Studenten bouwen detector voor ChatGPT-teksten

Prototype • Een project in Eindhoven biedt zicht op een programma dat kan zien of een tekst door AI-software is geschreven. Dat sluit bedrog niet uit.

Merijn van Nuland
redactie binnenland

Was kunstmatige intelligentie (AI) voorheen een nogal ongrijpbaar verschijnsel, inmiddels zit het als een onzichtbaar monster in de banken van ieder klaslokaal. Met online programma's als ChatGPT maak je immers in een handomdraai een vrij accuraat opstel over (noem eens wat) de Tachtigjarige Oorlog of Louis Couperus. En dat brengt fraude met zich mee.

Niek van Dam en enkele andere studenten van de Fontys-hogeschool in Eindhoven gaan nu de strijd aan met het programma. Zij ontwikkelden met hun bedrijfje OpenMaze een ChatGPT-detector die een inschatting maakt of een tekst door AI is gegenereerd. Een Engelstalig prototype is inmiddels beschikbaar.

Dat prototype lijkt al aardig te werken als je enkele voorbeelden invoert. Eerst een gegenereerde tekst: daarvan stelt de detector dat deze met 99,61 procent zekerheid is gemaakt door een AI-programma. Het omgekeerde is het geval als je een zelfgeschreven tekst invoert, dan blijkt de score namelijk 0,02 procent. Niek van Dam: "Mensen schrijven vaak vrij creatief en aritmisch. Computers zoeken juist naar een optimaal woordgebruik en zinsvolgorde. Dat maakt het best prettig leesbaar, maar ook op te sporen."

Met hun detector zijn de studenten niet helemaal uniek, geeft Van Dam grif toe. Zo lanceerde OpenAI, de maker van ChatGPT, deze week ook een opsporingstool, die overigens nog lang niet altijd accuraat blijkt. Het grote verschil is dat OpenMaze het programma binnen afzienbare tijd ook in het Nederlands wil aanbieden, en dat het 'betrappen' van fraudeurs voor de makers eigenlijk bijzaak is.

"Het gebruik van ChatGPT blijft", zegt Van Dam. "Het is een techniek waar we simpelweg niet omheen kunnen. Wij willen docenten vooral gereedschap geven om daar op een didactische manier mee om te gaan." Zo gaat het model van OpenMaze binnenkort automatische open vragen genereren waarmee de docent het gesprek kan aangaan met zijn studenten of scholieren. Kunstmatige intelligentie kan immers prima vertellen wie Pim Fortuyn was, maar krijgt het al moeilijker als je vraagt waarom zijn opkomst de Nederlandse politiek zo op zijn kop zette.

> 'De focus moet zijn: hoe zorgen we ervoor dat studenten niet willen frauderen?'

Is Van Dam niet bang dat docenten zijn website toch vooral zullen gebruiken om fraudeurs te straffen? Hij denkt even na. "Ik kan een docent natuurlijk niet verbieden om boos te worden. Maar idealiter is zo'n gegeneerde tekst het begin van een dialoog tussen student en docent." Bovendien is het systeem nog niet waterdicht. "Wie het echt wil, kan nog altijd *loopholes* vinden om aan detectie te ontsnappen."

Maar uiteindelijk is die wedloop tussen chatbots en opsporingsprogramma's bijzaak, zegt Serge Thill. Hij is universitair hoofddocent AI aan de Radboud Universiteit, en zelf niet betrokken bij het Eindhovense project. "Als iemand de boel wil bedriegen, dan lukt dat meestal. Een nieuwe chatbot verandert daar weinig aan. De focus in het onderwijs moet dus altijd zijn: hoe zorgen we ervoor dat studenten niet willen frauderen?"