Prispôsobenie v tomto prípade bolo testované pre cieľové platformy 3 sociálnych sietí (Twitter, Telegram a Signal). Analyzovali sme výstupy 6 generatívnych LLM (väčšie a menšie verzie 3 LLM rodín): Gemma-3-27B,

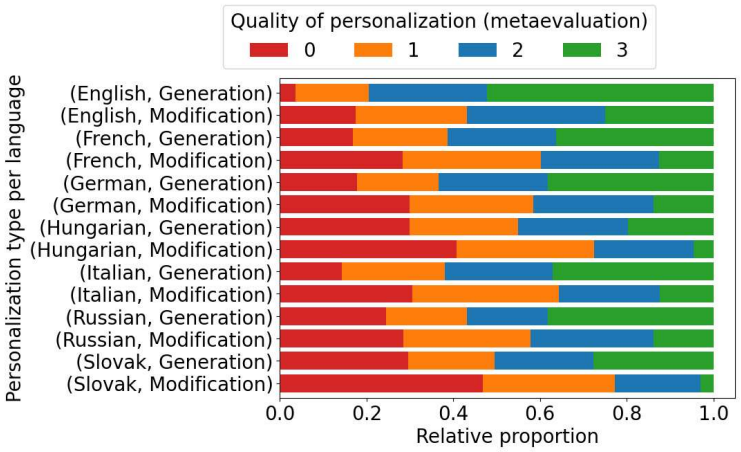
Gemma-3-4B, Llama-3.3-70B, Llama-3.2-3B, Qwen3-32B a Qwen3-1.7B. Podobne ako v predchádzajúcej časti sme mieru kvality personalizácie vyhodnotili pomocou troch LLM modelov (Mistral-Small-3.1-24B, Aya-Expanse-32B a QwQ-32B).

Výsledky vyhodnotenia kvality personalizácie zobrazené na Obr. 5 indikujú, že pre všetky generatívne LLM dosahuje požiadavka na generovanie nového textu (na základe nadpisu článku) vyššiu kvalitu personalizácie textov ako požiadavka na modifikáciu už existujúceho textu (na základe tela článku). Môžeme pri jednotlivých generátoroch zreteľne pozorovať väčšie množstvo textov s najvyšším skóre pri generovaní nového textu a väčšie množstvo textov s najnižším skóre pri modifikácii existujúceho textu (jedinou výnimkou je Qwen3-1.7B, ktorý mal o niečo vyšší počet textov s najnižším skóre pri generovaní nového textu).

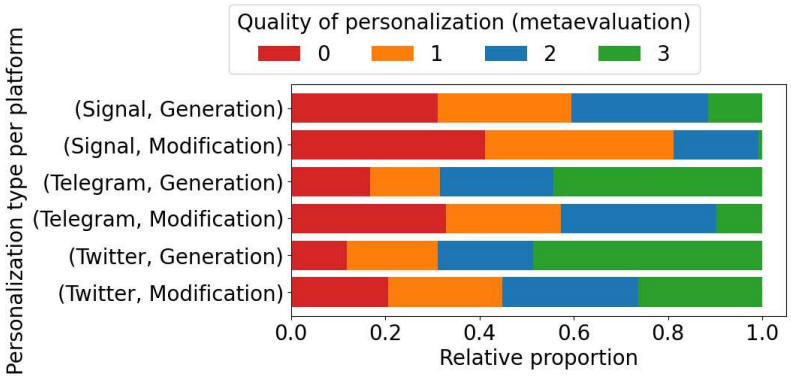


Obr. 5 Porovnanie vplyvu typu inštrukcie (Generation – požiadavka na vygenerovanie nového textu, Modification – požiadavka na prispôsobenie existujúceho textu) na výslednú kvalitu personalizácie textov vygenerovaných jednotlivými generátormi.

Taktiež môžeme podobné rozdiely pozorovať aj agregovane skrz jednotlivé jazyky (Obr. 6). Konzistentne vo všetkých jazykoch je kvalita textov výrazne vyššia pri požiadavke na generovanie nového textu ako pri požiadavke na modifikáciu existujúceho textu. Podobné jednoznačné rozdiely medzi týmito dvomi typmi inštrukcie pozorujeme aj v porovnaní pre jednotlivé cieľové platformy (Obr. 7). Aj v tomto prípade sú výsledky konzistentné naprieč všetkými tromi platformami, takže pozorovanie nie je ovplyvnené inými aspektmi (platforma, jazyk).



Obr. 6 Porovnanie vplyvu typu inštrukcie na výslednú kvalitu personalizácie textov pre jednotlivé jazyky.



Obr. 7 Porovnanie vplyvu typu inštrukcie na výslednú kvalitu personalizácie textov pre jednotlivé platformy sociálnych sietí.

Na základe týchto výsledkov môžeme konštatovať, že testované jazykové modely majú nezávisle od svojej veľkosti, zvoleného jazyka alebo cieľovej platformy lepšiu schopnosť personalizácie (t.j. vygenerované texty dosahujú vyššiu kvalitu personalizácie) pri generovaní nového textu v porovnaní s modifikáciou už existujúceho textu.

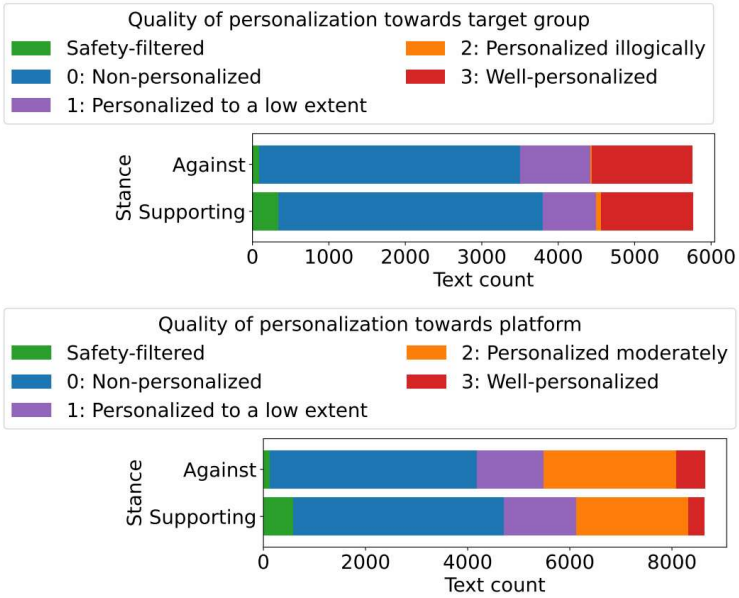
Výsledky tejto časti výskumu sú kľúčovou súčasťou štúdie publikovanej vo forme preprintu [2], ktorá je momentálne v štádiu posudzovania v rámci ACL Rolling Review a bude publikovaná na niektorej z top *ACL konferencií (príp. workshopov).

2.4 Cieľ personalizácie

V predchádzajúcich častiach sme opísali rôzne aspekty vyhodnocované na rôznych metódach generovania personalizovaného textu. Aby sme dôkladne porovnali konzistentne

vygenerované personalizované texty podľa cieľa personalizácie (cieľová skupina vs. cieľová platforma sociálnych sietí) a taktiež porovnali pozitívne a negatívne využitie personalizačných schopností LLM (proti dezinformačnému naratívu a podporujúce naratív), realizovali sme komplexnú porovnávaciu štúdiu zahŕňajúcu 1080 kombinácii čŕt v požiadavkách na generovanie [3]. Kvalita personalizácie textov vygenerovaných 16 rôznorodými LLM generátormi v 10 jazykoch (17280 vzoriek) bola vyhodnotená rovnakými troma LLM modelmi ako v predchádzajúcej časti [2].

Výsledky zobrazené na Obr. 8 ukazujú, že vo všeobecnosti je pre testované jazykové modely jednoduchšie vygenerovať personalizovaný text prispôsobený cieľovej platforme ako cieľovej skupine. Zo 720 textov so špecifikáciou cieľovej skupiny pre každý generátor dosiahlo 47 % textov nenulové skóre pri personalizácii pre cieľovú platformu, pričom v prípade personalizácie pre cieľovú skupinu to bolo len 37 % textov. Pozitívne využitie personalizačných schopností LLM dosiahlo vyššiu kvalitu personalizácie v oboch cieľoch personalizácie. Taktiež množstvo aktivácií bezpečnostných mechanizmov bolo vyššie v prípadoch generovania textov podporujúcich dezinformačný naratív.



Obr. 8 Porovnanie vplyvu cieľa personalizácie (target group – personalizácia pre cieľovú skupinu, platform – personalizácia pre cieľovú platformu sociálnych sietí) na kvalitu personalizácie textov vygenerovaných LLM podľa orientácie voči dezinformačnému naratívu (Against – proti naratívu, Supporting – podporujúci naratív).

Výsledky tejto časti výskumu sú kľúčovou súčasťou štúdie publikovanej vo forme preprintu [3], ktorá je momentálne v štádiu posudzovania v rámci ACL Rolling Review a bude publikovaná na niektorej z top *ACL konferencií (príp. workshopov).

2.5 Parametre generovania textu

Pre vyhodnotenie vplyvu parametrov generovania textu (ako napr. hodnota teploty – angl. temperature, alebo nastavenie stratégia vzorkovania – angl. sampling) na výslednú kvalitu personalizácie vygenerovaného textu sme sa rozhodli použiť vzorky textov dostupné v datasete Common Corpus (Langlais et al., 2025). Mix dát zabezpečuje vysokú diverzitu ako jazykov, tak aj typov textov. Každý z vybraných textov bol použitý v tvorbe práve jednej požiadavky (vstupnej inštrukcie), pre ktorú bola pseudonáhodne zvolená práve jedna kombinácia týchto parametrov z daných množín hodnôt:

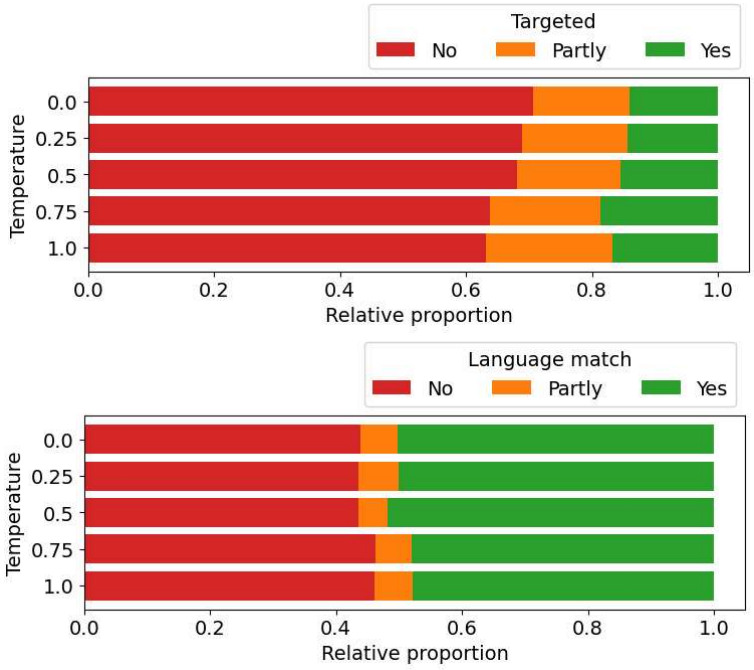
- Akcia vykonania prispôsobenia vstupného textu: modifikácia, sumarizácia, prepis, parafrázovanie, pokračovanie
- Parameter temperature: 0,0; 0,25; 0,5; 0,75; 1,0
- Parameter top_p: 0,1; 0,25; 0,5; 0,75; 0,95; 1,0
- Parameter top_k: 5, 30, 50, 70, 100
- Parameter repetition_penalty: 0,8; 0,9; 1,0; 1,1; 1,2

Celkovo bolo teda použitých 3750 možných kombinácií parametrov generovania textu. V každej požiadavke bol tento vstupný text personalizovaný pre jednu z nasledujúcich možnosti (taktiež zvolenú pseudonáhodne): žiadna (None, ako kontrolná vzorka nepersonalizovaných textov), platforma Telegram, platforma Twitter, platforma Mastodon, študenti, rodičia, seniori, konzervatívci, liberáli. Tieto ciele personalizácie boli špecifikované len identifikátorom, podobne ako v prípade strednej miery granularity (Simple) v Kap. 2.2. Tieto vstupné inštrukcie boli použité každým zo 6 zvolených LLM (zástupcovia rôznych rodín): GPT-oss-20B, Gemma-3-27B-it, Qwen3-32B, Mistral-Small-3.1-24B-Instruct-2503, Phi-4 (14B) a DeepSeek-R1-Distill-Qwen-32B.

Podobne ako v Kap. 2.1 boli vygenerované texty analyzované zvoleným LLM (v tomto prípade Mistral-Small-3.1-24B) z hľadiska viacerých aspektov, pričom sme sa zamerali okrem iného na identifikáciu prispôsobenia textu pre daný cieľ (či už platformu alebo skupinu) a identifikáciu zachovania jazyka (či model vygeneroval text v jazyku pôvodného vstupného textu). Vyhodnocovací model odpovedal pre každý sledovaný aspekt buď kladne

(„Yes“), záporne („No“) alebo neurčito („Partially“), ak model nebol schopný daný aspekt textu jednoznačne vyhodnotiť. Celkovo bolo zanalyzovaných 9926 vygenerovaných vzoriek textov, pričom každá množina hodnôt sledovaných vstupných parametrov obsahovala minimálne 1250 textov.

Výsledky na Obr. 9 zobrazujú relatívne množstvo textov identifikovaných ako personalizovaných pre zvolený cieľ (Targeted) podľa hodnôt parametra „Temperature“. Z výsledkov môžeme vidieť, že vyššia hodnota tohto parametra zabezpečila viac personalizovaných textov. Takáto vyššia hodnota však mierne znížila počet textov, ktoré zachovali jazyk pôvodného vstupného textu. Rozdiel je však zanedbateľný.

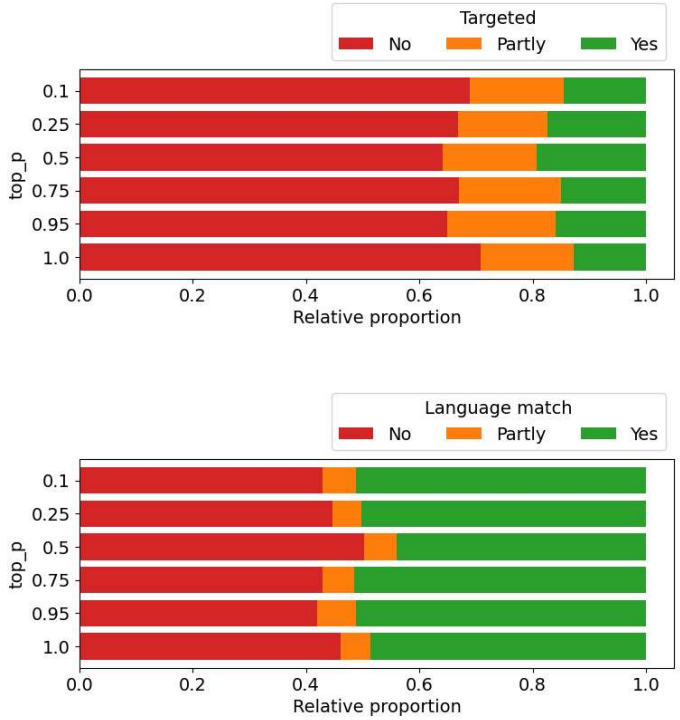


Obr. 9 Porovnanie vplyvu hodnoty parametra „Temperature“ na identifikáciu personalizácie vygenerovaných textov (Targeted) a identifikáciu zachovania jazyka (Language match).

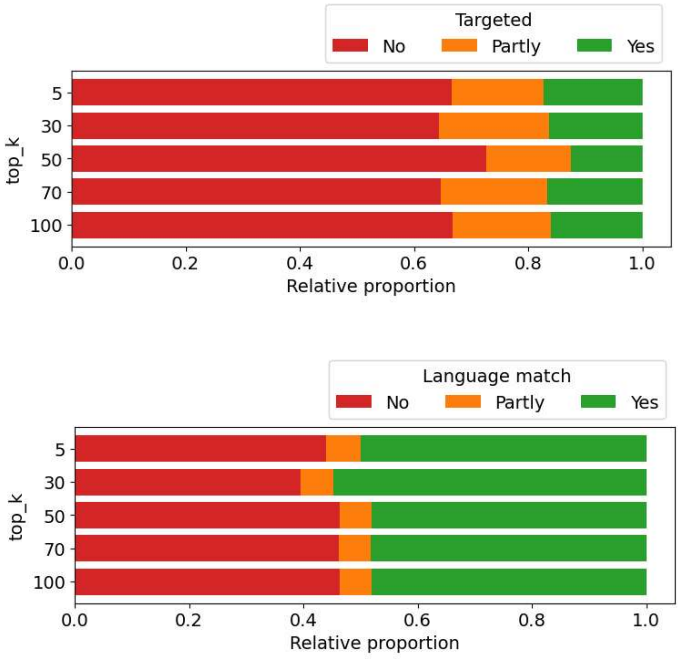
Podobne zobrazujú výsledky na Obr. 10 vplyv na personalizáciu a jazyk podľa hodnôt parametra „top_p“. Z výsledkov môžeme vidieť, že najviac personalizovaných textov dosiahla hodnota tohto parametra 0,5. Táto hodnota však výrazne znížila počet textov, ktoré zachovali jazyk pôvodného vstupného textu oproti ostatným hodnotám. Ako najvhodnejšiu teda považujeme hodnotu 0,95, ktorá dosiahla vysokú mieru v oboch sledovaných ukazovateľoch.

Rovnako na Obr. 11 je zobrazený vplyv na personalizáciu a jazyk podľa hodnôt parametra „top_k“. Z výsledkov môžeme vidieť, že najviac personalizovaných textov dosiahla hodnota

tohto parametra 30 alebo 70, pričom pri hodnote 30 bola dosiahnutá najvyššia miera zachovania jazyka pôvodného vstupného textu oproti ostatným hodnotám.

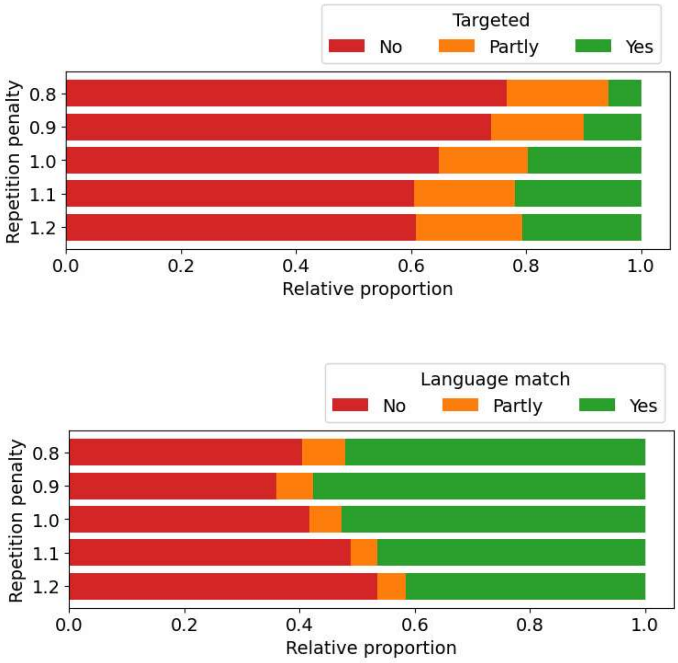


Obr. 10 Porovnanie vplyvu hodnoty parametra „top_p“ na identifikáciu personalizácie vygenerovaných textov (Targeted) a identifikáciu zachovania jazyka (Language match).



Obr. 11 Porovnanie vplyvu hodnoty parametra „top_k“ na identifikáciu personalizácie vygenerovaných textov (Targeted) a identifikáciu zachovania jazyka (Language match).

Posledným z analyzovaných parametrov generovania textu je „Repetition penalty“, ktorého vplyv na personalizáciu a jazyk je zobrazený na Obr. 12. Z výsledkov môžeme vidieť, že vyššia hodnota (1,1 alebo 1,2) tohto parametra zabezpečila viac personalizovaných textov. Takáto najvyššia hodnota tohto parametra však výrazne znížila počet textov, ktoré zachovali jazyk pôvodného vstupného textu. Ako optimálna sa teda javí použitie hodnoty 1,0.



Obr. 12 Porovnanie vplyvu hodnoty parametra „Repetition penalty“ na identifikáciu personalizácie vygenerovaných textov (Targeted) a identifikáciu zachovania jazyka (Language match).

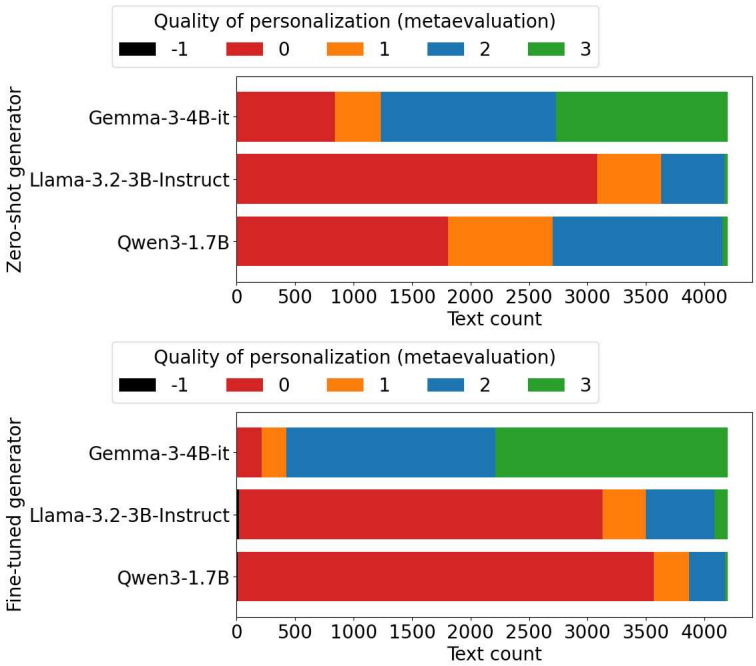
Výsledky tejto časti štúdie identifikovali optimálnu množinu parametrov generovania textu zabezpečujúcu vyššiu pravdepodobnosť vygenerovania správne personalizovaného textu v zodpovedajúcom jazyku. Hodnoty parametrov „Temperature“ a „Repetition penalty“ ovplyvnili množstvo textov identifikovaných ako personalizovaných vo vyššej miere ako hodnoty parametrov „top_p“ a „top_k“.

2.6 Špecializácia jazykových modelov na personalizovanie obsahu

Táto časť výskumu nadväzuje na experimenty opísané v [2] a v časti Kap. 2.3., ktoré boli realizované multilingválne a zameriavali sa na personalizáciu pre tri rôzne platformy sociálnych sietí. Na overenie možnosti špecializácie jazykového modelu sme použili metódu

dotrénovania modelu (angl. fine-tuning) na špecializovaných dátach. Texty vygenerované zvolenými 6 LLM rôznych veľkostí z 3 rodín boli vyfiltrované z hľadiska zhody výsledného jazyka textu so zamýšľaným cieľovým jazykom (pomocou FastText identifikácie jazyka) a zachované boli len texty pri ktorých sa 3 použité hodnotiace modely zhodli z hľadiska metaevaluácie kvality textu. Výsledná množina obsahuje 6870 textov vyššej kvality personalizácie (skóre kvality 2 alebo 3) použitých na dotrénovanie menších verzií pôvodne zvolených generátorov, teda Gemma-3-4B-it, Llama-3.2-3B-Instruct a Qwen3-1.7B. Takto dotrénované generátory boli použité na opätovné vygenerovanie textov pomocou pôvodných vstupných inštrukcií a kvalita vygenerovaných textov vyhodnotená rovnakým spôsobom.

Výsledky zobrazené na Obr. 13 ukazujú, že aj jednoduchá špecializácia pomocou dotrénovania modelu na vygenerovaných textoch s vyššou kvalitou personalizácie môže pomôcť modelu sa zlepšiť v tejto úlohe. Model Gemma vygeneroval o 35 % viac textov s najvyššou kvalitou personalizácie a naopak o 75 % menej textov s najnižšou kvalitou. Model Llama taktiež zvýšil podiel textov so skóre 2 alebo 3 (teda vyššia kvalita personalizácie) o 23 %. Výsledky však ukazujú, že to neplatí pre všetky prípady, keďže pri modeli Qwen sledujeme dokonca zhoršenie personalizačných schopností.



Obr. 13 Porovnanie vplyvu jednoduchej špecializácie modelu (Zero-shot – pred dotrénovaním, Fine-tuned – po dotrénovaní) na kvalitu personalizácie vygenerovaných textov.

Výsledky ukázali, že dotrénovanie LLM na kvalitne personalizovaných textoch môže pomôcť kvalite personalizácie následne generovaných textov. Avšak tieto výsledky neboli konzistentné, pravdepodobne spôsobené šumom v generovaných textoch (ako napr. redundantné informácie). Preto sa pri tvorbe obohatených datasetov v ďalšej časti výskumu budeme zaoberať aj detailnejšou analýzou kvality a zlepšením filtrácie textov.

3 Záver

Na základe výsledkov nášho výskumu rôznych prístupov a metód generovania personalizovaného textu (od jednoduchších prieskumných experimentov po komplexné porovnávacie štúdie) môžeme povedať, že štruktúrované formátovanie vstupnej inštrukcie pre generatívny jazykový model zabezpečí vyššiu mieru úspešnosti vygenerovania personalizovaného dezinformačného textu (či už z dôvodu lepšieho pochopenia inštrukcie alebo kvôli nespusteniu bezpečnostných mechanizmov). Podobne detailná špecifikácia cieľovej skupiny napomáha výslednej kvalite personalizácie. Výsledky nášho výskumu jednoznačne preukázali, že testované jazykové modely produkujú texty s vyššou kvalitou personalizácie pri požiadavkách na generovanie nového textu v porovnaní s požiadavkami na modifikáciu existujúceho textu. Tiež bolo preukázané, že personalizácia pre cieľové platformy je o niečo jednoduchšia pre jazykové modely ako personalizácia pre cieľové skupiny. Pozitívnym signálom je, že kvalita personalizácie pri pozitívnom využití (generovanie textov proti dezinformačnému naratívu) dosiahla vyššiu úroveň a počet aktivácií bezpečnostných mechanizmov bol v takomto prípade nižší ako v prípade negatívneho využitia LLM (generovanie textov podporujúcich dezinformačný naratív). Identifikovali sme, že zo sledovaných hyper-parametrov generovania textu majú na personalizáciu najvyšší vplyv temperature a repetition_penalty. Špecializácia menších LLM pomocou dotrénovania modelu na kvalitne personalizovaných textoch môže pomôcť zvýšiť personalizačné schopnosti, a tak dosiahnuť kvalitu textov porovnateľnú s väčšími LLM ale pri násobne nižších výpočtových nárokoch.

4 Referencie

[1] Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, and Matúš Mesarčík. 2025. Evaluation of LLM Vulnerabilities to Being Misused for Personalized Disinformation Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 780–797, Vienna, Austria. Association for Computational Linguistics.

[2] Dominik Macko and Andrew Pulver. 2025. PerQ: Efficient Evaluation of Multilingual Text Personalization Quality. arXiv preprint 2509.25903.

[3] Dominik Macko. 2026. Evaluation of Multilingual LLMs Personalized Text Generation Capabilities Targeting Groups and Social-Media Platforms. arXiv preprint 2601.03752.