Katarzyna Grabowska
NASK

CyberTalk, czyli cyberbezpieczeństwo oczami kobiet

dr Agnieszka Karlińska
UW

Rozmowa z dr Agnieszką Karlińską, partnerką i kierowniczką zespołu AI, Quantica Lab, adiunktką, Uniwersytet Warszawski.

KG: Co skłoniło Cię do pracy z dużymi modelami językowymi i jak wygląda Twoja praca na co dzień?

AK: Praca z dużymi modelami językowymi jest dla mnie naturalną kontynuacją wcześniejszych zainteresowań sztuczną inteligencją i przetwarzaniem języka naturalnego. Od lat zajmuję się analizą tekstu oraz badaniem relacji między językiem, technologią i kontekstem społecznym. Rozwój LLM-ów stał się dla mnie obszarem, w którym łączę refleksję badawczą nad językiem z projektowaniem i ewaluacją systemów AI funkcjonujących w realnych kontekstach społecznych i instytucjonalnych. W ubiegłym roku kierowałam projektem HIVE AI, którego celem był rozwój modeli PLLuM oraz ich pilotażowe wdrażanie w administracji publicznej. Znaczną część czasu zajmowało mi zarządzanie dużym konsorcjum – ponad 200 osób z ośmiu instytucji realizowało równolegle kilkadziesiąt zadań, od pozyskiwania i przetwarzania danych, przez trening modeli, po ich ewaluację i implementację. Wymagało to licznych decyzji strategicznych oraz koordynowania pracy zespołów o bardzo zróżnicowanych kompetencjach.

Równolegle starałam się wygospodarować czas na zaangażowanie w prace merytoryczne, zwłaszcza w obszarze danych. Koncentrowałam się na etapie posttreningu, czyli dostrajania i wychowania modeli (ang. alignment) – ich dostosowania do oczekiwań użytkowników oraz norm i wartości społecznych. To właśnie tutaj najlepiej widać, jak decyzje projektowe przekładają się na sposób komunikacji modelu

z użytkownikiem i to na tym etapie kluczowe znaczenie ma zaplecze humanistyczno-społeczne. Zależało nam, aby modele PLLuM odpowiadały nie tylko poprawnie – pod względem merytorycznym i językowym – ale również w sposób wyważony i bezpieczny. W praktyce oznaczało to m.in. tworzenie zbiorów danych bazujących na odpowiedziach wzorcowych oraz antyprzykładach. Opracowanie koncepcji takich danych i koordynacja ich tworzenia były dużym wyzwaniem, bo wciąż brakuje ugruntowanych dobrych praktyk, a cały proces musi być osadzony w lokalnym kontekście językowym i społecznym. Wiele decyzji podejmowaliśmy więc na bazie doświadczeń i eksperymentów, a nie gotowych schematów.

Obecnie łączę pracę badawczą z działalnością wdrożeniową: z jednej strony badam społeczno-kulturowe uwarunkowania działania modeli językowych i prowadzę analizy lingwistyczne, z drugiej – wdrażam LLM-y w środowiskach produkcyjnych, a nie wyłącznie pilotażowych. Doświadczenie zdobyte przy budowie modeli PLLuM pozwala mi świadomie projektować i wdrażać rozwiązania oparte na LLM-ach – rozumiem nie tylko możliwości i ograniczenia modeli, ale także decyzje projektowe, które kształtują ich zachowanie.

KG: Jak LLM-y mogą wspierać organizacje w obszarze związanym z cyberbezpieczeństwem?

AK: Duże modele językowe są szczególnie przydatne tam, gdzie organizacja pracuje z dużym wolumenem danych tekstowych – logami systemowymi, zgłoszeniami incydentów, raportami czy dokumentacją techniczną. Potrafią szybko porządkować i podsumowywać

informacje, wskazywać powtarzające się schematy oraz wyłapywać sygnały, które mogą wymagać dalszej analizy. Dzięki temu zespoły bezpieczeństwa mogą szybciej identyfikować zdarzenia wymagające reakcji. Dobrym przykładem jest wsparcie w analizie zgłoszeń incydentów: model może grupować podobne przypadki, wskazywać wzorce oraz sygnalizować, które zdarzenia wymagają pilnej interwencji, a które mają charakter rutynowy. To pozwala analitykom koncentrować się na realnych zagrożeniach zamiast na wstępnej selekcji informacji.

LLM-y mogą wspierać analizę polityk bezpieczeństwa i procedur wewnętrznych, identyfikując niejasności, niespójności czy potencjalne luki. Wykorzystuje się je do analizy prób phishingu, wykrywania podejrzanych treści i symulowania scenariuszy ataków w ramach testów odporności organizacji. Coraz częściej są także integrowane z tradycyjnymi rozwiązaniami, takimi jak systemy wykrywania i zapobiegania włamaniom (IDS), gdzie ich zdolność rozumienia kontekstu pozwala poprawiać rozpoznawanie anomalii i nowych typów ataków, ograniczając jednocześnie liczbę fałszywych alarmów.

Warunkiem ich bezpiecznego wykorzystania jest jednak praca w kontrolowanym środowisku, na zaufanych danych oraz z zachowaniem zasady „human-in-the-loop", czyli realnego nadzoru człowieka nad decyzjami systemu.

KG: Czy istnieje ryzyko, że modele językowe zostaną wykorzystane przez cyberprzestępców? Jak można temu przeciwdziałać?

AK: Duże modele językowe są technologią obosieczną. Te same mechanizmy, które pozwalają automatyzować analizę informacji

czy generować spójne treści, dostosowane do konkretnej grupy odbiorców, mogą zostać wykorzystane do działań szkodliwych, na przykład do tworzenia przekonujących wiadomości phishingowych. Nie są to jednak nowe zagrożenia. Modele obniżają próg wejścia i zwiększają skalę działań, które wcześniej również były możliwe.

Przeciwdziałanie nadużyciom polega przede wszystkim na odpowiedzialnym projektowaniu i wdrażaniu modeli. Obejmuje to zabezpieczenia na etapie treningu i posttreningu, stosowanie dodatkowych mechanizmów kontrolnych (tzw. guardrails), testowanie odporności modeli na próby manipulacji oraz jasne określenie zakresu i kontekstu ich zastosowania. Równie istotne są procedury organizacyjne oraz monitoring sposobu wykorzystania systemu w praktyce.

KG: Jak zapobiegać generowaniu szkodliwych treści przez LLM-y?

AK: Wymaga to podejścia kompleksowego i iteracyjnego. Kluczową rolę odgrywają decyzje podjęte na etapach dostrajania i wychowania modeli. W projekcie HIVE AI uczyliśmy modele nie tylko tego, jak odpowiadać, ale także kiedy odmówić odpowiedzi lub zasygnalizować brak wystarczających informacji.

Drugim istotnym elementem jest odporność na próby obejścia zabezpieczeń, takie jak wstrzykiwanie promptów (ang. prompt injection). Modele powinny być testowane w warunkach zbliżonych do rzeczywistych, z wykorzystaniem zróżnicowanych scenariuszy ataków, aby sprawdzić, jak reagują na prowokacyjne lub manipulacyjne polecenia.

Nasze analizy pokazały, że wiele popularnych otwartych modeli jest dobrze zabezpieczonych

w języku angielskim, natomiast znacznie gorzej radzi sobie z atakami w języku polskim. Modele PLLuM, trenowane również na polskojęzycznych scenariuszach ataków, są wyraźnie bardziej odporne na tego typu próby – choć oczywiście ryzyka nie da się całkowicie wyeliminować.

Uzupełnieniem wychowania są wspomniane wcześniej dodatkowe mechanizmy zabezpieczające (guardrails), działające na poziomie całego systemu. Mogą one filtrować wejście i wyjście, blokować określone typy treści lub ograniczać zakres odpowiedzi. Zapobieganie generowaniu szkodliwych treści nie jest jednorazowym działaniem, ale procesem wymagającym ciągłej ewaluacji aktualizacji zabezpieczeń.

KG: Jakie wyzwania etyczne i prawne wiążą się z wykorzystaniem dużych modeli językowych?

AK: Jednym z kluczowych wyzwań jest transparentność. Użytkownicy powinni wiedzieć, w jakim zakresie decyzje lub rekomendacje są wspierane przez systemy AI oraz gdzie przebiega granica odpowiedzialności między człowiekiem a technologią. Równie istotne jest to, aby rozumieli ograniczenia systemu i nie przypisywali mu kompetencji, których w rzeczywistości nie posiada. W tym kontekście problemem bywa pozorna wiarygodność LLM-ów – potrafią generować bardzo przekonujące odpowiedzi nawet wtedy, gdy są one niepełne lub błędne. Modele nie mają zdolności rozpoznawania własnych ograniczeń, dlatego konieczne jest świadome projektowanie interakcji z użytkownikiem oraz stosowanie dodatkowych zabezpieczeń.

Istotnym wyzwaniem pozostaje legalność danych treningowych oraz zgodność z regulacjami

krajowymi i europejskimi, w tym z nowymi ramami prawnymi dotyczącymi AI. Pozyskanie dużych wolumenów wysokiej jakości tekstów – liczonych w setkach miliardów słów – jest jednym z najtrudniejszych elementów budowy LLM-ów. W przypadku PLLuM jedynie niewielka część zgromadzonych danych (ok. 5%) mogła zostać wykorzystana do trenowania w pełni otwartych modeli, co wymagało rygorystycznej selekcji oraz dokumentowania pochodzenia tekstów. Praca z danymi oznacza również odpowiedzialność za ich bezpieczeństwo, zwłaszcza gdy modele operują na informacjach wrażliwych. Kluczowe jest pełne panowanie nad tym, gdzie i w jaki sposób dane są przetwarzane oraz kto ma do nich dostęp.

Równocześnie należy uwzględniać obciążenia i niereprezentatywność danych – modele uczą się na materiałach historycznych, które mogą odzwierciedlać istniejące nierówności czy stereotypy. Bez systematycznej ewaluacji i świadomej pracy nad danymi istnieje ryzyko ich utrwalania, a w konsekwencji wzmacniania nierówności zamiast ich ograniczania.

KG: Jakie kompetencje są dziś kluczowe, jeśli chodzi o specjalistów pracujących przy LLM-ach?

AK: Praca nad dużymi modelami językowymi wymaga bardzo zróżnicowanych kompetencji i trudno wskazać jeden „idealny” profil specjalisty. Niezbędni są eksperci techniczni odpowiedzialni za architekturę modeli, trening, optymalizację i wdrożenia. Równie istotną rolę odgrywają lingwiści i specjaliści od danych, którzy wnoszą wiedzę o strukturze języka, jego

odmianach i rejestrach. Coraz większe znaczenie ma także znajomość prawa, w tym prawa autorskiego, ochrony danych osobowych oraz regulacji dotyczących AI. W mojej ocenie szczególnie cenna jest umiejętność łączenia tych perspektyw. Potrzebne są osoby odgrywające rolę „tłumaczy" między światem technicznym, lingwistycznym, prawnymi organizacyjnym – takie, które rozumieją zarówno architekturę systemu, jak i jego konsekwencje społeczne.

KG: Jakie rady dałabyś kobietom, które chciałyby rozwijać się w obszarze AI i cyberbezpieczeństwa?

AK: AI i cyberbezpieczeństwo mogą sprawiać wrażenie zamkniętych światów z wysokim progiem wejścia. W rzeczywistości są to obszary zespołowe, wymagające bardzo różnych kompetencji. Podstawy techniczne są ważne, ale równie istotne jest znalezienie dla siebie roli zgodnej z własnymi zainteresowaniami i predyspozycjami, zamiast próby dopasowania się do jednego, wąsko rozumianego wzorca specjalisty. Na pewno warto jak najwcześniej pracować na realnych problemach – nawet w niewielkiej skali. Kontakt z praktyką pozwala szybciej zrozumieć, jak teoria przekłada się na działanie systemów i jakie pytania naprawdę mają znaczenie. Nie należy się też zrażać tym, że środowisko jest nadal zmaskulinizowane. Różnorodność perspektyw w zespołach pracujących nad AI przekłada się na jakość projektowanych systemów – pomaga lepiej identyfikować ryzyka, dostrzegać społeczne konsekwencje technologii i unikać uproszczeń.

KG: Dziękuję za rozmowę i podzielenie się doświadczeniami związanymi z pracami

z dużymi modelami językowymi.

AK: Dziękuję!

EUROPEJSKI MIESIĄC CYBER BEZPIECZEŃSTWA