Gmail

Marcin fic <marcin.fic@demagog.org.pl>

Prośba o komentarz – SI w polityce

Filip Biały <filip.bialy@amu.edu.pl> 31 marca 2025 11:35
Do: Marcin Fic <marcin.fic@demagog.org.pl>

Dzień dobry,

przesyłam poniżej krótki komentarz.

Grok - model stworzony przez xAI, firmę Elona Muska - jest zintegrowany z serwisem X (Twitter) w formie bota, co pozwala użytkownikom w prosty sposób, w konwersacji z innymi użytkownikami, wysyłać pytania do niego i uzyskiwać natychmiast odpowiedzi. Podobny mechanizm funkcjonuje np. w serwisie Telegram. Nie jest to rozwiązanie samo w sobie złe i można wyobrazić sobie wiele sytuacji, w których może być pomocne. Jednak jeśli Grok i inne modele językowe traktowane są jako źródła rzetelnej informacji, pojawia się problem. Przyczyn jest kilka.

Po pierwsze, powszechnie wiadomo już, że dostępne dziś modele językowe mają tendencję do halucynacji, czyli generowania sformułowanych bardzo przekonywająco, lecz w istocie nieprawdziwych treści. Modele te, w sytuacji, kiedy nie posiadają w danych uczących adekwatnych informacji, zastępują je treścią, którą - w przypadku człowieka - nazwalibyśmy zmyśloną. W przypadku modeli językowych treść generowana jest poprzez przewidywanie statystycznego prawdopodobieństwa, następowania po sobie słów w zdaniu - nie ma w tym procesie oceny zgodności generowanego tekstu z faktami.

Po drugie, modele językowe trenowane są na historycznych danych, co oznacza, że pytanie ich o weryfikację faktów na temat bieżących wydarzeń może być nieskuteczne. Wprawdzie najnowsze modele - w tym Grok - mogą przeszukiwać internet i na tej podstawie udzielać odpowiedzi, jednak wówczas ich odpowiedzi zależą od jakości źródeł, do których model ma dostęp. Co więcej, nawet posiadanie faktycznie prawdziwych danych nie oznacza, że model poprawnie je streści. Niektóre modele, nawet jeśli dostarczymy im rzetelne dane wejściowe do przetworzenia (np. treść strony internetowej wiarygodnego medium), mogą treści te zignorować i zastąpić je halucynacją. W innym przypadku sposób streszczenia źródła może wprowadzać w błąd: model może sparafrazować informację źródłową (np. czyjś cytat), zmieniają jej sens.

Po trzecie, wszystkie dostępne publicznie narzędzia oparte na modelach językowych (ChatGPT, DeepSeek, Claude, Gemini, Grok itd.) posiadają zabezpieczenia (guardrails), które albo uniemożliwiają wygenerowanie pewnych treści (np. instrukcji zbudowania bomby), albo też wskazują, co dana odpowiedź powinna zawierać. W przypadku Groka kilka tygodni temu okazało się, że model ten poinstruowany został przez inżynierów xAI (firmy, która odpowiada za stworzenie Groka), aby odpowiadając na pytania o dezinformację, pomijać nazwiska Elona Muska oraz Donalda Trumpa jako osób upowszechniających dezinformację. Ponieważ twórcy modeli językowych nie ujawniają szczegółów dotyczących opisywanych ograniczeń, nie wiemy, w jaki sposób może to, w konkretnych przypadkach, wpływać na uzyskiwane odpowiedzi.

Opisywane cechy są charakterystyczne dla wszystkich współczesnych modeli językowych. W przypadku Groka pojawia się pytanie o to, jakiego rodzaju ideologię model ten reprezentuje w swoich odpowiedziach i jak może być wykorzystywany politycznie. Dowodzi tego niedawna sytuacja w Indiach, gdzie Grok stał się istotnym narzędziem walki politycznej (https://www.bbc.co.uk/news/articles/cd65p1pv8pdo). Musimy przyzwyczaić się do myśli, że modele językowe stały się elementem dyskursu politycznego, pozwalając na generowanie i amplifikowanie komunikatów politycznych i ideologicznych. Jednak ten sposób ich wykorzystania nie ma wiele wspólnego z fact-checkingiem. Zwykli użytkownicy mogą odwoływać się do Groka także na zasadzie argument z autorytetu - znany jest mechanizm pokładania przez ludzi dużego zaufania w werdyktach maszyny, która ma rzekomo być bardziej obiektywna.

Nie twierdzę, że modele językowe nie będą stawały się coraz dokładniejsze, nawet jeśli ze względu na to, jak technicznie są skonstruowane, pewnych mechanizmów (takich jak halucynacje) nie uda się w pełni wyeliminować. Większą poprawność modeli zawdzięczać będziemy także temu, że to użytkownicy weryfikują uzyskiwane treści, zwrotnie informując te modele, kiedy popełniają błąd (est to mechanizm znany jako Reinforcement Learning from Human Feedback). W międzyczasie - a także i później - zalecałbym jednak duży sceptycyzm i ostrożność.

Z pozdrowieniami
Filip Biały
[Ukryto cytowany tekst]