PRELIFE: Adaptieve misconceptiedetectie en vraaggeneratie voor abiogenese-onderwijs

CESAR VAN HARDEVELD, Master Applied AI, Smart Education Lab, Nederland

Hoe leven kon ontstaan uit niet-levende materie is een van de grote open vragen van de wetenschap. Abiogenese (de studie van deze abiotische processen) is een verplicht onderdeel van het vwo-examenprogramma biologie, maar wordt in de praktijk oppervlakkig behandeld. Leerlingen ontwikkelen daarbij misconcepties en krijgen nauwelijks gerichte feedback op hun redeneerfouten, terwijl adaptieve leermiddelen voor dit domein ontbreken.

In dit onderzoek is via iteratieve ontwikkeling een adaptief intelligent tutorsysteem (ITS) ontworpen, geïmplementeerd en formatief geëvalueerd. Het systeem beoordeelt open antwoorden met een transparante, regelgebaseerde beoordelaar, registreert per concept welke denkfout mogelijk speelt, en genereert bij persistente denkfouten automatisch gerichte oefenvragen via een RAG-pipeline. De beoordeling verloopt lokaal en deterministisch, zonder doorgifte van leerlingdata aan externe servers.

De onderzoeksaanpak omvat een vergelijkende benchmark van beoordelingsmethoden, een validatie op los gehouden sets leerlingantwoorden, evaluatie van de retrieval- en misconceptiedetectiekwaliteit, een leesbaarheidsanalyse en een exploratieve gebruikerstest. De uiteindelijke beoordelaar velt een binair goed/fout-oordeel via een meerderheidsstem van drie gedecorreleerde deterministische maten: de gemiddelde semantische conceptdekking (cosine per concept met BGE-M3), de kernterm-aanwezigheid en de woordoverlap met het antwoordmodel. Op twee onafhankelijke sets van elk vijftig échte, handmatig gelabelde leerlingantwoorden (een kalibratie- en een los gehouden testset) haalt deze stem **92% (±7%)** bij vijfvoudige kruisvalidatie en 98% op de kleine held-out set (precisie 1,0). Ze combineert drie gedecorreleerde signalen die elk een ander type fout afvangen, is volledig reproduceerbaar en per concept herleidbaar. De misconceptiedetectie, een lokale LLM-stap, is scoped op de vooraf gelabelde denkfout per vraag: signaleert zij een denkfout, dan klopt dat in circa 79% van de gevallen (precisie). De recall van 47% is een ondergrens, doordat leerlingen op echte data vaak een ándere denkfout maken dan de ene voorziene; een leerlingbevestiging vangt de resterende valse detecties af. Een LiNT-leesbaarheidsanalyse van de feedback die leerlingen daadwerkelijk zien, laat zien dat de regelgebaseerde hint grotendeels binnen de vwo-streefzone valt (71%), terwijl de twee LLM-gegenereerde teksten te complex zijn (bevestiging 43%, misconceptie-uitleg 16% binnen de zone) ondanks een expliciete vwo-instructie in de prompt. De gebruikerstest bevestigde dat de basisinteractie zelfstandig doorlopen kan worden.

Het resultaat is een werkend prototype dat privacy-compliant opereert binnen de beperkingen van CPU-only hardware. Vervolgonderzoek richt zich op grootschaligere evaluatie met vakdocenten en leerlingen.

Additional Key Words and Phrases: adaptief tutorsysteem, regelgebaseerde antwoordbeoordeling, concept-coverage, Bayesian Knowledge Tracing, spaced repetition, abiogenese, privacy-compliant AI

Author's Contact Information: Cesar van Hardeveld, Master Applied AI, Smart Education Lab, Amsterdam, Nederland, cees.van.hardeveld@hva.nl.

1 Introductie

1.1 Context

Het overbrengen van wetenschappelijk complexe onderwerpen aan jongeren is een maatschappelijke uitdaging. Vakken als biologie en scheikunde behandelen steeds vaker vraagstukken die interdisciplinair zijn, alleen via indirecte aanwijzingen te onderbouwen en moeilijk intuïtief zijn te doorgronden [1]. Klassikale instructie biedt hiervoor beperkte ruimte: de leerstof wordt afgestemd op de gemiddelde leerling, terwijl de spreiding in voorkennis en tempo juist bij complexe onderwerpen groot is [2]. Vanuit die uitdaging opereert het PRELIFE-consortium (*Pathways, Reactions, and Environments leading to LIFE*), een NWO-gefinancierd samenwerkingsverband van Nederlandse onderzoeksgroepen naar de oorsprong van leven, met naast fundamenteel onderzoek een expliciete opdracht tot kennisoverdracht via het onderwijs [1], waaraan het Smart Education Lab van de HvA als deelnemende groep bijdraagt.

Abiogenese is het vraagstuk hoe leven kon ontstaan uit niet-levende materie. Het is in het Nederlandse voortgezet onderwijs opgenomen als subdomein F4 binnen het vwo-examenprogramma biologie [3]. De eindterm is smal: leerlingen moeten kunnen uitleggen welke theorie het ontstaan van leven verklaart, met aandacht voor chemische evolutie, zelforganisatie en de eerste levende cellen [3]. Dat de eindterm smal is, hangt samen met het feit dat abiogenese wetenschappelijk nog niet volledig is opgehelderd: de processen die tot het eerste leven leidden zijn nooit direct waargenomen. Toch doet dit weinig recht aan de complexiteit van het onderwerp, waarin chemische, geologische en biologische processen over miljarden jaren op elkaar inwerken [1]. Een begrip dat beperkt blijft tot het reproduceren van één theorie laat leerlingen achter zonder het conceptuele kader om verwante vraagstukken in biochemie, evolutiebiologie en geologie te doorgronden en om te redeneren over niet direct waarneembare processen [4]. Het beoogde begripsniveau blijft daarbij binnen de SLO-eindterm: leerlingen hoeven de scheikundige details van prebiotische synthese niet te beheersen, maar wel het causale redeneerpatroon te kunnen volgen en toepassen dat aan de gekozen theorie ten grondslag ligt, vergelijkbaar met hoe leerlingen leren redeneren over evolutie of celbiologie, zonder zelf op dat niveau onderzoek te doen.

Het onderwijzen van abiogenese stelt specifieke didactische eisen: het bewijs is indirect en de interdisciplinaire aard maakt het moeilijk toegankelijk [1]. Een deel van de leerlingen haalt bijvoorbeeld abiogenese en evolutietheorie door elkaar en beschouwt de oorsprong van leven als een enkelvoudige gebeurtenis [5, 6]. Zulke misconcepties zijn zonder gerichte aandacht moeilijk te corrigeren en vragen om individuele begeleiding die klassikale instructie niet kan bieden. Adaptieve intelligente tutorsystemen (ITS) kunnen dat mechanisme bieden: ze houden per leerling bij welke concepten beheerst zijn en selecteren de passende vervolgvraag, wat in eerder onderzoek tot een grotere leerwinst in minder tijd leidde [2]. Dit project onderzoekt de inzet daarvan voor abiogenese in de bovenbouw.

1.2 Probleem

Het smalle aandeel van abiogenese in het examenprogramma brengt het onderwijs in de praktijk terug tot reproductieve kennisoverdracht [3]. Het causaal redeneren en systeemdenken dat het onderwerp vraagt komen daarmee nauwelijks aan bod; internationaal is abiogenese vergelijkbaar ondervertegenwoordigd [6].

De denkfouten die hierdoor ontstaan, zijn zonder gerichte, individuele ondersteuning moeilijk te corrigeren [5, 6]. Die ondersteuning is voor een docent echter niet structureel uitvoerbaar: bij klassen van 25 tot 30 leerlingen is dat per leerling te tijdsintensief [7], waardoor maatwerk beperkt blijft tot grove niveaugroepen. Abiogenese is daarmee een representatief geval: de combinatie van wetenschappelijke complexiteit, veelvoorkomende misconcepties en beperkte

curriculumruimte geldt voor meer vwo-domeinen, waardoor deze proof of concept potentieel overdraagbaar is naar vergelijkbare vakgebieden.

1.3 Bestaand werk: adaptieve tutorsystemen en antwoordbeoordeling

Adaptieve, intelligente tutorsystemen zoals ASSISTments [7] en Carnegie Learning [8] schatten per leerling doorlopend in welke concepten al beheerst worden en stemmen de vervolgvraag af op die individuele antwoordgeschiedenis. De effectiviteit van deze aanpak is breed aangetoond, in het bijzonder in domeinen als wiskunde en taalvaardigheid [1]; de onderliggende techniek voor kennisschatting komt in sectie 2.2 aan bod. Voor het voortgezet onderwijs zijn daarnaast generieke oefenplatforms beschikbaar (Kahoot!, Quizlet, Snappet, Khan Academy, ALEKS).[1] Deze bieden soms beperkte adaptiviteit, maar zijn domeinonafhankelijk en bevatten geen vragenbank of conceptuele leerlijn voor abiogenese. Voor een onderwerp waarbij causaal redeneren centraal staat, volstaan gesloten vragen bovendien niet [4]: een leerling kan het juiste antwoord herkennen zonder de onderliggende redenering te beheersen, waardoor misconcepties in eigen formulering ongedetecteerd blijven. Open vragen zijn daarvoor noodzakelijk, maar worden door deze platforms doorgaans niet geanalyseerd.

Voor het beoordelen van open antwoorden zijn twee methoden relevant. Traditionele methoden (Automatic Short Answer Grading, ASAG) scoren een antwoord op oppervlakkige tekstkenmerken of semantische gelijkenis met een antwoordmodel, maar leggen niet uit *waarom* een antwoord fout is [9]. Recentere benaderingen zetten Large Language Models (LLMs) in als beoordelaar en kunnen wel misconcepties benoemen [10, 11, 12], maar kennen twee beperkingen: hallucinatie van plausibel ogende, onjuiste informatie [13] en een oordeel dat doorgaans niet herleidbaar is tot expliciete criteria [12]. Beide beperkingen zijn uitgewerkt in sectie 2.2.

Geen van deze benaderingen biedt een volledige oplossing. Platforms missen domeinspecifieke content voor abiogenese; traditionele ASAG schiet tekort in feedbackdiepgang en misconceptieherkenning; LLM-judges zijn lastig reproduceerbaar en ondoorzichtig; en cloudgebaseerde taalmodellen roepen privacybezwaren op onder de AVG. Dit onderzoek verkent of een combinatie van lokale, regelgebaseerde open-antwoordbeoordeling, adaptieve kennisschatting en kennisbank-gebaseerde vraaggeneratie een privacy-compliant en uitlegbaar alternatief kan vormen.

[1] https://kahoot.com, https://quizlet.com, https://nl.snappet.org, https://nl.khanacademy.org en https://www.aleks.com.

1.4 Gat in de kennis

De afzonderlijke bouwstenen voor adaptief, formatief leren zijn in de literatuur goed onderzocht, maar hun combinatie voor een domein als abiogenese ontbreekt. Voor het Nederlandse voortgezet onderwijs bestaat geen domeinspecifiek intelligent tutorsysteem dat open antwoorden over abiogenese adaptief verwerkt. De openstaande vragen liggen op zes vlakken; een volledige onderbouwing per vlak staat in de projectwiki.[2]

(1) **Abiogenese als domein.** ITS- en misconceptie-onderzoek richt zich vooral op wiskunde, taal en de evolutietheorie in brede zin [7, 8, 14]; de eigen denkfouten en leerlijn van abiogenese [6, 5] zijn nog niet in een adaptief systeem vertaald.

(2) **Transparante beoordeling binnen de adaptieve lus** (de cyclus van vraag, antwoord, beoordeling en een daarop bijgestelde vervolgvraag). Open-antwoordbeoordelaars zijn vooral als losstaande evaluatietool onderzocht, niet als sturende component in de vraagselectie, en geven zelden per concept herleidbaar inzicht [12, 14].

(3) **Privacy-compliant lokale beoordeling.** Bestaande oplossingen verwerken antwoorden doorgaans via externe cloud-API's [15]; hoe een volledig lokale, AVG-conforme beoordeling zich in kwaliteit verhoudt tot die cloudoplossingen is nauwelijks onderzocht. Dit onderzoek toont een bruikbare nauwkeurigheid voor de lokale aanpak (sectie 3.5.1); een uitputtende vergelijking met cloud-API's valt buiten de scope.

(4) **Nederlandstalige beoordeling zonder gelabelde data.** Methoden zijn vrijwel uitsluitend voor het Engels ontwikkeld [9], terwijl Engelstalige modellen het Nederlands onderpresteren [16].

(5) **Inzetbaarheid zonder GPU.** Of compacte, lokaal draaiende modellen [17, 18] op GPU-loze schoolhardware voldoende kwaliteit halen, is een open empirische vraag.

(6) **Misconceptie-gestuurde vraaggeneratie binnen één lus.** Een aanpak waarbij een hardnekkige misconceptie automatisch een gerichte oefenvraag uit een externe kennisbasis genereert [19], geïntegreerd met kennisschatting en beoordeling tot één privacy-compliant systeem, is nog niet als geheel onderzocht.

[2] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/onderzoek/kenniskloof-analyse.md

1.5 Voorstel

Dit onderzoek introduceert de PRELIFE AI-module: een adaptief intelligent tutorsysteem voor abiogenese, gericht op vwo-leerlingen in de bovenbouw. De centrale onderzoeksvraag luidt:

> *Hoe kan een privacy-compliant adaptief intelligent tutorsysteem worden ontworpen en geëvalueerd dat vwo-leerlingen in de bovenbouw ondersteunt bij het leren van abiogenese, door open antwoorden automatisch te beoordelen, misconcepties te detecteren en op basis van persistente misconcepties gerichte oefenvragen te genereren?*

Om deze vraag te beantwoorden zijn vier deelvragen opgesteld:

(1) Welke methoden worden in vergelijkbare adaptieve tutorsystemen toegepast voor open-antwoordbeoordeling, misconceptiedetectie en gerichte vraagselectie, en wat is bekend over hun effectiviteit en beperkingen?
(2) Welke requirements vloeien voort uit de onderwijskundige context van abiogenese, de behoeften van de stakeholders en de geldende juridische en ethische kaders?
(3) Hoe kan een privacy-compliant adaptief tutorsysteem worden ontworpen dat open antwoorden lokaal beoordeelt, misconcepties detecteert en bij persistente denkfouten gerichte oefenvragen genereert via een domeinspecifieke kennisbasis?
(4) In hoeverre voldoet het ontwikkelde systeem aan de gestelde vereisten, gemeten via automatische metrics (labelaccuracy, retrieval-kwaliteit, misconceptiedetectie) en kwalitatieve evaluatie door vakdocenten en studenten als proxy voor vwo-leerlingen?

De module is iteratief ontwikkeld: in opeenvolgende ontwikkelcycli is telkens één functionaliteit toegevoegd of verbeterd, gevolgd door een gerichte evaluatie waarvan de uitkomst de volgende stap stuurde.

2 Achtergrond

2.1 Misconcepties bij abiogenese en conceptuele verandering

Abiogenese is een abstract onderwerp: de processen die tot het eerste leven leidden zijn nooit direct waargenomen en spelen zich af over een tijdschaal ver buiten de menselijke ervaring. Juist zulke domeinen zijn vatbaar voor hardnekkige denkfouten. Leerlingen hebben moeite met de enorme tijdschalen die aan evolutie en abiogenese ten grondslag liggen [20] en grijpen bij het verklaren van natuurlijke processen vaak terug op doelgerichte, teleologische redeneringen [21, 22]; specifiek voor abiogenese verwarren zij bijvoorbeeld het ontstaan van leven met de evolutie van reeds bestaand leven [5, 6].

Volgens de Conceptual Change Theory [23] zijn zulke misconcepties geen losse vergissingen, maar stabiele alternatieve denkkaders die niet vanzelf verdwijnen: zij moeten eerst expliciet worden gemaakt voordat een leerling ze kan bijstellen. Een deel berust op ontologische categoriefouten, waarbij een geleidelijk, meerstappig proces ten onrechte als een enkelvoudige gebeurtenis wordt opgevat [24]. Een leeromgeving voor abiogenese moet deze denkfouten daarom niet alleen vermijden, maar actief zichtbaar maken en gericht aanpakken. Deze theoretische basis verklaart waarom misconceptiedetectie een kernfunctie van het systeem is; de vijf onderscheiden misconceptie-patronen staan in sectie 3.4.3.

2.2 Gerelateerd werk en basiskennis

Deze sectie behandelt de bouwstenen waarop het ontwerp berust. Per bouwsteen komt niet alleen de gekozen techniek aan bod, maar ook de afweging tegen de alternatieven en de randvoorwaarden die de doorslag geven: lokale verwerking, weinig data en uitlegbaarheid. Dezelfde onderbouwing staat in gestructureerde vorm (eis, alternatieven, keuze) in de projectwiki.[3]

2.2.1 Adaptieve leerondersteuning. Leerlingen verschillen sterk in voorkennis, tempo en misconcepties, terwijl klassikale instructie zich noodgedwongen op de gemiddelde leerling richt. Klassikale lessen en generieke oefenplatforms geven wel structuur, maar reageren niet op het individuele antwoordgedrag; juist bij een misconceptiegevoelig onderwerp als abiogenese is dat een gemis. Een **intelligent tutorsysteem (ITS)** vult die tekortkoming wél: ITS-instructie haalt volgens de meta-analyse van [2] een leerrendement dat vergelijkbaar is met één-op-éénbegeleiding. Die bevinding is aan het onderzochte domein gebonden en niet zonder meer naar formatief abiogenese-onderwijs door te trekken, maar maakt een ITS hier wel het meest kansrijke vertrekpunt.

2.2.2 Kennisschatting per concept. Het systeem moet per leerling en concept inschatten wat al beheerst is, en dat met een beperkte vragenbank en weinig responsdata. Die dataschaarste is hier de doorslaggevende randvoorwaarde. **Item Response Theory** is nauwkeuriger, maar vereist grote, gekalibreerde itembanken [25], en **Performance Factor Analysis** modelleert successen en fouten apart tegen hogere parametervereisten [26]; beide vragen meer data dan beschikbaar is. **Bayesian Knowledge Tracing (BKT)** schat de beheersingskans per concept uit de antwoordgeschiedenis via vier interpreteerbare parameters: initiële beheersing, leerkans, gokkans en slipkans [27]. BKT valt daarmee niet op nauwkeurigheid, maar op passendheid bij de context: interpreteerbaar per concept, transparant richting docenten en spaarzaam met data. De aanname dat concepten onafhankelijk zijn en niet vervagen is wel een vereenvoudiging die bij opschaling heroverweging verdient.

[3] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/onderzoek/bouwstenen-en-methodekeuzes.md

2.2.3 Cognitieve vraagordening. Vragen moeten op een passend cognitief niveau komen én herleidbaar zijn naar didactische leerdoelen. Een ordening op empirische moeilijkheidsgraad heeft als voordeel dat ze op responsdata rust, maar blijft beschrijvend en zegt niets over het beoogde leerdoel. De **taxonomie van Bloom** [4] ordent leerdoelen juist inhoudelijk, van reproductie naar analyse, en koppelt de vraagselectie zo aan curriculaire doelen; ze sluit bovendien aan op de eindtermstructuur van het SLO-examenprogramma [3], wat het niveau interpreteerbaar maakt voor vakdocenten. De keuze weegt didactische traceerbaarheid dus zwaarder dan empirische precisie. Kanttekening is dat de niveau-indeling per vraag deels een interpretatie blijft en docentvalidatie behoeft.

2.2.4 Herhalingsplanning. Leerstof moet op het juiste moment terugkomen voor langetermijnretentie, zonder beheerste stof onnodig te herhalen. Vaste herhalingsintervallen zijn eenvoudig te implementeren, maar houden geen rekening met de actuele beheersing en leiden zo tot onnodige herhaling én te weinig aandacht voor zwakke concepten. **Spaced repetition** sluit hier beter op aan: verspreid oefenen op toenemende intervallen geeft volgens het *spacing effect* een betere retentie dan geblokt oefenen [28]. Voor een systeem dat beheersing per concept bijhoudt is dat het passendste uitgangspunt. De huidige implementatie benadert het principe met een vaste cooldown en benut het beheersingsafhankelijk groeiende interval nog niet volledig; dat is een bewuste vereenvoudiging voor het prototype.

2.2.5 Antwoordbeoordeling. Open antwoorden moeten lokaal, reproduceerbaar en per concept uitlegbaar worden beoordeeld, zonder gelabelde Nederlandse trainingsdata. Die eisen wringen, want de semantisch sterkste methoden voldoen juist het slechtst aan reproduceerbaarheid en uitlegbaarheid. **Automatic Short Answer Grading (ASAG)** is licht en lokaal, maar beoordeelt via tekstgelijkenis [9] en kent daardoor een vocabulaire-overlap- en beknoptheidsbias zonder toelichting. Een **LLM-as-a-judge** beoordeelt semantisch het rijkst: het werkt op vraag én referentieantwoord zonder trainingsdata [12] en kan misconcepties benoemen [11]. Juist die kracht botst echter met de eisen, want de uitvoer is niet-deterministisch, lastig te verantwoorden en traag op CPU. Een **regelgebaseerde beoordelaar op basis van concept-coverage** splitst het antwoordmodel in losse concepten en toetst per concept de semantische dekking met meertalige sentence-embeddings [29, 16], met een aanvullende *kernterm-controle* zodat on-topic maar inhoudelijk onjuiste antwoorden niet meeliften. Deze laatste is niet per definitie de nauwkeurigste, maar wel de enige die determinisme, lokale verwerking en uitlegbaarheid tegelijk haalt. De prijs is dat inhoudelijke vrije-tekstfeedback vervalt en sterk geparafraseerde antwoorden lastiger blijven; of dat opweegt tegen de gewonnen transparantie is een empirische vraag, getoetst in sectie 3.5.1 met ROUGE [30] en BLEU [31] als vergelijkingsbasis.

2.2.6 Misconceptiedetectie. Naast een goed/fout-oordeel is het waardevol te weten wélke onderliggende denkfout meespeelde bij een fout antwoord. Een zuiver regelgebaseerde aanpak zou voor elke denkfout een uitputtende set patronen vergen en breekt bij onverwachte formuleringen. Een taalmodel benoemt een misconceptie juist flexibel op basis van het antwoord, de vraag en het antwoordmodel [11]. Die flexibiliteit heeft een prijs: de uitvoer is niet-deterministisch [12] en een compact model neigt naar overdetectie, waarbij het ook een denkfout meldt waar eigenlijk alleen een los kennisgat speelt. Anders dan bij de scoring, waar determinisme bindend was, weegt hier de flexibiliteit zwaarder, mits de overdetectie wordt ingedamd. Daarom is gekozen voor taalmodelgebaseerde detectie, verankerd op het antwoordmodel en met een leerlingbevestiging als precisievangnet, zodat alleen erkende denkfouten meetellen. De theoretische grond staat in sectie 2.1; de implementatie volgt in sectie 3.5.3.

2.2.7 Domeinspecifieke vraaggeneratie. Het systeem moet nieuwe, domeinspecifieke oefenvragen kunnen genereren zonder gelabelde trainingsdata en op hardware zonder GPU. Fine-tuning zou de domeinkennis het diepst in het

model verankeren, maar vereist een gelabelde domeinset en is zonder GPU rekenkundig onhaalbaar. **Retrieval-Augmented Generation (RAG)** verlegt die kennis naar inferentietijd: het model krijgt de benodigde passages uit een externe kennisbasis aangereikt [19] en hoeft zelf geen domeinkennis in de gewichten te dragen. Onder de gestelde randvoorwaarden is RAG daarmee de haalbare keuze, met als keerzijde dat de kwaliteit van de gegenereerde vraag staat of valt met de dekking van de kennisbasis: ontbreekt een onderwerp daarin, dan levert retrieval geen bruikbare context. Voor Nederlandstalige retrieval zijn bovendien meertalige sentence-embeddings nodig, omdat Engelstalig getrainde modellen slechter presteren op het Nederlands [16].

2.3 Stakeholders en gebruikersonderzoek

De twee primaire gebruikersgroepen zijn vwo-leerlingen bovenbouw en vakdocenten biologie; een volledige stakeholderanalyse met Power–Interest-matrix staat in de projectwiki.[4] **Vwo-leerlingen** komen abiogenese voor het eerst tegen in de bovenbouwstof biologie (subdomein F4) [3]; de combinatie van weinig voorkennis en verhoogd risico op misconcepties suggereert dat een leeromgeving die denkfouten actief diagnosticeert en corrigeert meerwaarde biedt [5, 6]. **Vakdocenten** zijn gebruiker en validator tegelijk: zij beoordelen via het docentpanel of AI-gegenereerde vragen correct en didactisch geschikt zijn. Hun behoefte is kwaliteitscontrole zonder veel tijdsinvestering; een systeem dat de inhoudelijke eindverantwoordelijkheid bij de docent legt, sluit daarop aan (sectie 3.3.1).

2.4 Huidige situatie

Figuur 1 toont de huidige situatie: abiogenese-onderwijs zonder adaptieve ondersteuning, met de voornaamste knelpunten en hun gevolgen voor leerlingen.

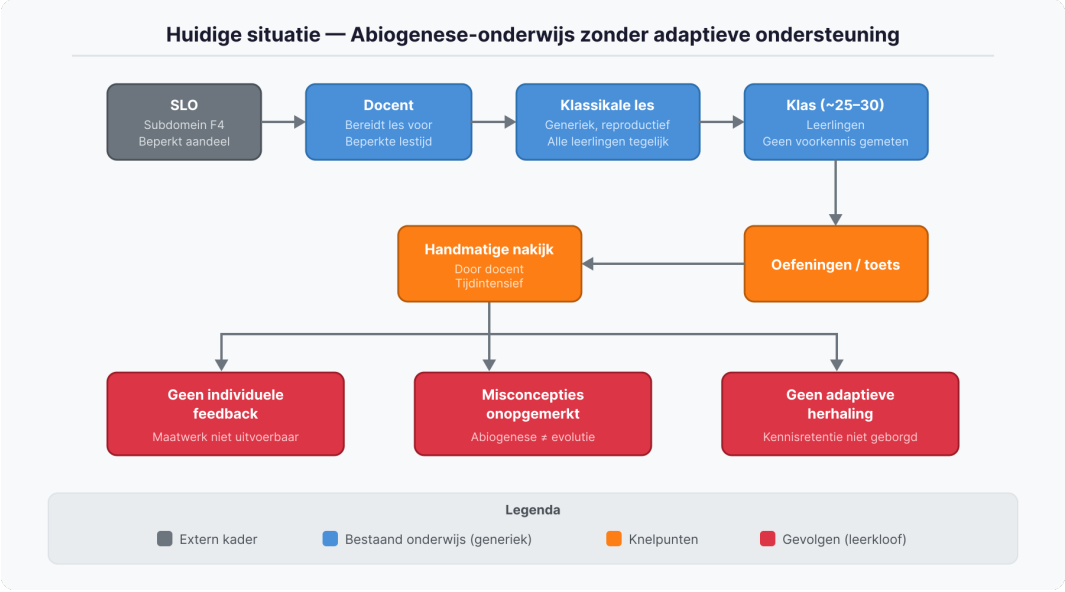


Fig. 1. Huidige situatie: abiogenese-onderwijs zonder adaptieve ondersteuning. Oranje blokken duiden knelpunten aan; rode blokken tonen de gevolgen voor leerlingen.

[4] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/onderzoek/context-en-stakeholderanalyse.md

2.5 Overwegingen voor het inzetten van AI

Het vraagstuk vereist vier functies tegelijk die geen niet-AI-alternatief combineert: (1) beoordeling van open antwoorden op betekenisniveau, (2) adaptieve vraagselectie op basis van een leerlingprofiel, (3) expliciete misconceptiedetectie, en (4) het genereren van gerichte oefenvragen bij persistente misconcepties, een combinatie die geen bestaand alternatief biedt (sectie 1.3).

AI wordt gericht ingezet, en bewust niet overal. De beoordeling draait op een regelgebaseerde, deterministische judge met meertalige sentence-embeddings: reproduceerbaar, per concept herleidbaar en niet aan een taalmodel uitbesteed. Een compact lokaal taalmodel komt pas in beeld voor de twee taken die zonder generatieve AI niet uitvoerbaar zijn: het flexibel benoemen van een misconceptie uit vrije tekst, en het genereren van nieuwe oefenvragen via retrieval (sectie 3.5.4). Doordat het antwoordmodel, de vraag en de opgehaalde passages als context in de prompt staan, is hiervoor geen domeinspecifiek trainingsproces nodig; voor abiogenese, zonder gelabelde Nederlandse trainingsdata, is dat een cruciale eigenschap.

Deze inzet kent reële beperkingen die het ontwerp mitigeert. De scoring is deterministisch; de resterende afhankelijkheid zit in de drempelwaarden en het embeddingmodel, geëvalueerd tegen menselijke ground truth (sectie 3.5.1). De BKT-parameters staan op standaardwaarden uit de literatuur [27] en zijn bij opschaling te kalibreren; de afhankelijkheid van de vragenbank (50 items) wordt gedeeltelijk opgevangen door RAG-vraaggeneratie met menselijke validatie. Dat generatie lokaal gebeurt volgt uit twee harde randvoorwaarden: verwerking van leerlingdata door externe clouddiensten is onder de AVG bezwaarlijk [15], en de productie-VPS (8 GB RAM, geen GPU) laat grotere modellen niet toe. Omdat de docent elke gegenereerde vraag valideert, raakt een incidentele generatiefout de leerling niet rechtstreeks.

Op basis hiervan is bewust gekozen voor een beperkte mate van autonomie: een *human-in-the-loop* conform mensgericht AI-ontwerp [32]. Het systeem automatiseert kennisschatting, antwoordbeoordeling en vraaggeneratie, maar laat beslissingen over leerdoelen en leerinhoud bij de docent. Gegenereerde vragen worden nooit automatisch aangeboden, maar doorlopen eerst de reviewworkflow in het docentpanel. Zo voldoet de module aan de eis van menselijk toezicht uit de EU AI Act [33]: leerlingen worden nooit op basis van AI-uitvoer beoordeeld, en gegenereerde inhoud vereist altijd goedkeuring. De bredere ethische afweging volgt in sectie 5.6.

3 Methodologie

3.1 Overzicht

Dit onderzoek volgt een *design science*-aanpak: ontwerpen, bouwen en evalueren van een AI-systeem voor een concreet onderwijsprobleem. De methodologie bestaat uit vijf fasen:

(1) **Literatuuronderzoek** naar adaptieve tutorsystemen, kennisschatting, automatische antwoordbeoordeling, misconceptiedetectie en RAG, als grondslag voor requirements en ontwerpkeuzes.
(2) **Requirements** volgens de MoSCoW-methode, op basis van probleemanalyse, literatuur en stakeholderanalyse.
(3) **Ontwerp en implementatie** van de module: vragenbank, kennisschatting (BKT), regelgebaseerde antwoordbeoordeling, misconceptieverwerking en RAG-vraaggeneratie.
(4) **Modelselectie en evaluatie**: benchmark en validatie van de regelgebaseerde beoordelaar op twee onafhankelijke sets échte leerlingantwoorden, aangevuld met evaluaties van misconceptiedetectie en retrieval-kwaliteit.
(5) **Gebruikers- en leesbaarheidsevaluatie**: een formatief gebruikersonderzoek met hardop-denkprotocol en interview toetst de bruikbaarheid, en een LiNT-analyse de begrijpelijkheid van de feedback.

3.2 Requirements

De requirements vloeien voort uit de probleemanalyse, het literatuuronderzoek en de stakeholderanalyse. Requirements die uitsluitend bij volledige productie-implementatie relevant zijn (governance, beheer, monitoring) vallen buiten de prototypescope en zijn niet opgenomen. De tabellen 1 tot en met 4 vatten de juridische, ethische, functionele en technische eisen samen.

Tabel 1. Juridische requirements

ID	Categorie	Requirement	MoSCoW	Bron
REQ 1.1	**Privacy & Databeheer**			
REQ 1.1.1		Leerlingantwoorden en masterprofiel blijven lokaal; geen opslag van herleidbare persoonsgegevens of overdracht extern. [15]	Must	AVG Art. 5
REQ 1.1.2		Het systeem maakt transparant dat AI wordt ingezet en gegevens lokaal worden verwerkt. [33]	Must	EU AI Act Art. 50
REQ 1.2	**Menselijke controle**			
REQ 1.2.1		Het systeem biedt menselijk toezicht: de docent ziet en corrigeert elk oordeel en blijft eindverantwoordelijk. [33]	Must	EU AI Act Art. 14
REQ 1.2.2		De beoordeling is uitsluitend formatief; AI-uitvoer draagt niet bij aan een summatief cijfer, als risicomitigatie ten opzichte van de classificatie in Bijlage III. [33]	Must	Eigen ontwerp (n.a.v. Bijlage III)

Tabel 2. Ethische requirements

ID	Categorie	Requirement	MoSCoW	Bron
REQ 2.1.1	**Fairness & Bias**	Open antwoorden met correcte betekenis maar afwijkende bewoording worden niet systematisch lager beoordeeld. [9, 12]	Must	Burrows 2015
REQ 2.2.1	**Uitlegbaarheid**	Het systeem geeft transparante feedback over waarom een antwoord incorrect is en informeert expliciet bij een gedetecteerde misconceptie. [33, 11]	Must	EU AI Act Art. 86 / Lit.
REQ 2.3.1	**Leerautonomie**	Het systeem stimuleert eigen redenering en ontmoedigt afhankelijkheid van AI-output. [34, 35]	Should	Black & Wiliam 1998

Tabel 3. Functionele requirements

ID	Categorie	Requirement	MoSCoW	Bron
REQ 4.1.1	**Adaptieve vraagselectie**	Vraagselectie volgt de BKT-beheersing per concept via drie Bloom-niveaus (drempels 0,40/0,70). [27, 4]	Must	Corbett 1994
REQ 4.1.2		Spaced repetition: cooldown van minimaal 5 vragen voor herhaling. [28]	Should	Cepeda 2006
REQ 4.1.3		Meerdere vraagtypes (meerkeuze, open, meerselectie, rangorde).	Could	Eigen ontwerp
REQ 4.2.1	**Misconceptieherkenning**	Gedetecteerde misconcepties worden per concept geregistreerd en sturen de vervolgvraagkeuze. [11, 5]	Must	McNichols 2024
REQ 4.2.2		Bij een actieve misconceptie krijgt de vraagkeuze prioriteit voor corrigerende vragen.	Must	Eigen ontwerp
REQ 4.2.3		Bij ≥ 2 bevestigde misconcepties per concept genereert het systeem een remediërende vraag via RAG.	Could	HvA Lab
REQ 4.2.4		Gegenereerde vragen worden alleen aangeboden na expliciete goedkeuring door de docent. [33]	Must	EU AI Act Art. 14
REQ 4.3.1	**Antwoordbeoordeling**	Open antwoorden worden op betekenisniveau beoordeeld: op semantische conceptdekking, niet op woordoverlap. [9, 12]	Must	Eigen ontwerp
REQ 4.3.2		Feedback uitsluitend in het Nederlands, beperkt tot maximaal drie zinnen. [36]	Must	Sweller 1988
REQ 4.4.1	**Vragenbank**	Minimaal 50 vragen over drie Bloom-niveaus en de PRELIFE-subthema's, getagd met subthema, niveau en misconceptielabel. [4, 3]	Must	SLO 2023

Tabel 4. Technische requirements

ID	Categorie	Requirement	MoSCoW	Bron
REQ 5.1.1	**Foutafhandeling**	Bij validatie- of runtimefouten retourneert de backend een gestructureerde foutcode; de beoordeling is deterministisch en vereist geen externe modelaanroep.	Must	Eigen ontwerp
REQ 5.2.1	**Beveiliging**	Communicatie via HTTPS; geen persoonsgegevens naar externe API's. [15]	Must	AVG Art. 25
REQ 5.3.1	**Performance**	Harde time-out op LLM-aanroepen voor de asynchrone vraaggeneratie; de UI toont een laadstatus tijdens het beoordelen van een antwoord. [37]	Must	Nielsen 1994
REQ 5.4.1	**Kwaliteitscontrole**	De beoordelaar levert een deterministisch, gestructureerd oordeel (label + gedekte concepten) dat bij gelijke invoer reproduceerbaar is. [9]	Must	Eigen ontwerp
REQ 5.4.2		Accuratesse wordt op een gelabelde testset geëvalueerd tegen losse tekstmetrieken en lokale LLM-judges. [12, 9]	Should	Eigen ontwerp
REQ 5.5.1	**Infrastructuur & RAG**	Backend via FastAPI; broncode openbaar op GitLab.	Must	Eigen ontwerp
REQ 5.5.2		RAG-generatie is asynchroon; kennisbasis-content vereist schriftelijke toestemming. [37, 38]	Must	Lit. / NEMO Kennislink

3.3 Concept

3.3.1 Value proposition. De module biedt meerwaarde voor drie partijen. De **leerling** krijgt op open vragen directe, inhoudelijke feedback, waar bestaande oefenmaterialen vrijwel uitsluitend gesloten vragen of statische tekst bieden; het cognitieve niveau van de vervolgvraag past zich aan op de beheersingshistorie (Bloom + BKT), en een gedetecteerde misconceptie krijgt een gerichte toelichting met bronanker. De **docent** ziet misconcepties per leerling via een gekleurde matrix met triggerende vraag en antwoord, en kan elke gegenereerde oefenvraag goedkeuren, afwijzen of aanpassen. Voor de **school** verloopt de beoordeling lokaal en regelgebaseerd: leerlinginvoer verlaat de schoolomgeving niet, en de module is AVG-conform en GPU-vrij inzetbaar.

3.3.2 Flow diagram. Figuur 2 toont de volledige pipeline in drie fasen: de quizcyclus met regelgebaseerde beoordeling en BKT-update (A), de RAG-pipeline bij een persistente misconceptie met docentvalidatie (B), en de adaptieve vraagkeuze uit de gecombineerde vragenbank (C).

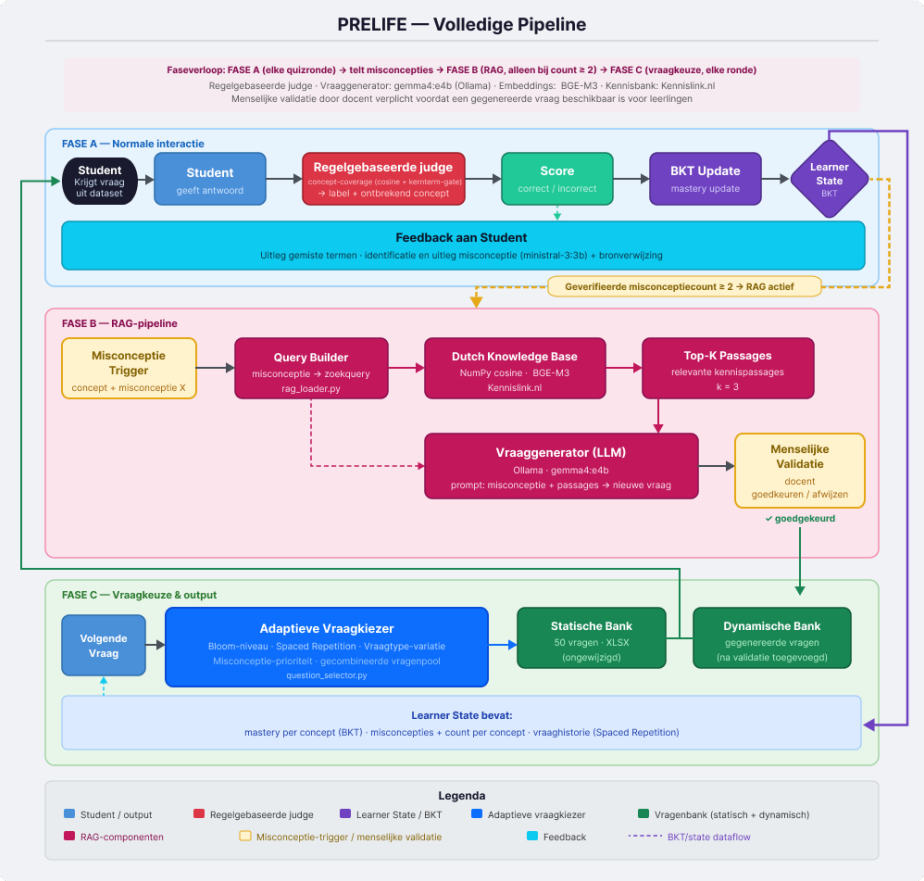


Fig. 2. Volledige pipeline in drie fasen. A: quizcyclus met regelgebaseerde beoordeling, BKT-update en misconceptieopslag. B: RAG-generatie bij twee bevestigde misconcepties met docentvalidatie. C: adaptieve vraagkeuze uit statische en dynamische vragenbank.

3.3.3 AI-breakdown. De AI-functionaliteit is opgesplitst in vier taken. Tabel 5 brengt per taak de mogelijke fouten, hun impact en de vereiste menselijke controle in kaart. De kolom *Level of Autonomy* volgt het mensgericht-AI-ontwerpkader van Kore [32]; kritieke onderdelen zoals vraaggeneratie zijn bewust *human-in-the-loop* ingericht.

In	Taak	Uit	Mogelijke fouten (FP/FN)	Impact	Toegestaan fout%	Mogelijke oplossingen	Level of Autonomy
Leerlingantwoord + vraag + antwoordmodel	Beoordeel binair via een meerderheidsstem van drie maten; lokaliseer zwakst gedekt concept	Label, gedekte concepten, feedbacktekst	**FP:** correct antwoord als incorrect bij afwijkende bewoording **FN:** incorrect antwoord als correct bij semantisch nabije fout	High (FN), Med (FP)	20%	Meerderheidsstem (cosine-dekking, kernterm-gate, unigram-BLEU); drempels gevalideerd op gelabelde testset; deterministisch	**Level 7:** regelgebaseerd; voert automatisch uit en toont het oordeel; herleidbaar per concept.
Gedetecteerde misconceptie + concept	Genereer uitleg waarom de denkfout onjuist is	Max. 3 zinnen uitleg op vwo-niveau	**FP:** uitleg op incorrect gedetecteerde misconceptie **FN:** uitleg te vaag of niet aansluitend	Med	30%	Systeemprompt met vwo-niveau; uitleg verankerd op antwoordmodel; leerling verifieert de denkfout met ja/nee	**Level 5:** AI genereert, leerling bevestigt; alleen bevestigde denkfouten leiden tot vervolgactie.
Concept + misconceptiebeschrijving	Haal kennispassages op uit de Kennislink-kennisbasis via dense retrieval	Top-3 relevante chunks (tekst + bron-URL)	**FP:** irrelevante chunks **FN:** relevante passage gemist	Med	25%	Gecombineerde concept + misconceptie-query; meertalig embedmodel	**Level 7:** intern proces; output alleen zichtbaar voor docent.
Top-3 chunks + misconceptie + concept	Genereer een nieuwe open oefenvraag gericht op de denkfout	Open vraag + antwoordmodel (status: *pending*)	**FP:** inhoudelijk onjuiste of ongeschikte vraag **FN:** vraag sluit niet aan; te generiek	High	0% (voor leerlingen)	Docent-in-the-loop: elke vraag status *pending*, pas beschikbaar na goedkeuring; tekst aanpasbaar	**Level 5:** AI genereert een voorstel; wordt pas ingezet na goedkeuring door de docent.

Tabel 5. AI-breakdown PRELIFE-systeem: per AI-taak mogelijke fouten, impact, toegestaan foutpercentage en mate van autonomie.

3.3.4 Design patterns. De ontwerpkeuzes zijn gebasseerd op de *People + AI Guidebook* van Google PAIR [39], een verzameling principes en patronen voor mensgerichte AI. Het sterkst herkenbaar is het principe *design for the appropriate level of user autonomy*: elke taak in tabel 5 kreeg een expliciet autonomieniveau, met een docent-in-the-loop voor vraaggeneratie. Daarnaast zijn onder meer *determine how to show model confidence, if at all* (de binaire judge toont bewust geen misleidend percentage), *make precision and recall trade-offs carefully* (recall-gerichte misconceptiedetectie met de docent als vangnet) en *provide a way forward* bij fouten (fallbacks wanneer het lokale model uitvalt) toegepast. De volledige toewijzing van patronen aan systeemonderdelen staat in de projectwiki.[5]

3.3.5 Testopzet en evaluatiemethoden. De evaluatie bestaat uit een kwantitatieve benchmark van de beoordelaar en een kwalitatieve gebruikerstest.

Benchmark van beoordelaars. Drie typen zijn binair vergeleken op labelaccuracy: losse tekstmetrieken (globale cosine, ROUGE-L [30], BLEU [31]), vier lokale LLM-judges en de regelgebaseerde beoordelaar. De vergelijking draait op een los gehouden testset van vijftig échte, handmatig gelabelde leerlingantwoorden (25 correct, 25 incorrect); de drempels zijn uitsluitend op een onafhankelijke kalibratieset van vijftig antwoorden gekozen. Omdat de regelgebaseerde judge geen getrainde parameters kent, vallen de trainings- en validatierol samen in één kalibratieset (A); set B fungeert

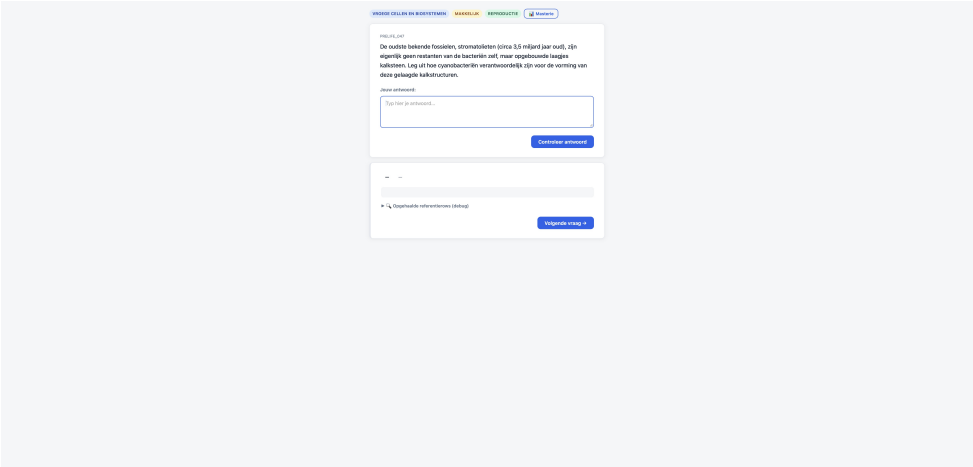
[5] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/ontwerp-en-uitwerking/design-patterns.md

als onafhankelijke testset. Als generalisatiemaat is bovendien een vijfvoudige kruisvalidatie over de honderd antwoorden uitgevoerd, met de drempels per fold op de kalibratiehelft (sectie 3.5.1).

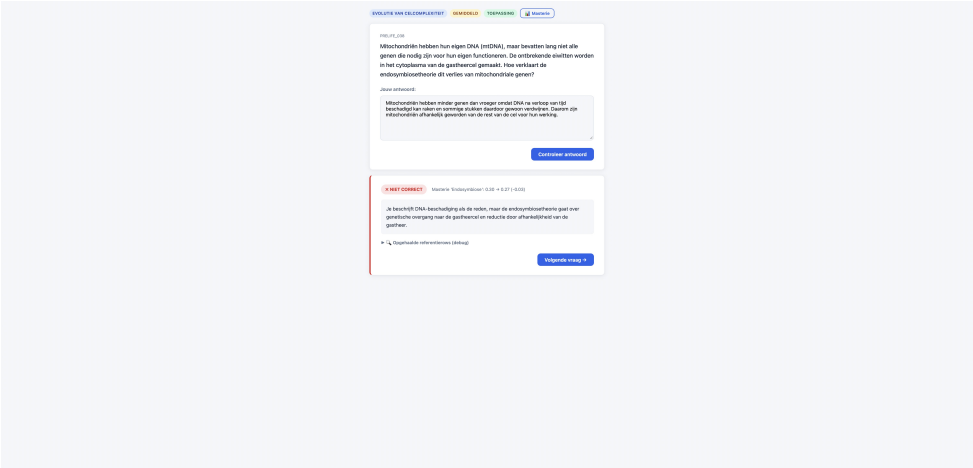
Gebruikerstest. De bruikbaarheid is getoetst via een exploratief hardop-denkprotocol: deelnemers voerden zonder instructie een quizsessie van tien à vijftien minuten uit, waarbij een observator noteerde waar zij vastliepen en hoe zij op feedback en misconceptiemeldingen reageerden, gevolgd door een semi-gestructureerd interview over de begrijpelijkheid van de feedback, het vertrouwen in de AI-beoordeling en de ervaren leermeerwaarde. Deelname was vrijwillig, zonder opslag van persoonsgegevens; de beoogde doelgroep (actieve vwo-leerlingen en vakdocenten) is niet bereikt (sectie 4.1).

3.3.6 Ontwerpiteraties. Het ontwerp is in vier iteraties tot stand gekomen; een volledige beschrijving per versie staat in de projectwiki.[6]

Versie 1: functioneel prototype. De eerste versie bewees de kernfunctionaliteit, maar miste een welkomstscherm, voortgangsoverzicht en docentinterface (figuur 3).



(a) Vraagscherm V1: subthema, moeilijkheid en Bloom-label bovenaan.

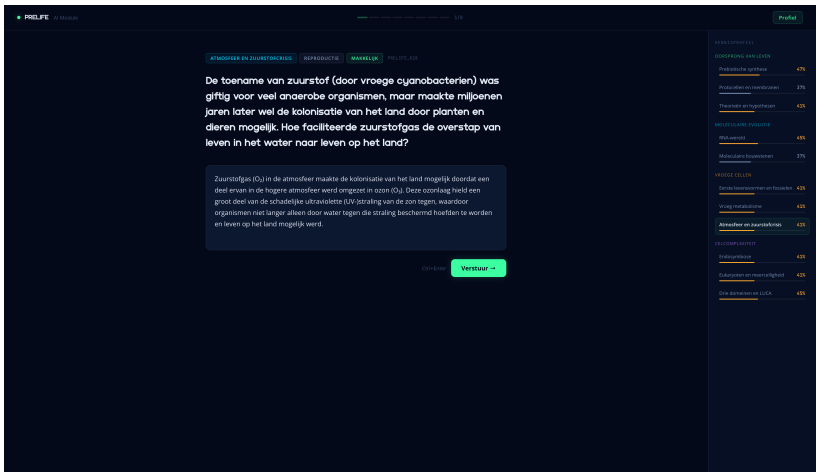


(b) Feedbackscherm V1 na een fout antwoord met beheersingswijziging.

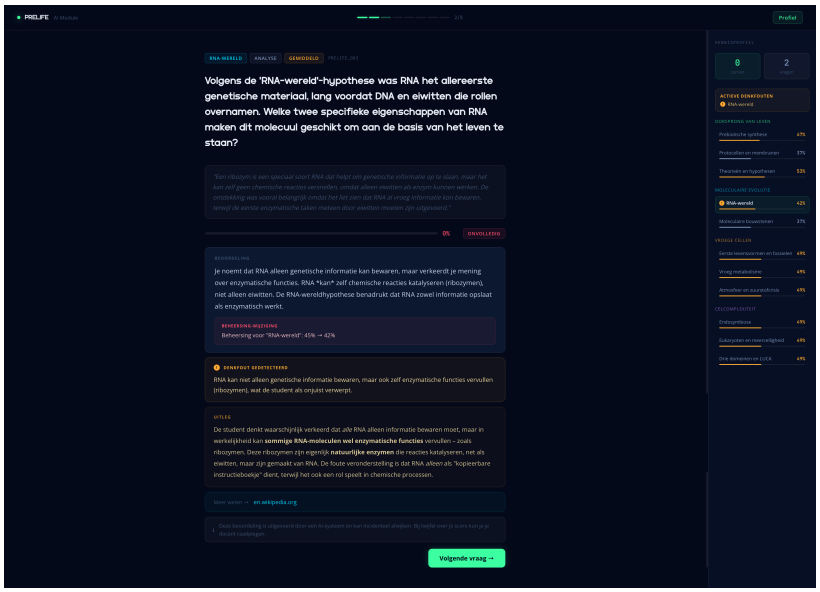
Fig. 3. Ontwerp versie 1 (V1): functioneel prototype.

[6] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/ontwerp-en-uitwerking/ontwerp-iteraties.md

Versie 2: herontwerp met docentpanel. De tweede versie voegde een welkomstscherm, een live beheersings-zijbalk en een volledig docentpanel toe, met reviewworkflow voor gegenereerde vragen (figuren 4 en 5).

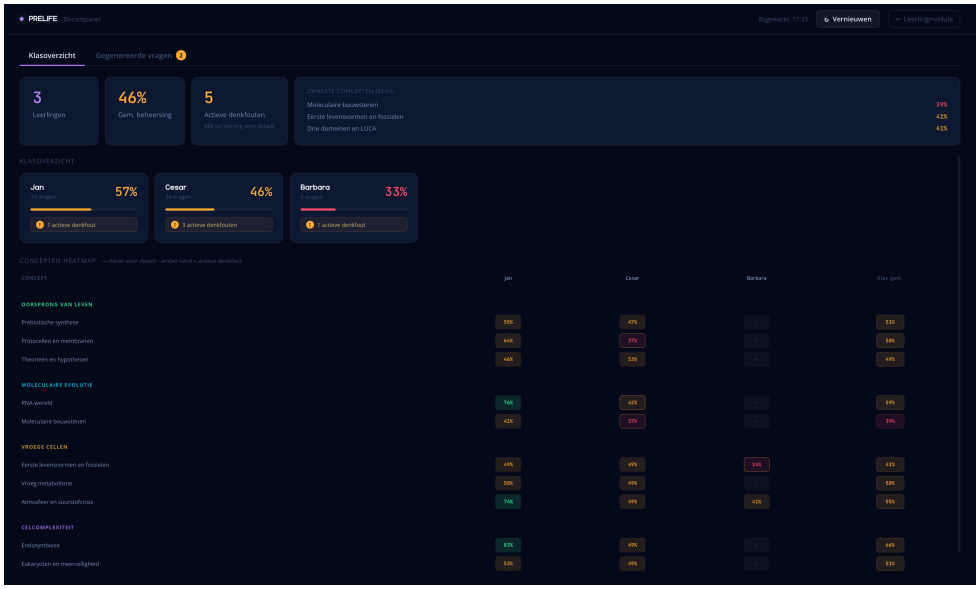


(a) Quizscherm V2: vraag links, live beheersings-zijbalk rechts.

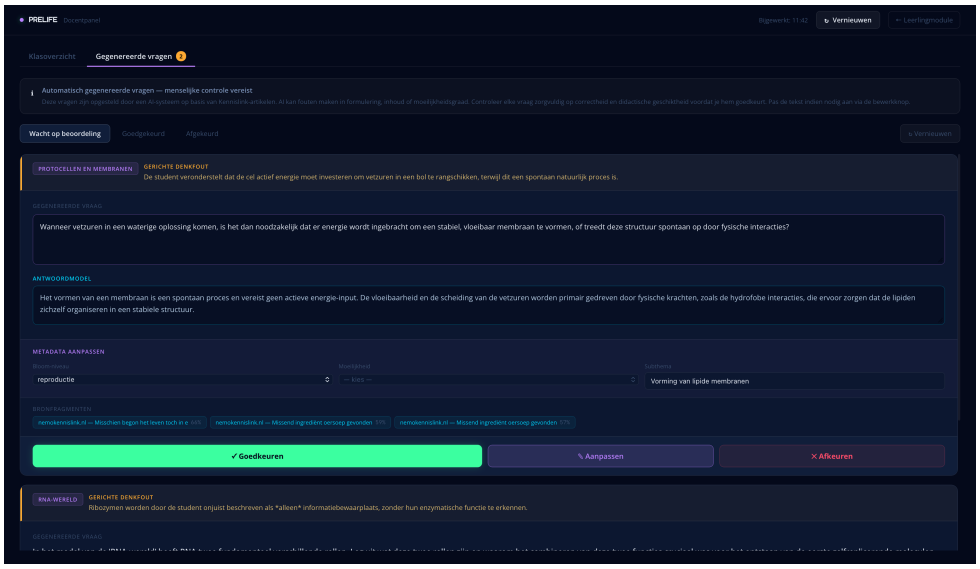


(b) Feedbackscherm V2: misconceptie zichtbaar in feedback en zijbalk.

Fig. 4. Ontwerp versie 2 (V2): leerlingmodule.



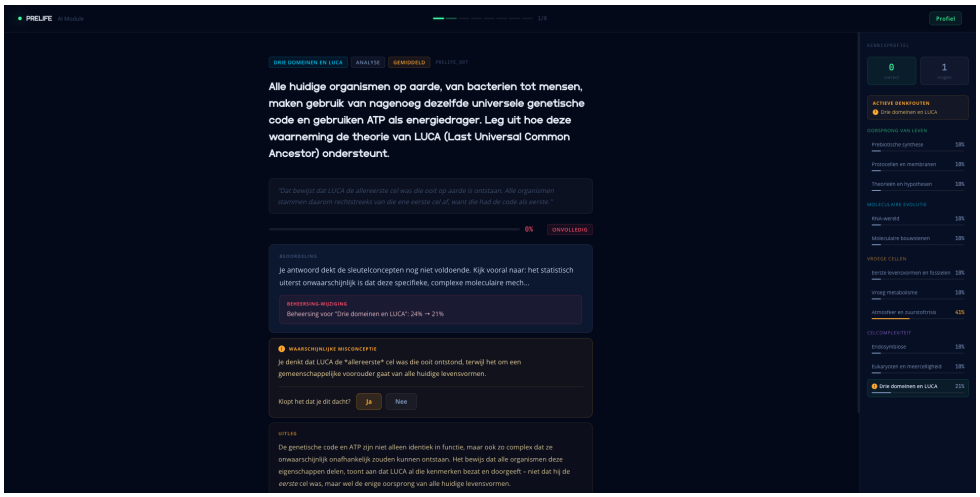
(a) Docentpanel, tabblad *Klasoverzicht*: leerlingkaarten met gemiddelde beheersing en actieve denkfouten, gevolgd door de conceptheatmap.



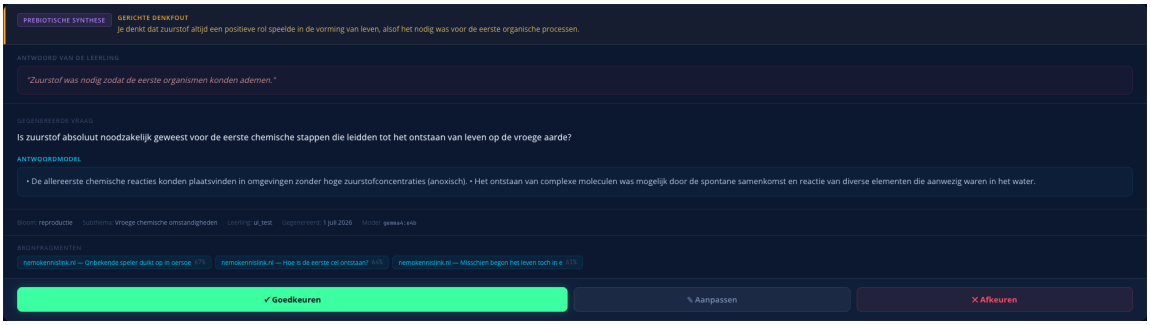
(b) Docentpanel, tabblad *Gegenereerde vragen*: vraagkaart met denkfout, gegenereerde vraag, antwoordmodel en goedkeur-/afkeurknoppen.

Fig. 5. Ontwerp versie 2 (V2): docentpanel.

Versie 3: regelgebaseerde beoordeling en leerlingverificatie. De derde versie verving de LLM-judge door de regelgebaseerde concept-coverage-beoordelaar (per concept herleidbaar, reproduceerbaar) en voegde leerlingverificatie toe: bij een gesignaleerde denkfout verschijnt een ja/nee-vraag, zodat alleen bevestigde denkfouten meetellen. Het docentpanel toont voortaan ook het triggerende leerlingantwoord (figuur 6).



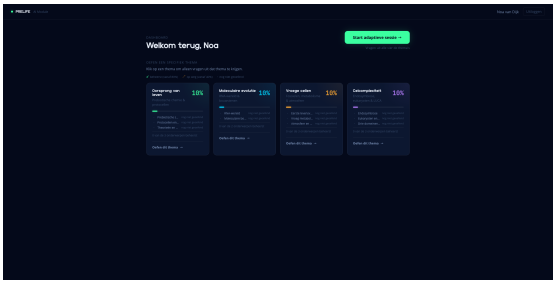
(a) Leerlingzijde: de herkende denkfout met de ja/nee-verificatie en een uitleg.



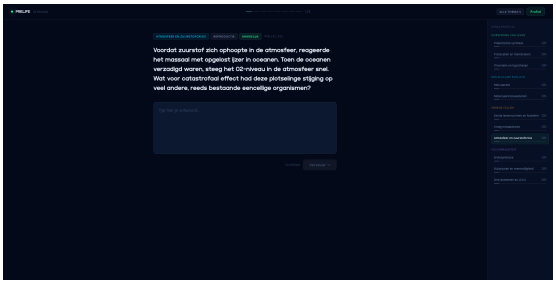
(b) Docentzijde: de reviewkaart met triggerend leerlingantwoord, antwoordmodel, bronfragmenten en goedkeuringsknoppen.

Fig. 6. Ontwerp versie 3 (V3): leerlingverificatie en uitgebreide docent-reviewkaart.

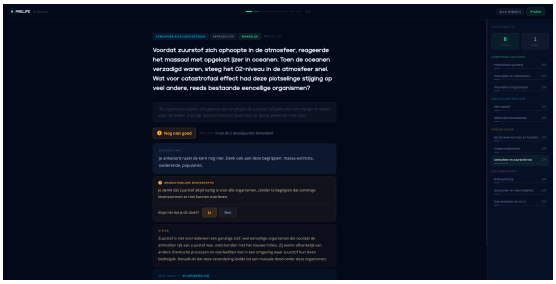
Versie 4: binaire beslisregel op echte leerlingantwoorden. Validatie van versie 3 legde een zwakte bloot: het per concept gaten en tellen tot drie labels bleek bros op vrij geformuleerde antwoorden, doordat één borderline-concept het hele oordeel deed kantelen. De vierde versie vereenvoudigt de beslisregel tot een binair goed/fout-oordeel, eerst op de *gemiddelde* conceptdekking (92%) en vervolgens op een meerderheidsstem van drie gedecorreleerde maten, met behoud van de per-concept transparantie, wat een sprong van 72% naar 98% accuratesse op de held-out set opleverde (sectie 4.2).



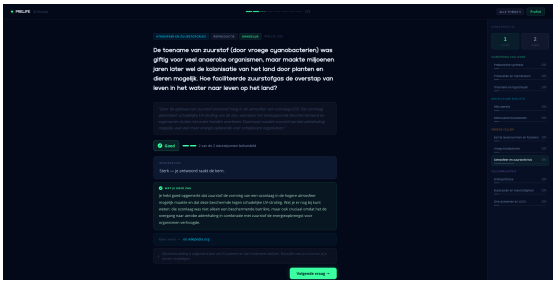
(a) Dashboard: voortgang en beheersing per thema.



(b) Quizscherm: open vraag met subthema-, Bloom- en moeilijkheidslabel en het kennisprofiel.



(c) Feedback bij een fout antwoord: regelgebaseerde hint, herkende misconceptie met ja/nee-verificatie en uitleg.



(d) Bevestiging bij een correct antwoord.

Fig. 7. Ontwerp versie 4 (V4): het werkende prototype in vier schermen — dashboard, quizscherm, feedback met misconceptieverificatie en bevestiging.

Werkend prototype. Versie 4 is volledig geïmplementeerd en bereikbaar via http://149.210.189.90:8000. De bescheiden VPS (8 GB RAM, geen GPU) is de bottleneck voor de asynchrone vraaggeneratie, niet voor de lichte, deterministische beoordeling; het prototype bedient meerdere leerlingen tegelijk en bewaart de learner state tussen sessies.

3.4 Data

Het systeem gebruikt twee databronnen: de **statische vragenbank** (50 vooraf opgestelde, AI-gegenereerde en handmatig gevalideerde open vragen over abiogenese, als basis voor BKT-kennisschatting en adaptieve vraagselectie) en de **RAG-kennisbasis** (Nederlandstalige populairwetenschappelijke artikelen van NEMO Kennislink voor het genereren van oefenvragen bij persistente misconcepties). De term *statisch* duidt op het vooraf vastliggen van de vragen, tegenover de *dynamisch* gegenereerde RAG-vragen, en niet op handmatige constructie.

3.4.1 Vragenbank: constructie en structuur. Binnen het consortium waren bij aanvang geen lesmateriaal of vragenbanken beschikbaar, en abiogenese op vwo-niveau is niet als gestructureerd vragenmateriaal via openbare bronnen te vinden; een eigen vragenbank was daarmee de enige optie. Voor de formulering van vragen en correctiemodellen is het taalmodel *Gemini* via aivoorleerlingen.nl [40] ingezet, uitsluitend als formuleringsondersteuning. Conceptselectie, SLO-koppeling, Bloom-niveau, moeilijkheidsclassificatie en misconceptiebeschrijving zijn handmatig bepaald op basis van de SLO-eindtermen [3] en de vakdidactische literatuur. Alle vragen zijn vervolgens tegen dezelfde vier criteria gelegd: (1) inhoudelijke correctheid van het correctiemodel, (2) aansluiting bij de SLO-eindterm, (3) consistentie van het Bloom-niveau met de cognitieve complexiteit, en (4) geschiktheid van de misconceptiebeschrijving als diagnostisch label. Een formele inhoudsvalidatie door een externe biologie-vakexpert heeft niet plaatsgevonden. Deze validatie is wel actief nagestreefd, maar het bleek binnen de projectkaders niet haalbaar om een gekwalificeerde vakexpert te betrekken. Dit vormt een beperking van de dataset (sectie 5.4).

De 50 vragen zijn ingedeeld in vier thema's en elf categorieën, elk gekoppeld aan een SLO-subdomein (F4.1–F4.4): **Oorsprong van leven: chemisch** (18), **Moleculaire evolutie** (7), **Vroege cellen en biosystemen** (9) en **Evolutie van celcomplexiteit** (16). F4.1 is zwaarder vertegenwoordigd als kern van het PRELIFE-consortium en de minste voorkennis bij leerlingen (verdeling in Tabel 6).

Tabel 6. Verdeling van items over Bloom-niveau, moeilijkheid en SLO-code ($n = 50$)

Bloom-niveau	**n**	**Moeilijkheid**	**n**	**SLO-code**	**n**
Begrijpen	16	makkelijk	16	F4.1	24
Toepassen	12	gemiddeld	19	F4.2	8
Analyseren	15	moeilijk	15	F4.3	16
Evalueren	7			F4.4	2
Totaal	**50**	**Totaal**	**50**	**Totaal**	**50**

3.4.2 Metadata en cognitieve niveaus. Elke vraag heeft elf metadatakolommen die de classificatie en de koppeling aan het adaptieve systeem vastleggen (Tabel 7). Het cognitieve niveau is per vraag handmatig bepaald conform de taxonomie van Bloom [4], op basis van de vereiste cognitieve handeling; het reproductieniveau (Onthouden) is bewust weggelaten omdat de module verdieping en redeneertraining beoogt, geen feitelijke reproductie. De Bloom- en moeilijkheidsclassificaties dienen als initiële priors voor het BKT-model en kunnen na gebruik empirisch worden bijgesteld [27] (verdeling in Tabel 6).

Tabel 7. Metadatakolommen in de PRELIFE Vragenbank v3

Kolom	Type	Beschrijving
`id`	string	Unieke vraagidentificator (`PRELIFE_001` e.v.)
`thema`	string	Overkoepelend thema (4 thema's)
`categorie`	string	Subcategorie binnen het thema (11 categorieën)
`SLO_code`	string	Subdomeincode examenprogramma (F4.1–F4.4)
`SLO_context`	string	Omschrijving van de bijbehorende eindterm
`niveau_bloom`	string	Cognitief niveau (Begrijpen / Toepassen / Analyseren / Evalueren)
`moeilijkheid`	string	Moeilijkheidsclassificatie (makkelijk / gemiddeld / moeilijk)
`vraag`	string	Vraagformulering
`antwoord/correctiemodel`	string	Inhoudelijk correcte kern en beoordelingscriteria
`misconceptie_doel`	string	Te diagnosticeren misconceptie
`bronwebsite`	string	Bronverwijzing voor de vraaginhoud

3.4.3 Misconceptiegerichte vraagconstructie. Elke vraag is gekoppeld aan een `misconceptie_doel`: de specifieke denkfout die de vraag beoogt te diagnosticeren (conform de Conceptual Change Theory, sectie 2.1). In de 50 beschrijvingen zijn vijf terugkerende patronen onderscheiden, toegepast op abiogenese:

(1) **Anachronistische projectie**: de huidige atmosferische samenstelling wordt op de vroege aarde geprojecteerd.
(2) **Teleologisch denken**: processen krijgen een intentie of doel toegeschreven, alsof moleculen *willen* samenwerken om leven te vormen.
(3) **Oversimplificatie**: graduele, meerstappige processen (RNA-wereld, endosymbiose) worden gereduceerd tot één enkelvoudige gebeurtenis.
(4) **Onterechte generalisatie**: een beperkt geldig principe wordt universeel toegepast, bijv. de aanname dat al het leven dezelfde oorsprong moet hebben.
(5) **Onvolledige domeinkennis**: specifieke kennisgebreken leiden tot onjuiste redeneringen, zonder een actieve denkfout.

Wanneer de beoordelaar een concept als niet-gedekt markeert, levert het `misconceptie_doel`-veld een concrete kandidaat-denkfout; of het antwoord die denkfout daadwerkelijk vertoont, bepaalt vervolgens een lokale taalmodelstap met leerlingbevestiging (sectie 3.5.3).

3.4.4 Curriculaire validiteit. Curriculaire validiteit steunt op de directe koppeling van elk item aan een SLO-eindterm [3] (`SLO_code`, `SLO_context`). Enkele items zijn gebaseerd op gepubliceerde vwo-centraalexamenvragen (2012–2025),

omgeschreven zonder de kern te wijzigen; de overige komen uit de SLO-eindtermen en vakdidactische literatuur, maar zijn niet formeel door een curriculumexpert gevalideerd.

3.4.5 RAG-kennisbasis: NEMO Kennislink. Voor de RAG-pipeline is een Nederlandstalige kennisbasis opgebouwd uit elf populairwetenschappelijke artikelen van NEMO Kennislink [38] (schriftelijke toestemming, correspondentie april 2026), binnen het thema *Oorsprong van Leven* en dekkend de vier SLO-subdomeinen van abiogenese, opgeknipt in 142 overlappende tekstfragmenten (chunking-parameters binnen aanbevolen literatuurbereik [13]). Volledige parameters en artikellijst staan in de projectwiki.[7]

[7] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/ontwikkeling/dataset.md

3.5 Model

De ontwerpkeuzes volgen uit de eerder vastgestelde requirements (sectie 3.2) en de beschikbare data: de eisen van betekenisgerichte, lokale en reproduceerbare beoordeling (REQ 4.3, 5.4, 1.1) en het ontbreken van gelabelde Nederlandse trainingsdata bepalen samen de keuze voor een regelgebaseerde judge met semantische embeddings boven een getraind model of een externe LLM.

3.5.1 Selectie van de beoordelaar. Om de meest geschikte beoordelaar te selecteren zijn de methoden uit de benchmark (sectie 3.3.5) binair vergeleken; de vijfvoudige kruisvalidatie geeft 92% (sectie 4.2). De ingezette stem combineert drie gedecorreleerde signalen die elk een ander type fout afvangen. Drie typen zijn vergeleken op labelaccuracy: losse tekstmetrieken, lokale LLM-judges en de regelgebaseerde concept-coverage-judge (Tabel 8).

Tabel 8. Benchmark van beoordelaars op binaire labelaccuracy (held-out set B, 50 antwoorden; drempels uitsluitend op de kalibratieset A gekozen).

Type	Methode	Accuracy
Losse metriek	Globale cosine	90%
	ROUGE-L	94%
	BLEU	98%
LLM-judge	`llama3.2:1b`	48%
	`gemma4:e2b`	94%
	`gemma4:e4b`	96%
	`ministral-3:3b`	**98%**
Regelgebaseerd	Gemiddelde conceptdekking	92%*
	Meerderheidsstem (ingezet)	**98%***

*Accuratesse op held-out set B. Bij vijfvoudige kruisvalidatie over de honderd antwoorden: 92% (meerderheidsstem), 91% (gemiddelde conceptdekking) en circa 96% voor de sterkste LLM-judge.

Losse metrieken zijn misleidend hoog. Globale cosine (90%), ROUGE-L (94%) en BLEU (98%) scoren hoog doordat leerlingen sommige woorden van het antwoordmodel ook gebruiken, stijl-afhankelijk gedrag dat bij anders formulerende leerlingen wegvalt en de vocabulaire-meeliftbias herintroduceert, waardoor deze maten als zelfstandige beoordelaar onbetrouwbaar zijn [41].

LLM-judges. De LLM-judge-familie is sterk: bij vijfvoudige kruisvalidatie haalt ze circa 96%. `gemma4:e4b` (96% held-out) en `ministral-3:3b` (98% held-out) zijn de sterkste; `ministral-3:3b` wint op de assen die er voor een tutor toe doen: precisie 1,0 (geen enkel fout antwoord onterecht goedgekeurd) en snelheid (circa 1 s per antwoord, tegen 4–8 s voor de gemma-modellen). `llama3.2:1b` (48%) is ongeschikt doordat het vrijwel elk antwoord goedkeurt. `ministral-3:3b` was daarmee de sterkste LLM-judge en fungeerde in eerdere iteraties dan ook als primaire beoordelaar. Met de overstap naar de regelgebaseerde judge (hieronder toegelicht) verviel die rol als scorer; het model blijft wel ingezet voor de lokale LLM-stappen buiten de beoordeling, namelijk de duiding van een gedetecteerde misconceptie en de bevestiging van een correct antwoord. De promptstrategie en operationele metrieken staan in de projectwiki.[8]

[8] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/onderzoek/llm-judge-onderzoek.md

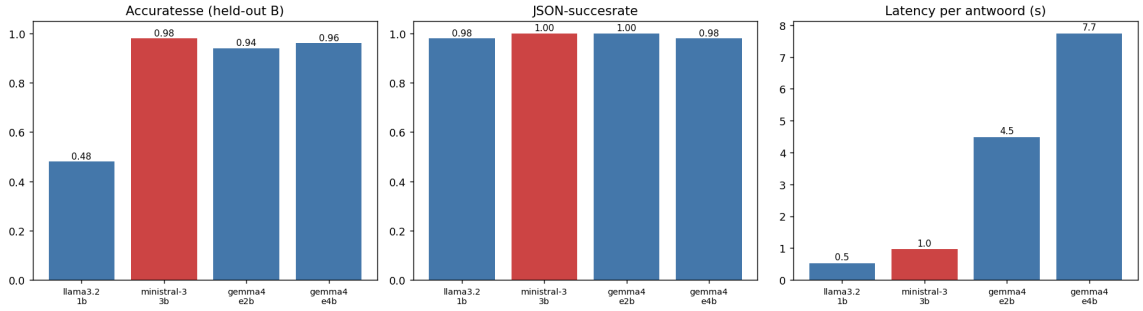


Fig. 8. Benchmarkresultaten per LLM op de held-out set: accuracy, JSON-succesrate en latency per antwoord.

Keuze: de regelgebaseerde judge. De ingezette judge is de **meerderheidsstem** van drie gedecorreleerde deterministische maten (gemiddelde conceptdekking, kernterm-gate en lexicale overlap): circa 92% accuratesse bij vijfvoudige kruisvalidatie, met precisie 1,0 op de held-out set. Dat ligt iets onder de beste lokale LLM-judge (circa 96%), maar zonder generatief model. De eenvoudiger drempel op alléén de gemiddelde conceptdekking bleef op 91%. De regelgebaseerde beoordelaar is als primaire scorer gekozen omdat zij die nauwkeurigheid combineert met volledige reproduceerbaarheid en per-concept-herleidbaarheid, zonder LLM-inferentie draait en zo tegelijk directe feedback aan de leerling en controle voor de docent oplevert, eigenschappen die in een hoog-risico-onderwijscontext (EU AI Act) [33] zwaarder kunnen wegen dan een fractie extra accuratesse.

3.5.2 Architectuur van de regelgebaseerde judge. Als primaire beoordelaar is gekozen voor een regelgebaseerde judge op basis van *concept-coverage*. De judge ontvangt per beoordeling de vraag, het antwoordmodel uit de vragenbank, en het antwoord van de student. Omdat het antwoordmodel altijd beschikbaar is, verschuift de taak van kennistoetsing naar het meten of de student de vereiste concepten benoemt [10].

Stap 1: Concepten afsplitsen. Het antwoordmodel bestaat uit losse rubric-punten; de judge splitst het op deze punten in afzonderlijke concept-claims en beoordeelt elk concept vervolgens onafhankelijk van de andere. Dat is een vereenvoudiging: rubric-punten kunnen inhoudelijk samenhangen, en die samenhang gaat in de huidige aanpak verloren (sectie 5.4).

Stap 2: Semantische dekking per concept. Elk concept en elke zin van het studentantwoord wordt ingebed met het meertalige model `BGE-M3` [42]. Voor elk concept geldt de hoogste cosine-gelijkenis met een zin uit het antwoord als maat voor semantische dekking. In een eerdere iteratie is hiervoor het lichtere `paraphrase-multilingual-MiniLM-L12-v2` [16, 43] gebruikt; BGE-M3 presteert beter op parafrases (sectie 4.2). Het embedmodel wordt gedeeld met de RAG-retrieval, zodat er één model in het geheugen staat.

Stap 3: Kernterm-controle. Een hoge cosine alleen is onvoldoende: een incorrect maar on-topic antwoord kan meeliften op het vocabulaire van het antwoordmodel [9]. Daarom telt een concept pas als gedekt wanneer het antwoord óók een minimaal deel van de kerntermen bevat (stam-matching, tolerant voor verbuigingen). Een concept telt langs twee routes: bij een heldere match ligt de cosine boven de drempel t_{cos} én zijn de kerntermen aanwezig; daarnaast telt een *grijze zone* ($t_{cos,\text{grijs}} \leq \text{cosine} < t_{cos}$) alsnog mee wanneer de lexicale bevestiging stevig is. Deze grijze zone vangt correcte parafrases op die net onder de cosine-drempel vallen, zonder de vangrail tegen vocabulaire-meeliften los te laten. De drempels ($t_{cos} = 0{,}68$, $t_{cos,\text{grijs}} = 0{,}56$) zijn niet uit een theoretisch model afgeleid, maar via een systematische sweep op de kalibratieset bepaald, over gevarieerde leerlingantwoorden, en blijken robuust over een brede band (figuur 9).

Stap 4: Binair label via meerderheidsstem. Het eindoordeel volgt uit een 2-van-3 stem van drie gedecorreleerde signalen: de gemiddelde cosine-dekking over de concepten ($t \geq 0{,}69$), de fractie concepten die door de kernterm-gate komt ($t = 1{,}0$) en de unigram-woordoverlap (BLEU) met het antwoordmodel ($t \geq 0{,}05$). Een antwoord is *correct* zodra minstens twee signalen hun drempel halen. Omdat de maten verschillende fouten afvangen (parafrase, vocabulaire-meeliften, willekeurige overlap), is de stem robuuster dan elke losse maat (sectie 4.2); de drempels zijn uitsluitend op de kalibratieset gekozen. Zij verving een eerder, brosser driewegs-ontwerp (sectie 3.3.6). De per-concept dekking blijft behouden als docentgerichte transparantie: de feedback benoemt het zwakst gedekte concept en welke signalen slaagden, en bij een misconceptie-gerichte vraag wordt de vooraf gelabelde `misconceptie_doel` gemarkeerd. Het oordeel is deterministisch en per concept herleidbaar. Een leeg antwoord wordt al bij invoer geweigerd. De labelwaarden gaan intern naar scorepunten voor de BKT-update: *correct* = 1,0 en *incorrect* = 0,0.

Voorbeeld. Vraag `PRELIFE_005` vraagt één moleculair bewijs voor de nauwere verwantschap tussen Archaea en Eukaryoten. Het antwoordmodel biedt twee gelijkwaardige concepten (verbonden met OF): (1) RNA-polymerase/ribosomen bij Archaea lijken structureel meer op die van eukaryoten dan van bacteriën, en (2) histonen (of vergelijkbare DNA-verpakkingseiwitten) komen voor bij Archaea én Eukaryoten, niet bij Bacteriën. Een leerling antwoordt: *"Archaea lijken gewoon een beetje op eukaryoten omdat ze ook geen echte bacteriën zijn."* Dit antwoord is on-topic, maar mist moleculair bewijs.

Tabel 9 toont de per-concept diagnostiek. Beide concepten worden als gedekt gemarkeerd: concept 1 via de heldere match (cosine boven t_{cos}, kerntermen aanwezig), concept 2 via de grijze zone (cosine tussen $t_{cos,\text{grijs}}$ en t_{cos}, met voldoende kernterm-overlap). Puur op conceptdekking zou dit antwoord dus onterecht *correct* scoren: de gedeelde onderwerpwoorden (*Archaea*, *eukaryoten*, *bacteriën*) raken toevallig de kerntermen van het concept, zonder dat de leerling het gevraagde molecuulbewijs noemt.

Tabel 9. Per-concept diagnostiek voor het voorbeeldantwoord ($t_{cos} = 0{,}68$, $t_{cos,\text{grijs}} = 0{,}56$).

Concept	Cosine	Kernterm-overlap	Gedekt?
(1) RNA-polymerase/ribosomen-structuur	0,714	0,40	Ja (heldere match)
(2) Histonen bij Archaea en Eukaryoten	0,630	0,40	Ja (grijze zone)

De meerderheidsstem corrigeert deze fout (Tabel 10): de gemiddelde conceptdekking (0,672) blijft onder de stemdrempel (0,691), en de unigram-woordoverlap met het antwoordmodel (0,018) is vrijwel nul: de leerling herformuleert het antwoord vrijwel volledig in eigen woorden, zonder de gevraagde vaktermen. Alleen de kernterm-gate (1,0) haalt zijn drempel. Met 1 van de 3 signalen gehaald blijft de meerderheid (2 van 3) uit, en velt het systeem het oordeel *incorrect*, gelijk aan het handmatige label.

Tabel 10. Meerderheidsstem (Stap 4) voor het voorbeeldantwoord.

Signaal	Waarde	Drempel	Gehaald?
Gemiddelde conceptdekking	0,672	0,691	Nee
Kernterm-gate (fractie)	1,000	1,000	Ja
Unigram-BLEU	0,018	0,048	Nee
Eindoordeel			**incorrect (1/3)**

3.5.3 Misconceptieverwerking. De beoordeling zelf is regelgebaseerd; de duiding van een fout antwoord op een misconceptie-gerichte vraag is een aparte, lokale LLM-stap. Wanneer de judge een antwoord niet als *correct* markeert, bepaalt een compact taalmodel (`ministral-3:3b`) of het antwoord de bij die vraag vooraf gelabelde denkfout (`misconceptie_doel`) daadwerkelijk vertoont. De LLM bedenkt dus geen nieuwe denkfout, maar bevestigt alleen een bekende. Het systeem doorloopt vervolgens zes stappen:

(1) **Detectie en uitleg.** De prompt geeft het taalmodel vier velden mee: de vraag, het correcte antwoordmodel, de vooraf gelabelde `misconceptie_doel` en het leerlingantwoord. Het model beslist daarmee binair of het antwoord *specifiek deze* denkfout toont ("match": true/false), expliciet onderscheiden van een antwoord dat wel fout is, maar om een andere reden of gewoon onvolledig. Dat onderscheid is precies waar de detectie op mist: leerlingen maken op echte data vaak een ándere denkfout dan de ene voorziene (sectie 4.3). Bij een match levert de aanroep in één keer ook een uitleg van maximaal drie zinnen op vwo-niveau. Twee robuustheidsmaatregelen zijn hier essentieel: de uitvoer wordt afgedwongen als geldige JSON (`format="json"`), en oordeel en uitleg worden verankerd in het antwoordmodel met de instructie geen feiten toe te voegen die daar niet in staan. Beide voorkomen dat een correcte detectie verloren gaat door onparseerbare uitvoer en dat de uitleg feitelijk afdwaalt (sectie 4.3). Vertoont het antwoord de verwachte denkfout niet, dan krijgt de leerling generieke terugkoppeling naast de per-concept feedback van de judge.

(2) **Opslaan.** De denkfout wordt in de learner state onder het betreffende concept opgeslagen, met een teller en een opgelost-status.

(3) **Leerlingbevestiging.** De herkende denkfout verschijnt met een ja/nee-vraag. Alleen door de leerling *bevestigde* denkfouten tellen mee. Deze stap is het precisievangnet tegen de neiging van een compact model om ook een los kennisgat als denkfout te bestempelen (sectie 4.3).

(4) **Prioritering.** Bij de volgende vraagselectie krijgen concepten met een actieve, niet- opgeloste misconceptie de hoogste prioriteit, ongeacht de BKT-beheersing van andere concepten.

(5) **Oplossen.** Beantwoordt de leerling een volgende vraag over hetzelfde concept correct, dan wordt de misconceptie als opgelost gemarkeerd en vervalt de prioriteit.

(6) **RAG-trigger.** Bereikt het aantal *bevestigde* denkfouten voor een concept de drempel van twee, dan activeert het systeem asynchroon de RAG-pipeline via een FastAPI `BackgroundTask`, zodat de leerling geen wachttijd ervaart. Dat alleen bevestigde denkfouten meetellen, voorkomt dat een eenmalige detectiefout onnodig een vraag triggert.

3.5.4 RAG-pipeline en kennisretrieval. Wanneer de RAG-trigger activeert, combineert de zoekquery concept én misconceptiebeschrijving en worden de drie best passende chunks opgehaald met hetzelfde `BGE-M3`-model [42] als de judge; deze gaan als prompt-context naar de vraaggenerator, elk voorzien van artikeltitel en bron-URL voor de docent. De volledige queryconstructie- en retrievalstappen staan in de projectwiki.[9]

[9] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/ontwikkeling/model.md

3.5.5 Vraaggeneratie en docentvalidatie. Op basis van de retrieval-context genereert het systeem een nieuwe oefenvraag gericht op de gedetecteerde misconceptie, met `gemma4:e4b` [44] als generatiemodel, asynchroon en incidenteel ingezet, zodat de leerling geen wachttijd ervaart en een krachtiger model inzetbaar is dan voor de realtime beoordeling haalbaar zou zijn.[10]

Gegenereerde vragen worden opgeslagen met de status *pending* en zijn pas beschikbaar na expliciete goedkeuring door de docent. Menselijke validatie is verkozen boven automatische LLM-beoordeling op drie gronden: een compact taalmodel kan zijn eigen output niet betrouwbaar op correctheid beoordelen (dezelfde blinde vlekken), de beoordeling van didactische geschiktheid vereist vakdidactische expertise, en de EU AI Act schrijft menselijk toezicht voor in onderwijscontexten [33]. Door de eindverantwoordelijkheid bij de docent te leggen voldoet het systeem aan die eis en vergroot het draagvlak.

3.5.6 Adaptieve vraagkeuze en herhalingsmechanisme. De vraagkeuze combineert drie beslissingen: concepten met een actieve misconceptie krijgen prioriteit, anders het concept met de laagste BKT-beheersing; het Bloom-niveau volgt uit de beheersing: naarmate de beheersing groeit verschuift de gevraagde cognitieve handeling naar een hoger niveau in de taxonomie van Bloom [4], wat de cognitieve belasting bij lage beheersing beperkt [36] (de drempels 0,40/0,70 zijn niet empirisch gekalibreerd maar initiële priors [27]); en elke vraag krijgt een cooldown van minimaal vijf beantwoorde vragen, een vereenvoudiging van het volledige spacing-model [28]. De volledige mechaniek staat in de projectwiki.[11]

De beoordelaar is ontworpen om reference bias te beperken: beoordelaars die op lexicale overlap leunen bevoordelen het antwoordmodel boven semantisch equivalente maar anders geformuleerde antwoorden [41]. De concept-coverage-judge meet semantische dekking en blijkt daardoor minder gevoelig; deze gevoeligheid is als evaluatiecriterium in de benchmark meegenomen (uitsplitsing naar labeltype).

De prestaties van de beoordelaar worden primair gemeten via **labelaccuracy** (per labeltype apart om classificatiebias zichtbaar te maken), aangevuld met een **vijfvoudige kruisvalidatie** als generalisatiemaat. Voor de LLM-judges zijn daarnaast de JSON-succesrate en de latency op CPU meegenomen als operationele indicatoren.

[10] Promptstrategie en modelkeuze: zie https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/ontwikkeling/model.md.
[11] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/ontwikkeling/model.md

4 Resultaten

4.1 Concept- en gebruikerstests

De gebruikerstest is uitgevoerd met drie volwassen deelnemers met een afgeronde vwo-achtergrond in bètavakken, die als proxy voor vwo-bovenbouwleerlingen dienden. Omdat de beoogde doelgroep niet is bereikt, zijn de bevindingen exploratief en indicatief; uitspraken over gebruikersgedrag gelden onder dit voorbehoud. Volledige resultaten per thema staan op de projectwiki.[12] De test leverde zes bevindingen op:

(1) **Zelfstandig gebruik.** De basisfunctionaliteit was zonder toelichting te doorlopen; er waren geen fundamentele usability-blokkades.

(2) **Open vragen en AI-feedback.** Het verwachte detailniveau was niet altijd duidelijk, wat tot te korte antwoorden leidde; feedback werd het nuttigst gevonden wanneer ze concreet aansloot op het antwoord; generieke feedback wekte twijfel of het systeem het antwoord had begrepen. In een latere ronde met dezelfde proxy-groep werd de huidige regelgebaseerde hint, die de ontbrekende kernbegrippen benoemt, als nuttigst beoordeeld, juist omdat die richting geeft zonder het volledige antwoord prijs te geven.

(3) **Vertrouwen en adaptiviteit.** Uitleg bij een herkende denkfout verhoogde het vertrouwen in de terugkoppeling; de adaptieve vraagkeuze zelf was voor deelnemers niet zichtbaar.

(4) **Docentpanel.** De proxy-deelnemers vonden het overzicht van conceptbeheersing potentieel waardevol, maar konden de bruikbaarheid voor de docentpraktijk zelf niet beoordelen; dat oordeel volgde uit de docentdemo (hieronder).

Aanvullend gaf vakdocent biologie en projectbegeleider Marco Kragten tijdens een docentdemo (21 mei 2026) feedback. Hij beoordeelde het docentpanel met het overzicht van conceptbeheersing als bruikbaar en waardevol voor de docentpraktijk, en gaf gerichte feedback op de gegenereerde uitleg: toon en complexiteit lagen voor vwo-leerlingen te hoog, met vakjargon en te ingewikkelde formuleringen. Een analyse bracht de oorzaak aan het licht: de prompt bevatte geen expliciete eis dat de tekst op vwo-niveau moest zijn. De prompt is daarop uitgebreid met de eis dat de uitleg helder, direct en zonder onnodig vakjargon op vwo-niveau wordt geformuleerd. De begrijpelijkheid na deze aanpassing is kwantitatief getoetst met de leesbaarheidsanalyse (sectie 4.4); die laat zien dat de vwo-instructie de complexiteit onvoldoende terugbrengt: de gegenereerde uitleg blijft met een gemiddelde LiNT-score van 67,2 de moeilijkste van de drie feedbackkanalen, zodat een sterkere maatregel dan promptinstructie nodig blijft.

[12] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/onderzoek/field-research.md

4.2 Prestaties in de context van het concept

De benchmark gebruikt echte antwoorden van voormalige vwo-leerlingen, die korter en vakmatiger formuleren dan huidige leerlingen. Deze sectie toetst de gekozen binaire beslisregel op de twee **onafhankelijke sets** van elk vijftig échte, handmatig gelabelde antwoorden: een kalibratieset (A) en een los gehouden testset (B). Set B is handmatig door de onderzoeker samengesteld uit échte leerlingantwoorden en bewust ingedeeld naar gevarieerde stijlen (compact, informeel, met spelfouten, parafraserend en denkfout-dragend). Aanvankelijk was het doel om driewegs te beoordelen (correct, gedeeltelijk correct, incorrect); set A is dan ook oorspronkelijk driewegs gelabeld en set B direct binair. Voor een regelgebaseerde judge bleek de tussencategorie echter lastig betrouwbaar te scheiden, waardoor de driewegs-aanpak onvoldoende presteerde (sectie 3.5.1). Daarom is voor een binair oordeel gekozen: set A is naar één binair label teruggebracht, waarbij *deels correct* als niet-correct telt.

Het oorspronkelijke driewegs-ontwerp bleef op deze antwoorden op 72% steken. De beslisregel is daarom eerst vereenvoudigd tot een enkele drempel op de *gemiddelde* cosine-dekking, die 92% (held-out, drempel uitsluitend op kalibratieset A gekozen) en 91% (vijfvoudige kruisvalidatie) haalt. Geen enkele losse maat domineert echter overtuigend: de kernterm-gate (98%) en lexicale maten (ROUGE-L 94%, globale cosine 90%) scoren rond of boven de gemiddelde conceptdekking, al liggen die verschillen binnen de ruis en zijn de lexicale maten stijl-afhankelijk. Daarom is de **ingezette** beslisregel een meerderheidsstem van drie gedecorreleerde maten (conceptdekking, kernterm-gate en lexicale overlap; sectie 3.5.2): doordat elk signaal een ander type fout afvangt, bepaalt geen enkel faalmechanisme het oordeel alleen. Bij vijfvoudige kruisvalidatie haalt de stem 92% (±7%) en op de held-out set 98% (precisie 1,0); de combinatie is robuuster dan elke losse maat, die stijl-afhankelijk kan wegvallen. Met 25 antwoorden per klasse verschuift zo'n held-out-cijfer al bij één misclassificatie met 4 procentpunt; de kruisvalidatie, met haar bredere foutmarge, is daarom de robuustere generalisatieschatting.

Als embedmodel is BGE-M3 [42] gekozen boven het lichtere MiniLM [16, 43]: het schat parafrases hoger in, waardoor minder correcte antwoorden onterecht onder de drempel vallen. Rond het operating point (drempel ≈ 0,71) haalt BGE-M3 een hogere piekaccuratesse dan MiniLM (0,84 tegen 0,80) en blijft daar stabiel (figuur 9).

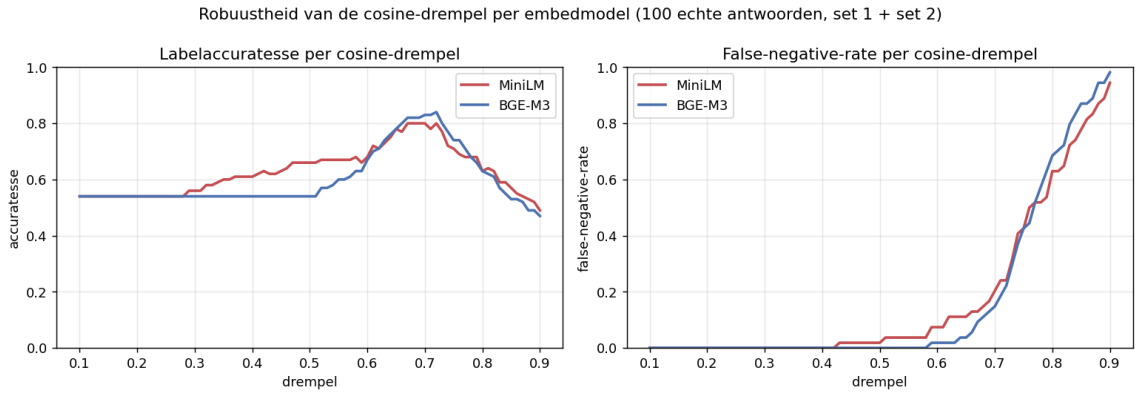


Fig. 9. Robuustheid van de cosine-drempel per embedmodel op honderd échte leerlingantwoorden: labelaccuratesse (links) en false-negative-rate (rechts). Rond het operating point (≈ 0,71) ligt BGE-M3 hoger dan MiniLM (piek 0,84 tegen 0,80).

De ingezette stem presteert daarnaast stabiel over de verschillende antwoordstijlen: denkfout-dragende, onvolledige, slordig gespelde en de meeste correcte antwoorden worden alle betrouwbaar beoordeeld (100%); alleen bij lange,

redenerende antwoorden (94%) viel één inhoudelijk correct antwoord net buiten de stem doordat uitweiding de gemiddelde conceptdekking verdunt. Beide sets bevatten authentiek leerlingmateriaal, maar van oud-leerlingen; validatie op actieve vwo-bovenbouwleerlingen blijft nodig (sectie 5.4). Een end-to-end evaluatie van de volledige pipeline en een formele kwaliteitsevaluatie van de RAG-vragen hebben binnen dit onderzoek niet plaatsgevonden.

4.3 Misconceptiedetectie op echte leerlingantwoorden

Om te beoordelen hoe betrouwbaar `ministral-3:3b` een denkfout herkent, is de detector via de volledige pijplijn geëvalueerd op de échte studentendata: twee sets, een kalibratieset A en een los gehouden set B, van elk vijftig antwoorden, samen honderd, waarvan tweeëndertig een aantoonbare denkfout dragen. De detector draait áchter de judge-gate, precies zoals in de applicatie: alleen niet-correct beoordeelde antwoorden gaan naar het taalmodel, dat binair bepaalt of de vooraf gelabelde denkfout (`misconceptie_doel`) aanwezig is. De evaluatie is reproduceerbaar vastgelegd.[13]

Omdat de detector bewust alleen de vooraf gelabelde denkfout bevestigt en nooit zelf een nieuwe verzint, zodat een leerling geen gehallucineerde denkfout te zien krijgt, is precisie de betekenisvolle maat: signaleert het systeem een denkfout, dan klopt dat in **79%** van de gevallen (gepoold; 78% en 80% per set). De gemeten **recall van 47%** is een *ondergrens* en geen zwakte van de detector. De waarheid telt namelijk élke denkfout, terwijl de detector alleen de éne voorziene denkfout herkent. Analyse van de gemiste gevallen laat zien dat vrijwel geen ervan door de gate verloren gaat (die liet slechts één denkfout-antwoord door), maar dat leerlingen op echte data juist vaak een ándere denkfout maken dan de ene die per vraag is geanticipeerd; die afwijkingen tellen hier als gemist.

Deze cijfers onderbouwen een centrale ontwerpkeuze. Bij een precisie van 79% is ongeveer een op de vijf gesignaleerde denkfouten geen echte denkfout, en het onderscheid tussen een actieve denkfout en een puur onvolledige redenering blijkt voor een 3B-model te subtiel om betrouwbaar te maken. Daarom draait de detectie niet op correcte antwoorden en bevestigt de leerling elke gesignaleerde denkfout voordat die meetelt voor de RAG-vraaggeneratie (sectie 3.5.3); zo blijven valse signaleringen zonder gevolg.

De recall bleek daarnaast kwetsbaar voor de *vorm* van de modeluitvoer: een uitgebreidere prompt gaf vaker onparseerbare JSON en drukte de recall op een controleset naar 57%, wat een tolerante JSON-parser zonder precisieverlies herstelde. Bij compacte modellen is de robuustheid van het uitvoerformaat dus even bepalend als de inhoudelijke kwaliteit van het oordeel. Honderd antwoorden blijft klein; de cijfers zijn indicatief.

[13] Notebook en resultaat-JSON in `code/evaluaties/volledige_pipeline/`.

4.4 Leesbaarheidsanalyse van de feedback (LiNT)

De leerling ziet drie soorten feedback: een regelgebaseerde hint en, gegenereerd door `ministral-3:3b`, een misconceptie-uitleg of een bevestiging. Een leesbaarheidsanalyse met LiNT (streefzone niveau 2–3, score 34–58) [45] op de 119 feedbackteksten uit set A en B (figuur 10) laat zien dat de hint grotendeels binnen de streefzone valt (**54,3**; 71%), terwijl beide LLM-teksten, ondanks een expliciete vwo-instructie in de prompt, te complex zijn (bevestiging **59,9**; 43%, misconceptie-uitleg **67,2**; 16%). Volledige methode en scoreverdeling per kenmerk staan in de projectwiki.[14]

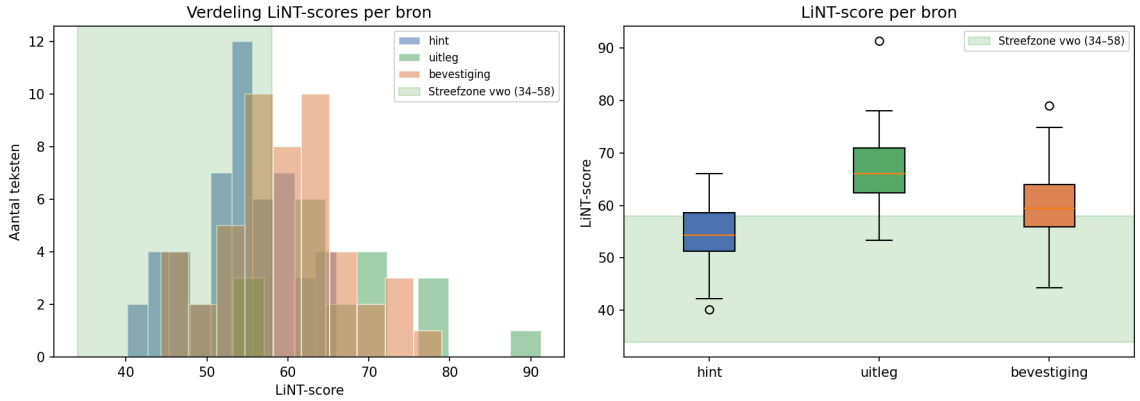


Fig. 10. Verdeling van LiNT-scores over de 119 feedbackteksten (links) en per bron: regelgebaseerde hint, misconceptie-uitleg en LLM-bevestiging (rechts). De groene band is de streefzone voor vwo-bovenbouw (score 34–58, niveau 2–3).

4.5 RAG-kwaliteitsevaluatie

Een steekproef van tien concept-misconceptie-combinaties (drempel: hit rate ≥ 70%) geeft een hit rate van **8/10 (80%)**, ruim boven de drempel. De twee missers (endosymbiose, homochiraliteit) wijzen op een gat in de kennisbasis, geen prestatiefout van het retrieval-algoritme. Methodologie en de verdeling per misconceptiecategorie staan in de projectwiki.[15]

4.6 Openstaande evaluaties

Een eerste stakeholderevaluatie met een vakdocent heeft plaatsgevonden tijdens de docentdemo met Marco Kragten (sectie 4.1). Een bredere evaluatie met meerdere onafhankelijke vakdocenten biologie staat nog open; via een semi-gestructureerd interview wordt dan beoordeeld of het docentpanel en de reviewworkflow aansluiten bij de werkpraktijk.

Daarnaast staat een validatie met de primaire doelgroep zelf nog open. De gebruikerstests zijn tot nu toe met een proxy-groep van oud-vwo-leerlingen uitgevoerd; een test met actieve vwo-bovenbouwleerlingen in de klas is nodig om te beoordelen of het systeem ook voor de beoogde doelgroep bruikbaar en begrijpelijk is en om de huidige, indicatieve bevindingen naar die groep te kunnen generaliseren.

[14] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/evaluatie-en-validatie/testresultaten.md
[15] https://gitlab.fdmci.hva.nl/maai/2025-2026-sem2/smart-education-lab/cesar-hardeveld-maai-portfolio/-/blob/main/wiki/evaluatie-en-validatie/testresultaten.md

5 Discussie

5.1 Evaluatie en interpretatie van de resultaten

De evaluatie is bewust per component uitgevoerd, zodat zichtbaar wordt welk deel van de pipeline welke kwaliteit levert. De regelgebaseerde beoordelaar presteert het sterkst van de vergeleken methoden en generaliseert met de binaire beslisregel en BGE-M3 naar twee onafhankelijke sets échte leerlingantwoorden; de retrieval haalt de vooraf gestelde hit rate-drempel. De LLM-misconceptiedetectie kent een precisiegrens die door de leerlingbevestiging wordt ondervangen, terwijl een robuustheidsfix de recall borgde. De leesbaarheidsanalyse en de exploratieve gebruikerstest zijn indicatief, niet bevestigend: de steekproeven zijn te klein voor harde conclusies.

De accuratesse vraagt om voorzichtige lezing: de benchmark en de validatiesets bestaan uit antwoorden van voormalige vwo-leerlingen, en een systematische validatie op actieve leerlingen in de klas staat nog uit. Dat de beoordelaar een volledig antwoord betrouwbaar als *correct* herkent, is verklaarbaar: zo'n antwoord dekt de rubric-punten ruim, waardoor de gemiddelde conceptdekking duidelijk boven de drempel uitkomt. De resterende fouten liggen vooral rond die drempel, met een concentratie bij antwoorden met spelfouten. Een winstpunt is dat de beoordelaar zijn oordeel per concept verantwoordt: grensgevallen kan het systeem als onzeker markeren en naar de docent escaleren, in plaats van een schijnzekere uitkomst te tonen.

5.2 Evaluatie van ontwerpkeuzes en iteraties

De bepalende ontwerpkeuze was de overgang van een LLM-judge naar een regelgebaseerde beoordelaar, tegen een kleine accuratesse-prijs (circa 92% tegen 96%) voor determinisme, herleidbaarheid en snelheid. De ontwerpkeuzes zijn iteratief tot stand gekomen: de lage precisie van de misconceptiedetectie leidde tot de leerlingbevestiging als precisievangnet voor de detector, terwijl de beoordelaar zelf een menselijk vangnet kent doordat de docent elk oordeel kan inzien en op goed of fout corrigeren (REQ 1.2.1).

5.3 Relatie tussen resultaten, requirements en doelstellingen

Van de negentien must-requirements zijn er zestien volledig gehaald; drie zijn op prototypeniveau gedeeltelijk ingevuld, omdat hun volledige realisatie een productiecontext vereist. Tabel 11 vat de status samen.

Tabel 11. Verificatie van de must-requirements

ID	Requirement (kort)	Status
Juridisch		
REQ 1.1.1	Lokale verwerking, geen herleidbare gegevens extern	Gehaald
REQ 1.1.2	Transparantie over AI-inzet en lokale verwerking	Gehaald
REQ 1.2.1	Geen autonome beslissing zonder menselijke bevestiging	Gehaald
REQ 1.2.2	Beoordeling uitsluitend formatief	Gehaald
Ethisch		
REQ 2.1.1	Gelijke beoordelingskwaliteit ongeacht schrijfstijl	Deels (validatie)
REQ 2.2.1	Transparante feedback en expliciete misconceptie-uitleg	Gehaald
Functioneel		
REQ 4.1.1	Adaptieve vraagselectie op basis van BKT-beheersing	Gehaald
REQ 4.2.1	Misconcepties geregistreerd en sturen de vraagkeuze	Gehaald
REQ 4.2.2	Prioriteit voor misconceptie-corrigerende vragen	Gehaald
REQ 4.2.4	Gegenereerde vragen alleen na docentgoedkeuring	Gehaald
REQ 4.3.1	Beoordeling op betekenisniveau, niet op woordoverlap	Gehaald
REQ 4.3.2	Feedback in het Nederlands, maximaal drie zinnen	Gehaald
REQ 4.4.1	Vragenbank van minimaal 50 getagde vragen	Gehaald
Technisch		
REQ 5.1.1	Gestructureerde foutcode, deterministische beoordeling	Gehaald
REQ 5.2.1	HTTPS en geen persoonsgegevens naar externe API's	Deels (productie)
REQ 5.3.1	Harde time-out op LLM-aanroep en UI-laadstatus	Deels (productie)
REQ 5.4.1	Deterministisch, reproduceerbaar oordeel	Gehaald
REQ 5.5.1	FastAPI-backend, broncode openbaar op GitLab	Gehaald
REQ 5.5.2	RAG asynchroon, content met schriftelijke toestemming	Gehaald

De drie gedeeltelijk ingevulde requirements delen één oorzaak: ze vergen een productie- of veldcontext buiten de prototypescope. REQ 5.2.1 (HTTPS) is een deploymenteis: het prototype is inmiddels op een externe server uitgerold, maar het afdwingen van TLS hoort bij een geharde productieopstelling en valt buiten de prototypescope; "geen persoonsgegevens naar externe API's" is wél geborgd. Bij REQ 5.3.1 toont de interface een laadstatus, maar een harde time-out op de vraaggeneratie hoort bij productie-inzet. REQ 2.1.1 (gelijke beoordelingskwaliteit ongeacht schrijfstijl) is op de twee gevarieerde sets grotendeels aangetoond, met spelfout-antwoorden als zwakste segment. De centrale onderzoeksvraag is daarmee op prototypeniveau grotendeels beantwoord: het systeem verwerkt open antwoorden adaptief, privacyvriendelijk en transparant, en voldoet aan de must-requirements voor zover zonder productiecontext realiseerbaar. Of het in de praktijk leerwinst oplevert en of de adaptieve lus end-to-end effectief is, vergt validatie met de werkelijke doelgroep.

5.4 Beperkingen van het onderzoek

De opzet en uitvoering kennen een aantal beperkingen die de geldigheid van de conclusies inperken.

Omvang van de vragenbank. Met 50 items is de vragenbank toereikend als proof-of-concept, maar voor stabiele BKT-parameterschatting zijn grotere itemsets aanbevolen [27]: de BKT-parameters en de handmatig ingeschatte moeilijkheidswaarden zijn daarmee initiële priors, geen empirisch onderbouwde schattingen, en zouden bij opschaling en na een pilot gekalibreerd moeten worden.

Ontbrekende domeinexpertvalidatie. Formele inhoudsvalidatie door een externe biologie-vakexpert heeft niet plaatsgevonden. Deze validatie is wel nagestreefd, maar het betrekken van een gekwalificeerde vakexpert bleek binnen de projectkaders niet haalbaar. Ook de vijf misconceptiepatronen (gebaseerd op literatuur en curriculumanalyse) zijn niet empirisch bij de doelgroep gevalideerd. Fouten in correctiemodellen of misconceptiebeschrijvingen kunnen daardoor niet worden uitgesloten, een risico dat zwaarder weegt omdat de regelgebaseerde judge volledig op de kwaliteit van het antwoordmodel steunt.

Onafhankelijke conceptbeoordeling en denkfout versus kennisgat. De judge beoordeelt concepten onafhankelijk van elkaar (sectie 3.5.2), terwijl rubric-punten inhoudelijk kunnen samenhangen; daarnaast onderscheidt de misconceptiedetectie niet expliciet tussen een actieve denkfout en het simpelweg niet weten van het antwoord; beide krijgen dezelfde generieke terugkoppeling. Een verfijnder onderscheid is toekomstig werk.

Automatische scoring en validatiepopulatie. De validatie op twee onafhankelijke sets, met drempels op de kalibratiehelft en vijfvoudige kruisvalidatie als generalisatiemaat, geeft een robuuste schatting, maar de antwoorden komen van oud-leerlingen en zijn door de onderzoeker zelf gelabeld. Een klassikale pilot met de actieve doelgroep en een onafhankelijke tweede beoordelaar blijft nodig om de drempel en de keuze voor BGE-M3 definitief te onderbouwen.

Hardware-afhankelijkheid van het embedmodel. BGE-M3 vergt circa 1,8 GB extra werkgeheugen ten opzichte van MiniLM; op de CPU-only VPS (8 GB) is dat budget krap maar werkbaar, met MiniLM als terugvaloptie.

Beperkte dekking RAG-kennisbasis. Elf Kennislink-artikelen dekken het abiogenese-domein niet uitputtend, waardoor de RAG-pipeline bij niche-misconcepties onvoldoende geschikte retrieval-resultaten kan opleveren.

Representativiteit gebruikerstest. De exploratieve gebruikerstest is uitgevoerd met drie voormalige vwo-leerlingen, buiten de primaire doelgroep: de bevindingen zijn indicatief voor usability-knelpunten, maar niet generaliseerbaar naar huidige vwo-leerlingen zonder aanvullend onderzoek. Voor vakdocenten is één evaluatie uitgevoerd, de docentdemo met Marco Kragten (sectie 4.1); bredere generalisatie naar vakdocenten vraagt om meerdere onafhankelijke docenten.

5.5 Reflectie op betrouwbaarheid, validiteit en generaliseerbaarheid

Dit onderzoek volgt een design-science-aanpak, passend bij het exploratieve, prototypische karakter. De **betrouwbaarheid** van de regelgebaseerde judge is hoog dankzij het deterministische ontwerp, met de eerder gerapporteerde kruisvalidatie- en held-out-cijfers (sectie 4.2) als kwantitatieve onderbouwing. De **interne validiteit** is begrensd doordat de testantwoorden door de onderzoeker zelf zijn gelabeld, zonder onafhankelijke tweede beoordelaar. De **constructvaliditeit** van de misconceptielabels is onzeker en de **generaliseerbaarheid** beperkt: alle evaluaties betreffen één domein en kleine steekproeven. De bevindingen zijn daarom het best te lezen als een onderbouwd proof of concept, niet als bewijs van effectiviteit in de klas.

5.6 Ethische en maatschappelijke reflectie

Impact op de leerling. De module geeft directe feedback op redeneerfouten en biedt individuele terugkoppeling die in een klas van 25 tot 30 leerlingen structureel ontbreekt [2]. Een keerzijde is afhankelijkheid: wie elke redeneer-twijfel aan het systeem overlaat, reproduceert AI-uitvoer in plaats van zelf te denken [35], gemitigeerd doordat de feedback tot drie zinnen beperkt blijft, alleen de ontbrekende kernbegrippen als hint benoemt zonder het volledige antwoordmodel te tonen, en pas na een eigen antwoordpoging verschijnt (REQ 2.3.1). Een tweede risico is foutieve feedback van het taalmodel, zoals bleek tijdens de docentdemo (sectie 4.1): de natuurlijke-taal uitleg is niet equivalent aan menselijke beoordeling. Omdat de beoordeling formatief is en de docent eindverantwoordelijk blijft, zijn de gevolgen beperkt, maar het blijft een structurele beperking.

Impact op de docent. Het systeem is een aanvulling op, geen vervanging van de docent. De docentdemo toonde dat vakdocenten de reviewworkflow waarderen: zij behouden eindverantwoordelijkheid zonder elke vraag zelf te formuleren. Een praktische zorg is de werkdruk bij grotere klassen; vervolgonderzoek moet bepalen of die acceptabel blijft.

Maatschappelijke implicaties. De keuze voor volledig lokale inferentie is bewust: leerlinggegevens blijven binnen de schoolinfrastructuur, wat de voorkeur verdient boven de gangbare cloudpraktijk [15]. Dit roept wel de vraag op naar gelijkwaardige toegang: scholen zonder eigen VPS of IT-capaciteit kunnen het systeem niet draaien, wat de kloof tussen scholen kan vergroten.

6 Conclusie

6.1 Conclusie en terugkoppeling op de requirements

Dit onderzoek onderzocht hoe een privacy-compliant, adaptief tutorsysteem vwo-leerlingen kan ondersteunen bij abiogenese door open antwoorden te beoordelen, misconcepties te signaleren en gericht oefenmateriaal te genereren. Het resultaat is een werkend prototype dat de must-have requirements grotendeels invult. Open antwoorden worden lokaal, betekenisgericht en reproduceerbaar beoordeeld (REQ 4.3, 5.4) door een regelgebaseerde, semantische beoordelaar die een binair goed/fout-oordeel velt via een meerderheidsstem van drie gedecorreleerde maten (92% ±7% bij vijfvoudige kruisvalidatie, tegen 72% voor het eerdere ontwerp). Leerlingdata blijven binnen de eigen infrastructuur (REQ 1.1), en elke gegenereerde vraag vereist docentgoedkeuring (REQ 1.2, 4.2.4). De adaptieve mechanismen (BKT, Bloom-progressie, spaced repetition) en de RAG-vraaggeneratie zijn geïmplementeerd en op componentniveau geëvalueerd, maar de volledige adaptieve lus is nog niet end-to-end op echte VWO leerdata getoetst. De deelvraag naar de prestatie van het model is daarmee positief, maar voorlopig, beantwoord; de praktijkvalidatie blijft open.

6.2 Implicaties, toekomstig werk en aanbevelingen

De belangrijkste implicatie is dat een transparante, regelgebaseerde beoordelaar voor rubric-gestuurde open vragen, in de vorm van een meerderheidsstem van gedecorreleerde maten, dicht bij de beste lokale LLM-judge komt (circa 92% tegen 96%), terwijl zij reproduceerbaar is, lokaal draait en per concept uitlegbaar is. Voor onderwijscontexten met privacy- en hardware-eisen kan dat een aantrekkelijk alternatief zijn voor cloudgebaseerde of zware lokale modellen. Toekomstig werk richt zich op vier punten: (1) validatie met de werkelijke doelgroep, met authentieke leerlingantwoorden en een stakeholderevaluatie met vakdocenten; (2) uitbreiding en expertvalidatie van de vragenbank, zodat de BKT-parameters empirisch gekalibreerd kunnen worden; (3) uitbreiding van de RAG-kennisbasis om gaten in de domeindekking (zoals endosymbiose) te dichten; en (4) een end-to-end-evaluatie van de volledige adaptieve lus op gesimuleerde en echte

leersessies. Voor inzet in de praktijk geldt als aanbeveling het systeem formatief en docentondersteunend te houden, met escalatie naar de docent bij onzekere beoordelingen.

Bronnenlijst

[1] PRELIFE Consortium. "PRELIFE – Pathways, Reactions, and Environments leading to LIFE: An interdisciplinary approach". NWO projectvoorstel, ongepubliceerd intern document. 2024.

[2] Kurt VanLehn. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems". In: *Educational Psychologist* 46.4 (2011), p. 197–221. DOI: 10.1080/00461520.2011.611369.

[3] SLO – Nationaal Expertisecentrum Leerplanontwikkeling. *P6: Ontstaan van het leven — vwo (subdomein F4), Handreiking schoolexamen biologie havo/vwo*. Website. Geraadpleegd april 2026. 2023. URL: https://www.slo.nl/handreikingen/havo-vwo/handreiking-se-bio-hv/examenprogramma/inhoud-examenprogramma/p6-ontstaan-leven-vwo-subdomein-f4/.

[4] Lorin W. Anderson en David R. Krathwohl. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001.

[5] J. W. Rice, D. A. Warner, C. D. Kelly e.a. "The Theory of Evolution is Not an Explanation for the Origin of Life". In: *Evolution: Education and Outreach* 3 (2010), p. 141–142. DOI: 10.1007/s12052-010-0225-1.

[6] Barend Vlaardingerbroek. "Abiogenesis in Upper Secondary Biology Curricula". In: *Evolution: Education and Outreach* 3 (2010), p. 432–435. DOI: 10.1007/s12052-010-0209-1.

[7] Neil T. Heffernan en Cristina Lindquist Heffernan. "The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching". In: *International Journal of Artificial Intelligence in Education* 24.4 (2014), p. 470–497. DOI: 10.1007/s40593-014-0024-x.

[8] Steven Ritter e.a. "Cognitive Tutor: Applied research in mathematics education". In: *Psychonomic Bulletin & Review* 14.2 (2007), p. 249–255. DOI: 10.3758/BF03194060.

[9] Steven Burrows, Iryna Gurevych en Benno Stein. "The Eras and Trends of Automatic Short Answer Grading". In: *International Journal of Artificial Intelligence in Education* 25.1 (2015), p. 60–117. DOI: 10.1007/s40593-014-0026-8.

[10] Gerd Kortemeyer. "Performance of the Pre-Trained Large Language Model GPT-4 on Automated Short Answer Grading". In: *Discover Artificial Intelligence* (2024). DOI: 10.48550/arXiv.2309.09338. arXiv: 2309.09338.

[11] Hunter McNichols e.a. "Can Large Language Models Replicate ITS Feedback on Open-Ended Math Questions?" In: *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*. 2024. DOI: 10.48550/arXiv.2405.06414. arXiv: 2405.06414.

[12] Lianmin Zheng e.a. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Deel 36. 2023.

[13] Yunfan Gao e.a. "Retrieval-Augmented Generation for Large Language Models: A Survey". In: *arXiv preprint* (2023). DOI: 10.48550/arXiv.2312.10997. arXiv: 2312.10997.

[14] Alexander Scarlatos, Ryan S. Baker en Andrew Lan. "Exploring Knowledge Tracing in Tutor-Student Dialogues using LLMs". In: *arXiv preprint* (2024). DOI: 10.48550/arXiv.2409.16490. arXiv: 2409.16490.

[15] Europees Parlement en Raad van de Europese Unie. *Verordening (EU) 2016/679 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens (Algemene Verordening Gegevensbescherming)*. Verordening L 119/1. Publicatieblad van de Europese Unie, 2016. URL: https://gdpr-info.eu.

[16] Nils Reimers en Iryna Gurevych. “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020. DOI: 10.48550/arXiv.2004.09813. arXiv: 2004.09813.

[17] Zechun Liu, Changsheng Zhao, Forrest Iandola e.a. “MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2024. DOI: 10.48550/arXiv.2402.14905. arXiv: 2402.14905.

[18] Elias Frantar e.a. “GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers”. In: *International Conference on Learning Representations (ICLR)*. 2023. DOI: 10.48550/arXiv.2210.17323. arXiv: 2210.17323.

[19] Patrick Lewis e.a. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Deel 33. 2020. DOI: 10.48550/arXiv.2005.11401. arXiv: 2005.11401.

[20] Kefyn M. Catley en Laura R. Novick. “Digging deep: Exploring college students’ knowledge of macroevolutionary time”. In: *Journal of Research in Science Teaching* 46.3 (2009), p. 311–332. DOI: 10.1002/tea.20273.

[21] Deborah Kelemen. “Why Are Rocks Pointy? Children’s Preference for Teleological Explanations of the Natural World”. In: *Developmental Psychology* 35.6 (1999), p. 1440–1452. DOI: 10.1037/0012-1649.35.6.1440.

[22] John D. Coley en Kimberly D. Tanner. “Common Origins of Diverse Misconceptions: Cognitive Principles and the Development of Biology Thinking”. In: *CBE—Life Sciences Education* 11.3 (2012), p. 209–215. DOI: 10.1187/cbe.12-06-0074.

[23] George J. Posner e.a. “Accommodation of a Scientific Conception: Toward a Theory of Conceptual Change”. In: *Science Education* 66.2 (1982), p. 211–227. DOI: 10.1002/sce.3730660207.

[24] Michelene T. H. Chi, James D. Slotta en Nicholas de Leeuw. “From Things to Processes: A Theory of Conceptual Change for Learning Science Concepts”. In: *Learning and Instruction* 4.1 (1994), p. 27–43. DOI: 10.1016/0959-4752(94)90017-5.

[25] Benjamin Deonovic e.a. “Learning meets assessment”. In: *Behaviormetrika* 45.2 (2018), p. 457–474. DOI: 10.1007/s41237-018-0070-z.

[26] Philip I. Pavlik Jr., Hao Cen en Kenneth R. Koedinger. “Performance Factors Analysis — A New Alternative to Knowledge Tracing”. In: *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED 2009)*. Amsterdam: IOS Press, 2009, p. 531–538. DOI: 10.3233/978-1-60750-028-5-531.

[27] Albert T. Corbett en John R. Anderson. “Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge”. In: *User Modeling and User-Adapted Interaction* 4.4 (1994), p. 253–278. DOI: 10.1007/BF01099821.

[28] Nicholas J. Cepeda e.a. “Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis”. In: *Psychological Bulletin* 132.3 (2006), p. 354–380. DOI: 10.1037/0033-2909.132.3.354.

[29] Nils Reimers en Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2019. DOI: 10.48550/arXiv.1908.10084. arXiv: 1908.10084.

[30] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, p. 74–81.

[31] Kishore Papineni e.a. “BLEU: A Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2002, p. 311–318. DOI: 10.3115/1073083.1073135.

[32] Akshay Kore. *Designing Human-Centric AI Experiences: Applied UX Design for Artificial Intelligence*. 1ste ed. Apress, 2022. DOI: 10.1007/978-1-4842-8088-1.

[33] Europees Parlement en Raad van de Europese Unie. *Verordening (EU) 2024/1689 betreffende geharmoniseerde regels inzake artificiële intelligentie (AI-verordening)*. Verordening L 2024/1689. Publicatieblad van de Europese Unie, 2024. URL: https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=OJ:L_202401689.

[34] Paul Black en Dylan Wiliam. "Assessment and Classroom Learning". In: *Assessment in Education: Principles, Policy & Practice* 5.1 (1998), p. 7–74. DOI: 10.1080/0969595980050102.

[35] Evan F. Risko en Sam J. Gilbert. "Cognitive Offloading". In: *Trends in Cognitive Sciences* 20.9 (2016), p. 676–688. DOI: 10.1016/j.tics.2016.07.002.

[36] John Sweller. "Cognitive Load During Problem Solving: Effects on Learning". In: *Cognitive Science* 12.2 (1988), p. 257–285. DOI: 10.1207/s15516709cog1202_4.

[37] Jakob Nielsen. *Response Times: The 3 Important Limits*. Nielsen Norman Group. Geraadpleegd april 2026. 1994. URL: https://www.nngroup.com/articles/response-times-3-important-limits/.

[38] NEMO Kennislink. *Oorsprong van leven — artikelenserie*. Website. Schriftelijke toestemming verkregen april 2026 voor gebruik binnen onderwijsonderzoek. 2024. URL: https://www.nemokennislink.nl.

[39] Google People + AI Research (PAIR). *People + AI Guidebook: Principles & Patterns*. Google. 2019. URL: https://pair.withgoogle.com/guidebook/patterns (bezocht op 12-08-2026).

[40] AI voor Leerlingen. *AI voor Leerlingen — leer met AI*. Website. Geraadpleegd april 2026. 2024. URL: https://aivoorleerlingen.nl.

[41] Aameer Badshah en Hassan Sajjad. "Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form QA". In: *arXiv preprint* (2024). DOI: 10.48550/arXiv.2408.09235. arXiv: 2408.09235.

[42] Jianlv Chen e.a. "BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation". In: *arXiv preprint* (2024). arXiv: 2402.03216.

[43] Wenhui Wang e.a. "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers". In: *arXiv preprint* (2020). DOI: 10.48550/arXiv.2002.10957. arXiv: 2002.10957.

[44] Google DeepMind. *Gemma 4*. Official model page; model card updated 2 April 2026. 2026. URL: https://deepmind.google/models/gemma/gemma-4/.

[45] Henk Pander Maat, Suzanne Kleijn en Servaas Frissen. "LiNT: een leesbaarheidsformule en een leesbaarheidsinstrument". In: *Tijdschrift voor Taalbeheersing* 45.3 (2023), p. 2–39. DOI: 10.5117/TVT2023.3.002.MAAT.