Pilot GenAI

Onderzoeksrapport

Lectoraat Learning Technology & Analytics

2025

let's change
YOU. US. THE WORLD.

DE HAAGSE HOGESCHOOL

Pilot GenAI

Onderzoeksrapport

Auteurs*
Manika Garg, Aslam Tanjung, Sunčica Bruck, Theo Bakker

Afdeling
Lectoraat Learning Technology & Analytics

Datum
7 oktober 2025

Type project
Onderzoeksproject

Versie
1.0

*Manika Garg was projectleider (onderzoek), ontwierp de studie, begeleidde de pilot en leidde het schrijven van het rapport. Theo Bakker zorgde voor supervisie, beoordeelde het rapport en gaf kritische feedback. Aslam Tanjung trad op als data-engineer en voerde de gegevensverwerking en -analyse uit. Sunčica Bruck voerde het literatuuronderzoek uit en leverde een bijdrage aan de ontwikkeling van de Theory of Change. Alle auteurs hebben de definitieve versie van het rapport goedgekeurd.

Voorwoord

Op verzoek van de dienst Facilitaire Zaken & IT (FZ/IT) heeft het lectoraat Learning Technology & Analytics (LTA) van De Haagse Hogeschool (De HHs) het onderzoek naar de pilot GenAI ontworpen en uitgevoerd. Het onderzoek onderzocht de invloed van generatieve AI op de ervaren bruikbaarheid, kwaliteit en productiviteit van onderwijs-, onderzoeks- en ondersteunend personeel.

Dit rapport presenteert de onderzoeksvragen, methoden, bevindingen en reflecties uit de pilot. Het combineert literatuuronderzoek, veranderingstheorie, vragenlijstgegevens, interviews en een experimenteel onderzoek om aan te tonen waar GenAI waarde toevoegt, waar beperkingen blijven bestaan en welke voorwaarden nodig zijn voor verantwoorde opschaling.

Het rapport is bedoeld om de AI-stuurgroep, de bredere HHs-gemeenschap en alle belanghebbenden te informeren bij hun beslissingen over het toekomstige gebruik van GenAI binnen onze onderwijsinstelling.

Dankwoord

Wij danken de AI-stuurgroep van De HHs voor de opdracht naar het onderzoek van de Pilot GenAI. Speciale dank gaat uit naar Eva Willems, projectleider, en het FZ/IT-team voor hun ondersteuning. Wij danken ook het AI-expertteam voor het ontwikkelen van de persona’s, het HCTL-team voor het geven van training aan de deelnemers en Npuls voor het beschikbaar stellen van het EduGenAI-platform. Wij zijn alle medewerkers die aan de pilot hebben deelgenomen zeer dankbaar.

Inhoudsopgave

Managementsamenvatting

Het doel van de GenAI-pilot was om te onderzoeken hoe generatieve AI de bruikbaarheid, kwaliteit en productiviteit beïnvloedt van routinematige, professionele taken die worden uitgevoerd door onderwijs-, onderzoeks- en ondersteunend personeel binnen De HHs. De pilot was bedoeld om empirisch bewijs te genereren voor een routekaart voor de verantwoorde en duurzame integratie van GenAI in de hele instelling.

Uit de bevindingen van het proefproject blijkt dat GenAI de potentie heeft om de professionele taken van HHs-medewerkers te ondersteunen, mits GenAI zorgvuldig en gestructureerd wordt geïntegreerd.

Aanbevelingen voor opschaling van GenAI binnen De HHs

Op basis van de bevindingen van de pilot is een reeks aanbevelingen opgesteld om de AI-stuurgroep te begeleiden bij het stellen van prioriteiten voor verantwoorde en duurzame integratie van GenAI. De aanbevelingen zijn gebaseerd op het volledige bewijsmateriaal van de pilot, waaronder literatuur, veranderingstheorie, vragenlijstgegevens, interviews en een micro-RCT, die in de volgende paragrafen van dit rapport worden gepresenteerd.

Verschillen in digitale geletterdheid, geschiktheid van taken voor GenAI-ondersteuning en houding van medewerkers in verschillende domeinen (onderwijs, onderzoek en ondersteuning) onderstrepen dat het opschalen van GenAI binnen De HHs niet volgens een uniform model kan gebeuren, maar gebaseerd moet zijn op gedifferentieerde ondersteuning en zorgvuldige governance.

1. Strategie

- **Gefaseerde opschaling.** Begin in domeinen die daar het meest klaar voor zijn, zoals onderwijs en onderzoek, en blijf ondertussen relevante *use cases* in ondersteunende dienstverlening verkennen. De opschaling moet geleidelijk verlopen, met voortdurende monitoring en feedback om de training, de governance en de integratie bij te sturen.
- **Zorg voor duurzaamheid.** Pak de door medewerkers geuite bezorgdheid over het milieu aan door bewust gebruik aan te moedigen en een energie-efficiënte infrastructuur te onderzoeken.
- **Zorg voor menselijk toezicht.** Blijf medewerkers eraan herinneren dat door AI gegenereerde output moet worden geverifieerd en dat het vertrouwen op GenAI niet ten koste mag gaan van kritisch denken.

2. HRM en training

- **Gerichte training.** Digitale geletterdheid had een grote invloed op de resultaten: ervaren gebruikers behaalden betere resultaten, terwijl beginners zich onzeker voelden. Training moet daarom gelaagd zijn, met inleidende modules voor nieuwkomers en geavanceerde sessies voor ervaren gebruikers.
- **Ingebouwde ondersteuning.** Tutorials, promptbibliotheken en contextuele voorbeelden binnen het platform kunnen de leercurve verlagen en de afhankelijkheid van leren door vallen en opstaan verminderen.
- **Peer learning.** Een helpdeskfunctie en *communities* waar medewerkers ervaringen uitwisselen, zouden de acceptatie verder ondersteunen.

3. Processen

- **Cultuur van veilig experimenteren.** Medewerkers hebben ruimte nodig om GenAI te proberen, fouten te maken en vragen te stellen zonder bang te zijn voor oordelen. Kleine pilots binnen faculteiten of teams, die worden gezien als kansen voor collectief leren, kunnen helpen om verantwoord experimenteren gangbaar te maken.
- **Deel use cases.** Het benadrukken van positieve en praktische voorbeelden uit verschillende domeinen kan scepsis verminderen en laten zien hoe GenAI waarde creëert.

4. Procedures en IT

- **Versterk vertrouwen en governance.** AVG-compliance en institutioneel toezicht waren belangrijke redenen waarom medewerkers EduGenAI meer vertrouwden dan externe platforms. De HHs moet duidelijk communiceren welke gegevens worden verzameld, hoe deze worden beveiligd en wat ethisch gebruik inhoudt met betrekking tot het gebruikte GenAI-platform.
- **Ethisch toezicht.** Een adviesmechanisme of ethische commissie zou begeleiding kunnen bieden bij kwesties als *bias*, milieu-impact en gepaste grenzen van het gebruik van GenAI.
- **Workflow integratie.** Sommige ervaren gebruikers vonden EduGenAI minder handig dan externe tools omdat ze van systeem moesten wisselen. Om opschaling te laten slagen, moet GenAI naadloos integreren met tools die al dagelijks worden gebruikt.

Samenvatting

De pilot GenAI onderzocht hoe generatieve AI het dagelijkse werk van onderwijs-, onderzoeks- en ondersteunend personeel kan ondersteunen, waarbij de nadruk lag op drie dimensies: bruikbaarheid, productiviteit en kwaliteit. Dit rapport presenteert de resultaten van het onderzoek naar de Pilot GenAI die bij De HHs is uitgevoerd.

Het onderzoek is opgezet op basis van het Nederlandse 3E-raamwerk (ontwikkeld door LTA), dat drie niveaus van bewijs combineert om onderwijstechnologie te evalueren.

- Bewijsmateriaal op bronsniveau omvatte een literatuuronderzoek en een Theory of Change om verwachtingen in kaart te brengen.
- Het bewijs op zilverniveau werd verzameld door middel van vragenlijsten en semigestructureerd interviews, waarbij de percepties van medewerkers voor en na het gebruik van GenAI in kaart werden gebracht.
- Bewijsmateriaal op goudniveau werd gegenereerd door middel van een micro-gerandomiseerde gecontroleerde studie, waarin causale effecten op taakprestaties werden getest.

In totaal schreven 205 medewerkers zich in voor de pilot. Hiervan voldeden 159 aan de inclusiecriteria, vulden 106 beide vragenlijsten in, namen 67 deel aan de RCT en werden 10 geïnterviewd. De deelnemers vertegenwoordigden onderwijs-, onderzoeks- en ondersteunende functies, met verschillende leeftijden, digitale geletterdheid en eerdere AI-ervaring.

Bevindingen op basis van dimensies (bruikbaarheid, kwaliteit en productiviteit):

- Uit de bevindingen blijkt dat de bruikbaarheid positief werd beoordeeld. Het gebruiksgemak bleef hoog in alle domeinen (voor = 4,77; na = 4,65; niet significant [ns]). Het waargenomen nut nam echter af na praktische ervaring (voor = 5,33; na = 4,20; $p < 0,001$), omdat medewerkers kritischer werden nadat ze de tool in de praktijk hadden gebruikt.
- Kwaliteit had het sterkste effect. Outputs die met GenAI werden geproduceerd, kregen hogere scores in GPT-gebaseerde evaluaties (M = 4,62 vs. 3,68; $p < .001$), terwijl zelfgerapporteerde scores geen significant verschil lieten zien (M = 4,07 vs. 3,82; $p = .120$). Tegelijkertijd benadrukten de deelnemers de noodzaak van menselijk toezicht vanwege bezorgdheid over nauwkeurigheid en hallucinaties.
- De productiviteitsresultaten waren gemengd. GenAI-ondersteunde taken werden sneller voltooid in de RCT (M = 38,8 minuten vs. 45,9 minuten; $p = .015$). De op GPT gebaseerde productiviteitsscores waren significant hoger ($p < 0,001$), terwijl de zelfgerapporteerde productiviteitsstijgingen minder uitgesproken waren ($p = 0,008$), wat een weerspiegeling is van de extra inspanning die nodig was om de nieuwe technologie te leren, prompts te verfijnen en outputs te beoordelen.

Bevindingen op basis van domein (onderwijs, onderzoek en ondersteuning): De verschillen werden bepaald door variërende niveaus van digitale geletterdheid, eerdere ervaring met AI en de geschiktheid van taken voor GenAI-ondersteuning.

- Docenten waardeerden GenAI voor het structureren en opstellen van lesmateriaal, maar benadrukten dat pedagogisch inzicht essentieel was om de output aan te passen aan de behoeften van de studenten.
- Onderzoekers, die over het algemeen over een hogere digitale geletterdheid beschikken, rapporteerden de grootste voordelen, vooral voor schrijfopdrachten zoals samenvattingen en het opstellen van rapporten. Zij uitten echter ook hun bezorgdheid over hallucinaties en verzonnen referenties.
- Ondersteunend personeel had de meest uiteenlopende ervaringen. Sommigen erkenden dat ze tijd bespaarden bij taken zoals het samenvatten van vergaderingen, terwijl anderen moeite hadden om GenAI zinvol te verbinden aan hun contextspecifieke rollen, waarbij vaak persoonlijke

gegevens betrokken waren. Hun lagere digitale geletterdheid droeg bij aan hun aarzeling om met de technologie te experimenteren.

Omstandigheden die het gebruik bepalen: Ondersteuning door het management, AVG-compliance, technologische training en helpdeskbronnen maakten acceptatie mogelijk, terwijl scepsis over de nauwkeurigheid, ongelijke digitale geletterdheid, bezorgdheid over het milieu en angst voor overmatige afhankelijkheid een bredere acceptatie in de weg stonden.

Conclusie: De Pilot GenAI laat zien dat GenAI de efficiëntie en kwaliteit in het hoger onderwijs kan verbeteren, maar dat de waarde ervan afhangt van de bereidheid van de gebruiker, culturele omstandigheden en voortdurend menselijk toezicht. De opschaling bij De HHs moet in fasen plaatsvinden, met gedifferentieerde training, sterk bestuur, integratie in bestaande werkprocessen en open aandacht voor ethiek en duurzaamheid.

1 Inleiding

Het **Instellingsplan 2023-2028 van De HHs** legt de nadruk op digitalisering, co-creatie, inclusiviteit en onderzoeksgestuurde innovatie [1]. Net als veel andere academische instellingen staat De HHs echter voor uitdagingen op het gebied van digitalisering in de praktijk. Medewerkers van De HHs (onderwijs/docenten, onderzoekers en ondersteunend personeel) verschillen sterk in hun digitale geletterdheid. Sommigen zijn zelfverzekerd in het experimenteren met en integreren van nieuwe technologieën, terwijl anderen meer begeleiding en ondersteuning nodig hebben.

In een tijd waarin generatieve AI (GenAI) snel aan belang wint en veel medewerkers het gebruik ervan al verkennen, wil De HHs meer inzicht krijgen in de impact ervan. De instelling wil toekomstige beslissingen baseren op bewijs, zodat de integratie van GenAI in professionele werkprocessen verantwoord, effectief en in lijn met de institutionele doelstellingen verloopt. In lijn hiermee heeft de **AI-stuurgroep** opdracht gegeven voor het project **Pilot GenAI** om empirisch bewijs te verzamelen over hoe generatieve AI HHs-medewerkers kan ondersteunen, waar de beperkingen liggen en onder welke voorwaarden het op verantwoorde wijze in de instelling kan worden geïntegreerd.

In de volgende subparagrafen worden GenAI en de onderzoeksvragen die aan deze studie ten grondslag liggen, geïntroduceerd.

1.1 Generatieve AI

Generatieve kunstmatige intelligentie of GenAI verwijst naar modellen die zijn getraind op grote datasets en die nieuwe inhoud kunnen genereren, zoals tekst, afbeeldingen of code, wanneer gebruikers daarom vragen [2]. Deze modellen herkennen patronen in bestaande gegevens en gebruiken probabilistische methoden om samenhangende outputs te voorspellen en te construeren. Hoewel AI al tientallen jaren in ontwikkeling is, hebben generatieve modellen sinds 2017 een snelle ontwikkeling doorgemaakt met de introductie van *transformer* architecturen. De algemene acceptatie ervan versnelde in 2022 na de release van grootschalige toepassingen zoals ChatGPT [3]. Tegenwoordig wordt GenAI toegepast in vele sectoren, waaronder het bedrijfsleven, de gezondheidszorg, de rechtspraak en het onderwijs, waardoor het een snelgroeiend onderdeel is geworden van de professionele en academische wereld.

In het onderwijs heeft het gebruik van GenAI tot discussie geleid. Enerzijds ondersteunt GenAI de efficiëntie door repetitieve taken te automatiseren, te helpen bij het schrijven en educatieve inhoud te genereren. Anderzijds blijven er zorgen bestaan over overmatige afhankelijkheid, mogelijk verlies van kritische vaardigheden, de nauwkeurigheid van de output en het ethisch gebruik van institutionele gegevens [4]. Deze discussies zijn direct relevant voor De HHs, waar veel medewerkers al zijn begonnen met het experimenteren met GenAI in hun dagelijkse professionele taken.

1.2 Onderzoeksvragen

Het project Pilot GenAI had tot doel inzicht te krijgen in hoe het gebruik van GenAI door HHs-medewerkers van invloed is op drie kerndimensies: bruikbaarheid, kwaliteit en productiviteit.

1. **Bruikbaarheid** verwijst naar hoe medewerkers GenAI ervaren in relatie tot hun professionele taken. Dit omvat zowel gebruiksgemak (of het platform intuïtief en toegankelijk is) als bruikbaarheid (of de tool medewerkers ondersteunt bij het effectiever uitvoeren van hun taken).
2. **Kwaliteit** verwijst naar de standaard van het werk dat wordt geproduceerd bij het uitvoeren van professionele taken. Het weerspiegelt zowel de technische nauwkeurigheid als de professionele relevantie van de output die met GenAI wordt gegenereerd.
3. **Productiviteit** verwijst naar de mate waarin medewerkers hun professionele taken efficiënter kunnen uitvoeren wanneer ze worden ondersteund door GenAI. Het geeft inzicht in de vraag of GenAI medewerkers heeft geholpen tijd te besparen, taken sneller te voltooien of effectiever te werken.

Bruikbaarheid richt zich op percepties en ervaringen, de kwaliteit van de resultaten en de productiviteit, waarmee de efficiëntie van werkprocessen wordt aangepakt.

Het project heeft deze dimensies bekeken in drie domeinen: **onderwijzend personeel/docenten** (onderwijs- en pedagogische functies), **onderzoekspersoneel** (academische en toegepaste onderzoeksfuncties) en **ondersteunend personeel** (administratieve, beleids- en dienstverlenende functies).

Samengevat zijn er drie **onderzoeksvragen** die als leidraad dienen voor deze studie:
1. Hoe ervaren medewerkers de bruikbaarheid van GenAI?
2. Op welke manieren beïnvloedt GenAI de kwaliteit van de output van medewerkers?
3. Op welke manieren beïnvloedt GenAI de productiviteit van medewerkers?

De antwoorden op deze vragen helpen bij het identificeren van zowel kansen als beperkingen voor de invoering en leveren input voor een routekaart naar een verantwoorde en duurzame integratie van GenAI binnen De HHs.

De rest van het rapport is als volgt opgebouwd: in hoofdstuk 2 wordt de methodologie beschreven, in hoofdstuk 3 worden de bevindingen gerapporteerd, in hoofdstuk 4 worden de ervaringen met het EduGenAI-platform beschreven en in hoofdstuk 5 worden de resultaten besproken.

2 Methodologie

In dit hoofdstuk wordt beschreven hoe het Pilot GenAI-project is opgezet en uitgevoerd. Het geeft een overzicht van het onderzoeksontwerp op basis van het Nederlandse 3E-raamwerk, het EduGenAI-platform dat binnen het pilotproject is gebruikt, legt de volgorde van de activiteiten tijdens het pilotproject uit en geeft details over de deelnemers en ethische overwegingen.

2.1 Onderzoeksopzet

De Pilot GenAI is ontworpen volgens **het Nederlandse (Evidence-informed Evaluation of EdTech) 3E-raamwerk** [5]. Het raamwerk is ontwikkeld door het lectoraat Learning Technology & Analytics van De HHs in samenwerking met Npuls. Het is opgezet om instellingen een gestructureerde manier te bieden om onderwijstechnologische hulpmiddelen te evalueren, waarbij verschillende niveaus van bewijs worden gecombineerd om zowel de potentiële als de daadwerkelijke impact te beoordelen.

Het Nederlandse 3E-kader onderscheidt drie niveaus van bewijs:
(1) Bronzen bewijs (theoretische inzichten om verwachtingen vast te stellen),
(2) Zilveren bewijs (op de praktijk gebaseerde inzichten die percepties weergeven) en
(3) Gouden bewijs (causaal bewijs voor meetbare impact).

Elk niveau biedt een ander perspectief, variërend van theorie tot praktijk tot causaliteit. Samen gaven deze verschillende perspectieven een uitgebreid beeld van hoe GenAI de bruikbaarheid, kwaliteit en productiviteit van het werk bij De HHs beïnvloedt.

1. **Brons –** Op dit niveau zijn twee methoden gebruikt om verwachtingen vast te stellen:

    - Aan het begin van het project werd een literatuuronderzoek uitgevoerd (gedetailleerd beschreven in paragraaf 3.1.1). Relevante internationale en nationale studies over het gebruik van GenAI in het onderwijs, onderzoek en ondersteunend personeel binnen onderwijsinstellingen werden

geanalyseerd. Het onderzoek bracht de potentiële voordelen en risico's in kaart van de integratie van GenAI in de professionele werkprocessen van een kennisinstelling.

- Vervolgens werd een *Theory of Change* (ToC) ontwikkeld (paragraaf 3.1.2). De ToC maakte de aannames achter de pilot expliciet en diende zowel als een reeks hypothesen voor deze studie als basis voor het plannen van de toekomstige opschaling en integratie van GenAI in de werkprocessen van De HHs.

2. **Zilver –** Op dit niveau werden twee methoden gebruikt om de percepties van medewerkers in kaart te brengen:

   Er werden twee online **vragenlijsten** gehouden. Beide vragenlijsten volgden het *Technology Acceptance Model* (TAM) [6]. TAM is een veelgebruikte vragenlijst voor het bestuderen van de acceptatie van nieuwe technologieën. Het meet twee kernconstructen: (1) *Perceived Ease of Use* (PEoU), de mate waarin gebruikers verwachten dat het systeem geen inspanning vereist, en (2) *Perceived Usefulness* (PU), de mate waarin gebruikers geloven dat het systeem hun werk zal verbeteren.

   De eerste basis-vragenlijst (Pre-Pilot TAM) werd afgenomen vóór de toegang tot het EduGenAI-platform (besproken in paragraaf 2.2) om de verwachtingen te meten. De vervolgvragenlijst (Post-Pilot TAM) werd na vier weken gebruik afgenomen om de daadwerkelijke ervaringen vast te leggen. De antwoorden van beide vragenlijsten werden vervolgens geanalyseerd om de veranderingen in perceptie als gevolg van het daadwerkelijke gebruik van het platform te beoordelen.

   - Er werden **interviews** gehouden met 10 deelnemers uit de drie domeinen op verschillende niveaus van digitale geletterdheid. Om vooringenomenheid door alleen enthousiaste deelnemers uit te sluiten, werd ook contact opgenomen met medewerkers die waren afgehaakt tijdens de pilot en hen gevraagd hun mening te geven. In de interviews werd gedetailleerd onderzocht hoe deelnemers GenAI hadden ervaren, welke voordelen zij hadden geconstateerd en welke uitdagingen zij waren tegengekomen.

     Deze interviews werden online gehouden via MS Teams en werden met toestemming opgenomen, getranscribeerd en geanonimiseerd. Vervolgens werden interviewverslagen opgesteld en teruggekoppeld aan de deelnemers ter verificatie van de interpretatie. De inzichten werden geanalyseerd door middel van thematische codering, waarbij terugkerende patronen en verschillen tussen de domeinen werden geïdentificeerd. Vervolgens werden deze thema's vergeleken met de andere bevindingen om een rijkere context te bieden en observaties te verklaren.

3. **Goud** – Op dit niveau werden causale effecten gemeten door middel van een experiment.

   - Na de toegangsperiode van vier weken tot EduGenAI werd een **micro-RCT** georganiseerd. RCT is een methode waarbij deelnemers willekeurig in groepen worden ingedeeld om het effect van een interventie te testen [7]. In deze studie voltooide groep A (de experimentele groep) taken met GenAI, terwijl groep B (de controlegroep) dezelfde taken zonder GenAI voltooide. De randomisatie werd afzonderlijk uitgevoerd binnen elk van de drie domeinen (onderwijs, onderzoek en ondersteuning). Er werd gebruikgemaakt van een gestratificeerde steekproefmethode, eerst op basis van het niveau van digitale geletterdheid van de deelnemers en vervolgens op basis van hun leeftijd. Binnen elk domein werden de deelnemers vervolgens willekeurig ingedeeld in de experimentele groep (gebruik van GenAI) of de controlegroep (geen gebruik van GenAI). Een momentopname van de deelnemers op de dag van de RCT is te zien in figuur 1.

Er wordt gesproken van een micro-RCT omdat deze over een kort tijdsbestek (één dag in juni 2025) met een relatief klein aantal deelnemers werd uitgevoerd. Voor deze methode werd gekozen om causaal bewijs te verkrijgen over productiviteit en kwaliteit, inzichten die niet alleen uit vragenlijsten of interviews konden worden afgeleid.

Tijdens de RCT voltooide elke deelnemer een taak die was ontworpen om hun beroepspraktijk te weerspiegelen. Leraren stelden een lesplan op, onderzoekers schreven een korte samenvatting van hun onderzoek en ondersteunend personeel maakte een vergaderverslag of beleidsupdate. De uitkomstmaten omvatten de duur van de taak, de zelfbeoordeelde kwaliteit en de door GPT geëvalueerde kwaliteit. Dit laatste werd uitgevoerd door twee LLM's (GPT-4.1 en mistral-large-2411). Deze LLM's werden geselecteerd vanwege hun sterke benchmarkprestaties en complementaire benaderingen. Voor de evaluatie van de output kregen de LLM's gestructureerde instructies om de tekstkwaliteit te beoordelen. Elk model werd gevraagd om de output te beoordelen aan de hand van vooraf gedefinieerde criteria (duidelijkheid, structuur, relevantie en coherentie) op een schaal van 1 tot 5. Er werden verschillende sets prompts op maat gemaakt voor het type taak om ervoor te zorgen dat de evaluatie een realistische weerspiegeling was van de normen voor het werk van HHs-medewerkers. Ten slotte werd de interbeoordelaarsbetrouwbaarheid voor beide modellen gecontroleerd om de consistentie te waarborgen.

Figuur 1. De deelnemers tijdens de RCT-dag

2.2 EduGenAI-platform

De pilot werd uitgevoerd met behulp van het EduGenAI-platform [8], ontwikkeld door Npuls [9], het nationale digitale transformatieprogramma voor hoger en beroepsonderwijs in Nederland. Npuls heeft tot doel de digitalisering in het onderwijs te versnellen door gezamenlijke infrastructuren, kennis en tools voor instellingen te ontwikkelen. Het EduGenAI-platform maakt deel uit van deze inspanning en is ontworpen om een veilige en AVG-conforme omgeving te bieden voor het verkennen van de toepassing van generatieve AI in het onderwijs. Het platform biedt een interface die vergelijkbaar is met andere grote taalmodellen, maar werkt binnen een gecontroleerde institutionele context.

> Het EduGenAI-platform werd gekozen voor de pilot op basis van de beschikbaarheid ervan binnen het nationale Npuls-initiatief, de naleving van wettelijke vereisten zoals de AVG en de geschiktheid voor gebruik in het hoger onderwijs.

Als onderdeel van de pilot heeft het **AI-expertteam** van De HHs drie **persona's** voor drie domeinen ontwikkeld. Een persona is in deze context een rolspecifieke configuratie van de AI, vooraf geladen met instructies die de institutionele prioriteiten en beperkingen weerspiegelen. Elke persona stuurde de AI aan om te reageren op een manier die relevant is voor het dagelijkse werk van medewerkers in dat domein.

De persona Onderwijs bevatte bijvoorbeeld prompts over lesontwerp, leerresultaten en feedback voor studenten; de persona Onderzoek bevatte prompts over literatuuronderzoek, samenvattingen en vragenlijsten; en de persona Ondersteuning bevatte prompts voor vergaderverslagen, communicatie en beleidsteksten. Deze persona's bevatten ook HHs-specifieke nalevingsvereisten en ethische grenzen, waardoor de output in overeenstemming was met het beleid en de waarden van de instelling. De persona's werden geïntroduceerd om de toegangsdrempel voor medewerkers te verlagen.

2.3 Volgorde van activiteiten

De pilot volgde een gestructureerde volgorde van voorbereiding tot rapportage (figuur 2). De voorbereidingsfase (april-mei 2025) omvatte het literatuuronderzoek en de ontwikkeling van de ToC. Deelnemers konden zich vanaf 15 april 2025 inschrijven en de pilot ging officieel van start met een kick-off op 15 mei. Tijdens de kick-off-sessie werd informatie gegeven over de pilot, waarna op 16 mei *informed consent* werd verzameld en de eerste vragenlijst (Pre-Pilot TAM) werd verspreid. Van 15 mei tot 15 juni hadden de deelnemers toegang tot het EduGenAI-platform. De tweede vragenlijst (Post-Pilot TAM) werd op 16 juni, direct na de gebruiksperiode, verstuurd. Op 19 juni werd een micro-RCT uitgevoerd voor alle domeinen. Na de RCT werden semigestructureerd interviews gehouden met medewerkers uit verschillende domeinen en met verschillende niveaus van digitale geletterdheid, waaronder enkele die waren afgehaakt. De analyse- en rapportagefase vond plaats tussen juli en september 2025.

Figuur 2. Tijdlijn van activiteiten

2.4 Deelnemers

De deelnemers aan dit onderzoek waren medewerkers van De HHs die werkzaam waren op het gebied van onderwijs, onderzoek en ondersteuning. Deelname aan het onderzoek was voorbehouden aan medewerkers die de *HCTL AI-basiscursus* met succes hadden afgerond, het toestemmingsformulier hadden ondertekend en toestemming van hun leidinggevende hadden gekregen om deel te nemen.

In totaal schreven 205 medewerkers zich in voor de pilot, maar slechts 159 deelnemers voldeden aan de inclusiecriteria. Deze groep vormde de kernsteekproef voor het onderzoek.

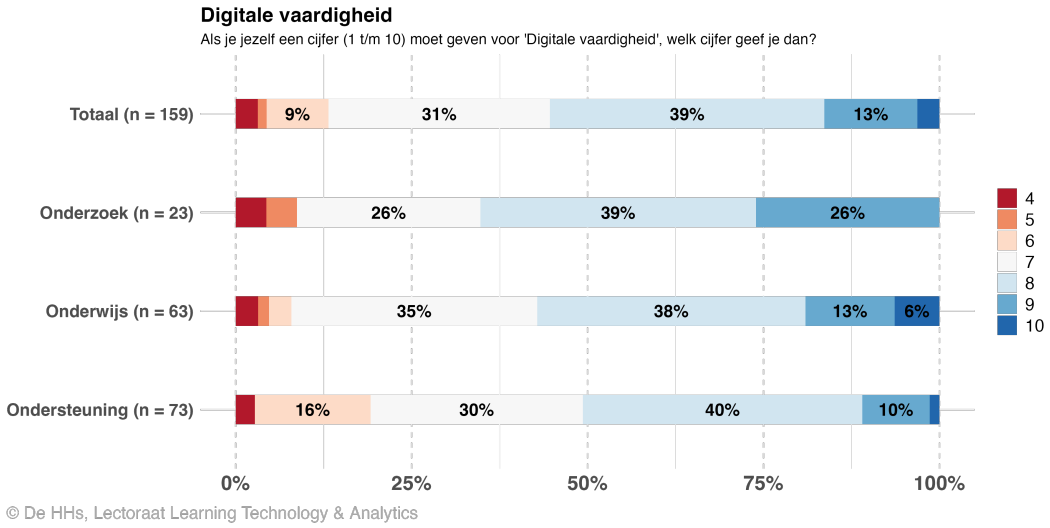
De deelnemers vertegenwoordigden een brede dwarsdoorsnede van het HHs-personeel: 63 uit het onderwijs, 23 uit onderzoek en 73 uit ondersteuning. De demografische informatie (geslacht en leeftijd) wordt weergegeven in tabel 1.

Tabel 1. Demografische kenmerken van de deelnemers

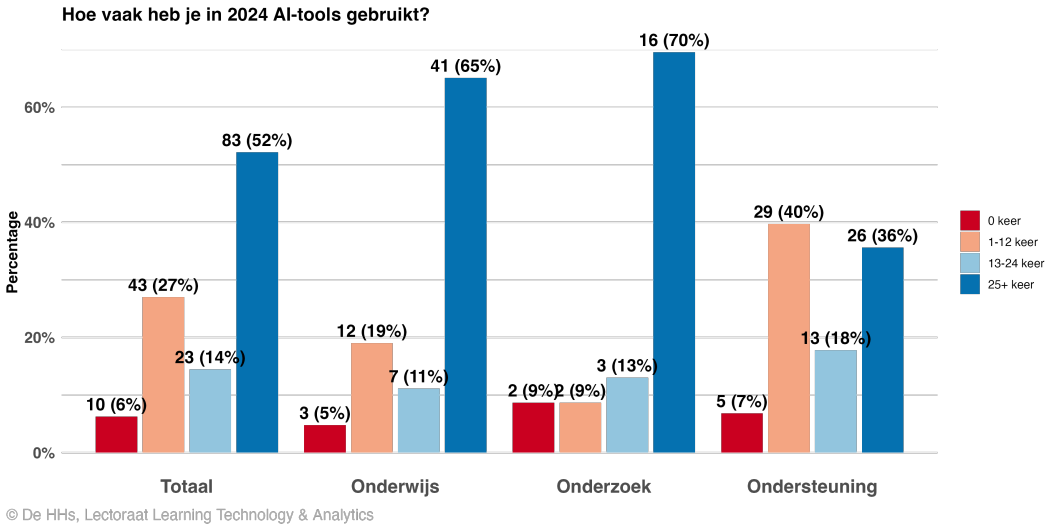
Karakteristiek	N = 159[1]
Geslacht	
Man	65 (41%)
Vrouw	92 (58%)
Onbekend	2 (1,3%)
Leeftijdsgroep	
18 - 25	2 (1,3%)
26 - 35	30 (19%)
36 - 45	35 (22%)
46 - 55	54 (34%)
> 55	35 (22%)
Onbekend	3 (1,9%)

[1] n (%)

Het niveau van digitale geletterdheid was over het algemeen hoog (figuur 3), waarbij de meeste respondenten zichzelf (in het registratieformulier) een score tussen 7 en 9 gaven op een schaal van 10. De eerdere ervaring met AI-tools varieerde (figuur 4), maar meer dan de helft (52%) gaf aan in 2024 minstens 25 keer AI-tools te hebben gebruikt, wat wijst op een sterke basiskennis.



Figuur 3. Zelfgerapporteerde digitale geletterdheid van deelnemers



Figuur 4. Zelfgerapporteerde eerdere ervaring met AI-tools in 2024

De deelname varieerde tussen de verschillende fasen van het onderzoek, zoals weergegeven in tabel 2.

Tabel 2. Deelname in de verschillende fasen van het Pilot GenAI

Fase	Aantal deelnemers
Geregistreerd voor de pilot	205
Informed consent en basis-vragenlijst (Pre-Pilot TAM)	159
Deelname aan RCT	67 (2 uitgesloten: geen toestemming)
Voltooide vervolg-vragenlijst (post-pilot TAM)	106 (2 uitgesloten: geen toestemming)
Geïnterviewd	10

2.5 Ethische overwegingen

Deelname aan het onderzoek was vrijwillig en gebaseerd op *informed consent*. Alle deelnemers ontvingen informatie over de doelstellingen, methoden en te verzamelen gegevens, en kregen de garantie dat ze zich op elk moment zonder gevolgen konden terugtrekken. De gegevensverzameling en -opslag voldeden aan de AVG-voorschriften, er werd geanonimiseerd gerapporteerd en alle gegevens zullen tien jaar lang worden bewaard in overeenstemming met de Nederlandse Gedragscode Wetenschappelijk Onderzoek. Het ethisch toezicht op de pilot werd verzorgd door de interne beoordelingsprocedures van De HHs.

3 Resultaten

De resultaten worden gepresenteerd volgens de drie bewijsniveaus in het 3E-raamwerk: brons (literatuuronderzoek en ToC), zilver (vragenlijsten en interviews) en goud (RCT). Binnen elk niveau worden de bevindingen gerapporteerd voor de drie dimensies bruikbaarheid, productiviteit en kwaliteit.

3.1 Bronsniveau bewijs

Het bronzen niveau van het 3E-raamwerk vormt de theoretische basis voor de pilot. Dit niveau omvatte een literatuuronderzoek (paragraaf 3.1.1), waarin internationale en nationale studies over GenAI in het onderwijs werden samengevat, en de ontwikkeling van een ToC (paragraaf 3.1.2), waarin in kaart werd gebracht hoe GenAI de werkwijzen van medewerkers bij De HHs zou kunnen beïnvloeden. Samen

vormden deze elementen de hypothesen en aannames die als leidraad dienden voor het empirische onderzoek in latere fasen.

3.1.1 Literatuuronderzoek

Het literatuuronderzoek laat zien dat GenAI weliswaar steeds vaker wordt bestudeerd in het hoger onderwijs, maar dat de meeste bevindingen zich richten op onderwijs- en onderzoekspersoneel, terwijl ondersteunende functies duidelijk ondervertegenwoordigd zijn. Desondanks worden administratieve taken vaak in bredere zin besproken onder verschillende domeinen.

Bruikbaarheid. GenAI wordt over het algemeen als nuttig en gemakkelijk toe te passen beschouwd, vooral naarmate gebruikers meer ervaring opdoen. Automatisering van routinetaken, het genereren van inhoud en schrijfhulp waren de meest genoemde use cases.

Kwaliteit. GenAI zou de kwaliteit van de output in alle functies verbeteren, maar er waren vaak zorgen over bijvoorbeeld verkeerde informatie en academische integriteit. Menselijk toezicht blijft essentieel om de normen te handhaven.

Productiviteit. GenAI zou de productiviteit consequent verhogen door routinetaken te automatiseren en tijd vrij te maken voor meer strategisch en creatief werk in alle personeelsdomeinen.

Dit literatuuronderzoek bestudeert recente studies naar de impact van GenAI op drie uitkomsten in het hoger onderwijs: ervaren bruikbaarheid, kwaliteit en arbeidsproductiviteit, voor drie personeelsdomeinen: onderwijs, onderzoek en ondersteunend personeel. Het onderzoek omvat Engelstalige studies die sinds 2023 zijn gepubliceerd, uitgevoerd in Europa en gericht op het gebruik van GenAI in het hoger onderwijs. Van de 195 geïdentificeerde artikelen zijn er na screening 15 opgenomen. De onderzochte literatuur richt zich voornamelijk op onderwijs- en onderzoekspersoneel, terwijl ondersteunend personeel duidelijk ondervertegenwoordigd is. Slechts één studie [10]richt zich uitsluitend op administratieve en onderwijsondersteunende taken. Taken die verband houden met ondersteunend personeel (bijv. administratieve taken) worden echter ook vaak besproken in verband met de andere twee domeinen.

Voor alle personeelsdomeinen zijn **de ervaren bruikbaarheid en gebruiksgemak** cruciale factoren die van invloed zijn op de acceptatie van GenAI. De opgenomen studies richten zich op de manier waarop deze percepties de acceptatie van GenAI binnen de drie domeinen beïnvloeden en hoe de percepties veranderen door voortdurend gebruik. Volgens de TAM-vragenlijsten spelen de ervaren bruikbaarheid en gebruiksgemak een cruciale rol bij het vormgeven van toekomstige intenties voor gebruik. Voor onderwijzend personeel is de ervaren bruikbaarheid een sterke positieve bepalende factor voor acceptatie. Onderwijzers melden dat GenAI hen helpt taken sneller uit te voeren, hun werkprestaties te verbeteren en hun algehele effectiviteit bij administratieve taken te vergroten, en dat het een hulpmiddel is voor zelfstudie en intellectuele betrokkenheid [11–13]. Bovendien, naarmate ze meer vertrouwd raken met GenAI, vinden ze het gemakkelijker te gebruiken en nuttiger [14]. Soortgelijke observaties kunnen worden gedaan bij onderzoekspersoneel, waar de ervaren bruikbaarheid en gebruiksvriendelijkheid van GenAI cruciale drijfveren zijn voor acceptatie. Tegelijkertijd wordt opgemerkt dat ervaring met GenAI de zelfredzaamheid bevordert en de bevestiging van de waarde ervan versterkt [14]. Interessant is dat ervaring met GenAI ook de beperkingen ervan blootlegt, zoals hallucinerende citaten en onbetrouwbare outputs. Dit verandert vaak de perceptie van onderzoekers over de bruikbaarheid [15,16]. Voor ondersteunend personeel is de ervaren bruikbaarheid sterk gericht op de verwachte efficiëntie van het verminderen van de administratieve werklast. Er werd ook waargenomen dat ondersteunend personeel, in vergelijking met onderwijs- en onderzoekspersoneel, aanzienlijk positievere opvattingen heeft over de mogelijkheden van AI en over het algemeen een hogere zelfeffectiviteit ten aanzien van AI uitdrukt [13].

De impact van GenAI op **de kwaliteit van het werk** wordt met meer voorzichtigheid beschreven vanwege de potentie om de output te verbeteren en de risico's die het kan opleveren voor de integriteit en verlies van fundamentele vaardigheden. Voor onderwijzend personeel kan GenAI de kwaliteit van beoordelingsprocessen verbeteren door directe, gedetailleerde en degelijke feedback te geven, waardoor zij consistente standaarden voor toetsing kunnen toepassen [17,18]. Tegelijkertijd vormt GenAI een bedreiging voor de integriteit van toetsing, omdat docenten hun methoden moeten aanpassen en dit mogelijk leidt tot een hogere werklast voor controles [19,20]. Voor onderzoekers kan GenAI de kwaliteit van output verbeteren door teksten duidelijker en grammaticaal beter te maken en andere soorten van redactionele ondersteuning [13,18,21]. Een opvallende beperking is echter de onnauwkeurigheid van de output (hallucinaties), met name wat betreft referenties en biografische informatie. Dit creëert een risico op desinformatie en een mogelijke achteruitgang van de onderzoekskwaliteit als gevolg van "plausibele maar gebrekkige resultaten" [13,15,18]. Ten slotte wordt de kwaliteit van het werk van ondersteunend personeel verbeterd doordat zij tijd kunnen besteden aan hogere, strategische en creatieve activiteiten als GenAI repetitieve taken automatiseert. Bovendien zou GenAI de kwaliteit van taalkundige en administratieve output verbeteren [13,18,20]. In alle drie domeinen hangt het behoud van de kwaliteit af van het feit dat mensen verantwoordelijk blijven voor alle door GenAI geproduceerde content.

Ten slotte wordt verbeterde **productiviteit** consequent in verband gebracht met GenAI, grotendeels vanwege het vermogen om routinetaken te automatiseren, intellectuele werkprocessen te stroomlijnen en tijd vrij te maken voor hoger orde denken. De productiviteit van onderwijzend personeel wordt verbeterd door repetitieve taken te automatiseren, zoals het schrijven van mededelingen, het genereren van beoordelingsonderdelen en het beheren van taaltaken (bijv. teksten proeflezen, controleren op grammaticale fouten, enz.) [11,19,22]. Ook onderzoekers profiteren van AI-ondersteunde concepten en technische ondersteuning, wanneer processen die traditioneel veel tijd en moeite kosten, worden gestroomlijnd [14,15,21]. Het verhoogt ook de productiviteit door de kwaliteit van output te verbeteren en het genereren van ideeën en het interpreteren van code te vergemakkelijken [16,23]. Voor ondersteunend personeel verbetert GenAI de productiviteit door werkprocessen te stroomlijnen en repetitieve taken zoals het opstellen van e-mails, rapporten en procesdocumentatie te automatiseren [10,13]. Deze verschuivingen vergroten het volume aan werk dat binnen hetzelfde tijdsbestek kan worden voltooid en verbeteren het vermogen van medewerkers om bij te dragen aan bredere institutionele doelen door middel van meer doordacht en impactvol werk.

Over het algemeen suggereren studies tot nu toe dat GenAI een waardevol hulpmiddel is voor het verbeteren van de efficiëntie, het verbeteren van de kwaliteit van output en het ondersteunen van complexer werk voor alle functies. De acceptatie en impact van GenAI worden bepaald door factoren zoals functiespecifieke behoeften, ervaringsniveaus en zorgen over betrouwbaarheid.

3.1.2 Theorie van verandering

Op basis van het literatuuronderzoek is voor deze studie een Theory of Change (ToC; figuur 5) opgesteld. Deze theorie bevat hypothesen over de mechanismen waarmee GenAI naar verwachting de bruikbaarheid, de kwaliteit van het werk en de productiviteit van medewerkers van De HHs zal beïnvloeden.

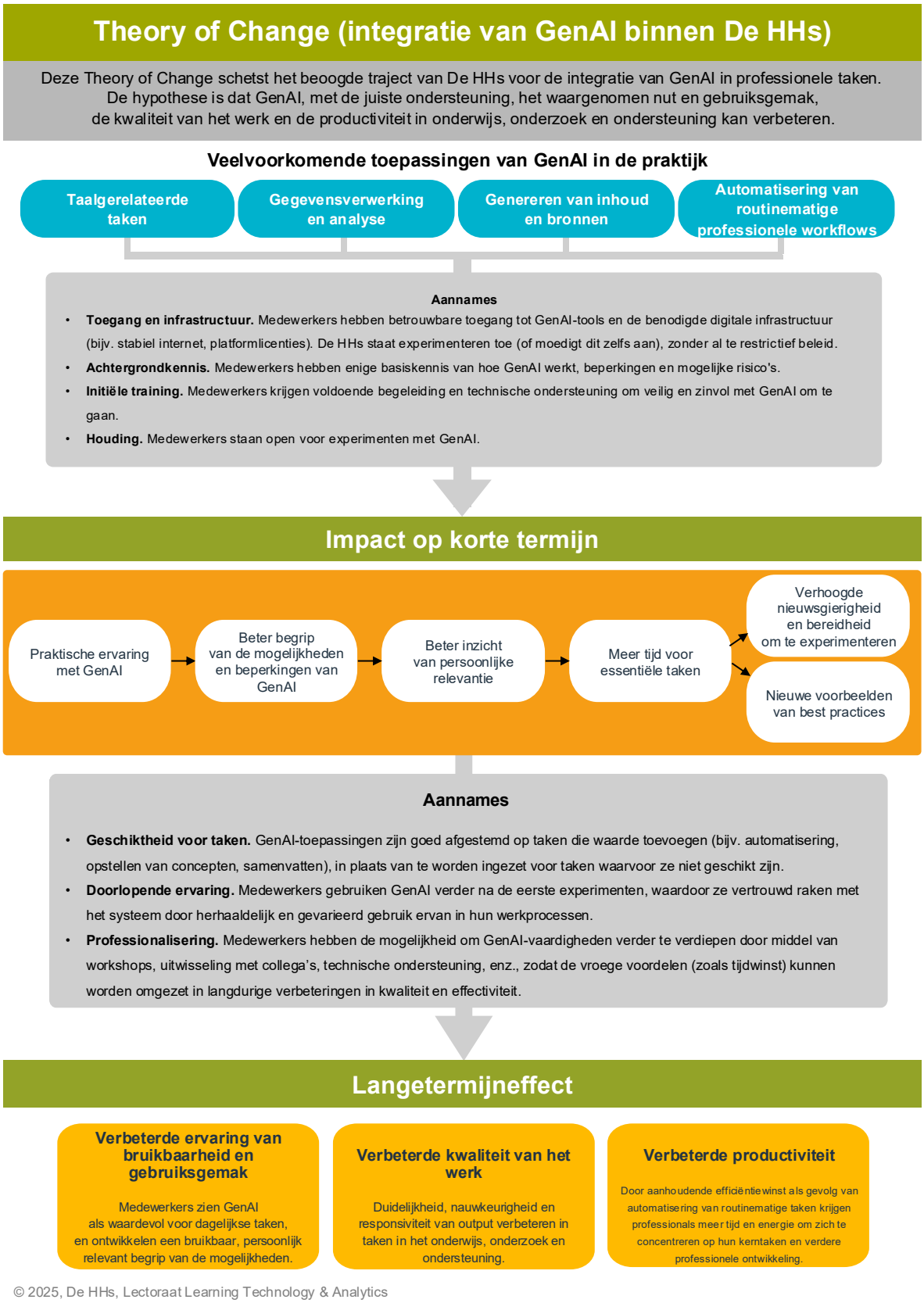
De centrale hypothese van dit onderzoek is dat, wanneer medewerkers toegang hebben tot GenAI, zij dit geleidelijk zullen integreren in hun dagelijkse routines, als zij worden ondersteund door training, duidelijke begeleiding en een cultuur waarin experimenteren mogelijk is zonder angst voor beoordeling. Verwacht wordt dat dit proces de bruikbaarheid zal verbeteren (door GenAI intuïtiever en relevanter te maken), de kwaliteit van de output zal verhogen (door meer gestructureerde en vloeiender conceptteksten) en de productiviteit zal verhogen (door minder tijd te besteden aan repetitieve taken). De omvang van deze effecten zal echter waarschijnlijk verschillen per domein, afhankelijk van de kenmerken van de taken, eerdere ervaringen en digitale volwassenheid.

Een ToC is een gestructureerd kader dat verduidelijkt hoe specifieke activiteiten naar verwachting tot de gewenste resultaten zullen leiden, onder bepaalde aannames [24]. Een ToC is voor deze studie ontworpen op basis van het literatuuronderzoek dat in het vorige hoofdstuk is beschreven. Het bevat een hypothese over hoe de invoering van GenAI door onderwijs-, onderzoeks- en ondersteunend personeel bij De HHs kan bijdragen aan verbeteringen op korte en lange termijn in de ervaren bruikbaarheid, de kwaliteit van de werkoutput en de productiviteit.

Het proces begint met het integreren van GenAI in de routinematige professionele taken van medewerkers (zoals taalkundige taken, data-analyse en het creëren van content). In het onderwijs kan het bijvoorbeeld helpen bij het ontwerpen van lesmateriaal, het beoordelen van toetsen en het genereren van feedback. In onderzoek kan het ondersteuning bieden bij het voorbereiden van data, het samenvatten van literatuur en het opstellen van onderzoeksartikelen. Voor ondersteunend personeel kan het het plannen van telefoongesprekken, het schrijven van correspondentie, het documenteren van notulen van vergaderingen enz. stroomlijnen.

Nu GenAI voor het eerst in de praktijk is geïntegreerd, wordt verwacht dat HHs-medewerkers meer open zullen staan voor experimenten, mogelijkheden zullen zien om de kwaliteit van hun werk te verbeteren en zullen bijdragen aan nieuwe voorbeelden van effectief gebruik. Deze eerste kennismaking kan medewerkers helpen een duidelijker beeld te krijgen van de mogelijkheden en algemene beperkingen van GenAI, en, nog belangrijker, van de relevantie ervan voor hun werk. Dit zou het gemakkelijker kunnen maken om minder inspannende en routinematige taken aan GenAI te delegeren, waardoor de cognitieve en operationele werklast wordt verminderd, wat zou leiden tot meer tijd voor het uitvoeren van hun kerntaken. De kortetermijneffecten zullen naar verwachting leiden tot langetermijneffecten (verbeterde bruikbaarheid, kwaliteit en productiviteit) bij verdere blootstelling.

Er zijn verschillende aannames die ten grondslag liggen aan de hypothetische ToC. Ten eerste moet De HHs blijvend toegang bieden tot GenAI-tools en ondersteunende infrastructuur, basisopleidingen op het gebied van AI en workshops. Bovendien is een stimulerende cultuur van vertrouwen, experimenteren en passende afstemming van taken nodig om de daadwerkelijke voordelen van AI te realiseren. Deze voorwaarden weerspiegelen ook de externe variabelen die in de TAM worden beschreven als factoren die van invloed zijn op de perceptie van gebruikers van de bruikbaarheid [6]. Ten slotte moeten medewerkers openstaan voor experimenten met nieuwe technologie, leren deze te gebruiken en de opgedane kennis delen met collega’s.



Figuur 5. Theory of Change voor de pilot GenAI

3.2 Zilveren niveau bewijs

In dit hoofdstuk worden de resultaten gepresenteerd van het zilveren niveau van het 3E-raamwerk, dat zich richtte op de percepties en ervaringen van medewerkers tijdens de pilot. Het bewijs werd verzameld door middel van TAM-vragenlijsten die voor en na de pilot werden afgenomen, aangevuld met interviews. Deze methoden bieden inzicht in de bruikbaarheid, kwaliteit en productiviteit in verschillende domeinen. De bevindingen van de TAM-vragenlijsten worden gepresenteerd in paragraaf 3.2.1 en die van de interviews in paragraaf 3.2.2.

3.2.1 Vragenlijsten

De TAM-vragenlijsten werden online uitgevoerd en gaven een kwantitatief beeld van de houding van medewerkers. De TAM vóór de pilot legde de verwachtingen van de deelnemers vast, terwijl de TAM na de pilot hun ervaringen na vier weken toegang tot het EduGenAI-platform meette.

Bruikbaarheid

De algehele bruikbaarheid werd binnen de hele instelling als positief ervaren. In alle domeinen bleef de ervaren gebruiksvriendelijkheid stabiel, wat bevestigt dat GenAI als technisch gemakkelijk te gebruiken werd ervaren. De ervaren bruikbaarheid nam echter aanzienlijk af, vooral op het gebied van onderwijs en ondersteuning, wat erop wijst dat medewerkers kritischer werden nadat ze de tool in de praktijk hadden getest.

TAM wordt meestal gebruikt om de bruikbaarheid of acceptatie van nieuwe technologie te beoordelen door middel van metingen van de ervaren gebruiksvriendelijkheid en het ervaren nut. De items van de vragenlijst worden gemeten op een schaal van 7. De domeinspecifieke en geaggregeerde resultaten worden weergegeven in tabel 3.

Tabel 3. Ervaren gebruiksgemak (PEoU) en bruikbaarheid (PU) voor en na de pilot

	Onderwijs			Onderzoek			Ondersteuning			Totaal		
Waarde	**Pre** N = 63[1]	**Post** N = 37[1]	**p-waarde**[2,3]	**Pre** N = 23[1]	**Post** N = 17[1]	**p-waarde**[2,3]	**Pre** N = 73[1]	**Post** N = 52[1]	**p-waarde**[2,3]	**Pre** N = 159[1]	**Post** N = 106[1]	**p-waarde**[2,3]
Gebruiksgemak	4,75 (1,28)	4,70 (1,43)	>0,9	4,92 (1,23)	5,17 (1,35)	0,2	4,73 (1,08)	4,44 (1,31)	0,2	4,77 (1,18)	4,65 (1,37)	0,8
Gebruiksnut	5,34 (1,24)	4,37 (1,50)	<0,001***	5,46 (0,78)	4,66 (1,37)	0,026*	5,29 (1,08)	3,92 (1,52)	<0,001***	5,33 (1,10)	4,20 (1,50)	<0,001***

[1] Gemiddelde (SD)
[2] Gepaarde t-toets
[3] *p<0.05; **p<0.01; ***p<0.001

De gemiddelde PEoU-score vóór de pilot was 4,77 (SD = 1,18) en na de pilot 4,65 (SD = 1,37). Het verschil was niet statistisch significant, wat aangeeft dat de verwachtingen van medewerkers ten aanzien van de bruikbaarheid grotendeels werden bevestigd door hun ervaringen. Daarentegen daalde de PU aanzienlijk, van 5,33 (SD = 1,10) in de fase vóór de pilot tot 4,20 (SD = 1,50) in de fase na de pilot (p < .001). Hoewel het nut nog steeds relatief hoog werd beoordeeld, blijkt uit deze verschuiving dat medewerkers kritischer werden na praktische ervaring met GenAI. De analyse op domeinniveau laat verschillen zien:

- **Docenten** rapporteerden een stabiel gebruiksgemak (Pre = 4,75; Post = 4,70; p > .900), maar een significante daling in bruikbaarheid (Pre = 5,34; Post = 4,37; p < .001).
- **Onderzoekers** gaven iets meer gebruiksgemak aan (Pre = 4,92; Post = 5,17; p = 0,200) en een bescheiden maar significante daling in bruikbaarheid (Pre = 5,46; Post = 4,66; p = 0,026).
- **Ondersteunend personeel** liet meer uiteenlopende resultaten zien. Het gebruiksgemak daalde licht (Pre = 4,73; Post = 4,44; p = 0,200), terwijl het nut sterk daalde (Pre = 5,29; Post = 3,92; p < 0,001).

Kwaliteit

De kwaliteit van het werk werd over het algemeen als matig ervaren in alle domeinen. Docenten en onderzoekers rapporteerden matige tot positieve effecten van GenAI op de kwaliteit, terwijl ondersteunend personeel lagere scores gaf.

In de TAM-vragenlijst na de pilot werd de deelnemers gevraagd om aan te geven in hoeverre zij het eens waren met twee uitspraken over kwaliteit (prestaties en effectiviteit), zoals weergegeven in tabel 4. Deze zelfgerapporteerde percepties bieden een domeinspecifiek perspectief op de invloed van GenAI op de kwaliteit van het werk.

Tabel 4. Domeinspecifieke kwaliteitspercepties.

Waarde	Onderwijs N = 37[1]	Onderzoek N = 17[1]	Ondersteuning N = 52[1]
Gen AI hielp me om beter te presteren in mijn werk.	4,16 (1,69)	4,18 (1,47)	3,71 (1,59)
Gen AI maakte me effectiever in mijn werk.	4,38 (1,60)	4,47 (1,55)	3,81 (1,62)

[1] Gemiddelde (SD)

Over alle domeinen heen varieerden de gemiddelde scores voor kwaliteit van 3,71 tot 4,47, wat wijst op een over het algemeen gematigde perceptie. Analyse op domeinniveau laat verschillen zien:

- **Docenten** rapporteerden matige tot positieve scores (4,16 en 4,38).
- **Onderzoekers** rapporteerden de hoogste waarden van alle domeinen, met een gemiddelde van 4,18 voor prestaties en 4,47 voor effectiviteit.
- **Ondersteunend personeel** gaf lagere beoordelingen, met gemiddelde scores van 3,71 voor prestaties en 3,81 voor effectiviteit.

Productiviteit

De productiviteit werd over het algemeen als enigszins verbeterd beschouwd, maar niet als algemeen hoog. Onderzoekers rapporteerden consequent de sterkste productiviteitsvoordelen, docenten gaven aan dat er sprake was van een matige verbetering en ondersteunend personeel rapporteerde de minste impact.

De productiviteit werd beoordeeld aan de hand van twee TAM-items over snelheid en algehele productiviteit (tabel 5).

Tabel 5. Op TAM gebaseerde percepties van productiviteit per domein.

Waarde	Onderwijs N = 37[1]	Onderzoek N = 17[1]	Ondersteuning N = 52[1]
Dankzij Gen AI kon ik mijn werk sneller doen dan met andere hulpmiddelen.	4,27 (1,82)	4,65 (1,54)	4,10 (1,72)
Gen AI verhoogde mijn productiviteit.	4,30 (1,82)	4,65 (1,77)	3,54 (1,61)

[1] Gemiddelde (SD)

Over alle domeinen heen varieerden de gemiddelde scores voor productiviteit van 3,54 tot 4,65. Analyse op domeinniveau laat verschillen zien:

- **Docenten** rapporteerden gemiddelde scores van 4,27 voor snelheid en 4,30 voor productiviteit.
- **Onderzoekers** gaven de hoogste scores, met een gemiddelde van 4,65 voor beide items.
- **Ondersteunend personeel** scoorde 4,10 voor snelheid en 3,54 voor productiviteit, de laagste scores van alle domeinen.

3.2.2 Interviews

De interviews gaven kwalitatieve inzichten in hoe medewerkers GenAI in de praktijk ervoeren. In tegenstelling tot de vragenlijsten, die algemene patronen in kaart brachten, boden de interviews gedetailleerde verslagen van de reflecties van medewerkers op bruikbaarheid, productiviteit en kwaliteit in hun dagelijkse werk.

Bruikbaarheid

De perceptie van bruikbaarheid werd bepaald door digitale geletterdheid, geschiktheid voor taken en eerdere ervaring met GenAI. Docenten en onderzoekers vonden duidelijke toepassingen in schrijfgerichte taken, terwijl ondersteunend personeel vaak moeite had om de tool te koppelen aan hun diverse en contextgebonden werkprocessen.

De ervaringen met de bruikbaarheid hingen nauw samen met digitale geletterdheid en eerdere blootstelling aan GenAI. Medewerkers met beperkte digitale vaardigheden hadden vaak extra begeleiding nodig om prompts te formuleren en outputs te interpreteren, terwijl zelfverzekerde gebruikers zelfstandig experimenteerden. Nieuwe gebruikers beschreven een mengeling van nieuwsgierigheid en aarzeling, terwijl ervaren gebruikers EduGenAI vaak vergeleken met tools zoals ChatGPT en het beschouwden als een uitbreiding van hun vaste routines.

“Ik had nog nooit GenAI gebruikt, maar ik was nieuwsgierig en stond open voor experimenten.”
(Interview, ondersteuningsdomein)

“Ik gebruik ChatGPT bijna dagelijks, vooral om de structuur en duidelijkheid van teksten te controleren.”
(Interview, onderzoeksdomein)

Docenten benadrukten de waarde ervan voor het ontwerpen van lessen, leerresultaten en het genereren van ideeën, hoewel sommigen opmerkten dat de output te algemeen was en te weinig rekening hield met de context van de studenten.

“Het is een nuttige brainstormpartner, maar vervangt mijn professionele oordeel niet.”
(Interview, onderwijsdomein)

Onderzoekers beschreven het platform als gebruiksvriendelijk voor het samenvatten, structureren en verfijnen van concepten. Hun grootste zorgen hadden betrekking op de betrouwbaarheid van de output en verzonnen referenties.

“Het is alsof je een taalcoach hebt. Ik gebruik het om mijn teksten te verfijnen, niet om mijn schrijfwerk te vervangen.”
(Interview, onderzoeksdomein)

Ondersteunend personeel vertoonde de grootste variatie. Zij vonden het nuttig voor veel routinetaken, maar hadden moeite om de relevantie ervan in te zien voor hun contextspecifieke taken, vooral wanneer het om persoonsgegevens ging. Onzekerheid over wat ‘toegestaan’ was, beperkte ook de verkenning.

“Het is bruikbaar, maar ik wist niet wat ik eraan moest vragen. Mijn werk is soms te contextspecifiek.”
(Interview, ondersteuningsdomein)

Kwaliteit

In alle domeinen werd een kwaliteitsverbetering waargenomen, maar deze bleef voorwaardelijk. Docenten en onderzoekers profiteerden het meest van gestructureerde concepten, terwijl ondersteunend personeel ongelijkmatige resultaten ervoer, afhankelijk van het type taak en de kwaliteit van de gegevens.

Medewerkers brachten kwaliteitsverbeteringen in verband met het vermogen van GenAI om gestructureerde concepten en duidelijkheid in de output te bieden, maar benadrukten ook de noodzaak van professionele beoordeling.

Docenten benadrukten vaak het nut van GenAI bij het produceren van gestructureerde concepten, maar benadrukten dat deze output moest worden verfijnd om vakspecifieke nauwkeurigheid te garanderen.

“Als je het onderwerp niet kent, kun je te snel tevreden zijn. Je moet je vragen verfijnen.”
(Interview, onderwijsdomein)

Onderzoekers merkten consequent verbeteringen op in vroege concepten, vragenlijsten en academische teksten. Tegelijkertijd benadrukten ze het belang van het controleren van de inhoud op nauwkeurigheid en correcte verwijzingen.

“Het heeft me echt geholpen bij het eerste concept van de vragenlijst... het gaf me een voorsprong.”
(Interview, onderzoekdomein)

Ondersteunend personeel rapporteerde gemengde resultaten. Sommigen waardeerden duidelijkere samenvattingen, terwijl anderen benadrukten dat de kwaliteit van de output sterk afhankelijk was van de kwaliteit van de input. Taken die genuanceerde contextuele kennis vereisten, waren moeilijker effectief te ondersteunen.

“Maak een uittreksel van deze tekst... dat werkt. Maar als de gegevens onder de AI zwak zijn, is garbage in garbage out.”
(Interview, ondersteuningsdomein)

Productiviteit

Onderzoekers rapporteerden de sterkste productiviteitswinst, gevolgd door docenten, terwijl ondersteunend personeel de meest gemengde ervaringen had, waarbij ze een afweging maakten tussen tijdwinst en de inspanning om nieuwe werkprocessen te leren en zich daaraan aan te passen.

Uit de interviews bleek dat GenAI tijd kon besparen bij routinematige of repetitieve taken, hoewel de omvang van de waargenomen winst per domein verschilde.

Docenten merkten op dat GenAI de voorbereidingstijd verkortte en hielp bij het snel structureren van lesmateriaal, vooral wanneer taken werden herhaald. Toch benadrukten ze dat de output moest worden aangepast aan de pedagogische doelstellingen.

“Toen ik iets voor de derde of vierde keer moest doen, hielp het me om het snel te herformuleren.”
(Interview, onderwijsdomein)

Onderzoekers beschreven een aanzienlijke tijdwinst bij het schrijven van academische artikelen, het ontwerpen van vragenlijsten en het ondersteunen van codering. Voor sommigen versnelde GenAI taken die normaal gesproken weken in beslag zouden nemen.

“Normaal gesproken kost deze onderzoekstaak me een maand... met GenAI kostte het me maar een week.”
(Interview, onderzoeksdomein)

Ondersteunend personeel was meer verdeeld. Sommigen beschreven een duidelijke tijdwinst bij professioneel werk, zoals het schrijven van vergaderverslagen, terwijl anderen benadrukten dat de tool extra inspanning vereiste om nieuwe werkprocessen te leren, prompts te formuleren en outputs te verfijnen.

“Een paar minuten later had ik mijn vergadering samengevat... Ik dacht: ja, dit bespaart me veel tijd.”
(Interview, ondersteuningsdomein)

“O nee, ik met GenAI? Ik ben niet handig genoeg... Ik heb alle blokken apart gedaan – ik had ze in één goede prompt moeten combineren.”
(Interview, ondersteuningsdomein)

3.3 Gouden niveau bewijs

Het bewijs van goudniveau biedt causale inzichten in de effecten van GenAI op het werk van medewerkers. Dit bewijs is gegenereerd door middel van een micro-RCT. In de volgende subparagraaf worden de bevindingen van de micro-RCT in detail gepresenteerd.

3.3.1 RCT

De RCT werd uitgevoerd om te meten hoe GenAI de kwaliteit en productiviteit beïnvloedde wanneer medewerkers domeinspecifieke taken uitvoerden.

Kwaliteit

De kwaliteit van de output werd op twee manieren beoordeeld:
1. door GPT geëvalueerde kwaliteit (LLM: GPT-4.1 en mistral-large-2411), en
2. Door deelnemers zelfgerapporteerde kwaliteit.

Over het algemeen beoordeelde GPT-4.1 de output van GenAI in alle domeinen consistent als van hogere kwaliteit (M = 4,62 vs. 3,69; $p < 0,001$), terwijl de zelfgerapporteerde scores geen significant verschil lieten zien (M = 4,07 vs. 3,82; $p = 0,12$).

De domeinspecifieke en geaggregeerde resultaten worden weergegeven in tabel 6.

Tabel 6. Kwaliteitsscores voor alle maatregelen (GPT-gebaseerde evaluaties en zelfrapportages)

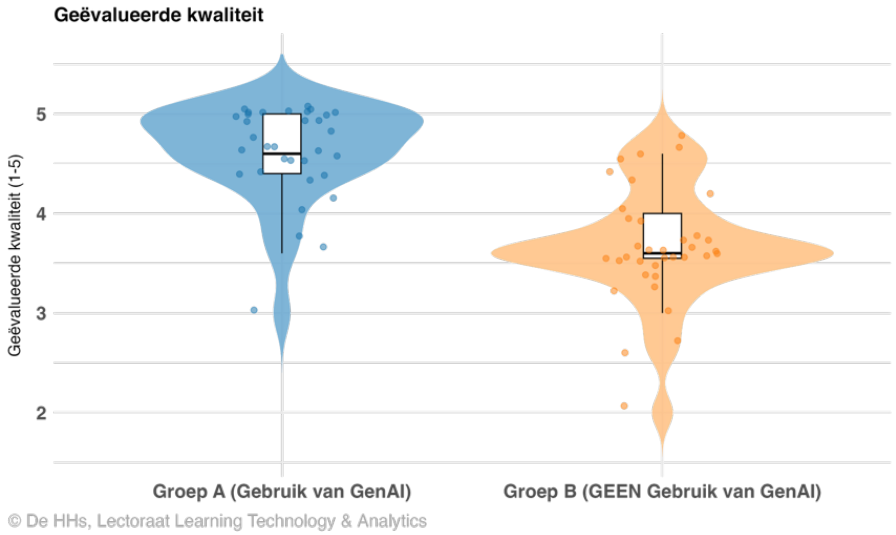
	Onderwijs			Onderzoek			Ondersteuning			Totaal		
Waarde	**Groep A (met GenAI)** N = 11[1]	**Groep B (zonder GenAI)** N = 11[1]	**p-waarde**[2]	**Groep A (met GenAI)** N = 3[1]	**Groep B (zonder GenAI)** N = 7[1]	**p-waarde**[2]	**Groep A (met GenAI)** N = 19[1]	**Groep B (zonder GenAI)** N = 16[1]	**p-waarde**[2]	**Groep A (met GenAI)** N = 33[1]	**Groep B (zonder GenAI)** N = 34[1]	**p-waarde**[2]
GPT-4.1	4,56 (0,62)	3,67 (0,66)	0,006	4,27 (0,23)	3,77 (0,18)	0,034	4,72 (0,38)	3,66 (0,67)	<0,001	4,62 (0,47)	3,69 (0,59)	<0,001
Mistral Large 2411	4,56 (0,31)	3,49 (0,77)	0,003	4,87 (0,23)	3,77 (0,24)	0,020	4,34 (0,47)	3,60 (0,61)	<0,001	4,46 (0,43)	3,60 (0,61)	<0,001
Zelfevaluatie	3,87 (0,47)	3,75 (0,52)	0,4	4,73 (0,46)	3,80 (0,97)	0,2	4,08 (0,73)	3,88 (0,82)	0,5	4,07 (0,66)	3,82 (0,75)	0,12

[1] Gemiddelde (SD)

[2] Mann-Whitney-Wilcoxontoets

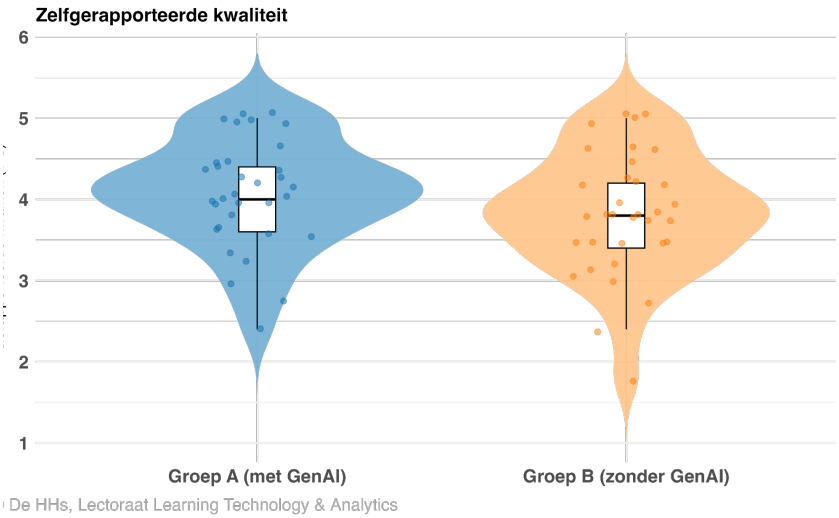
1. **GPT-geëvalueerde kwaliteit.** Volgens GPT-4.1 (figuur 6) behaalden de outputs van groep A (met GenAI) significant hogere scores (M = 4,62, SD = 0,47) dan groep B (zonder GenAI) (M = 3,68, SD = 0,57; $p < 0,001$). Mistral-large leverde vergelijkbare resultaten op, waarbij groep A (M = 4,46, SD = 0,43) hoger scoorde dan groep B (M = 3,60, SD = 0,59; $p < .001$). De interbeoordelaarsbetrouwbaarheid tussen GPT-4.1 en mistral-large was goed (ICC = 0,807), wat wijst op een sterke consistentie tussen de twee LLM's.

   Aangezien GPT-4.1 als het meer geavanceerde model wordt beschouwd, werden de resultaten ervan gebruikt als belangrijkste referentie voor verdere analyse.



Figuur 6 . GPT-4.1 beoordeelde de kwaliteitsscores van de verschillende groepen.

2. **Zelfgerapporteerde kwaliteit.** De eigen beoordelingen van de deelnemers lieten geen significant verschil tussen de groepen zien. Groep A rapporteerde M = 4,07 (SD = 0,66), terwijl groep B M = 3,84 (SD = 0,76; $p = 0,17$) rapporteerde.



Figuur 7. Zelfgerapporteerde kwaliteitsscores voor alle groepen.

Bij beide LLM's werden taken die met GenAI waren voltooid, significant hoger beoordeeld op kwaliteit dan taken die zonder GenAI waren voltooid. Deze verbetering kwam echter niet tot uiting in de zelfrapportages van de deelnemers.

De RCT-resultaten laten duidelijke domeinspecifieke patronen zien in de kwaliteitsresultaten.

- In **het onderwijs** werden de door GenAI ondersteunde outputs significant hoger beoordeeld door GPT-4.1 (M = 4,56 vs. 3,67; p = 0,006) en Mistral-large (M = 4,56 vs. 3,49; p = 0,003), hoewel de zelfrapportages geen significant verschil lieten zien.
- Op het gebied van **onderzoek** kregen de GenAI-outputs opnieuw hogere scores (GPT-4.1: M = 4,27 vs. 3,77; p = 0,034; Mistral-large: M = 4,87 vs. 3,77; p = 0,020), terwijl de zelfbeoordelingen niet significant verschilden.
- De grootste verschillen kwamen naar voren bij **ondersteuning**, waar GPT-4.1 de GenAI-output veel hoger beoordeelde (M = 4,72 vs. 3,66; p < 0,001), bevestigd door Mistral-large (M = 4,34 vs. 3,60; p < 0,001), maar ook hier lieten zelfrapportages geen significante verandering zien.

Overeenstemming tussen deelnemers en GPT. De overeenstemming tussen de zelfevaluaties van de deelnemers en de GPT-4.1-scores was laag (ICC = 0,184, niet significant). Verschillende factoren kunnen deze divergentie verklaren. *Ten eerste* werd de rubriek die in de RCT werd gebruikt in algemene termen beschreven, waardoor er ruimte was voor subjectieve interpretatie. *Ten tweede* worden zelfevaluaties vaak beïnvloed door persoonlijke vooringenomenheid, aangezien sommige deelnemers hun eigen werk onderschatten, terwijl anderen juist genereuzer zijn. De LLM's daarentegen richtten zich consequent op structuur, duidelijkheid en coherentie.

Zelfbeoordelingen van deelnemers en geautomatiseerde evaluaties hebben verschillende kwaliteitsaspecten gemeten en kwamen niet erg nauw overeen met elkaar, wat de noodzaak onderstreept om beide perspectieven mee te nemen.

Productiviteit

De productiviteit van de output van de deelnemers werd op twee manieren beoordeeld:
1. **Taakduur**, waarbij de tijd werd gemeten die nodig was om een taak te voltooien.

2. **Productiviteitsscores**, berekend als: $\frac{Kwaliteitsscore}{Taakduur}$

Er werd een productiviteitswinst waargenomen wanneer medewerkers GenAI gebruikten. Taken werden sneller afgerond en wanneer rekening werd gehouden met de kwaliteit, waren de productiviteitsscores aanzienlijk hoger bij zowel zelfbeoordelingen als GPT-gebaseerde metingen.

De domeinspecifieke en geaggregeerde resultaten worden weergegeven in tabel 7.

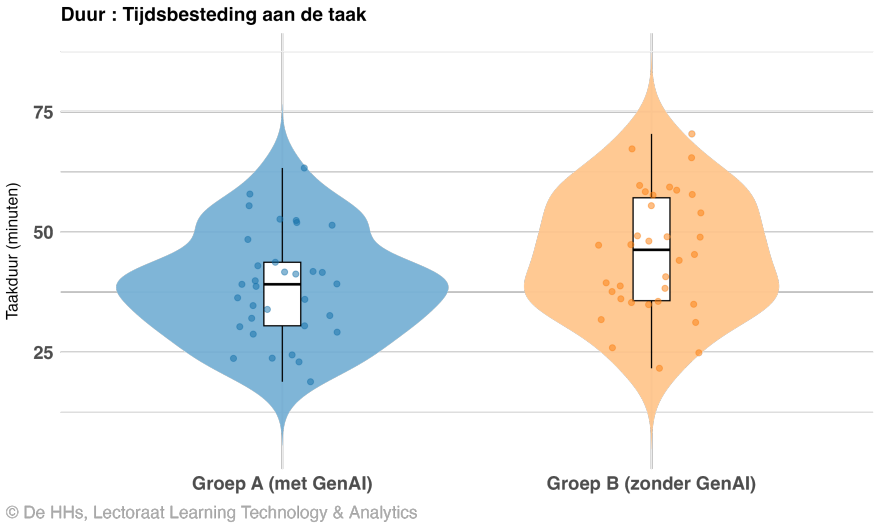
Tabel 7. Productiviteitsscores voor alle metingen

	Onderwijs			Onderzoek			Ondersteuning			Totaal		
Waarde	**Groep A (met GenAI)** N = 11[1]	**Groep B (zonder GenAI)** N = 11[1]	**p-waarde**[2]	**Groep A (met GenAI)** N = 3[1]	**Groep B (zonder GenAI)** N = 7[1]	**p-waarde**[2]	**Groep A (met GenAI)** N = 19[1]	**Groep B (zonder GenAI)** N = 16[1]	**p-waarde**[2]	**Groep A (met GenAI)** N = 33[1]	**Groep B (zonder GenAI)** N = 34[1]	**p-waarde**[2]
Taakduur (minuten)	39,74 (11,64)	41,46 (14,43)	0,8	32,50 (8,14)	50,91 (12,19)	0,067	39,29 (11,25)	46,11 (11,21)	0,088	38,82 (11,04)	45,59 (12,61)	0,033
Productiviteit: GPT-4.1	0,13 (0,05)	0,10 (0,04)	0,12	0,14 (0,04)	0,08 (0,02)	0,067	0,13 (0,04)	0,08 (0,03)	<0,001	0,13 (0,04)	0,09 (0,03)	<0,001
Productiviteit: Zelfevaluatie	0,11 (0,04)	0,10 (0,04)	0,7	0,15 (0,05)	0,08 (0,03)	0,067	0,11 (0,04)	0,09 (0,03)	0,056	0,12 (0,04)	0,09 (0,03)	0,017

[1] Gemiddelde (SD)

[2] Wilcoxon rank sum exact test

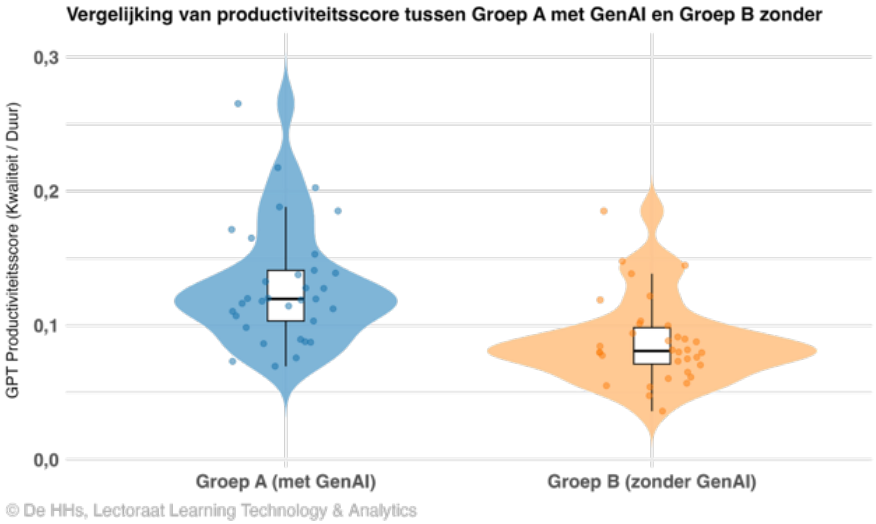
Taakduur. Zoals te zien is in figuur 8, voltooiden deelnemers in de GenAI-groep (groep A) taken in aanzienlijk minder tijd (M = 38,8 min, SD = 11,04) in vergelijking met de controlegroep zonder GenAI (groep B, M = 45,9 min, SD = 12,61; p = 0,033). Dit bevestigt dat GenAI de tijd die nodig was om taken te voltooien, heeft verkort.



Figuur 8. Taakduur (minuten) voor alle groepen

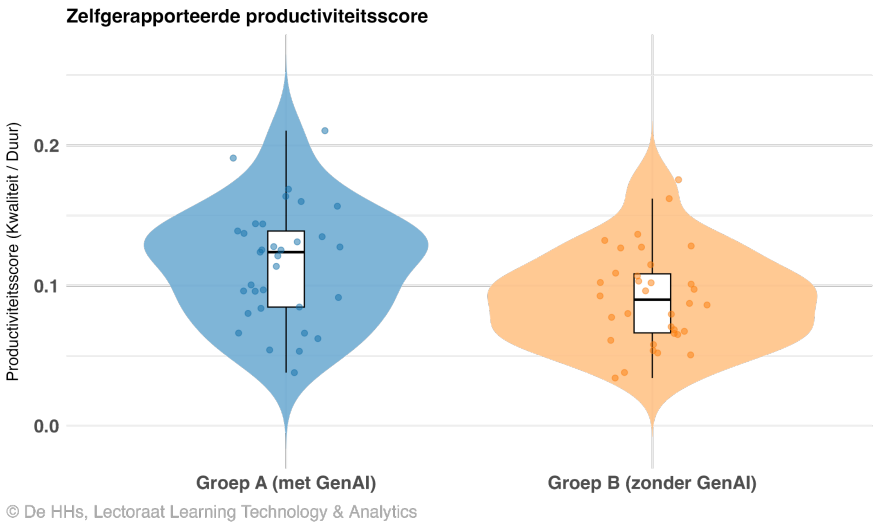
Productiviteitsscores. De duur van de taak alleen geeft geen volledig beeld van de productiviteit, aangezien ook de kwaliteit van belang is. Daarom werd de productiviteit berekend aan de hand van zowel de door GPT geëvalueerde kwaliteit als de zelfgerapporteerde kwaliteit.

- **Door GPT geëvalueerde productiviteit.** Berekend als GPT-4.1-kwaliteitsscore gedeeld door de taakduur. Groep A scoorde opnieuw significant hoger dan groep B ($p < 0{,}001$), wat bevestigt dat GenAI de productiviteit verhoogde (figuur 9).



Figuur 9. Door GPT geëvalueerde productiviteitsscores voor alle groepen.

- **Zelfgerapporteerde productiviteit.** Berekend als zelfgerapporteerde kwaliteit gedeeld door de duur van de taak. Groep A rapporteerde een hogere productiviteit dan groep B ($p = 0{,}017$). Figuur 10 laat zien dat deelnemers zich productiever voelden bij het gebruik van GenAI.



Figuur 10. Zelfgerapporteerde productiviteitsscores voor alle groepen.

Domeinspecifieke resultaten.

- **Docenten** in groep A voltooiden taken iets sneller (M = 39,7 min, SD = 11,6) dan groep B (M = 41,5 min, SD = 14,4), hoewel het verschil niet significant was (p = 0,800). De GPT-productiviteit was hoger in de GenAI-groep (M = 0,13, SD = 0,05 vs. 0,10, SD = 0,04; p = 0,12), maar ook dit verschil was niet significant.
- **Onderzoekers** boekten de grootste winst. Groep A voltooide taken veel sneller (M = 32,5 min, SD = 8,1) dan groep B (M = 50,9 min, SD = 12,2; p = 0,067). De GPT-productiviteit (M = 0,14 vs. 0,08; p = 0,067) en de zelfgerapporteerde productiviteit (M = 0,15 vs. 0,08; p = 0,067) waren hoger, hoewel de steekproeven klein waren.
- **Ondersteunend personeel** in groep A voltooide taken sneller (M = 39,3 min, SD = 11,2) dan groep B (M = 46,1 min, SD = 11,2; p = 0,088). De GPT-productiviteit was significant hoger met GenAI (M = 0,13 vs. 0,08; p < 0,001) en de zelfgerapporteerde productiviteit vertoonde ook een stijgende trend (M = 0,11 vs. 0,09; p = 0,056).

GenAI-ondersteuning leidde tot een snellere afronding van taken op alle domeinen, met aanzienlijk hogere productiviteitsscores wanneer rekening werd gehouden met kwaliteit. De winst was het grootst in onderzoek, matig in onderwijs en gemengd voor ondersteunend personeel, waar de gemeten verbeteringen duidelijker waren dan de zelfgerapporteerde verbeteringen.

4 Ervaringen met het EduGenAI-platform

Het EduGenAI-platform werd gewaardeerd om zijn veiligheid, persona's en toegankelijke ontwerp. De impact ervan werd echter beperkt door onduidelijke functies, de noodzaak van een betere onboarding en een zwakkere integratie in vergelijking met veelgebruikte externe platforms.

De deelnemers reflecteerden op hun ervaringen met het EduGenAI-platform met betrekking tot veiligheid, bruikbaarheid en integratie in hun dagelijkse werk. Ze rapporteerden gemengde meningen die zowel de sterke punten als de verbeterpunten weerspiegelden. Veel deelnemers beschreven het platform als AVG-conform en veilig voor gebruik binnen de institutionele omgeving, wat hen vertrouwen gaf in vergelijking met andere bestaande tools. Verschillende gebruikers vonden de interface eenvoudig en het gebruik van vertrouwde terminologie maakte het intuïtiever. Functies zoals persona's hielpen de deelnemers ook om snel vertrouwd te raken met het platform.

"De persona's maakten het voor mij gemakkelijker om te beginnen met het ontwerpen van lessen. Het bespaarde me veel tijd om na te denken over waar ik moest beginnen."
(Interview, onderwijsdomein)

Anderen vonden delen van de interface onduidelijk en vertrouwden op *trial and error* om te leren hoe ze het systeem moesten gebruiken. Verschillende deelnemers spraken de wens uit voor een duidelijkere onboarding en ingebouwde begeleiding. Sommige ervaren gebruikers gaven de voorkeur aan andere platforms zoals ChatGPT of Copilot, omdat ze daar vertrouwd mee waren en deze beter konden integreren in hun dagelijkse software. In vergelijking met die tools werd EduGenAI soms als minder handig beschouwd omdat het nodig was om van systeem te wisselen. Sommigen aarzelden om gevoelige interne gegevens in te voeren, ondanks de AVG-conformiteit.

"Het is bruikbaar, maar ChatGPT staat altijd open in mijn browser – overschakelen naar EduGenAI voelde minder handig."
(Interview, onderzoeksdomein)

5 Discussie

In dit gedeelte wordt teruggeblikt op de bevindingen van de pilot en wordt gekeken wat deze betekenen voor het gebruik en de opschaling van GenAI bij De HHs. De discussie combineert bewijs uit literatuur, vragenlijsten, interviews en de RCT met bredere reflecties op institutionele en culturele factoren.

5.1 Bevindingen en per dimensie: bruikbaarheid, kwaliteit en productiviteit

De pilot bracht verschillende perspectieven aan het licht voor de drie dimensies bruikbaarheid, kwaliteit en productiviteit.

De bruikbaarheid werd over het algemeen positief ervaren. Het gebruiksgemak bleef stabiel gedurende de pilotperiode, wat overeenkomt met de literatuur waarin wordt opgemerkt dat LLM-interfaces over het algemeen intuïtief zijn. Het ervaren nut nam echter af zodra medewerkers EduGenAI in de praktijk ging gebruiken. Dit patroon suggereert dat de eerste indrukken weliswaar optimistisch waren, maar dat het praktische gebruik leidde tot meer kritische beoordelingen naarmate medewerkers de mogelijkheden en beperkingen van het platform ontdekten. Uit interviews bleek dat persona's en vertrouwde terminologie de toegangsdrempels weliswaar verlaagden, maar dat de deelnemers toch tijd nodig hadden om te experimenteren met prompts en om de output aan te passen. Dit suggereert dat bruikbaarheid niet alleen een kwestie is van technisch ontwerp, maar ook van begeleiding van gebruikers en vertrouwen in de toepassing van GenAI bij dagelijkse taken.

Kwaliteit was het gebied waarop GenAI het meest duidelijk effect had. In alle domeinen werden de outputs die met GenAI werden geproduceerd hoger beoordeeld door GPT-gebaseerde evaluaties. Het personeel beoordeelde de kwaliteit echter voorzichtiger. Docenten en onderzoekers beschreven GenAI als nuttig voor het produceren van gestructureerde concepten en repetitieve taken waarbij snelheid en duidelijkheid belangrijk waren. Tegelijkertijd benadrukten ze dat deze concepten onvoldoende diepgang hadden en verfijning behoefden. Ondersteunend personeel was meer verdeeld, wat de moeilijkheid weerspiegelt om generieke output toe te passen op hun zeer contextspecifieke taken. De literatuur over GenAI in het hoger onderwijs sluit aan bij deze bevindingen: studies melden vaak winst in efficiëntie en structuur, maar benadrukken ook terugkerende problemen met feitelijke nauwkeurigheid, hallucinaties en onvolledige verwijzingen. Daarom benadrukken veel auteurs de noodzaak van systematische menselijke controle en verfijning voordat de output in de beroepspraktijk kan worden gebruikt. De hier waargenomen discrepantie tussen de zelfbeoordelingen van medewerkers en de op modellen gebaseerde kwaliteitsscores weerspiegelt hetzelfde probleem: hoewel de tekstuele output er misschien goed uitziet, vereisen professionele normen contextuele nauwkeurigheid, betrouwbaarheid en verantwoordingsplicht die GenAI alleen niet kan garanderen.

De productiviteit liet gemengde resultaten zien. Uit vragenlijsten bleek dat er sprake was van matige voordelen, met name voor onderzoekers en docenten, maar ondersteunend personeel was voorzichtiger. De RCT bevestigde dat taken sneller werden voltooid met GenAI en dat de productiviteitsscores aanzienlijk hoger waren wanneer rekening werd gehouden met de kwaliteit. Het personeel merkte deze voordelen echter niet altijd zelf op, wat het verschil tussen gemeten en ervaren productiviteit benadrukt. Deze spanning wordt ook beschreven in eerdere studies, waarin de cognitieve inspanning van het leren en verifiëren van AI-outputs de tijdwinst tenietdoet. Bij De HHs kwam dit tot uiting in het feit dat deelnemers een afweging maakten tussen de snelheid waarmee concepten werden gegenereerd en de inspanning die nodig was om deze te verfijnen, in context te plaatsen en te valideren.

5.2 Bevindingen per domein: onderwijs, onderzoek en ondersteunend personeel

De pilot bracht verschillen in perspectieven per domein aan het licht:

Docenten benadrukten de rol van GenAI bij het verminderen van de voorbereidingstijd en het genereren van lesstructuren, wat overeenkomt met literatuur waarin AI wordt gepositioneerd als een ondersteunend hulpmiddel voor repetitieve en tijdrovende taken. Zij merkten echter consequent op dat pedagogisch inzicht essentieel was om te voorkomen dat de output te algemeen zou worden. De afname van de ervaren bruikbaarheid in vragenlijsten suggereert dat docenten kritischer werden toen zij de output in de praktijk testten, waarbij zij evenwicht zochten tussen efficiëntie en bezorgdheid over de diepgang van het onderwerp en de individuele behoeften van studenten.

Onderzoekers rapporteerden de meest consistente voordelen. GenAI werd gewaardeerd voor het opstellen van samenvattingen, het structureren van teksten en het versnellen van literatuuronderzoek. Deze toepassingen sluiten aan bij studies die aantonen dat GenAI academisch schrijven kan versnellen, maar ook vragen oproept over betrouwbaarheid en bronvermelding. De RCT bevestigde dat onderzoekstaken sneller werden voltooid en hoger werden beoordeeld op kwaliteit met ondersteuning van GenAI. Toch bleven er zorgen bestaan over de feitelijke nauwkeurigheid en transparantie van bronnen.

Ondersteunend personeel was het meest verdeeld. Sommigen waardeerden GenAI voor het samenvatten van vergaderingen of het opstellen van communicatie, terwijl anderen het moeilijker vonden om een verband te leggen met hun gevarieerde en contextspecifieke rollen. Uit vragenlijstgegevens bleek de sterkste daling in de ervaren bruikbaarheid, en uit interviews bleek aarzeling in verband met onzekerheid over welke taken ‘toegestaan’ waren, tijd om te verkennen en angst voor overmatige afhankelijkheid. Literatuur over AI in administratieve functies merkt eveneens op dat de voordelen minder direct zijn wanneer taken contextspecifieke kennis of gevoelige gegevens vereisen.

5.3 Omstandigheden die het gebruik ondersteunen of belemmeren

Verschillen in resultaten werden bepaald door digitale geletterdheid, de geschiktheid van taken, institutionele ondersteuning en persoonlijke houding ten opzichte van technologie.

Verschillende omstandigheden bepaalden hoe HHs-medewerkers tijdens de pilot met GenAI omgingen.

De digitale geletterdheid was het hoogst onder onderzoekers en het laagst onder ondersteunend personeel, wat tot uiting kwam in hun vertrouwen in en perceptie van het gebruik van GenAI.

De geschiktheid van de taak voor ondersteuning door GenAI was een andere essentiële factor. Docenten en onderzoekers, die vaak meer ervaring hadden, konden GenAI gemakkelijk koppelen aan professionele taken zoals het ontwerpen van lessen, schrijfondersteuning en analyse. Ondersteunend personeel had daarentegen moeite om directe toepassingen te vinden, wat een weerspiegeling was van het meer contextspecifieke karakter van hun werk en het gebruik van persoonlijke gegevens, waardoor GenAI moeilijker toe te passen is.

Ook **institutionele ondersteuning** speelde een rol. Aanmoediging door het management, trainingsworkshops, hulp van de helpdesk en duidelijke handleidingen hielpen medewerkers om vertrouwen op te bouwen. Ze waardeerden bovendien de garantie van een AVG-conform platform voor veilig gebruik. Naast deze praktische ondersteuning benadrukten ze dat bruikbaarheid ook een cultureel aspect had. Ze hadden behoefte aan een omgeving waarin ze veilig konden experimenteren. Zoals een deelnemer opmerkte:

> *“AI kan ons werk sneller maken, maar alleen als mensen zich veilig genoeg voelen om het te verkennen zonder bang te zijn om iets verkeerds te doen.”*
> (Interview, onderzoeksdomein)

Ook **de houding** van medewerkers was van invloed op de bereidheid. Sommige medewerkers waren sceptisch over de nauwkeurigheid van de output van GenAI. Onzekerheid over best practices en ethische

grenzen maakte deze uitdagingen nog groter. Een aantal deelnemers uitte hun bezorgdheid over de milieukosten van grootschalig AI-gebruik. Anderen maakten zich zorgen over mogelijk verlies van schrijfvaardigheid of te grote afhankelijkheid van de technologie.

"Ik ben bang dat als ik het te veel gebruik, ik niet meer zelfstandig ga denken."
(Interview, ondersteuningsdomein)

"Ik denk na over de energie die GenAI verbruikt. Dat vind ik belangrijk."
(Interview, onderwijsdomein)

5.4 Beperkingen van het onderzoek

Ondanks de sterke punten van het ontwerp, meldt het onderzoek bepaalde beperkingen.

Ten eerste werd de TAM-vragenlijst versie 4 [6] gebruikt. Hoewel deze versie geschikt werd geacht voor een korte pilot, zijn er uitgebreidere versies beschikbaar. Voor onderzoek op langere termijn zou het waardevol zijn om andere versies te testen die een breder scala aan gebruikerspercepties beslaan.

Ten tweede was de reikwijdte van de RCT beperkt, aangezien deze over een korte periode met een beperkt aantal deelnemers werd uitgevoerd. Hoewel de RCT waardevolle causale inzichten opleverde, kunnen de resultaten niet volledig representatief zijn voor het langdurige gebruik van GenAI of de integratie ervan in complexe werkprocessen. We hebben geprobeerd deze beperking te compenseren door andere kwantitatieve en kwalitatieve methoden te combineren, maar verder onderzoek met grotere groepen en over langere perioden is nodig.

Ten derde was er sprake van een zekere mate van zelfselectiebias. De deelnemers meldden zich vrijwillig aan voor de pilot, wat kan betekenen dat ze nieuwsgieriger, gemotiveerder of digitaal zelfverzekerder waren dan het bredere personeelsbestand. Hoewel ook de uitvallers werden geïnterviewd om de bias te verminderen, moet bij de interpretatie van de resultaten met deze beperking rekening worden gehouden.

Ten slotte werd het onderzoek uitgevoerd in de specifieke context van De HHs. De bevindingen zijn mogelijk niet direct toepasbaar op andere instellingen met een ander niveau van digitale volwassenheid, andere werkprocessen of een andere organisatiecultuur. Toekomstig onderzoek moet gericht zijn op bevindingen die in verschillende contexten kunnen worden gereproduceerd of hergebruikt, aangezien dit de generaliseerbaarheid en waarde ervan zou versterken.

Lijst met afkortingen

- **3E-raamwerk** – Evidence-Informed Evaluation of EdTech Framework
- **AI** – Artificial Intelligence
- **AVG** – Algemene Verordening Gegevensbescherming
- **De HHs** – De Haagse Hogeschool
- **FZ/IT** – Facilitaire Zaken & IT
- **GenAI** – Generatieve kunstmatige intelligentie
- **ICC** – Intraclass Correlation Coefficient
- **P&O** – Personeel en Organisatie
- **LLM** – Large Language Model
- **LTA** – Learning Technology & Analytics
- **PEoU** – Perceived Ease of Use
- **PU** – Perceived Usefulness
- **RCT** – Randomized Control Trial
- **TAM** – Technology Acceptance Model
- **ToC** – Theory of Change

Referenties

1. De haagse hogeschool. Onderzoekend leren met impact. Accessed October 7, 2025. https://www.dehaagsehogeschool.nl/sites/hhs/files/documents/Instellingsplan%202023-2028.pdf

2. Sengar SS, Hasan A Bin, Kumar S, Carroll F. Generative artificial intelligence: a systematic review and applications. *Multimed Tools Appl*. 2024;84(21):23661-23700. doi:10.1007/s11042-024-20016-1

3. ChatGPT. Accessed September 30, 2025. https://chatgpt.com

4. Giannakos M, Azevedo R, Brusilovsky P, et al. The promise and challenges of generative AI in education. *Behaviour & Information Technology*. 2025;44(11):2518-2544. doi:10.1080/0144929X.2024.2394886

5. Garg M, Baker T. *The Dutch 3E Framework: Evidence-Informed Evaluation of EdTech*.; 2025. doi:10.5281/zenodo.15070789

6. Lewis JR. Comparison of four TAM item formats: effect of response option labels and order. *J Usability Stud*. Published online 2019.

7. Braga LH, Farrokhyar F, Dönmez Mİ, et al. Randomized controlled trials – The what, when, how and why. *J Pediatr Urol*. 2025;21(2):397-404. doi:10.1016/j.jpurol.2024.11.021

8. EduGenAI. Accessed September 30, 2025. https://npuls.nl/edugenai

9. Npuls. Accessed September 30, 2025. https://npuls.nl/

10. Kutty S, Chugh R, Perera P, et al. Generative AI in higher education: Perspectives of students, educators and administrators. *Journal of Applied Learning and Teaching*. 2024;7(2):47-60. doi:10.37074/JALT.2024.7.2.27

11. Castillo-Martínez IM, Flores-Bueno D, Gómez-Puente SM, Vite-León VO. AI in higher education: a systematic literature review. *Front Educ (Lausanne)*. 2024;9. doi:10.3389/feduc.2024.1391485

12. McDonald P, Hay S, Cathcart A, Feldman A. *Apostles, Agnostics and Atheists: Engagement with Generative AI by Australian University Staff*.; 2024. doi:10.5204/rep.eprints.252079

13. Viruel SR, Rivas ES, Palmero JR. The Role of Artificial Intelligence in Project-Based Learning: Teacher Perceptions and Pedagogical Implications. *Educ Sci (Basel)*. 2025;15(2). doi:10.3390/educsci15020150

14. Andersen JP, Degn L, Fishberg R, et al. Generative Artificial Intelligence (GenAI) in the research process - A survey of researchers' practices and perceptions. *Technol Soc*. 2025;81. doi:10.1016/j.techsoc.2025.102813

15. Fecher B, Hebing M, Laufer M, Pohle J, Sofsky F. Friend or foe? Exploring the implications of large language models on the science system. *AI Soc*. 2025;40(2):447-459. doi:10.1007/s00146-023-01791-1

16. Kallunki V, Kinnunen P, Pyörälä E, Haarala-Muhonen A, Katajavuori N, Myyry L. Navigating the Evolving Landscape of Teaching and Learning: University Faculty and Staff Perceptions of the Artificial Intelligence-Altered Terrain. *Educ Sci (Basel)*. 2024;14(7). doi:10.3390/educsci14070727

17. Jukiewicz M. The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Think Skills Creat*. 2024;52. doi:10.1016/j.tsc.2024.101522

18. Sáez-Velasco S, Alaguero-Rodríguez M, Delgado-Benito V, Rodríguez-Cano S. Analysing the Impact of Generative AI in Arts Education: A Cross-Disciplinary Perspective of Educators and Students in Higher Education. *Informatics*. 2024;11(2):37. doi:10.3390/informatics11020037

19. Marchena Sekli G, Godo A, Carlos Véliz J. Generative AI Solutions for Faculty and Students: A Review of Literature and Roadmap for Future Research. *Journal of Information Technology Education: Research*. 2024;23:014. doi:10.28945/5304

20. Stan MM, Dumitru C, Bucuroiu F. Investigating teachers' attitude toward integration of ChatGPT in language teaching and learning in higher education. *Educ Inf Technol (Dordr)*. Published online 2025. doi:10.1007/s10639-025-13396-w

21. Enang E, Christopoulou D. Exploring academic's intentions to incorporate ChatGPT into their teaching practice. *Journal of University Teaching and Learning Practice*. 2025;21(08). doi:10.53761/rn5y5614

22. Cambra-Fierro JJ, Blasco MF, López-Pérez MEE, Trifu A. ChatGPT adoption and its influence on faculty well-being: An empirical research in higher education. *Educ Inf Technol (Dordr)*. 2025;30(2):1517-1538. doi:10.1007/s10639-024-12871-0

23. Tran TM, Bakajic M, Pullman M. Teacher's pet or rebel? Practitioners' perspectives on the impacts of ChatGPT on course design. *High Educ (Dordr)*. Published online 2024. doi:10.1007/s10734-024-01350-7

24. Belcher BM, Bonaiuti E, Thiele G. Applying Theory of Change in research program planning: Lessons from CGIAR. *Environ Sci Policy*. 2024;160:103850. doi:10.1016/j.envsci.2024.103850

Gebruik van AI in het project

In verschillende fasen van het onderzoeksproces zijn AI-tools gebruikt. In de analysefase werden AI-ondersteunde tools gebruikt om interviewgegevens te transcriberen en te structureren, en om de resultaten in de RCT te vergelijken aan de hand van evaluaties op basis van grote taalmodellen. In de rapportagefase werd GenAI gebruikt om het opstellen en verfijnen van delen van de tekst te ondersteunen en om het rapport in het Nederlands te vertalen (de oorspronkelijke versie van het rapport is in het Engels geschreven). Het onderzoeksteam nam alle inhoudelijke beslissingen en beoordeelde, bewerkte en valideerde elke AI-ondersteunde output.