

AI kijkt nauwkeuriger na dan een docent

Een analyse van 20.000 beoordelingen: waarom de kwaliteit van het correctievoorschrift bepalend is voor het succes van AI in het onderwijs.

Gepubliceerd door: ToetsPers

Geschreven door: Abel Stam

Datum: februari 2026

Samenvatting

Generatieve Kunstmatige Intelligentie (Gen AI) belooft de werkdruk van docenten te verlagen, met name bij het nakijken van open vragen. Maar hoe nauwkeurig is een Large Language Model (LLM) vergeleken met het menselijk oordeel? En wat zijn de factoren die deze nauwkeurigheid beïnvloeden?

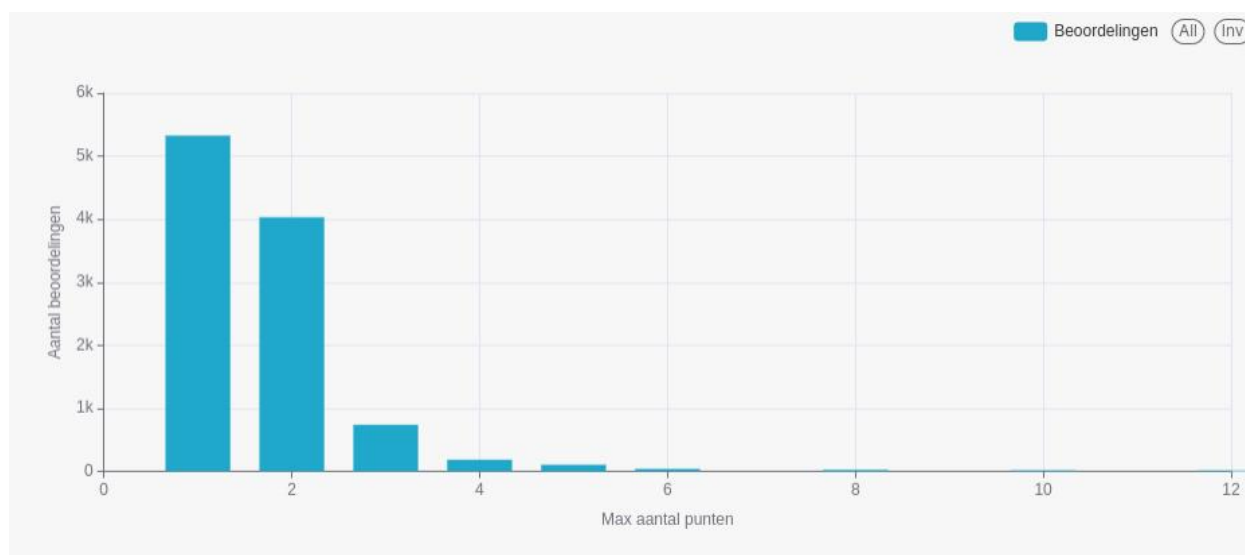
ToetsPers analyseerde 20.000 beoordelingen van docenten en vergeleek deze met scores van onze AI-nakijkmodule. De resultaten zijn opvallend: AI voorspelt in 80% van de gevallen exact dezelfde score als de docent. Echter, in de gevallen waarin de scores afwijken, blijkt de AI vaak 'nauwkeuriger' in het strikt volgen van de regels, terwijl de docent vaker de context van het vak en het niveau meeneemt in de beoordeling. Dit whitepaper toont aan dat de bottleneck voor AI-implementatie niet de technologie is, maar de reproduceerbaarheid van het correctievoorschrift.

Hoe verhouden AI en docent zich tot elkaar?

Eén van de maatregelen op door de AI-Act geclassificeerde hoogrisico systemen, is het in kaart brengen van de accurate van het systeem. Om de accurate van onze AI-module te toetsen, hebben we een dataset van 20.000 antwoorden op open vragen over diverse vakken geanalyseerd. We definieerden 'accuratie' als de mate waarin de AI-score overeenkomt met de door de docent gegeven score.

De dataset

De dataset bestaat uit 20.000 door docenten beoordeelde antwoorden op open vragen. De beoordeling heeft plaatsgevonden aan de hand van een correctievoorschrift. Voor de vragen zijn 1 tot 12 punten te verdienen, met het overgrote deel tussen de 1 en 3 punten. De beoordeelde antwoorden zijn verdeeld over beta-, gamma-, alfa-vakken en talen.



Figuur 1: *Verdeling van aantal beoordelingen ten opzichte van het maximaal aantal punten.*

De beoordelingen van de docent zijn vergeleken met de beoordelingen van een AI thinking model. Uit de afwijkende beoordelingen is een steekproef genomen van 150 resultaten die zijn nagekeken door een tweede menselijke beoordelaar. Ook is voor deze resultaten aangegeven of het bijbehorende correctievoorschrift toereikend is om het antwoord te beoordelen.

De basiscijfers

Uit de data blijkt een sterke correlatie:

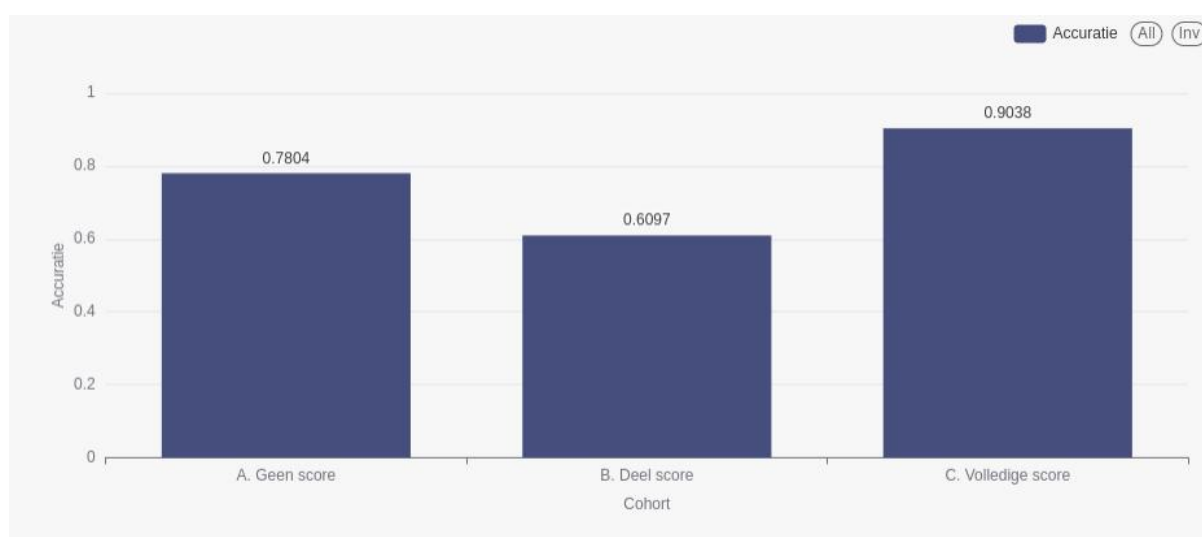
- 80% Overeenkomst: In 4 op de 5 gevallen geeft de AI exact hetzelfde aantal punten als de docent.
- 20% Afwijking: Waar de scores afwijken, is de verdeling als volgt:
 - 13%: AI is strikter dan de docent.
 - 7%: AI is coulanter dan de docent.

Dit suggereert direct een trend: AI neigt ernaar strenger te zijn dan de menselijke beoordelaar.

Zekerheid in extremen

Een belangrijk inzicht voor de implementatie van AI in het nakijkproces is de verdeling van de accuratie. De accuratie is het hoogst bij de extremen van de puntentelling.

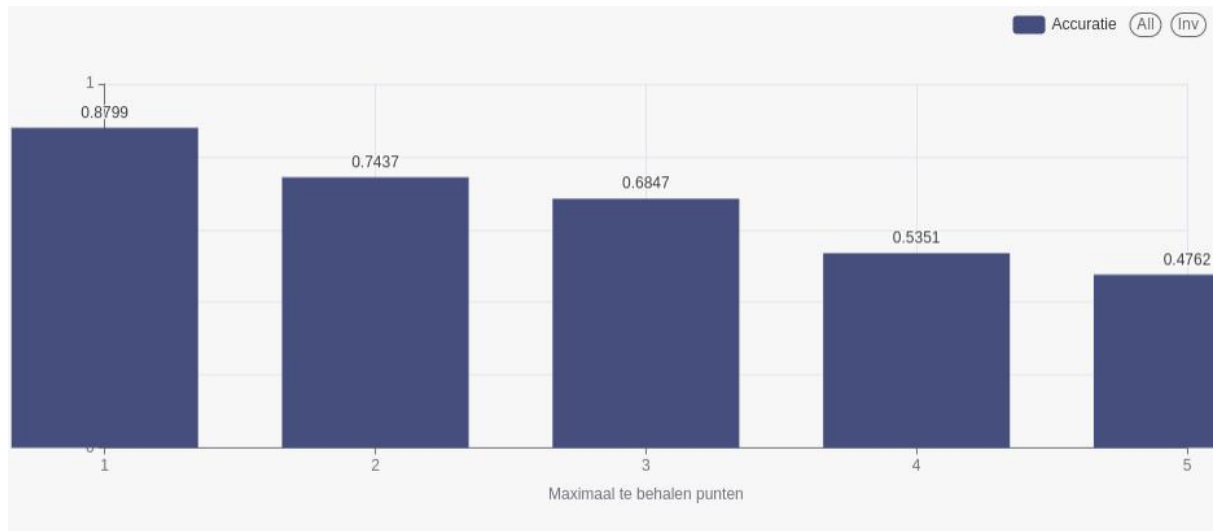
- Volledige score: Als AI het maximale aantal punten toekent, is de docent het daarin 90% van de gevallen mee eens.
- Geen score: Als AI geen punten toekent, is de docent het daar in 78% van de gevallen mee eens.
- Deelscores: Als AI een deel van het maximaal te behalen punten toekent, dan is de accuratie lager.



Figuur 2: Een staafdiagram dat de accuratie toont per score-categorie (Geen score, deelscore, volledige score).

Meer punten, minder accuraat

Ook is er een duidelijke negatieve correlatie tussen het aantal te behalen punten en de accuraat van AI ten opzichte van de docent.



Figuur 3: relatie tussen het maximaal aantal te behalen punten en de accuraat.

Het correctievoorschrift

De zekerheid in extremen kan verklaard worden door het correctievoorschrift. Het is aannemelijk dat het makkelijker is om een antwoord te classificeren als helemaal goed of fout, dan als deels goed. Ook wijst de relatie tussen het maximaal te behalen aantal punten en de accuraatheid naar het correctievoorschrift. Bij meer punten zou er een complexer correctievoorschrift moeten zijn.

Een goed correctievoorschrift moet functioneren als een herhaalbaar experiment: verschillende beoordelaars (of dit nu mensen of machines zijn) zouden op basis van de instructies tot dezelfde score moeten komen. In de praktijk blijkt dit vaak niet het geval.

Reproduceerbaarheid

Docenten nemen bij het nakijken onbewust en bewust veel context mee: het niveau van de klas, de inhoud van de toetsstof en eerdere antwoorden in de toets. Een AI doet dit niet; deze kijkt puur naar de tekst van het CV. Wanneer het CV niet compleet of niet deterministisch genoeg is, zorgt dit voor afwijkingen in de beoordeling.

Een goed correctievoorschrift:

- Wijst punten toe per onderdeel van het antwoord;
- Is abstract genoeg gedefinieerd zodat het punten toe kan kennen voor meerdere formuleringen;
- Beschrijft duidelijk eventuele afwijkingen die niet goed gerekend moeten worden;
- Geeft aan hoe er met spelfouten omgegaan moet worden.

Voorbeeld van een zwak correctievoorschrift:

Vraag: Wat is de meest gebruikte maatsoort in popmuziek? Leg ook uit wat de eigenschap van deze maat is in relatie tot muzieknoden.

CV: 4/4 + uitleg (Max score: 2)

Probleem: Het is onduidelijk wanneer de punten worden toegekend. Krijgt de leerling 1 punt voor "4/4" en 1 punt voor de uitleg? Wat zijn de eisen aan de uitleg? De docent weet welke eisen hij of zij stelt aan de uitleg, maar dit is niet duidelijk voor AI.

Voorbeeld van een verbeterd correctievoorschrift:

CV:

- *4/4 of vierkwartsmaat of een andere juiste benaming van de maat (1 punt)*
- *Uitleg waaruit blijkt dat de leerling begrijpt dat er 4 kwartnoten in een maat zitten (1 punt)*

Max score: 2

In het tweede voorbeeld zijn de criteria expliciet en de puntenverdeling helder. Hier stijgt de accurate van de AI aanzienlijk.

Waarom AI "nauwkeuriger" nakijkt dan de docent, maar de docent beter

We hebben de 20% afwijkende gevallen onderworpen aan een audit. Hierbij beoordeelde een derde partij het antwoord én het correctievoorschrift.

Uit deze audit kwamen twee belangrijke conclusies naar voren:

1. Het 'Slechte' CV (64% van de afwijkingen):

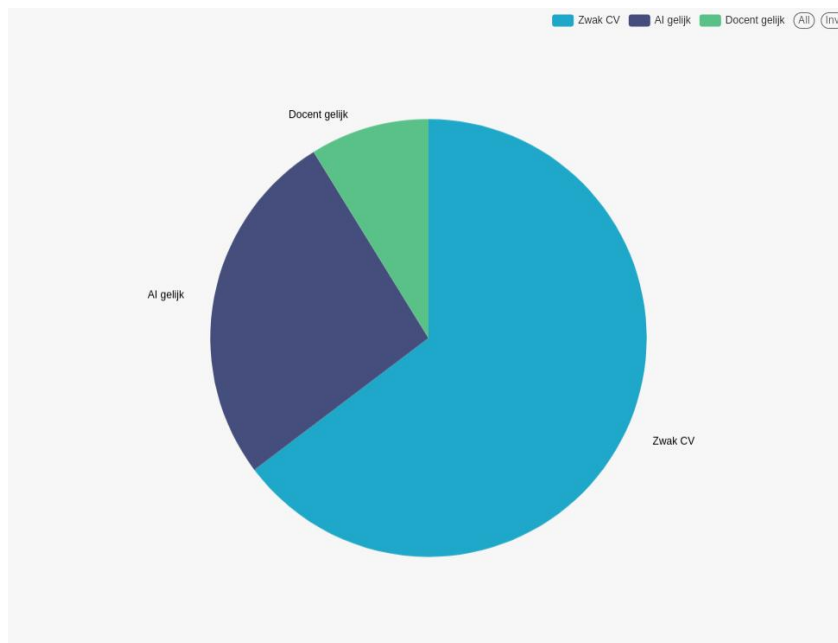
In meer dan de helft van de afwijkende gevallen was het correctievoorschrift niet toereikend. De AI hield zich hier vaak strikt aan de regels, terwijl de docent afweek van het CV om een 'eerlijkere' score te geven.

Conclusie: De docent kijkt hier 'menselijker' na, maar de AI kijkt technisch gezien 'nauwkeuriger' volgens de regels.

2. De 'Beter' Nakijker (36% van de afwijkingen):

Bij de overige afwijkingen, waar het CV wel duidelijk was, wees de audit uit dat de AI vaker de juiste score had toegekend dan de docent. De docent had namelijk volgens de audit in ongeveer 10% van de afwijkingen gelijk, en AI in 26% van de afwijkingen.

Docenten maakte in deze dataset vaker foute beslissingen zoals "juist" of "onjuist" omwisselen. Mogelijk kan dit verklaard worden door het verliezen van concentratie aan het einde van de beoordeling.



Figuur 4: Een cirkeldiagram van de afwijkingen (20% van de beoordelingen), verdeeld in 'Zwak CV' en 'Docent gelijk vs AI gelijk'.

Conclusie en advies

De titel "AI kijkt nauwkeuriger na dan een docent" bevat een essentiële nuance. Zolang correctievoorschriften goed geformuleerd zijn, is AI consistent en nauwkeuriger dan de mens. Zijn de voorschriften vaag? Dan wint de docent op basis van menselijke interpretatie en context.

Wat betekent dit voor de toekomst van toetsing?

De resultaten van dit onderzoek vormen de basis voor de doorontwikkeling van ToetsPers. Wij zien een toekomst waarin de docent en AI samenwerken:

1. Investeren in het correctievoorschrift

Het schrijven van eenduidige, 'AI-proof' correctievoorschriften verhoogt niet alleen de accuraatheid, maar ook de kwaliteit van toetsing in het algemeen. Wij zien kansen bij het (met AI) assisteren van de docent met het schrijven van goede correctievoorschriften.

2. Slimme Workflows

We kunnen voorspellen wanneer de AI het waarschijnlijk goed heeft.

- *Geeft AI 0 of de volle punten?* Grote kans dat dit klopt (automatiseren).
- *Geeft AI een deelscore?* Hier is menselijke controle van toegevoegde waarde.

3. Close the loop

Analyseer de toetsresultaten en gebruik *p*-, *rit*- en *rir*-waarden om afwijkende vragen te vinden. Controleer de antwoorden van de leerlingen en verbeter het cv voor de volgende keer.

Door transparant te maken wanneer een docent moet ingrijpen, transformeren we het nakijkwerk van een repetitieve taak naar een kwaliteitscontrole.

Over ToetsPers

ToetsPers ontwikkelt productiviteitssoftware voor toetsing. Onze missie is het verlagen van de werkdruk voor docenten door slimme inzet van technologie, zonder de menselijke maat uit het oog te verliezen.

Meer weten of een demo aanvragen? Ga naar www.toetspers.nl