De duistere kant van LLM's: hoe AI-zoekresultaten een goudmijn voor phishers zijn

De komst van generatieve AI en large language models (LLM's) verandert razendsnel hoe mensen informatie zoeken. Waar we voorheen vertrouwden op Google-resultaten, vragen steeds meer gebruikers direct aan ChatGPT, Claude of Perplexity: *"Wat is de inlogpagina van mijn bank?"* of *"Waar kan ik mijn account resetten?"*

Kevin van Beek, specialist in SEO en AI-search, ziet dagelijks hoe deze verschuiving niet alleen kansen creëert, maar ook een minder besproken keerzijde heeft. Wat efficiënt lijkt, kan namelijk ook een deur openzetten voor misleiding en phishing via AI-zoekresultaten.

AI die nep-domeinen of nep-content aanbeveelt

LLM's vertonen regelmatig informatie uit verkeerde of zelfs niet-bestaande domeinen als antwoord op simpele login-vragen. Denk aan: *"Ik ben mijn bookmark kwijt, kun je me de officiële loginpagina van bedrijf-x geven?"*

De resultaten zijn zorgwekkend:

- Een aanzienlijk deel van de domeinen die LLM's noemden, bleek niet geregistreerd of had geen content.
- Sommige verwijzingen leidden naar bestaande, maar ongerelateerde bedrijven.
- Een groot aantal domeinen kon direct door kwaadwillenden worden opgekocht en misbruikt.

Dat betekent dat een gebruiker op zoek naar de juiste loginpagina onbewust door AI naar een phishingsite kan worden gestuurd.

Van SEO-poisoning naar GEO-manipulatie

Wat ik in mijn werk als SEO-specialist steeds vaker zie, is dat bedrijven oude of dubieuze optimalisatietactieken inzetten om door LLM's zoals ChatGPT geciteerd te worden. Dit gebeurt met de beste bedoelingen en eerlijk is eerlijk: het werkt verrassend eenvoudig. Je hoeft niet eens zichtbaar te zijn in Google om toch via AI-zoekresultaten onder de aandacht te komen. Voor sommige merken is dat een kans, maar het legt ook iets zorgelijks bloot: de kwaliteits- en vertrouwenssignalen van LLM's liggen momenteel veel lager dan in Google.

In LLM's zoals ChatGPT zie je duidelijke verschillen met Google. Het model citeert regelmatig sites die in Google nauwelijks of niet meer zichtbaar zijn. Denk aan kleine websites met weinig autoriteit of sites die getroffen zijn door updates, of bronnen met lage E-E-A-T die Google normaal gesproken zou filteren.

Op het eerste gezicht lijkt dat positief: niet elke zoekmachine hanteert immers dezelfde strenge filters als Google, en dat kan nieuwe of kleinere spelers eerlijke kansen geven. Maar er schuilt ook een risico in. Zodra dit soort zwakke plekken bewust worden misbruikt, verandert een kans naar een kwetsbaarheid.

LLM's gedragen zich veel meer als een open deur

En precies daar loopt de lijn door naar de duistere kant. Want de kwetsbaarheden die bedrijven soms zien als kans, worden door kwaadwillenden gebruikt als wapen met phishing en misleiding als direct gevolg.

Waar Google streng filtert op betrouwbaarheid, presenteren modellen vaak onbekende of twijfelachtige sites alsof het gezaghebbende bronnen zijn. Soms zelfs inclusief hallucinaties of phishing-achtige links die met grote overtuiging worden gebracht.

Kwaadwillenden spelen hier actief op in door AI-geoptimaliseerde content te publiceren, bijvoorbeeld via GitHub-projecten, handleidingen of blogposts. Daarmee geven ze valse domeinen een vals aura van legitimiteit in de databronnen waar LLM's hun kennis uit putten. Vervolgens belanden deze domeinen met overtuiging in AI-antwoorden, waardoor nietsvermoedende gebruikers ze gepresenteerd krijgen alsof het betrouwbare bronnen zijn.

Daar zit direct een waarschuwing in. Bedrijven die deze kwetsbaarheid misbruiken, lopen het risico dat ze op korte termijn zichtbaar zijn, maar op langere termijn verdwijnen zodra LLM's hun kwaliteitsstandaarden aanscherpen. De geschiedenis leert dat spamachtige praktijken, die in Google hard afgestraft werden door algoritme-updates, ook in LLM's hun houdbaarheidsdatum zullen hebben.

Geen theorie, maar praktijk

Dit gevaar is niet hypothetisch. Er zijn al genoeg campagnes ontdekt waarbij duizenden AI-gegenereerde phishingpagina's werden ingezet, bijvoorbeeld in de cryptosector, bankieren of reisbranche. Deze sites zien er professioneel uit, laden snel, en zijn geoptimaliseerd voor zowel mens als machine. Precies wat AI-modellen nodig hebben om ze als betrouwbaar te classificeren.

Phishers hoeven in zo'n scenario niet eens meer gebruikers te lokken via advertenties of zoekresultaten, want de AI zelf raadt hun domein aan.

Waarom dit extra gevaarlijk is

De risico's zijn tweeledig:

1. Vertrouwen in AI: gebruikers ervaren AI-antwoorden als direct, helder en betrouwbaar. Een foutieve of malafide link wordt daardoor sneller aangeklikt.

2. Zichtbaarheid: AI-antwoorden staan vaak bovenaan in de zoekmachines en krijgen dus een voorkeurspositie ten opzichte van reguliere zoekresultaten.

Wat merken en ontwikkelaars kunnen doen

- LLM-ontwikkelaars zouden URL-verificatiesystemen moeten inbouwen, gebaseerd op officiële registries of trusted sources.

- Merken moeten proactief high-risk lookalike domeinen registreren en monitoren welke links AI-modellen suggereren.

- Threat intelligence kan helpen om verdachte domeinen vroegtijdig te identificeren en uit de lucht te halen.

Van kans naar risico

De verschuiving van zoekmachines naar AI-modellen opent ongekende kansen, maar ook nieuwe aanvalsvectoren. Als SEO-specialist richt ik me dagelijks op de zichtbaarheid in AI-search voor mijn klanten. Van Google's AI-overviews tot modellen als ChatGPT. De uitdaging is om dit op een duurzame manier te doen, zonder te vervallen in kortstondige, spamachtige tactieken die telkens aangepast moeten worden omdat ze niet meer werken.

Tegelijkertijd laat de praktijk zien dat dezelfde zwakke plekken die bedrijven nu kansen bieden, door kwaadwillenden worden misbruikt. LLM's zijn eenvoudig te manipuleren en kunnen met grote overtuiging verkeerde of zelfs gevaarlijke resultaten presenteren. Wat vandaag nog een onschuldige vergissing van een model lijkt, kan morgen de basis vormen van een grootschalige phishingcampagne.