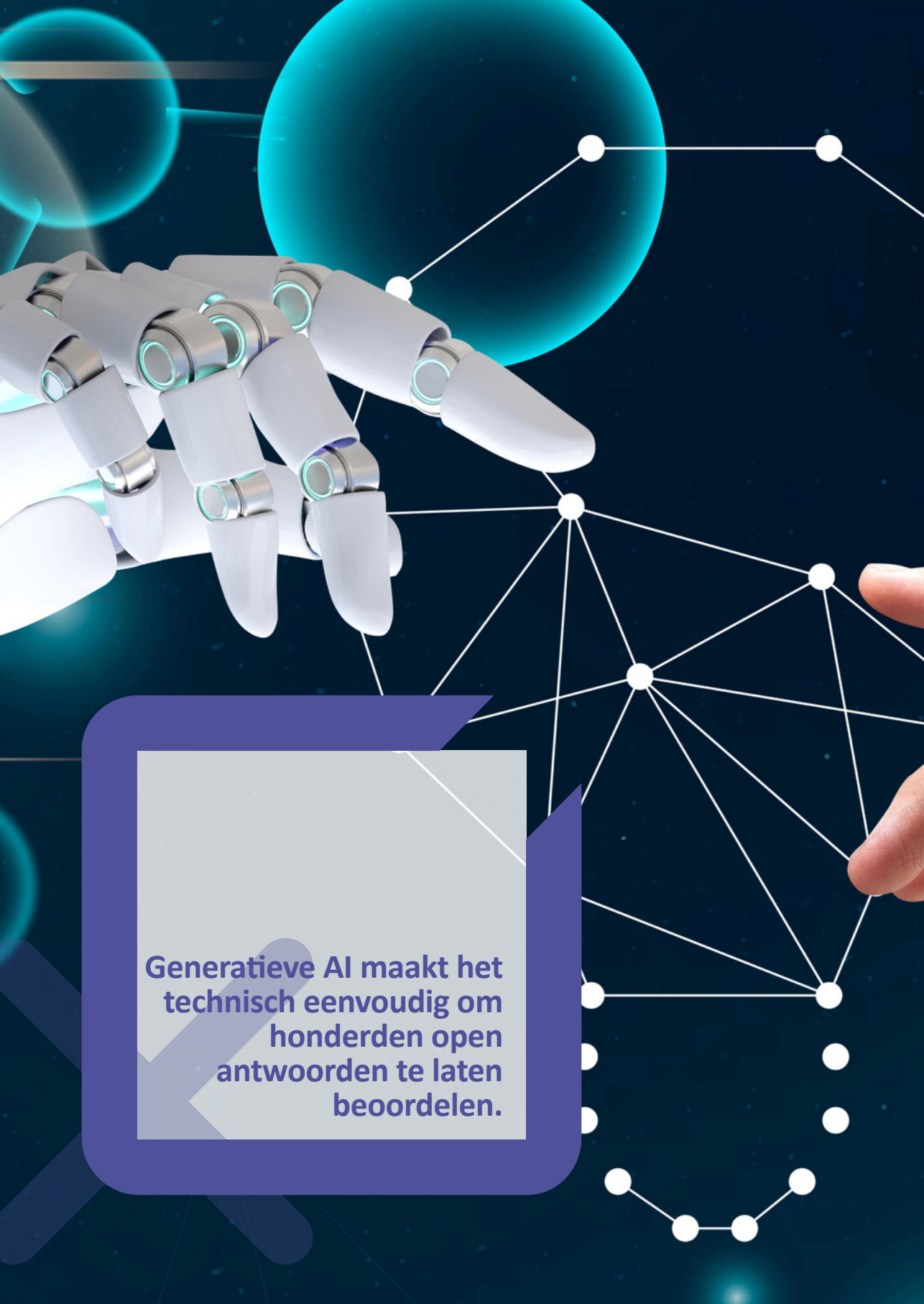


Generatieve AI maakt het technisch eenvoudig om honderden open antwoorden te laten beoordelen.

AI als tweede beoordelaar: wanneer kan de docent stoppen met alles zelf nakijken?

■ **Arnoud Engelfriet**
Mr. ir. Arnoud Engelfriet is Chief Knowledge Officer bij adviesbureau ICTRecht en docent aan de Vrije Universiteit. E-mail: a.engel@vu.nl

■ **Dr. ir. Silvester Draaijer**
Dr. ir. Silvester Draaijer is programma manager van het VU Education Lab van de Vrije Universiteit Amsterdam. E-mail: silvester.draaijer@vu.nl

■ **Luke Korthals**
Luke Korthals MSc is promovendus aan de afdeling Psychologie van de Universiteit van Amsterdam (UvA). E-mail: l.korthals@uva.nl

Zeshonderd studenten maken een tentamen met open vragen. De docent heeft een duidelijk correctievoorschrift, maar het nakijken kost dagen. Een taalmodel kan dezelfde antwoorden in korte tijd beoordelen. De voor de hand liggende werkwijze lijkt dan: laat de docent eerst nakijken, vergelijk zijn beoordelingen met de beoordeling van het model, verbeter waar nodig het correctievoorschrift en laat het model daarna steeds meer van het werk doen.

Maar wanneer mag de docent daadwerkelijk stoppen met ieder antwoord zelf te lezen? Dat is niet alleen een juridische vraag. Het is in de eerste plaats een toetskundige vraag: wanneer is er voldoende reden om erop te vertrouwen dat het model dezelfde beoordelingsstandaard toepast als de docent, en hoe organiseer je menselijke controle op de gevallen waarin dat waarschijnlijk niet zo is?

De vraag ná de AI Act

Joosten-ten Brinke en Beekman (2026) bespraken in het vorige nummer van Examens wat de EU AI Act betekent voor de toetspraktijk. Hun kernboodschap: AI bij toetsing is niet verboden, maar bij toepassingen die gevolgen hebben voor cijfers, studievoortgang of diplomering gelden zwaardere eisen en blijft betekenisvolle menselijke betrokkenheid noodzakelijk.
Wij onderzoeken in dit artikel één specifieke manier om die menselijke betrokkenheid vorm te geven: AI als tweede corrector. De docent kijkt eerst zelf een deel van de open antwoorden na aan de hand van het gebruikelijke correctievoorschrift. Een taalmodel beoordeelt daarna dezelfde antwoorden onafhankelijk. De verschillen worden vervolgens gebruikt om te onderzoeken waar het model, het correctievoorschrift of de vraag zelf de oorzaak is van de gevonden verschillen. Waar nodig wordt het

voorschrift aangescherpt en krijgt het model voorbeelden van eerder beoordeelde antwoorden.

Pas daarna komt de stap waar de mogelijke tijdwinst zit: als bij een vraag voldoende overeenkomst in scoring bestaat tussen de docent en het model, kan het model de overige antwoorden beoordelen. De docent leest dan niet meer ieder antwoord opnieuw, maar controleert vooraf geselecteerde gevallen: bijvoorbeeld antwoorden waarbij het model instabiel beoordeelt, vragen waarop eerder veel afwijkingen voorkwamen, studenten rond de zak-slaaggrens en daarnaast eventueel een willekeurige steekproef ter kwaliteitscontrole.

Het beoogde eindmodel is dus niet: AI kijkt na en de docent zet een vinkje. Het is een geleidelijke verschuiving van volledig menselijk nakijken, via dubbel beoordelen en vergelijken, naar AI-beoordeling met gerichte menselijke controle. De docent blijft verantwoordelijk voor de beoordelingsstandaard, maar hoeft niet noodzakelijk ieder individueel antwoord zelf te lezen.

Daarmee ontstaat de praktisch nog nauwelijks beantwoorde vraag: wanneer is er voldoende grond om die stapte maken? Als de docent iedere AI-beoordeling volledig opnieuw moet uitvoeren, levert automatisering immers nauwelijks tijdwinst op. Maar alleen afspreken dat een docent “in the loop” moet blijven, zegt evenmin hoeveel controle nodig is en waarop die controle zich moet richten.

De uitdaging is daarom een procedure te ontwerpen waarin menselijke aandacht terechtkomt op de antwoorden waar zij werkelijk waarde toevoegt. Dat is tegelijk een juridische en een toetskundige vraag.

Niet ‘leren zoals de docent’, maar kalibreren

Voor die eerste fase gebruiken we bewust het woord kalibreren. Het is verleidelijk te zeggen dat het taalmodel eerst “leert hoe de docent nakijkt”. Dat suggereert meer dan er technisch meestal gebeurt. Bij gebruik van een generatief taalmodel wordt doorgaans geen nieuw model getraind. Het systeem krijgt het correctievoorschrift, instructies en eerder door de docent beoordeelde antwoorden als voorbeelden.

De docent en het taalmodel beoordelen vervolgens dezelfde antwoorden onafhankelijk van elkaar. Hun verschillen zijn daarbij geen ruis die zo snel mogelijk moet verdwijnen, maar juist de belangrijkste informatiebron. Ze kunnen laten zien dat het model een criterium anders interpreteert, dat het correctievoorschrift ruimte laat voor meerdere lezingen of dat een vraag zich überhaupt slecht leent voor consistente beoordeling.

Dat proces heeft al waarde voordat er één minuut nakijktijd wordt bespaard. Het dwingt namelijk tot explicitering van de beoordelingsstandaard. Een rubric die voor een ervaren docent “vanzelf spreekt”, blijkt voor een tweede beoordelaar soms helemaal niet eenduidig. Dat probleem is niet uniek voor AI. Twee menselijke beoordelaars kunnen eveneens verschillend interpreteren hoeveel punten een bepaalde redenering verdient. De introductie van een taalmodel maakt die beoordelaarsvariantie alleen sneller zichtbaar. Vergelijk Engelfriet (2026, te verschijnen) over een vergelijkbare aanpak bij juridisch-wetenschappelijk onderzoek.

Daarom stellen wij voor om een taalmodel in eerste instantie niet te behandelen als een automatische nakijkmachine, maar als een tweede beoordelaar. Voor het vergelijken van beoordelaars beschikt de toetskunde al lang over begrippen en methoden en het vergelijken met beoordelingen van taalmodellen voegt daar informatie aan toe.

Wat een casus laat zien

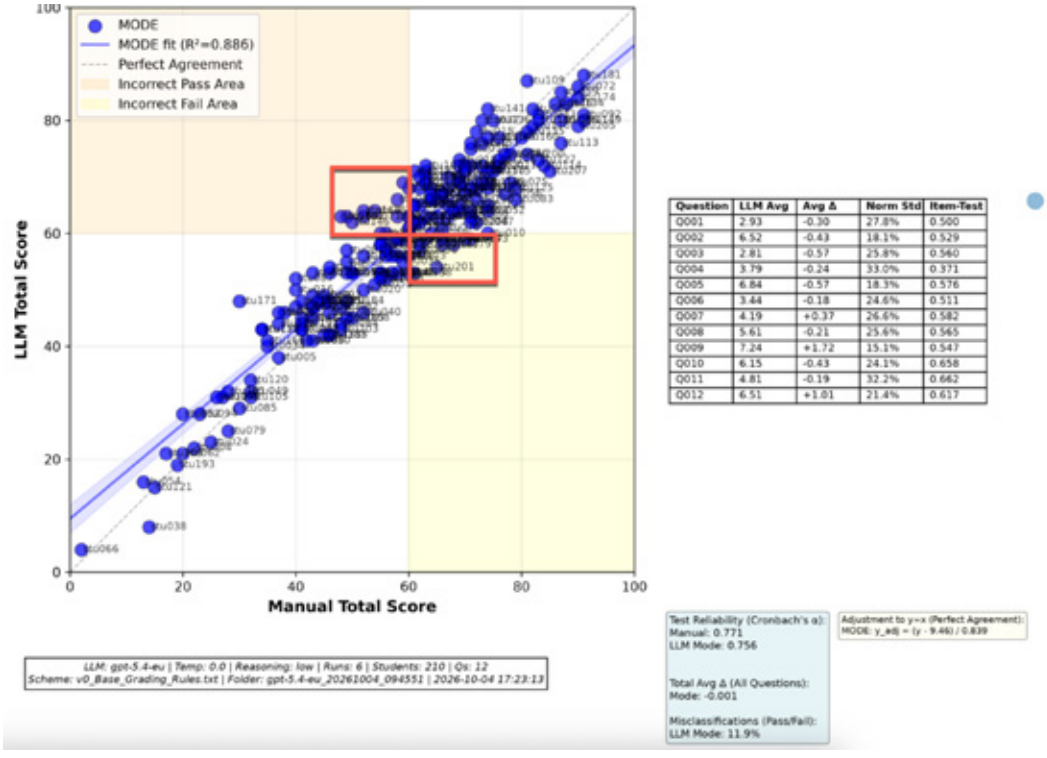
Bij een bedrijfskundig tentamen aan een universiteit in Amsterdam zijn handmatige beoordelingen vergeleken met beoordelingen door een

taalmodel. De analyse omvatte twaalf open vragen van 191 studenten. Ieder antwoord werd meerdere malen door het model beoordeeld; de moduswaarden van de taalmodelscores werden vervolgens met de docentbeoordelingen vergeleken.

Voor de **toets als geheel** (toetsniveau) levert dit een hoge samenhang op tussen de handmatige score en de score van het taalmodel (r=0,94) en eenzelfde gemiddelde score, zoals te zien is in onderstaande scatterplot van Figuur 1. Dit betekent dat het taalmodel heel consistent en met dezelfde strengheid scoort, vergelijkbaar met de handmatige beoordeling.

De analyse laat verder zien dat ongeveer 11 procent van de studenten op basis van de modelscore aan de andere kant van de zakslaaggrens terechtkomen dan op basis van de handmatige score (misclassificatie - de studenten in de rode vakken in Figuur 1). Een dergelijk percentage misclassificaties is gebruikelijk bij tentaminering, maar wordt in de praktijk verder genegeerd als indicator voor toetskwaliteit. We hebben gezien dat we aan de hand van een steekproef, en met in het oog houden van de onderliggende aannames en analyse, al een redelijke schatting kunnen maken van het aantal studenten dat binnen het bereik van de foutieve classificatie tussen 'geslaagd' en 'gezakt' valt. Deze studenten moeten op zijn minst handmatig worden beoordeeld. Als zo'n schatting niet mogelijk is, is het een verstandige keuze om álle studenten met een cijfer tussen een 5 en een 6 handmatig na te kijken.

Op **vraagniveau** komt een gevarieerd beeld naar voren. In onderstaande Figuur 2 van een methodologietentamen tonen we twee



Figuur 1. Scatterplot van handmatige scores en taalmodelscores op een bedrijfskunde tentamen.

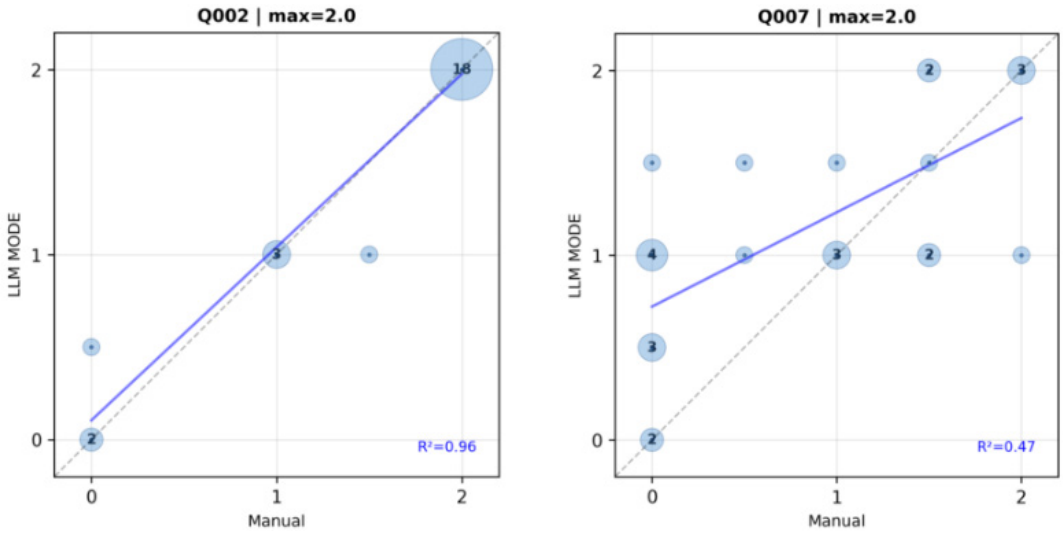
typische voorbeelden. In de linkerfiguur zien we een vraag waarbij de handmatige score en taalmodel sterk correleren en dezelfde gemiddelde score opleveren. Rechts zien we een vraag waarbij deze correlatie laag is en de gemiddelde score afwijkt. Dit betekent dat de linkervraag door het model al bijna identiek gescoord wordt als door de docent. Hier zou de docent kunnen stoppen met nakijken van die vraag en het model elke student scoren. Bij de rechtervraag scoort het model heel anders dan de docent. Hier is aanvullende inspectie nodig. De docent moet deze vraag naar alle waarschijnlijkheid zelf scoren.

Dat betekent ook dat een systeem niet simpelweg na een algemene proef "goedgekeurd" of "afgekeurd" kan worden voor alle open vragen. Een vraag kan geschikt blijken voor AI-ondersteund nakijken, terwijl de volgende vraag volledig bij een menselijke beoordelaar moet blijven. Daar past onmiddellijk een belangrijke beperking bij. In deze analyse fungeerde één menselijke corrector als referentie. Die corrector is zelf geen foutloze gouden standaard. Een verschil tussen docent en model is dus niet automatisch een fout van het model. Voor steviger vervolgonderzoek is vergelijking met een tweede menselijke beoordelaar noodzakelijk. Dit zal meer inzicht geven in mogelijke verklaringen van de verschillen tussen menselijke beoordelaars en AI-beoordelingen.

Eerst meekijken zonder gevolgen

Een verstandige invoering begint daarom met een fase waarin het model meekijkt zonder dat zijn oordeel gevolgen heeft.

De docent beoordeelt de antwoorden van een steekproef van studenten hiertoe op de gebruikelijke wijze. Het model beoordeelt dezelfde antwoorden parallel, maar de docent ziet die beoordeling niet en baseert zijn score er niet op. Zo ontstaat voor iedere vraag een dataset waarmee de overeenstemming kan worden onderzocht. Het voordeel daarvan is dat het



Figuur 2. Voorbeeld bubbelplots van twee vragen van een ander tentamen waarbij de handmatige score (manual) en het taalmodel (LLM MODE) sterk correleren (links) en een vraag waarbij de correlatie laag is. De linker vraag suggereert dat het model en de docent zeer vergelijkbaar beoordelen. De rechtervraag laat zien dat de score-range van het taalmodel veel lager is dan van de handmatige score en er heel weinig samenhang is.

risico voor studenten in de steekproef geheel wordt losgekoppeld van scoren met het taalmodel. Tegelijk ontstaat precies de informatie die later nodig is om te bepalen welke delen van het nakijkwerk aan het model kunnen worden toevertrouwd.

Die fase geeft inzicht in toevallige en structurele verschillen. Een structureel verschil kan volgens ons vier oorzaken hebben: de vraag is onduidelijk geformuleerd en staat uiteenlopende redelijke beoordelingen toe, het correctievoorschrift is onvoldoende precies, de handmatige beoordeling bevat fouten of het taalmodel is niet in staat om goed te evalueren. In de fase na de steekproef kan een docent alleen nog de eventuele incorrect handmatige beoordelingen aanpassen of het correctievoorschrift verbeteren om een grotere overeenstemming te bereiken tussen de handmatige score en de taalmodel score.

Mocht het aanpassen van incorrecte handmatige scores en het correctievoorschrift geen verbetering laten zien, dan is de juiste conclusie: deze vraag laten we niet door AI beoordelen.

Daarna selectief meekijken

Na voldoende kalibratie op basis van de steekproef kan de werkwijze worden omgedraaid. Het model beoordeelt dan de overige antwoorden van de vragen die het model goed kan scoren en menselijke beoordeling wordt gericht ingezet.

Bij deze beoordeling moet de meeste menselijke aandacht gaan naar antwoorden waarvan vooraf aannemelijk is dat beoordeling lastig is. Daarvoor kan een geavanceerde routering worden gebruikt. Er zijn verschillende signalen denkbaar die vragen weer voorleggen aan docenten om te beoordelen.

Een eerste signaal is de vraag zelf. Als tijdens de kalibratie blijkt dat docent en model op een bepaalde vraag vaak uiteenlopen, moet deze vraag volledig handmatig worden gescoord.



Als tijdens de kalibratie blijkt dat docent en model op een bepaalde vraag vaak uiteenlopen, moet deze vraag volledig handmatig worden gescoord.

Een tweede signaal is instabiliteit van het model. Een generatief model kan hetzelfde antwoord meerdere malen beoordelen en daarbij verschillende scores geven. Die variatie is geen perfecte maat voor onzekerheid, maar wel een bruikbaar routeringssignaal: als het model het kennelijk niet met zichzelf eens is, moet een mens kijken. Korthals et al. (2026) onderzoeken juist deze combinatie van zelfconsistentie en selectieve menselijke beoordeling.

Wanneer is een score instabiel?
Een taalmodel geeft niet noodzakelijk iedere keer dezelfde beoordeling. Als dezelfde vraag, hetzelfde correctievoorschrift en hetzelfde studentenantwoord

meerdere malen onafhankelijk aan het model worden voorgelegd, kan bijvoorbeeld achtereenvolgens 4, 4, 3, 4 en 2 punten worden toegekend.
Die herhaling kan worden gebruikt als onzekerheidssignaal. De meest voorkomende score — de modus, in dit voorbeeld 4 — kan als modelscore worden genomen. Tegelijk kijken we hoe vaak het model tot diezelfde score kwam en hoe sterk de overige scores daarvan afwijken. Vijfmaal dezelfde score geeft meer vertrouwen dan een reeks sterk uiteenlopende beoordelingen.Bij onvoldoende overeenstemming wordt het antwoord niet automatisch afgehandeld maar doorgestuurd naar een docent. Welke grens daarvoor geschikt is — bijvoorbeeld als minder dan 60% van de beoordelingen gelijk is — moet vooraf worden vastgesteld en in de kalibratiefase worden getoetst. Self-consistency is daarmee geen bewijs dat een score juist is, maar een manier om antwoorden te herkennen waarbij menselijke beoordeling waarschijnlijk de meeste waarde toevoegt (Korthals e.a. 2026).

Een derde signaal is de beslisgrens. Bij een student die ruim voldoende of ruim onvoldoende staat, heeft een klein beoordelingsverschil per vraag geen gevolg voor de uitslag. Rond de cesuur ligt dat anders. Daar hoort menselijke aandacht geconcentreerd te worden. Zoals we aangaven bij de figuur 1, kan met de steekproef al een goede indicatie worden gegeven van het aantal studenten rond de zak-slaaggrens die extra aandacht behoeven. Die studenten kunnen dan gerouteerd worden naar de nakijker.

Daarbovenop blijft een willekeurige steekproef aan te bevelen. Niet omdat ieder AI-oordeel nogmaals moet worden nagekeken, maar om te controleren of het systeem zich buiten de bekende risicogevallen nog steeds gedraagt zoals tijdens de kalibratie.

Het resultaat is dus geen simpele overgang van “100 procent mens” naar “10 procent mens”. Het is een routeringsmodel: sommige antwoorden kunnen door, sommige worden vanwege vooraf bepaalde signalen naar een mens gestuurd en een klein willekeurig deel wordt ter controle beoordeeld.

Laat de menselijke beoordelaar eerst zelf kijken

Bij die menselijke herbeoordeling via het routeringsproces ontstaat een ander probleem: de modelscore zelf kan het oordeel van de docent beïnvloeden.

Wie eerst ziet dat de AI vier van de vijf punten toekent, beoordeelt het antwoord niet meer volledig onafhankelijk. De vier punten vormen een anker. Juist bij twijfel bestaat vervolgens de neiging om het voorstel van het systeem te bevestigen.
Daarom zou de menselijke beoordelaar bij gerouteerde antwoorden in beginsel niet eerst de AI-score moeten zien. Eerst volgt een zelfstandige beoordeling op basis van het antwoord en het correctievoorschrift. Pas daarna kunnen beide beoordelingen worden vergeleken.
Hier komen toetskunde en de juridische eis van effectief menselijk toezicht dicht bij elkaar. De AI Act noemt expliciet het risico dat mensen automatisch of overmatig op de uitvoer van een hoogrisicosysteem gaan vertrouwen. Een menselijke klik op “akkoord” is geen zinvol toezicht wanneer die mens feitelijk door het eerdere AI-oordeel is gestuurd.

Het juridisch kader: AI Act en AVG

De AI Act noemt systemen die worden gebruikt om leerresultaten te beoordelen expliciet als hoogrisico. Dat verandert niet doordat een docent alleen twijfelgevallen en steekproeven controleert: menselijk toezicht is juist een van de eisen die aan zo'n hoogrisicosysteem worden gesteld. Dit toezicht moet effectief zijn. Effectief wil niet zeggen “iedere beoordeling van iedere vraag moet uiteindelijk door een mens gebeuren”. Het gaat erom dat de

mens ingrijpt daar waar het uitmaakt, zoals bij studenten wiens beoordeling rond de cesuur uitkomt. In combinatie met de reeds bestaande waarborgen bij nakijken (zoals inzage of de gang naar de examencommissie) zijn de rechten van studenten zo afdoende geborgd.

Naast de AI Act verdient ook de AVG aandacht. Deze verbiedt volledig geautomatiseerde besluitvorming, en dat raakt ook het nakijken van examens en toetsen. De hier beschreven werkwijze behoudt echter menselijke tussenkomst; niet als slaafs nakijken wat de AI heeft beoordeeld, maar juist controle van die antwoorden waar tussenkomst betekenisvol is. Daarnaast vereist de AVG een menselijke herbeoordeling op verzoek, iets dat bij toetsen eigenlijk altijd al standaard aanwezig is. De analyse van de vorige alinea is hier evenzo van toepassing.
Een open aandachtspunt is wat te doen bij beoordelingen van studenten die ruim buiten de cesuur zitten. Soms kan ook het verschil tussen een 8 en een 9 van belang zijn, zoals bij een cum laude beoordeling, selectie voor een vervolgopleiding of een andere regeling waarin het precieze cijfer doorwerkt. Menselijke controle kan daarom niet uitsluitend worden gekoppeld aan de zak-slaaggrens. Beter is vooraf vast te stellen bij welke beslisgrenzen een afwijkende AI-score gevolgen kan hebben, en de menselijke controle daarop te richten. Daarnaast blijft een willekeurige steekproef nodig om te controleren of ook buiten deze zones geen onverwachte systematische afwijkingen ontstaan.

Conclusie: niet minder verantwoordelijkheid, maar gerichtere aandacht

Generatieve AI maakt het technisch eenvoudig om honderden open antwoorden te laten beoordelen. De moeilijke stap is bepalen wanneer we die beoordeling voldoende vertrouwen om menselijk werk daadwerkelijk te verminderen. Daarvoor is niet alleen een beter taalmodel nodig. Er is vooral een uitgekiend proces nodig. Begin met AI als tweede beoordelaar. Laat het systeem eerst in de schaduw meedraaien. Meet waar mens en model van elkaar verschillen. Verbeter vragen en correctievoorschriften waar dat kan. Houd ongeschikte vragen buiten de automatisering. En verplaats daarna menselijke aandacht naar de antwoorden waar een beoordelingsverschil daadwerkelijk risico oplevert, met name rond de zak-slaaggrens.

We bedanken de docenten van de VU voor het deelnemen aan het wetenschappelijk onderzoek en de studenten voor het geven van toestemming voor het gebruiken van hun data. ■

Referenties

- Engelfriet, A. (2026). *Systematisch jurisprudentieonderzoek met 'artificial intelligence'*, RM Themis (te verschijnen).
- Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). *Can large language models make the grade? An empirical study evaluating LLMs' ability to mark short answer questions in K-12 education.* Proceedings of the Eleventh ACM Conference on Learning @ Scale, 300–304.
- Joosten-ten Brinke, D., & Beekman, K. (2026). Toetsing in het AI-tijdperk: de betekenis van de EU AI Act voor de praktijk. *Examens, tijdschrift voor de toetspraktijk*, 23(3), 32–39.
- Korthals, L., Akrong, E., Geller, G., Rosenbusch, H., Grasman, R., & Visser, I. (2026). Towards reliable LLM grading through self-consistency and selective human review: Higher accuracy, less work. *Machine Learning and Knowledge Extraction, 8*(3), 74.
- Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence, 4(*1), 47.
- Van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (Red.). (2014). *Toetsen in het hoger onderwijs.* Bohn Stafleu van Loghum.
- Verordening (EU) 2024/1689 tot vaststelling van geharmoniseerde regels betreffende artificiële intelligentie.
- Verordening (EU) 2026/1744 tot wijziging van onder meer Verordening (EU) 2024/1689.