PŘÍPADOVÁ ANALÝZA SELHÁNÍ · s AI o AI

Anatomie rozpadu výstupu jazykového modelu: případová analýza selhání nástroje Gemini Deep Research

O čem tato analýza je a o čem není

Tato analýza zkoumá **technické selhání generativního nástroje**, nikoli věcný obsah, který nástroj vyprodukoval. Předmětem zkoumání je tedy chyba samotná, její průběh a její příčiny, nikoli pravdivost, přesnost ani úplnost tvrzení, která se v generovaném textu objevila. Žádná data, statistika ani citace z napadeného výstupu zde nejsou ověřovány a nemají být brány jako potvrzené. Ověření faktické správnosti původního obsahu je samostatná práce, která s touto analýzou nesouvisí a bude případně provedena odděleně novým, čistým během.

Cílem je zachytit konkrétní, doložený případ rozpadu dlouhého výstupu komerčního nástroje v čase a v kontextu tak, aby mohl sloužit jako referenční materiál pro další zkoumání a pro porovnání s budoucími pokusy.

1. Popis incidentu

Nástroj Gemini Deep Research dostal zadání na hloubkový průzkum toho, čeho se nejvíc obávají lidé a zejména podnikatelé a živnostníci, kteří nepoužívají umělou inteligenci. Šlo o rozsáhlé zadání s explicitním požadavkem na velmi podrobný a strukturovaný přehled. Výstup byl vygenerován jako jeden dlouhý dokument. Časové razítko dokončení bylo 24. 6. ve 21:27.

Konkrétní podmínky běhu, doložené snímky obrazovky, jsou tyto. Použitý model byl Gemini 3.1 Pro v režimu Deep Research, s úrovní myšlení nastavenou na Standard. Konverzace probíhala ve webovém rozhraní na adrese s identifikátorem c6245e7ed3f8ed54. Tyto údaje zaznamenáváme záměrně, protože jsou pro pozdější porovnání zásadní a backendová verze modelu se časem mění.

První přibližně dvě třetiny až tři čtvrtiny výstupu byly věcné a strukturované. Text plynule procházel jednotlivými kapitolami, od demografických dat přes takzvanou past operativy, datová sila, halucinace, fenomén Shadow AI, GDPR, autorská práva, evropský AI Act a návratnost investic až po psychologické bariéry a nedůvěru spojenou s dezinformacemi. Kapitoly obsahovaly číslované odkazy na zdroje a držely konzistentní strukturu.

Zlom nastal postupně. U sekce o manuálních a řemeslných profesích se věty začaly nápadně natahovat a hromadit přívlastky. V sekcích o školství a zdravotnictví byla degradace výraznější. V závěrečné části nazvané Závěr a syntéza se výstup zhroutil úplně. Souvětí ztratila smysl, text se propadl do opakování stále stejných slov a nakonec do opakování pouhých spojek a předložek. Rozsah kolapsu je dobře patrný z exportu do PDF. Z přibližně třiatřiceti stran dokumentu je zhruba poslední třetina, řádově strany sedmnáct až třicet, zaplněna takřka výhradně opakováním shluků typu "po a po a a". Nejde tedy o drobný defekt na konci, ale o rozsáhlé pole rozpadlého textu. Na samém konci dokumentu se pak objevil blok textu v ukrajinštině a v přepisu, který s tématem neměl nic společného. Šlo o hádku z internetového fóra o polámaných policích ve vlakových kupé na trasách Lvov-Bachmut a Lvov-Odesa.

Po tomto výstupu následoval ještě druhý, samostatný projev selhání. Na zcela banální následné dotazy, například aby model přečetl vlastní vygenerovaný text, odpovídal nástroj generickými odmítavými frázemi. Doslovný přepis tohoto dialogu je rozebrán v sekci 2.3.

2. Tři vrstvy selhání

Incident není jedna porucha, ale řetězec tří rozlišitelných jevů. Jejich oddělení je hlavní analytický přínos tohoto materiálu, protože první dva jevy spolu mechanicky souvisí a třetí je jiného druhu.

2.1 Neurální textová degenerace a repetiční smyčka

Jazykový model generuje text autoregresivně. V každém kroku predikuje jen následující token na základě dosavadního kontextu a tento token se okamžitě vrací zpět na vstup pro predikci dalšího tokenu. Model přitom neoptimalizuje žádný globální cíl typu "vyhni se opakování v celé odpovědi", rozhoduje vždy jen lokálně o dalším kroku.

Degenerace výstupu do fádního a repetitivního textu je u tohoto způsobu generování dlouho popsaný jev. Holtzman a kolektiv ukázali už v roce 2020, že maximalizační dekódovací strategie vedou k výstupu, který je nezvykle repetitivní, a to i u jinak velmi kvalitního modelu. Zásadní je, že repetice má samoposilující charakter. Xu a kolektiv v roce 2022 doložili, že čím vícekrát se nějaká sekvence v kontextu zopakuje, tím vyšší je pravděpodobnost, že model bude v jejím opakování pokračovat. Jakmile tedy smyčka jednou vznikne, kontext se začne plnit opakovaným vzorem, model ten vzor čte jako silný signál a sám sebe v něm utvrzuje.

Tomu odpovídá i pozorovaná eskalace v napadeném souboru. Repetice obvykle začíná nenápadně a stupňuje se, model nejdřív opakuje slovo, pak frázi a nakonec celé věty. Přesně tuto trajektorii soubor ukazuje: nejprve přebujelé hromadění přívlastků, poté stovky opakování jednoho slova a nakonec opakování pouhých spojek a předložek. Celý systém přitom nemá žádnou vestavěnou brzdu. Jediný mechanismus, který by mohl smyčku přerušit, je samotná pravděpodobnostní distribuce, ze které se vybírá další token. Aby se model ze smyčky dostal, musel by ve vzorkování zvítězit nějaký jiný token než ten opakovaný.

2.2 Divergence a únik memorizovaného obsahu

Druhá vrstva vysvětluje ten zdánlivě nevysvětlitelný ukrajinský text na konci. Není to ani náhoda, ani vložení zvenčí, ani napadení nějakým skriptem. Je to únik memorizovaného trénovacího obsahu, který je s repeticí přímo svázaný.

Jazykové modely si část trénovacích dat zapamatovávají a za určitých podmínek je doslovně reprodukují. Carlini a kolektiv tuto schopnost popsali v roce 2021 a ukázali, že náchylnost k memorizaci roste s velikostí modelu, s tím, kolikrát se daný úsek v trénovacích datech opakoval, a s délkou poskytnutého kontextu. Na to navázali Nasr, Carlini a kolektiv v roce 2023 takzvaným divergence attack. Zjistili, že pokud se model přiměje pořád dokola opakovat jedno slovo, způsobí to, že model diverguje od svého obvyklého chování, a v okamžiku divergence začne vypisovat text zkopírovaný přímo z trénovacích dat.

Repetiční smyčka tedy funguje jako spouštěč úniku memorizovaného obsahu. A co je pro tento případ nejdůležitější, jev je doložen přímo u modelů řady Gemini. V experimentech Nasra a kolektivu u modelu Gemini 1.5 Pro divergovalo 44 procent z 3 750 generování a divergenci se podařilo vyvolat i u novějšího Gemini 2.5 Flash.

Rozdíl mezi laboratorním útokem a tímto incidentem je v jediné věci. Ve výzkumu je smyčka vyvolána záměrně, adversariálním promptem. Zde vznikla smyčka spontánně, sama od sebe, v

rámci jednoho dlouhého a postupně se rozpadajícího výstupu. Mechanismus za ní je ale stejný. Text se propadl do repetice, repetice vyústila v divergenci a po divergenci vyplaval na povrch memorizovaný úsek z nějakého webového fóra. Jde tedy o neúmyslnou, spontánní verzi přesně toho jevu, který je v literatuře popsán jako cílený útok.

Pro samotné určení, že jde o vložený cizí materiál a ne o vlastní pokračování textu modelu, máme i drobnou typografickou stopu. V uniklém bloku se v názvech tras objevují dlouhé pomlčky, tedy "Lviv—Bakhmut" a "Lviv—Odesa", zatímco okolní český text modelu dlouhé pomlčky nikde nepoužívá. Tato typografická nespojitost je nezávislou indicií, že blok pochází z jiného zdroje s odlišnou typografií, a nevznikl tedy plynulým generováním navazujícím na český text.

Co se týče obsahu samotného úniku, jde o neformální ukrajinskojazyčnou internetovou diskuzi o stavu spacích vozů ukrajinských drah (UZ), konkrétně o polámaných lůžkách a počmáraných sedačkách ve vlacích na trasách Lvov-Bachmut a Lvov-Odesa, doplněnou o zcela nesouvisející poznámku, jak si do telefonu a do počítače přidat ukrajinskou klávesnici. Kombinace latinkového přepisu a azbuky spolu s tou poznámkou o klávesnici je typická pro neformální diskuzní vlákno, kde jeden z účastníků nemá k dispozici azbukovou klávesnici.

Pokus o dohledání přesného zdroje podle nejvýraznějších frází exaktní původní vlákno nenašel. To je samo o sobě informativní zjištění, které dobře odpovídá literatuře o memorizaci. Emitované úseky bývají často konverzační, pocházejí z málo navštěvovaných stránek nebo jsou sloučené z více zdrojů, a nelze je tedy vždy přiřadit k jediné konkrétní adrese. Hledání nicméně potvrdilo dvě věci. Za prvé, že diskuze o stavu polic a lůžek ve vozech UZ jsou na ukrajinském internetu velmi hojný a běžný žánr, takže právě takový obsah se v trénovacích datech vyskytuje opakovaně. Za druhé, že přesně tento typ tematicky nesourodého, surově nasbíraného ukrajinského webového textu se reálně objevuje ve velkých veřejně dostupných webových crawl korpusech, které slouží k trénování modelů (například korpusy typu ParaCrawl a OPUS). Únik tedy nesměřuje k žádné záhadě, ale k běžné vrstvě trénovacích dat.

2.3 Následné fallback odmítání

Třetí vrstva je jiného druhu a je třeba ji prezentovat opatrněji, protože pro ni nemáme tvrdou akademickou oporu. Po rozpadlém výstupu se konverzace dostala do stavu, kdy nástroj na triviální požadavky reagoval generickými odmítnutími. Doslovný přepis tohoto dialogu vypadal takto.

Nástroj nejprve sám od sebe oznámil "na tohle nejsem naprogramován". Na žádost o pomoc se závěrem a syntézou odpověděl "s tím vám pomoct nemůžu, jsem pouze jazykový model a nemám potřebné informace a schopnosti". Na otázku, zda nedokáže přečíst text, který se vytvořil, odpověděl "jsem jazykový model a tohle ještě neumím". A na otázku, zda tedy neumí číst ani psát, uzavřel "jsem textová umělá inteligence a tohle je mimo moje možnosti".

Tyto odpovědi téměř jistě neznamenají, že by model "zapomněl" číst nebo psát. Pravděpodobnější výklad je, že po kolapsu se kolem modelu aktivovala záložní nebo bezpečnostní vrstva, která vrací předpřipravené odmítavé šablony ve chvíli, kdy se vstup nebo stav konverzace vyhodnotí jako problematický. Formulace tohoto typu mají charakter takových šablon. Jde tedy spíše o selhání obalu kolem modelu než o ztrátu jeho jazykové schopnosti. Tuto vrstvu uvádíme jako pozorování s hypotézou o příčině, nikoli jako prokázaný mechanismus.

3. Proč je dlouhý výstup typu deep research obzvlášť rizikový

Rozpad nastal na konci dlouhého výstupu a to není náhoda. Čím delší výstup, tím víc generovaných kroků a tím víc příležitostí, aby se generování ve kterémkoli z nich svezlo do repetice. Jakmile se tak stane, akumulovaný kontext začne pracovat proti modelu. Rozpadlý a repetitivní text postupně dominuje kontextu a táhne další generování stejným směrem, takže se porucha sama zesiluje.

Jedno z technických vysvětlení, proč repetice u produkčních modelů vede až k divergenci, souvisí s tím, jak se modely trénují. Při tréninku se více dokumentů spojuje do jednoho vstupu a oddělují se speciálním tokenem, který značí hranici dokumentu. Model se naučí, že u tohoto tokenu má takzvaně resetovat, tedy ignorovat předchozí kontext. Dlouhé opakování jednoho slova může takový reset napodobit, a model se pak chová, jako by začínal od nuly, což otevírá dveře k vypsání nesouvisejícího memorizovaného obsahu.

Pro praxi je podstatné, že agentní nástroje typu deep research řetězí velmi dlouhé generování bez průběžné lidské kontroly. Uživatel typicky dostane až finální dokument. Pokud se rozpad odehraje na konci, snadno zůstane přehlédnut, protože začátek a podstatná část textu vypadají naprosto v pořádku.

4. Praktický dopad

Z tohoto případu plyne několik praktických zásad, které navazují na obecnější princip obezřetného používání AI nástrojů.

Na jediný dlouhý výstup se nelze spoléhat jako na hotový produkt. Delší výstup má vyšší pravděpodobnost, že někde sklouzne, a kolaps se navíc často objevuje až na konci, tedy v místě, které se čte nejméně pozorně. Kontrola celého textu, a obzvlášť jeho závěru, by měla být samozřejmostí před jakýmkoli dalším použitím.

Adversariální audit, tedy ověření výstupu jiným nástrojem nebo druhým během, má smysl nejen kvůli faktické správnosti, ale i kvůli odhalení takovýchto strukturálních rozpadů. Lidská kontrola ve smyčce není formalita, je to jediná spolehlivá pojistka proti tomu, aby se rozpadlý nebo memorizovaný obsah dostal dál nepovšimnut.

5. Metodická poznámka pro budoucí referenci

Pro reprodukovatelnost a pro hodnotu tohoto materiálu do budoucna je třeba zaznamenat, co přesně bylo zachyceno a za jakých podmínek.

Zachyceny byly tři vzájemně se doplňující formy evidence. Za prvé syrový, neupravený výstupní soubor v markdownu i jeho export do PDF, oba včetně celé rozpadlé části. Za druhé snímky obrazovky webového rozhraní, které dokládají zadání, výběr modelu a závěrečný odmítavý dialog. Za třetí přepis celé interakce. Společně s tím je zaznamenáno původní zadání, časové razítko 24. 6. ve 21:27, identifikace modelu jako Gemini 3.1 Pro v režimu Deep Research s úrovní myšlení Standard a identifikátor konverzace c6245e7ed3f8ed54. Rozpadlá část výstupu se zde záměrně neopravuje ani nezkracuje, protože ta porucha je tím vlastním předmětem zkoumání. Korupce ve výstupu je data, ne vada, kterou bychom měli odstranit.

Zároveň je nutné počítat s omezenou reprodukovatelností. Backend komerčních nástrojů se průběžně mění a konkrétní verze modelu, která tento výstup vyprodukovala, nemusí být za nějaký čas dostupná ani identifikovatelná. Právě proto má smysl tento snímek pořídit a

uchovat nyní. Případný nový běh nad stejným zadáním je vhodné chápat jako samostatný pokus, který se s tímto zachyceným stavem porovná, nikoli jako jeho náhradu.

A znovu pro úplnost, tato analýza nezkoumá pravdivost ani kvalitu obsahu, který nástroj vygeneroval. Zkoumá výhradně technické selhání generování.

6. Citované zdroje

Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y. (2020). The Curious Case of Neural Text Degeneration. ICLR 2020. arXiv:1904.09751.

Xu, J. a kolektiv (2022). Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation. arXiv:2206.02369.

Carlini, N. a kolektiv (2021). Extracting Training Data from Large Language Models. USENIX Security Symposium 2021. arXiv:2012.07805.

Nasr, M., Carlini, N. a kolektiv (2023). Scalable Extraction of Training Data from (Production) Language Models. arXiv:2311.17035.

Příloha

Primární evidence k této analýze je syrový, neupravený výstupní soubor nástroje Gemini Deep Research, zachovaný včetně celé rozpadlé části a následného dialogu.