TNO innovation for life

LLM's in de zorg: Kwaliteitsevaluatie en automatisering

Bouwstenen voor een gevalideerd en geautomatiseerd evaluatieproces voor LLM-toepassingen

TNO Publiek › TNO 2025 R10722
28 maart 2025

Health & Work
www.tno.nl
+31 88 8667500
info@tno.nl

TNO 2025 R10722 – 28 maart 2025

LLM's in de zorg: Kwaliteitsevaluatie en automatisering

Bouwstenen voor een gevalideerd en geautomatiseerd evaluatieproces voor LLM-toepassingen

Auteurs	Robin van Stokkum Ilse Hellemans Roos Boereboom

Rubricering rapport	TNO Publiek
Titel	TNO Publiek
Rapporttekst	TNO Publiek
Aantal pagina's	33 (excl. voor- en achterblad)
Aantal bijlagen	7

Inhoudsopgave

1 Management samenvatting 4

2 Inleiding 7
2.1 Context
2.2 Doel van het rapport
2.3 Opbouw van het rapport

3 Relevante en bruikbare kwaliteitscriteria voor inzet van LLM's in de zorg 11
3.1 Methode
3.2 Geselecteerde kwaliteitscriteria
3.3 Evaluatiemethoden
3.4 Automatisch evalueren criteria

4 Praktijkvoorbeeld: LLM-toepassing voor automatisering ontslagbrieven 22
4.1 Inleiding en relevantie voor kwaliteitscriteria
4.2 Context en achtergrond
4.3 Validatiemethodiek
4.4 Bevindingen gerelateerd aan kwaliteitscriteria
4.5 Uitdagingen voor automatische evaluatie
4.6 Vervolgstappen en lessons learned

5 Conclusie en aanbevelingen 29
5.1 Conclusie
5.2 Aanbevelingen

6 Bijlagen 32
Bijlage 1. Lijst mogelijk relevante en bruikbare kwaliteitscriteria
Bijlage 2. Memorandum
Bijlage 3. Digitale vragenlijst
Bijlage 4. Aanpak onderzoek
Bijlage 5. Aangevulde lijst mogelijk relevante en bruikbare kwaliteitscriteria
Bijlage 6. Geselecteerde kwaliteitscriteria
Bijlage 7. Bronnenlijst

1 Management samenvatting

De belofte van AI in een sector onder druk

De Nederlandse zorgsector staat voor aanzienlijke uitdagingen: een toenemende zorgvraag gecombineerd met oplopende personeelstekorten. Dit zorgt ervoor dat zorgprofessionals een significant deel van hun werktijd aan administratieve taken besteden, tijd die zij niet voor directe patiëntenzorg kunnen inzetten. Large Language Models (LLMs) zouden mogen een rol kunnen spelen bij het ondersteunen van administratieve taken zoals medische documentatie, het genereren van ontslagbrieven en andere administratie werkzaamheden. In opdracht van het ministerie van Volksgezondheid, Welzijn en Sport (VWS) heeft TNO onderzoek gedaan naar relevante en bruikbare kwaliteitscriteria, specifiek voor het evalueren van LLM-toepassing voor de zorg met het doel de administratieve lasten van zorgverleners te verlichten. Het onderzoek van TNO is een verkenning en biedt een eerst inzicht in de bouwstenen voor verantwoorde evaluatie van LLM-toepassing in de zorg. De vraag die hierbij centraal staat is: Hoe kunnen we op een gestructureerde wijze de kwaliteit, betrouwbaarheid en veiligheid van LLM-technologie in de zorgcontext beoordelen?

De basis: negen (technische) kwaliteitscriteria voor verantwoorde inzet

Uit verkennend literatuuronderzoek, interactieve focusgroepen en verschillende expertgesprekken zijn negen kwaliteitscriteria geselecteerd die een eerste basis vormen voor een gestructureerde evaluatie, oftewel evaluatiekader:

- Accuraatheid van de gegeneerde uitkomst
- Betrouwbaarheid (consistentie)
- Privacy, gegevensbescherming & aanvullende wetgeving
- Relevantie van uitkomst
- Menselijke controle (en geautomatiseerde output controle)
- Gebruiksvriendelijkheid van de applicatie
- Bias & discriminatie
- Traceerbaarheid
- Transparantie & uitlegbaarheid

Van theorie naar praktijk: evaluatiemethoden

Het evalueren van LLM's in de zorg vraagt om een genuanceerde aanpak. In dit rapport wordt aandacht besteed aan twee (complementaire) methoden: (1) **Menselijke evaluatie**: de handmatige kwaliteitsbeoordeling van zorgprofessionals, cruciaal voor aspecten als accuraatheid en klinische relevantie en (2) **Automatische evaluatie**: de algoritmische methoden en benchmarks voor efficiënte evaluatie.

Het onderzoekt toont aan dat op dit moment een volledig geautomatiseerde evaluatie nog een utopie is doordat representatieve benchmarks en gouden standaarden die de complexiteit van de zorgcontext weerspiegelen ontbreken. Een hybride aanpak, waarbij automatische methoden menselijke beoordelingen ondersteunen, lijkt het meest kansrijk.

Lessen uit de praktijk: praktijkvoorbeeld van het UMC Utrecht

Het UMC Utrecht heeft in een pilot de inzet van een LLM-toepassing voor het genereren van ontslagbrieven geëvalueerd. De pilot bestond uit twee fasen: (1) de handmatige analyse van door AI-gegenereerde ontslagbrieven door medische studenten en door artsen, en (2) een vergelijkende studie waarin artsen blind moesten kiezen door AI-gegenereerde ontslagbrieven en door artsen geschreven briefen. Dit praktijkvoorbeeld van kwaliteitsevaluatie voor LLM-toepassingen in de zorgcontext biedt methodologische inzichten voor de gestructureerde beoordeling van AI-gegenereerde content, inclusief handvatten (geleerde lessen) voor toekomstige validatiestrategieën voor de zorg.

Er zijn drie belangrijke geleerde lessen uit het praktijkvoorbeeld op te halen:

1. **Multidisciplinaire evaluatie is essentieel**: In de pilot werd duidelijk dat de beoordelingen door medische studenten en de beoordelen door artsen opvallend uiteenliepen. Waar studenten 30% van de AI-gegenereerde inhoud als ‘hallucinatie’ bestempelden, herkenden ervaren artsen hierin klinisch relevante interpretaties.
2. **Contextafhankelijkheid van kwaliteitscriteria**: Het praktijkvoorbeeld toont aan dat criteria zoals accuraatheid zwaar wegen voor LLM-toepassingen waar foutieve informatie directie gezondheidsimplicaties kan hebben. Voor andere toepassingen kunnen bijvoorbeeld consistentie of gebruiksvriendelijkheid relevanter zijn.
3. **Balans tussen controle en efficiëntie**: Het UMC Utrecht koos voor een pragmatische benadering waarbij artsen altijd een eindcontrole uitvoeren. De uitdaging ligt in het vinden van de balans tussen menselijke betrokkenheid en administratieve efficiëntie. De netto tijdbesparing is een belangrijke succesfactor voor de adoptie van LLM-technologie voor administratieve lastenverlichting in de zorg.

De weg vooruit: uitdagingen en aanbevelingen

Verantwoorde inzet en succesvolle implementatie van LLM-toepassingen in de zorg kent een aantal uitdagingen:

- Het gebruik van LLM-toepassingen zonder goede validatie gaat gepaard met significante risico’s, mogelijk met directe implicaties voor de gezondheid van de patiënt als gevolg van bijvoorbeeld hallucinaties en foutieve informatie.
- De huidige validatiemethoden zijn arbeidsintensief én vragen specialistische kennis, waarmee opschaling wordt belemmerd.
- De grote verscheidenheid aan zorgcontexten vereist gedifferentieerde evaluatiemethoden die recht doen aan de specifieke kenmerken en risico's van elk zorgdomein.
- Er is een gebrek aan door het zorgveld gedragen evaluatiekaders, gouden standaarden en benchmarks. Dit belemmert de verantwoorde adoptie van AI-toepassingen en specifiek LLMs in de zorg.

Op basis van deze inzichten zijn vijf aanbevelingen opgesteld:

1. Breng de mogelijkheden om met benchmarks de adoptie van LLM-toepassingen te ondersteunen in kaart

2. Scherp de geselecteerde kwaliteitscriteria aan om tot een passend en gedragen evaluatiekader(s) te kunnen komen en ontwikkel collectieve benchmarks die vergelijking en automatische evaluatie mogelijk maken, en adoptie ondersteunen
3. Investeer in de ontwikkeling juridische kaders die veilig delen van data mogelijk maken
4. Stimuleer kennisuitwisseling en samenwerking tussen zorginstellingen
5. Ontwikkel hybride evaluatiemethoden die de meerwaarde van menselijke betrokkenheid geautomatiseerde efficiëntie combineren

Het ministerie van VWS kan een sleutelrol spelen als facilitator en aanjager van deze ontwikkelingen, waarbij expertise uit het ministerie, het zorgveld en technologieontwikkelaars wordt gebundeld voor een verantwoorde inzet van LLM's in de Nederlandse zorg.

2 Inleiding

2.1 Context

De Nederlandse zorg staat onder grote druk door een toenemende zorgvraag in combinatie met oplopende personeelstekorten. Hierdoor ontstaat een situatie waarin zorgprofessionals steeds minder tijd beschikbaar hebben voor directe patiëntenzorg. Administratieve taken nemen daarbij een significant en toenemend deel van hun werktijd in beslag. Dit zorgt niet alleen voor extra werkdruk, maar beperkt ook de capaciteit voor daadwerkelijke zorgverlening aan patiënten.

Generatieve AI (GenAI), en specifiek Large Language Models (LLM's), bieden veel potentie om de administratieve lasten van zorgprofessionals te verlagen. Door gebruik te maken van data en taaltechnologie kunnen taken als het samenvatten van medische documentatie, opstellen van ontslagbrieven, en ondersteunen van medische codering efficiënter worden uitgevoerd. Op dit moment experimenteren verschillende Nederlandse zorginstellingen met LLM-toepassingen om zorgprofessionals te ondersteunen in hun verslaglegging, waarbij een aantal ziekenhuizen de eerste toepassingen inmiddels geïmplementeerd hebben. Deze eerste praktijkvoorbeelden lijken aan te tonen dat dergelijke toepassingen kunnen bijdragen aan een efficiëntere inrichting van administratieve processen en daarmee potentieel de zorg toegankelijker en betaalbaarder maken.

Ondanks deze beloftevolle ontwikkelingen bestaan er grote uitdagingen rond het beoordelen van de kwaliteit, betrouwbaarheid en veiligheid van LLM's (zie ook TNO Rapport "Generatieve AI in de Nederlandse zorg"). Deze modellen functioneren vaak als zogenaamde 'black boxes'. Wat in het model gebeurt is voor de gebruiker namelijk niet helder, modellen zijn gevoelig voor onnauwkeurigheden en 'hallucinaties', en het ontbreekt aan inzichtelijke en eenduidige beoordelingsmethoden voor de zorgcontext. Dit maakt het voor zorginstellingen moeilijk te bepalen of en hoe zij LLM-technologie veilig en verantwoord kunnen inzetten.

Daarbij komt dat bestaande (juridische) kaders en richtlijnen voor het evalueren van AI-systemen, zoals die vanuit de Nederlandse Zorgautoriteit (NZa, 'Zes handvatten voor AI'), Zorginstituut Nederland (ZIN, 'Artificiële intelligentie en Passende zorg'), Nederlandse AI Coalitie (NLAIC, 'Hulmiddel Handelingsruimte'), vanuit het zorgveld ('Leidraad kwaliteit AI in de zorg') en internationale instanties en verdragen(Europese AI verordening (AI act), Amerikaanse FDA) grotendeels zijn opgesteld vóórdat GenAI-technologie breed werd toegepast. Deze kaders houden onvoldoende rekening met specifieke eigenschappen en risico's van LLM's, zoals variabele output, bias, en mogelijke hallucinaties. Hierdoor ontbreekt het aan een breed gedragen, eenduidige en praktisch toepasbaar evaluatiekader dat specifiek aansluit bij de behoeften van het zorgveld.

Om te voorkomen dat de toepassing van LLM-technologie in de zorg vastloopt door onduidelijkheid over kwaliteit en veiligheid, is het essentieel om een door alle stakeholders gedragen overzicht van de belangrijkste kwaliteitscriteria op te stellen en hoe hieraan te toetsen vast te leggen. Een dergelijk overzicht biedt zorginstellingen, beleidsmakers en technologieontwikkelaars heldere richtlijnen om LLM's verantwoord en effectief te kunnen inzetten.

Het huidige beoordelingsproces, waarbij de uitkomsten van LLM-toepassingen handmatig worden vergeleken met menselijke uitkomsten, is een praktisch knelpunt. Dit proces is tijdrovend, kostbaar en belast het schaarse zorgpersoneel. Daarnaast is er kans op 'automation bias', het principe dat als er eenmaal een paar keer is vastgesteld dat de uitkomst correct is, de uitkomst de volgende keer niet meer gecontroleerd wordt. Daarom is het belangrijk ook de mogelijkheden van automatische evaluatie te onderzoeken, zodat het evaluatieproces efficiënter kan worden ingericht en de schaalbaarheid wordt vergroot.

Er bestaan verschillende medische, automatische evaluatiemethoden (benchmarks) zoals MedHELM en het Huggingface Medical Leaderboard, maar de ervaren kwaliteit van deze methoden voor concrete zorgapplicaties is vaak laag omdat de benchmarks meestal generiek en taak-overstijgend zijn ontworpen. Ze beoordelen doorgaans alleen algemene eigenschappen zoals taalkundige overeenkomsten, woordoverlap of semantische gelijkenis, terwijl zorgprofessionals juist sterk letten op aspecten als klinische nauwkeurigheid, volledigheid van medische informatie en context specifieke relevantie. Hierdoor scoren modellen soms goed op generieke benchmarks, terwijl ze in specifieke zorgpraktijken juist onvoldoende presteren of cruciale informatie missen.

Dit rapport geeft een overzicht van wat er nodig is voor een gedragen, praktisch toepasbaar evaluatiekader en bespreekt mogelijkheden voor automatische evaluatie van LLM-toepassingen in de zorg.

2.2 Doel van het rapport

Het ministerie van Volksgezondheid, Welzijn en Sport (VWS) heeft TNO gevraagd om, in samenwerking met het zorgveld, onderzoek te doen naar en een overzicht op te stellen van relevante en bruikbare kwaliteitscriteria, specifiek voor het evalueren van LLM-toepassingen in de zorg en specifiek het verlichten van administratieve lasten. Met de resultaten van het onderzoek wil TNO onderzoekers, ontwikkelaars en beleidsmakers in de zorg bouwstenen bieden die gebruikt kunnen worden om de kwaliteit van LLM-toepassingen gestructureerd te beoordelen (evalueren).

De focus van dit onderzoek ligt op technische kwaliteitscriteria, waar het nog te ontwikkelen evaluatiekader een balans dient te bieden tussen technische kwaliteitscriteria en relevante zorguitkomsten. Hierbij spelen organisatorische criteria, juridische kaders en de risicoclassificatie van een specifieke LLM-toepassing bijvoorbeeld ook een belangrijke rol.

Daarnaast richten we ons op kwaliteitscriteria die relevant en bruikbaar zijn gedurende de volledige levenscyclus van een LLM-toepassing, van ontwikkeling tot implementatie, en uiteindelijk monitoring. In Figuur 1 - een aangepaste versie van het 'Hulpmiddel Handelingsruimte Waardevolle AI voor gezondheid en zorg' (NL AIC, 2022) – zijn de verschillende fasen van de levenscyclus weergeven die iteratief doorlopen worden bij de ontwikkeling en implementatie van AI-toepassingen in de zorg.

Figuur 1. AI-levenscyclus, aangepast van Hulpmiddel Handelingsruimte Waardevolle AI voor gezondheid en zorg (2022)

De fase van de levenscyclus evenals de specifieke toepassing van een LLM beïnvloeden de relevantie en de weging van een kwaliteitscriterium. Fase specificaties en toepassingsafhankelijkheid vallen buiten de kaders van dit onderzoek, maar zullen deel moeten worden van een toepasbaar evaluatiekader.

Evaluatie is in elke fase van de levenscyclus belangrijk en het verschilt per fase welke evaluatiemethode het meest effectief (en gewenst) is. De idee-, verkennings-, ontwikkel-, pilot B-, implementatie- en productiefase worden tot op heden gekenmerkt door handmatige evaluatie door experts, zorgverleners en patiënten, waarbij menselijke beoordeling essentieel is voor het valideren van concepten, gebruikerservaringen en de klinische waarde. Dit betekent dat het merendeel van de evaluaties in de AI-levenscyclus handmatig plaats vindt, hetgeen het een tijdrovend en kostbaar proces maakt.

We beschrijven de mogelijkheden om op basis van relevante en bruikbare kwaliteitscriteria delen van het evaluatieproces te automatiseren. In pilot A van de levenscyclus, waarbinnen de technische validatie zonder patiënten plaatsvindt, kunnen doormiddel van automatische evaluatie grote hoeveelheden outputscenario's systematisch worden doorlopen om zwakke plekken, hallucinaties en inconsistenties te identificeren alvorens de technologie met patiënten wordt getest. In de laatste fase, waarbij continue monitoren en opschaling centraal staat, kan automatische evaluatie ingezet worden om (subtiele) verandering in modelgedrag te detecteren en signaleren om zo gebruikers te waarschuwen voor onverwachte veranderingen in functioneren.

Daarnaast is een praktijkcasus uitgewerkt waarin wordt beschreven hoe een technische validatie van een LLM-toepassing plaats kan vinden. Het praktijkvoorbeeld betreft een validatietraject uitgevoerd door de afdeling Digital Health/ AI for Health en het Julius Centrum, beiden onderdeel van het UMC Utrecht, in samenwerking met de afdelingen NICU en IC. De casus dient als illustratie van hoe kwaliteitscriteria in een concrete zorgcontext kunnen worden toegepast.

2.3 Opbouw van het rapport

In het volgende hoofdstuk (hoofdstuk 3) van dit rapport wordt allereerst beschreven hoe een lijst van mogelijk relevante en bruikbare kwaliteitscriteria tot stand is gekomen en vervolgen is aangevuld en aangescherpt tot een lijst van geselecteerde kwaliteitscriteria. Daarnaast worden de geselecteerde kwaliteitscriteria beoordeeld op mogelijkheden voor automatische evaluatie. Vervolgens wordt in hoofdstuk 4 aan de hand van een

praktijkvoorbeeld geïllustreerd hoe kwaliteitscriteria en evaluatiemethoden voor de evaluatie van LLM-toepassingen in de praktijk kunnen worden ingezet. In het voorlaatste hoofdstuk (hoofdstuk 5) worden de bevindingen van het onderzoek samengebracht in conclusies en aanbevelingen voor doorontwikkeling en vervolg. De bijlagen toehorende aan dit rapport zijn gebundeld in hoofdstuk 6.

3 Relevante en bruikbare kwaliteitscriteria voor inzet van LLM's in de zorg

Het beoordelen van de kwaliteit van LLM-toepassingen voor administratieve taken in de zorg vraagt om een gestructureerde aanpak op basis van heldere, en bruikbare criteria. Dit hoofdstuk presenteert een set aan kwaliteitscriteria, samengesteld op basis van literatuuronderzoek, expertgesprekken, focusgroepen en vragenlijstonderzoek onder zorgprofessionals, ervaringsdeskundige en experts op het gebied van AI in de zorg. De focus ligt daarbij op de technische evaluatie van een LLM-toepassing (Pilot A in Figuur 1). Deze criteria zijn van belang voor de gehele AI-levenscyclus. Deze set kan worden gebruikt als input voor een uiteindelijk evaluatiekader.

Aanvullend wordt er in dit hoofdstuk aandacht besteed aan de verschillen tussen menselijke en automatische evaluatiemethoden, en worden geselecteerde kwaliteitscriteria beoordeelt op de mogelijkheden voor automatische evaluatie.

3.1 Methode

Om tot een eerste set van relevante en bruikbare kwaliteitscriteria te komen (zie hoofdstuk 2.2. Doel van het onderzoek) zijn er drie onderzoeksactiviteiten uitgevoerd:

Onderzoeksactiviteiten (deel A)
1. Inventariseren van bestaande methoden en benchmarks
Er is een vooronderzoek uitgevoerd bestaand uit literatuuronderzoek en gesprekken met AI-experts van TNO, binnen én buiten de zorg, om aan te sluiten bij het snel ontwikkelende AI-landschap. De inzichten verkregen uit het vooronderzoek zijn uitgewerkt in een lijst van mogelijk relevante en bruikbare kwaliteitscriteria (zie Bijlage 1) en een memorandum waarin het onderwerp, de uitvraag en de projectkaders verder werden toegelicht (zie Bijlage 2).

2. Aanvullen en prioriteren van mogelijk relevante en bruikbare kwaliteitscriteria
In afstemming met het ministerie van VWS zijn professionals, AI-experts en ervaringsdeskundigen (hierna: experts) uit het zorgveld benaderd om deel te nemen aan een tweetal focusgroepen. De deelnemende experts zijn actief in verschillende zorgsectoren en hebben in diverse mate specifiek ervaring op het gebied van AI in de zorg (zie Bijlage 4 voor een specificatie van functie, rol en ervaring van de experts). Doel van de focusgroepen was het toetsen van de lijst van mogelijk relevante en bruikbare kwaliteitscriteria, en het komen tot een (voor)selectie van de belangrijke criteria. Om de experts voor te bereiden op de focusgroep is voorafgaand aan de focusgroep het memorandum met de aanwezigen gedeeld en is bij de start ruim aandacht besteed aan het toelichten van de context en aanpak onderzoek, en hetgeen wel én niet onderdeel is van het onderzoek. De experts zijn gevraagd welke criteria voor hen ontbraken in de opgestelde lijst van kwaliteitscriteria, om vervolgens vijf criteria te selecteren die voor hen als meest relevant worden beschouwd over de volledige AI-levenscyclus.
3. Valideren van geprioriteerde kwaliteitscriteria
Om de resultaten van het vooronderzoek en de opbrengsten uit de focusgroepen te valideren is er een digitale vragenlijst opgesteld (zie bijlage 3. Vragenlijst) en verspreid onder de experts van de focusgroepen (zowel de aanwezige als experts die uitgenodigd waren maar niet aanwezig konden zijn), onder enkele experts uit het netwerk van TNO en onder het netwerk van het ministerie van VWS (door het ministerie van VWS). De resultaten van het vragenlijstonderzoek zijn samengevoegd met de inzichten uit het vooronderzoek en opbrengsten uit de focusgroepen en verwerkt tot een lijst van kwaliteitscriteria. Daarnaast bood de vragenlijst ruimte voor aanvullende input, onder andere op de organisatorische aspecten van ontwikkeling, evaluatie en implementatie van LLM-toepassingen in de zorg. Deze input is meegenomen in de aanbevelingen voor doorontwikkeling van het evaluatiekader en vervolg (zie hoofdstuk 5).

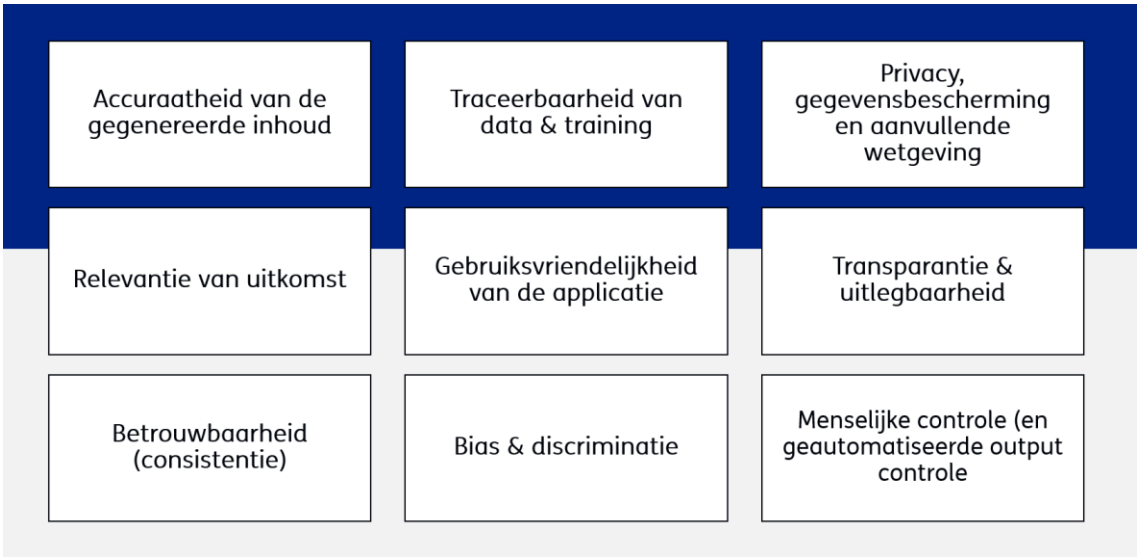
Zie Bijlage 4 voor een uitgebreidere toelichting op de aanpak, onderzoeksmethoden, activiteiten en resultaten.

3.2 Geselecteerde kwaliteitscriteria

Op basis van het vooronderzoek is een lijst van twaalf mogelijk relevante en bruikbare kwaliteitscriteria opgesteld (zie Bijlage 1) waarna de lijst van criteria door experts in de focusgroepen is aangevuld tot twintig mogelijk relevante en bruikbare kwaliteitscriteria (zie Bijlage 5). Vervolgens kregen de experts de opdracht vijf voor hen belangrijkste criteria te selecteren. De opbrengsten uit de focusgroepen zijn vervolgens gevalideerd en aangevuld aan de hand van een digitale vragenlijst en resulteerde in een selectie van negen kwaliteitscriteria die door het veld als meer belangrijk (relevant en bruikbaar) worden ervaren (zie Bijlage 6 voor de door het veld geselecteerde kwaliteitscriteria). **Accuraatheid**, **betrouwbaarheid**, **privacy & gegevensbescherming** en **relevantie** zijn door de experts als meest belangrijk geselecteerd. Daarnaast werden ook **menselijke controle**, **gebruiksvriendelijkheid**, **bias & discriminatie**, **traceerbaarheid van data & training** en **transparantie & uitlegbaarheid** door experts als belangrijke kwaliteitscriteria geselecteerd.

De kwaliteitscriteria zijn voorzien van een nummer ten hoeve van de leesbaarheid van dit rapport. De nummering en volgorde van kwaliteitscriteria houdt geen verband met de relevantie van de individuele kwaliteitscriteria of de onderlinge relatie. De relevantie van een kwaliteitscriterium hangt onder andere af van de specifieke toepassing van de LLM, de zorgcontext waarin de LLM-toepassing wordt ingezet, en de gekozen evaluatiemethode.

Het is belangrijk om op te merken dat deze kwaliteitscriteria niet volledig onafhankelijk van elkaar zijn. Een duidelijk voorbeeld is de afhankelijkheid tussen bias & discriminatie en accuraatheid: een AI-systeem dat is getraind op niet-representatieve data (bijvoorbeeld voornamelijk gegevens van mannelijke patiënten) kan uitstekende accuraatheidsscores behalen voor de vertegenwoordigde groep, maar systematisch minder nauwkeurige resultaten opleveren voor ondervertegenwoordigde groepen (zoals vrouwelijke patiënten). In dit geval is het onmogelijk om accuraatheid volledig te beoordelen zonder ook bias in overweging te nemen. Deze onderlinge verwevenheid van de kwaliteitscriteria vraagt om een geïntegreerde benadering bij de implementatie en evaluatie van AI-systemen in de zorg.



Figuur 2. Overzicht van de geselecteerde kwaliteitscriteria

1. Accuraatheid van de gegenereerde inhoud

Accuraatheid is essentieel omdat zorgverleners erop moeten kunnen vertrouwen dat (medische) relevante informatie en administratieve documenten correct en aanwezig zijn. Zelfs kleine fouten in medische gegevens kunnen ernstige gevolgen hebben voor de zorg voor patiënten. Daarnaast leidt een inaccurate toepassing tot extra tijdsbesteding voor gebruikers, omdat zij fouten moeten vinden, en corrigeren, wat administratieve lasten juist zou vergroten. Een bijkomend aandachtspunt is dat LLM's geen intrinsiek begrip hebben van de betekenis van woorden en taal, waardoor het realiseren van hoge accuratesse niet vanzelfsprekend is. Het belang van accuraatheid varieert afhankelijk van de toepassing maar is altijd hoog; bij medische diagnose of medicatieadvies is het essentieel, terwijl bij oriënterende taken een lagere accuratesse acceptabel kan zijn.

2. Betrouwbaarheid (consistentie)

Betrouwbaarheid verwijst naar het vermogen van LLM's om consistente resultaten te leveren over verschillende sessies heen. Voor zorgverleners is consistentie cruciaal om dezelfde reden waarom accuratesse belangrijk is: foute informatie kan leiden tot foute behandelingen en risico's voor de gezondheid van de patiënt. Als uitkomsten wisselend en onvoorspelbaar zijn, leidt dit tot onzekerheid en extra controletijd, wat contraproductief werkt voor het verminderen van administratieve lasten. Het belang van consistentie kan per toepassing verschillen: bij het genereren van medische en juridische documenten is consistentie bijvoorbeeld belangrijker dan algemene informatieverstrekking.

3. Privacy, gegevensbescherming & aanvullende wetgeving
De bescherming van patiëntgegevens is niet alleen wettelijk verplicht en ethisch verantwoord maar ook essentieel om het vertrouwen van patiënten en zorgverleners te waarborgen. LLM-toepassingen moeten strikt voldoen aan regelgeving zoals de AVG en andere relevante wetgeving rondom gegevensbeheer. Goede naleving voorkomt dat zorginstellingen zich organiseren en leveren van goede zorg bezig moeten houden met privacy incidenten. Dit criterium is met name van groot belang voor toepassingen die gevoelige patiëntinformatie verwerken, zowel voor het trainen als het gebruiken van LLM's.

4. Relevantie van uitkomst
De gegenereerde uitkomsten moeten direct aansluiten bij de behoeften en specifieke zorgcontext waarin ze worden gevraagd. Als de resultaten niet relevant of direct bruikbaar zijn, leidt dit tot extra werk in plaats van administratieve verlichting. Daarom moeten LLM's worden afgestemd op specifieke contexten en behoeften, zoals het genereren van patiënt specifieke ontslagbrieven of samenvattingen. Relevantie is essentieel bij toepassingen zoals gepersonaliseerde medische documenten, terwijl dit bij generieke content minder zwaar kan wegen. Wat relevant is, hangt sterk samen met de specifieke context van de toepassing.

5. Menselijke controle en geautomatiseerde output controle
Menselijke controle over de resultaten van LLM-toepassingen stelt zorgprofessionals in staat om gegenereerde informatie kritisch te beoordelen en waar nodig te corrigeren. Dit is cruciaal omdat zorgverleners eindverantwoordelijk blijven voor de inhoudelijke juistheid en toepasbaarheid van de informatie. Uit ons onderzoek bleek dat zorgprofessionals menselijke controle in alle administratieve toepassingen noodzakelijk vinden, mits dit eenvoudig en gebruiksvriendelijk kan gebeuren, zodat het netto leidt tot minder administratieve lasten.

Daarnaast ligt er ook een verantwoordelijkheid bij ontwikkelaars en leveranciers om door middel van geautomatiseerde begrenzingen en filters te voorkomen dat LLM's ongewenste of onveilige uitkomsten genereren. Dit verlaagt het risico op fouten, hallucinaties en overtredingen van privacyregels al voordat zorgverleners de output controleren. Een combinatie van menselijke controle binnen het zorgproces en automatische begrenzing van input en output door ontwikkelaars is daarmee essentieel voor veilige, effectieve inzet van LLM's in de zorg.

6. Gebruiksvriendelijkheid van de applicatie
Een gebruiksvriendelijke interface verlaagt de drempel voor zorgverleners om LLM-toepassingen effectief te gebruiken. Wanneer een toepassing intuïtief en eenvoudig te bedienen is, vereist dit minimale training en voorkomt het frustratie en extra tijdsbesteding. Gebruiksvriendelijkheid ondersteunt daarmee het hoofddoel van administratieve verlichting, doordat zorgverleners hun taken efficiënt kunnen uitvoeren.

7. Bias & discriminatie
LLM's worden getraind op grote hoeveelheden data. Wanneer deze data bias bevat- oftewel niet representatief is voor de volledige doelgroep- kan dit ertoe leiden dat het model niet voor iedereen even betrouwbaar of toepasbaar is. Bias en discriminatie moeten in LLM-gegenereerde informatie actief voorkomen worden, omdat deze negatieve gevolgen kunnen hebben voor patiënten en zorgbeslissingen. Ongelijke behandeling of vooringenomenheid in gegenereerde gegevens kan leiden tot verminderde zorgkwaliteit voor bepaalde groepen. Door te zorgen voor een zo goed mogelijke vertegenwoordiging van relevante groepen in de trainingsdata én de modeluitkomsten te controleren op bias, kan worden vastgesteld of de uitkomsten vrij zijn van vooringenomenheid. Dit biedt zorgverleners vertrouwen in de toepassing en voorkomt dat zorgverleners onnodig extra controlewerkzaamheden moeten

uitvoeren. Het belang van dit criterium is met name groot voor toepassingen die voor specifieke patiëntgroepen worden ingezet.

8. Traceerbaarheid van data & training
Traceerbaarheid maakt inzichtelijk hoe en op basis van welke data een LLM tot een bepaald resultaat komt. Dit vergroot de betrouwbaarheid en biedt de mogelijkheid om fouten snel op te sporen en te corrigeren. Transparantie rondom databronnen en trainingsmethoden ondersteunt vertrouwen, vergemakkelijkt compliance en vermindert het risico op administratieve lasten door onduidelijkheden of foutieve informatie, en geeft inzicht over eventuele bias. Veel leveranciers geven geen informatie over de gebruikte trainingsdata. Toch zijn er technische oplossingen die helpen de herkomst van informatie te achterhalen. Een voorbeeld is Retrieval Augmented Generation (RAG), waarbij een LLM niet alleen nieuwe informatie genereert op basis van beschikbare data, maar ook informatie ophaalt uit een gekoppelde bron (zoals een document of database) en dit in het antwoord weergeeft of citeert. Toch blijft het risico op hallucinatie bestaan: het model kan namelijk zelf een bron 'maken'. Dit maakt dat de echtheid van een bron direct verifieerbaar moet zijn voor veilig gebruik van de gegenereerde informatie.

9. Transparantie & uitlegbaarheid
De werking van LLM-toepassingen moet begrijpelijk en uitlegbaar zijn voor zorgverleners. Wanneer gebruikers inzicht hebben in hoe modellen functioneren, neemt het vertrouwen toe en is er minder behoefte aan extra controle of verificatie. Transparantie draagt bij aan de acceptatie en het effectief gebruik van LLM's, wat de administratieve efficiëntie in zorginstellingen ten goede komt. Vooral bij toepassingen die directe invloed hebben op klinische beslissingen of behandelingen is transparantie van belang. Transparantie en uitlegbaarheid zijn echter lastig te realiseren bij LLM's, aangezien deze modellen functioneren als 'black boxes' en niet altijd duidelijk is hoe bepaalde uitkomsten tot stand komen. Er is verder onderzoek nodig om transparante LLM's mogelijk te maken.

3.3 Evaluatiemethoden

Kwaliteitscriteria kunnen op twee manieren (evaluatiemethoden) worden ingezet om LLM-toepassingen voor de zorg te evalueren, namelijk via menselijke evaluatie en via automatische evaluatie.

Menselijke evaluatie: Bij menselijke evaluatie worden experts (bijvoorbeeld zorgverleners) gevraagd om de gegenereerde informatie handmatig te beoordelen. Dit kan door de oorspronkelijke tekst of de uitkomsten van een menselijk-gegenereerde voorbeeldtekst (blind) te vergelijken met een LLM-gegenereerde tekst. Het voordeel is dat de beoordelingen aansluiten bij wat de gebruikers aan kwaliteit verwachten. Het nadeel is dat menselijke evaluatie kostbaar en tijdrovend is. Dit kan een barrière kan vormen voor grootschalige adoptie van LLM's voor verlichting van administratieve werkdruk (en zorgt mogelijk zelfs eerst voor een verhoging ervan). In hoofdstuk 4 wordt aan de hand van een casestudie beschreven hoe menselijke evaluatie in de praktijk gebeurt.

Automatische evaluatie: Bij automatische evaluatie wordt gebruikgemaakt van algoritmen en tools die door een LLM-gegenereerde informatie automatisch beoordelen op basis van vooraf vastgestelde gouden standaarden ('ground truths'). Gouden standaarden zijn datasets met gevalideerde opdrachten, teksten en uitkomsten voor specifieke taken, zoals samenvattingen van patiëntdossiers, of het genereren van antwoorden op patiënt vragen per e-mail. Automatische metrieken zoals BLEU, ROUGE en METEOR worden gebruikt om te bepalen in welke de mate een LLM-toepassing erin slaagt de opdracht tot een goed einde te brengen. Deze metrieken richten zich hoofdzakelijk op semantische overeenkomst tussen de gouden standaard en de LLM-output. De combinatie van gouden standaarden en automatische evaluatiemethoden wordt ook wel een 'benchmark' genoemd. Bekende benchmarks voor de zorg zijn de eerdergenoemde MedHELM van Stanford en de Open Medical-LLM Leaderboard van Hugging Face. Een belangrijk aspect waar rekening mee moet worden gehouden bij het gebruik van deze benchmarks is dat ze domein en toepassing afhankelijk zijn. Hoe meer de gouden standaard overeenkomt met de gewenste toepassing, hoe beter de benchmark het gewenste resultaat weergeeft.

Dankzij automatische evaluatie is het eenvoudig(er) om verschillende LLM's te vergelijken voor een verzameling van taken. De hoge kosten die gepaard gaan met het opstellen van een gouden standaard dataset en de mogelijk lage correlatie tussen de uitkomst van een benchmark en de ervaren kwaliteit van een gebruiker (door feitelijke onjuistheden ondanks hoge semantische overeenkomst en de generieke opzet van bestaande benchmarks) zijn belangrijke nadelen van deze evaluatiemethode. In de praktijk kunnen modellen die goed scoren op benchmarks toch onvoldoende presteren in de (zorg)praktijk, omdat de resultaten onvoldoende aansluiten bij de reële behoeften en eisen van zorgverleners.
Onderstaande tabel (Tabel 1) geeft een overzicht van voor- en nadelen van de verschillende evaluatiemethoden met en zonder gouden standaard dataset:

	Zonder gouden standaard dataset	Met gouden standaard dataset
Zonder menselijke evaluatie	**Hoe werken dit soort metrieken?** Er wordt een LLM gebruikt om de tekst die gegenereerd is door een andere LLM te beoordelen. **Voordelen:** Er is geen menselijke beoordeling nodig en ook geen referentietekst nodig. **Nadelen:** a. De metrieken kunnen onvoorspelbaar zijn. Hoewel de metrieken gemiddeld goed correleren met menselijke annotatoren, is dat niet altijd zo. b. Het degelijk implementeren van deze metrieken is veel werk.	**Hoe werken dit soort metrieken?** Deze metrieken vergelijken een LLM-gegenereerde tekst met een referentietekst die geverifieerd is (een gouden standaard). **Voordelen:** Het gebruik van dergelijke metrieken kost weinig resources. **Nadelen:** a. Sommige metrieken hebben een lage correlatie hebben met menselijke annotatoren. b. Al deze metrieken hebben een referentietekst nodig, wat kosten voor dataverzameling met zich meebrengt. c. Ontwikkelaars kunnen naar de gouden standaard toe gaan werken, dit kan betekenen dat ze op andere echte datasets minder goed werken dan op de gouden standaard/

	Zonder gouden standaard dataset	Met gouden standaard dataset
Met menselijke beoordeling	**Hoe werken dit soort metrieken?** Menselijke evaluatoren geven een LLM-gegenereerde tekst een beoordeling. **Voordelen:** a. Ervan uitgaande dat je evaluatoren geschikt zijn voor de taak, krijg je beoordelingen van hoge kwaliteit. **Nadelen:** a. Als er experts nodig zijn voor een beoordeling, is het lastig om die beschikbaar te stellen. b. Als er een tweede evaluatie nodig is, moet die opnieuw uitgevoerd worden. Bij geautomatiseerde metrieken is dat niet zo.	

Tabel 1. De verschillende evaluatiemethoden voor LLM-toepassingen in de zorg.

Op basis van bovenstaande inzichten (voor- en nadelen) is het duidelijk dat op dit moment de validatie van LLM-toepassingen in de zorg niet zonder menselijke evaluatie kan en gebruik van automatische evaluatie op specifieke momenten in het adoptieproces menselijke valuatie (slechts) kan ondersteunen. Hoe specifieker de context van de evaluatie, hoe lastiger om automatisch te evalueren. De verwachting is dan ook dat automatische evaluatie met name vroeg in het adoptieproces ondersteunend kan zijn.

De mate waarin automatische evaluatie daadwerkelijk ondersteunend is aan menselijke evaluatie hangt sterk samen kwaliteit en betrouwbaarheid van het automatisch evalueren van specifieke kwaliteitscriteria. Door de geselecteerde kwaliteitscriteria hierop te beoordelen zal duidelijk worden dat enkele criteria zich redelijk lenen voor automatisch evalueren, terwijl andere criteria menselijke evaluatie vereisen.

Tot slot is verder onderzoek nodig naar gouden standaarden, LLM's die automatische evaluatie voor toepassingen in de Nederlandse zorg mogelijk maken en benchmarking. Individuele zorginstellingen verzamelen namelijk waardevolle data in verschillende validatiestudies, maar data blijven vaak binnen de organisatie. Een gecoördineerde aanpak voor het verzamelen en delen van data zou het opstellen van gouden standaarden en het ontwikkelen robuuste benchmarks die representatief zijn voor de Nederlandse zorgcontext kunnen bevorderen. Verschillende experts uit het zorgveld, waaronder specialisten op het gebied van AI en datadeling, benadrukken het belang hiervan in de focusgroepen en expertgesprekken. Een collectieve inspanning draagt daarnaast bij aan zowel de kwaliteit, als het verminderen van het aantal (herhalende) kostbare evaluatietrajecten.

3.4 Automatisch evalueren criteria

Om te kunnen beoordelen welk van de door experts geselecteerde kwaliteitscriteria geschikt zijn voor automatische evaluatie zijn aanvullende onderzoeksactiviteiten uitgevoerd:

Onderzoeksactiviteiten (deel B)
4. Ophalen inzichten op basis van expertgesprekken
TNO doet uitgebreid onderzoek naar de ontwikkeling en toepassing van GenAI in verschillende sectoren, waaronder de zorg. Binnen TNO zijn onderzoekers, ervaringsdeskundige en experts binnen en buiten de zorgsector actief en hun kennis en expertise delen in onderzoek naar veilige inzet van AI-toepassingen. Aan de hand van interviews met TNO-collega's zijn de mogelijkheden voor automatische evaluatie voor de geprioriteerde criteria opgehaald. De mogelijkheden zijn beoordeeld als: [niet mogelijk] [nauwelijks mogelijk] [gedeeltelijk mogelijk] en [goed mogelijk]. Alle beoordelingen zijn met een korte toelichting onderbouwd. Daarnaast is voor de vijf kwaliteitscriteria die door de experts als meest belangrijk werden geselecteerd (accuraatheid, betrouwbaarheid, relevantie, menselijke controle en privacy) een uitgebreidere toelichting op mogelijkheden voor automatische evaluatie uitgevraagd.

Onderstaande tabel biedt een eerste inzicht in de mogelijkheden voor automatische evaluatie per geprioriteerd criterium:

Criterium	Beschrijving	Automatische evaluatie
Accuraatheid van de gegenereerde inhoud	**De mate waarin de door LLM's gegenereerde informatie correct en foutloos is.** In de zorg betekent dit dat medische gegevens en administratieve documenten nauwkeurig moeten worden weergegeven om fouten in de patiëntenzorg te voorkomen.	**Redelijk mogelijk**. Er bestaan verschillende metrieken waarmee de kwaliteit van gegenereerde informatie gemeten kan worden. Deze metrieken zijn echter niet 100% betrouwbaar. Juist in de zorg kan een kleine fout enorme gevolgen hebben, waardoor het voor sommige zorgtoepassingen niet of nauwelijks mogelijk is.
Betrouwbaarheid (Consistentie)	**De consistentie van de gegenereerde antwoorden over verschillende sessies en toepassingen heen.** Voor de zorgsector betekent dit dat de LLM's betrouwbare en consistente resultaten moeten leveren, ongeacht de variatie in input of context, om vertrouwen bij zorgprofessionals op te bouwen.	**Redelijk mogelijk.** Het is relatief eenvoudig om een LLM meerdere malen hetzelfde te vragen en de variatie tussen de antwoorden te meten. Specifiek focussen op een consistente inhoud wanneer de zorgcontext variabel is, is lastiger.
Menselijke controle (en geautomatiseerde output controle)	**De mogelijkheid om de gegenereerde output te controleren en bij te sturen indien nodig.** Dit is belangrijk om ervoor te zorgen dat de informatie die door de LLM's wordt gegenereerd, voldoet aan de kwaliteitsnormen van de zorginstelling.	**Niet mogelijk.** Om te meten of gegenereerde output gecontroleerd en bijgestuurd wordt, kan er bijvoorbeeld een gebruikersstudie uitgevoerd worden met de eindgebruikers.
Relevantie van uitkomst	**De mate waarin de gegenereerde uitkomsten aansluiten bij de specifieke zorgcontext en behoeften van de gebruiker**. Dit houdt in dat de gegenereerde informatie direct toepasbaar moet zijn op de administratieve processen binnen zorginstellingen.	**Gedeeltelijk mogelijk.** Er bestaan metrieken om de relevantie van LLM-gegenereerde teksten te meten. Voor echt maatwerk in de zorgcontext kan er een gebruikersstudie uitgevoerd worden, of kan er een dataset met correcte vraag-antwoordparen verzameld worden.
Gebruiksvriendelijkheid van de applicatie	**De mate waarin de applicatie intuïtief en gemakkelijk te gebruiken is voor zorgprofessionals**. Dit omvat een eenvoudige interface, duidelijke instructies en minimale training, zodat zorgverleners efficiënt kunnen werken zonder extra administratieve belasting.	**Beperkt mogelijk.** De gebruiksvriendelijkheid van bepaalde applicaties dient gemeten te worden met gebruikersstudies. Er zijn mogelijkheden om deze studies gedeeltelijk te automatiseren maar dat is buiten scope van dit rapport.
Privacy, gegevensbescherming en aanvullende wetgeving	**De naleving van privacywetgeving en bescherming van patiëntgegevens**. Dit betekent dat de LLM's moeten voldoen aan de Algemene Verordening Gegevensbescherming (AVG) en andere relevante wet- en regelgeving, en dat er strikte protocollen moeten zijn voor gegevensbeheer en -beveiliging.	**Niet mogelijk.** De evaluatie van naleving van privacywetgeving en de bescherming van patiëntgegevens is te evalueren aan de hand van bureau-onderzoek.

Bias & discriminatie	**De mate waarin de gegenereerde inhoud vrij is van vooroordelen en discriminatie.** In de zorg is het essentieel dat de LLM's geen bias bevatten die de kwaliteit van de zorg kan beïnvloeden of bepaalde patiëntengroepen kan benadelen.	**Nauwelijks mogelijk.** Het meten van bias en discriminatie in vrije tekst is erg complex, contextgevoelig en niet heel betrouwbaar. Er zijn ontwikkelingen die dit op termijn mogelijk kunnen maken, maar nog niet breed toepasbaar.
Traceerbaarheid van data & training	**De mogelijkheid om de bron en de totstandkoming van de gegenereerde inhoud te achterhalen.** Dit is belangrijk voor het verifiëren van de juistheid van de informatie en het kunnen corrigeren van eventuele fouten.	**Redelijk mogelijk.** Door gebruik te maken van Retrieval Augmented Generation, kun je verwijzen naar de bron van een antwoord. Er bestaat echter altijd de mogelijkheid tot hallucinatie, waarvan de bron en totstandkoming niet achterhaald kan worden. Er moeten oplossingen komen waarmee de gebruiker de legitimiteit van de vermelde bron kan toetsen.
Transparantie & uitlegbaarheid	**De mate waarin de werking van de LLM's begrijpelijk en uitlegbaar is.** Zorgprofessionals moeten kunnen begrijpen hoe de modellen tot hun conclusies komen om deze te kunnen vertrouwen en verantwoorden.	**Nauwelijks mogelijk.** Het vakgebied dat zich bezighoudt met de uitlegbaarheid van LLM's staat nog in de kinderschoenen. De nieuwe methodieken die ontwikkeld worden zijn veelal nog onbetrouwbaar, en vereisen veel technische expertise om te implementeren en interpreteren. Onderzoek is nodig om dit mogelijk te maken.

Tabel 2: Geprioriteerde kwaliteitscriteria inclusief indicatie van mogelijkheid tot automatische evaluatie.

Voor de kwaliteitscriteria die door de experts zijn geselecteerd als meest belangrijk is een uitgebreidere toelichting en onderbouwing van de mogelijkheden voor automatische evaluatie opgesteld.

1. Accuraatheid

Het meten van accuraatheid is door middel van een automatische evaluatie is redelijk goed mogelijk, maar de meetinstrumenten zijn niet perfect. Zo zijn er bijvoorbeeld de traditionele metrieken (bijv. BLEU, ROUGE en METEOR), die vooral naar de tekstuele overlap kijken en in beperkte mate naar de betekenis van de tekst. Nieuwere methodes (bijv. G-EVAL) kijken niet enkel naar de vorm maar ook naar de inhoud. Deze methodes zijn echter wel onvoorspelbaarder, aangezien er ook een LLM wordt gebruikt om mee te evalueren. Voor de accuraatheid binnen het zorgdomein zijn beide methoden daarmee op dit moment beperkt toepasbaar, omdat een kleine fout een enorm effect kan hebben. Verder onderzoek kan uitwijzen of met een hoogwaardige gouden standaard een beter resultaat kan worden behaald.

2. Betrouwbaarheid (consistentie)

De afgelopen jaren is er onderzoek gedaan naar de consistentie van LLM's. Om consistentie te meten kan een LLM meerdere keren bevraagd worden. Het is dan mogelijk om de variatie in antwoorden te meten. Een complicerende factor is de creatieve aard van LLM's. Een

antwoord dat verschilt in vorm, kan qua inhoud hetzelfde zijn. Als de antwoorden qua vorm veel verschillen, maakt dat het moeilijker om de variatie en consistentie te meten.

3. Relevantie

De relevantie van de uitkomst is te meten aan de hand van metrieken, die dezelfde kanttekeningen vereisen als de metrieken voor het meten van accuraatheid. Relevantie zoals in zoals beschreven in tabel 2 gaat specifiek over de behoeften van de gebruiker en een specifieke zorgcontext. Voor dergelijk maatwerk zijn bestaande metrieken maar beperkt geschikt. Om te meten of er voorzien wordt in de behoeften van de gebruiker, is een gebruikersstudie aan te raden. Om aan te sluiten op een specifieke zorgcontext, kan een dataset die specifiek is ontwikkeld voor de context helpen. Bijvoorbeeld: voor een samenvattingstool kan er een dataset gecreëerd worden die bestaat uit oorspronkelijke teksten met daarbij een samenvatting die goedgekeurd is door experts (de gouden standaard).

4. Privacy, gegevensbescherming en aanvullende wetgeving

De naleving van privacy en bescherming van patiëntgegevens is deels te meten met een geautomatiseerde evaluatie, door te toetsen of de oorspronkelijke persoonsdata lekken in de uitkomsten van de LLM. Het is aan te raden om naleving van wetgeving ook te evalueren in een bureau-onderzoek, waarbij verschillende disciplines moeten samenwerken om tot een antwoord te komen. In elk geval moeten juridische, privacy- en data-experts samenwerken met ICT'ers en IT'ers om te bepalen of de privacywetgeving wordt nageleefd en de patiëntgegevens voldoende beschermd worden in alle stappen van het ontwikkelen en gebruik van de LLM.

5. Menselijke controle & output controle

Om te meten of gegenereerde output gecontroleerd en bijgestuurd wordt, kan een gebruikersstudie uitgevoerd worden met de eindgebruikers. Automatische evaluatie is hiervoor niet geschikt omdat het beoordelen van menselijke controle vraagt om observatie van hoe professionals in de praktijk omgaan met de AI-output en wanneer ze ingrijpen. Dit kan alleen worden vastgesteld door directe observatie en feedback van gebruikers. De juiste evaluatiemethode hangt af van de specifieke zorgcontext en de fase waarin de LLM-toepassing zich bevindt. Praktijkvoorbeelden helpen zorginstellingen te begrijpen welke methoden passend zijn voor hun situatie. Tot slot gebeurt output controle van een LLM vaak achter de schermen door de leverancier en is daarmee niet automatisch te evalueren.

De 'keuze' voor een passende evaluatiemethode en de relevantie van een specifiek kwaliteitscriterium hangt samen met de context - waaronder kaders en wensen van een unieke zorgorganisatie – waarbinnen een specifieke LLM-toepassing wordt ingezet en geëvalueerd. Door aan de hand van praktijkvoorbeelden te illustreren welke kwaliteitscriteria en evaluatiemethoden kunnen worden ingezet voor LLM-toepassingen in de zorg, wordt het voor zorginstellingen concreet hoe (in iedere fase van de levenscyclus) evaluatie een essentiële rol speelt bij het waarborgen van kwaliteit en veiligheid.

4 Praktijkvoorbeeld: LLM-toepassing voor automatisering ontslagbrieven in UMC Utrecht

4.1 Inleiding en relevantie voor kwaliteitscriteria

Om de geselecteerde kwaliteitscriteria voor evaluatie LLM's in de zorg in de context van de zorgpraktijk te kunnen plaatsen hebben we aanvullende onderzoeksactiviteiten uitgevoerd:

Onderzoeksactiviteiten (deel C)
4. Ophalen en uitwerken praktijkvoorbeeld op basis van expertgesprekken
Op basis verkennende gesprekken en inzichten verkregen uit de focusgroepen is het praktijkvoorbeeld van het UMC Utrecht geselecteerd om de praktische toepassing van kwaliteitscriteria en evaluatiemethoden te illustreren. In het praktijkvoorbeeld wordt een LLM-toepassing ten behoeve van administratieve lastenverlichting geëvalueerd. Daarnaast biedt het praktijkvoorbeeld inzicht in de uitdagingen die zorginstellingen ervaren bij de validatie van LLM's in de zorg en onderschrijft het de behoefte aan efficiëntere evaluatiemethoden. De betrokken expert van het UMC Utrecht is direct betrokken bij de evaluatie van de LLM-toepassing. Aan de hand van een interview is opgehaald wat het UMC Utrecht beoogde te evalueren (doel), hoe zijn dit hebben aangepakt (methode), welke resultaten zijn hebben opgehaald, welke uitdagingen zij hebben ervaren en hoe het vervolg voor het UMC Utrecht eruitziet.

4.2 Context en achtergrond

In het UMC Utrecht is een LLM-toepassing ontwikkeld voor het automatiseren van ontslagbrieven. De toepassing is ontwikkeld door het AI for Health team binnen het UMC Utrecht, gebruikmakend van een licentie-versie van GPT. Het doel van deze toepassing is om de administratieve last van artsen te verlichten door het genereren van conceptontslagbrieven op basis van informatie uit het elektronisch patiëntendossier (EPD).

4.3 Validatiemethodiek

Het UMC Utrecht bevindt zich in de Pilot A fase van de AI-levenscyclus, waarin evaluatie zonder patiënt interactie centraal staat (zie Figuur 1 voor de AI-levenscyclus). In deze fase

voert het UMC Utrecht een intensieve menselijke evaluatie uit waarbij zorgprofessionals door een LLM-gegenereerde ontslagbrieven beoordelen aan de hand van een aantal kwaliteitscriteria. Die kwaliteitscriteria zijn: relevantie, accuraatheid, samenhang/structuur en relevantie. De aanpak van deze evaluatiestudie omvat op een combinatie van handmatige en vergelijkende evaluatie, opgesteld op basis van een door het UMC Utrecht eerder uitgevoerd literatuuronderzoek.

Het kwaliteitscriterium samenhang/structuur ontbreekt in de selectie van relevante en bruikbare kwaliteitscriteria opgesteld in het onderzoek van TNO. Het kwaliteitscriterium is relevant voor de specifieke LLM-toepassing die in de pilot van het UMC Utrecht wordt geevalueerd en niet (of beperkt) voor toepassingen van LLMs ten behoeve van administratie lastenverlichting in het algemeen. De relevantie van het kwaliteitscriterium is toepassing-afhankelijk.

Literatuuronderzoek als basis

De validatiemethode van het UMC Utrecht is gebaseerd op een uitgebreide literatuurstudie naar bestaande evaluatiemethoden voor samenvattende NLP-taken. Hieruit zijn vier kernaspecten (kwaliteitscriteria) geïdentificeerd voor evaluatie:

1. **Relevantie:** Bevat de ontslagbrief alle essentiële informatie uit het patiëntendossier?
2. **Accuraatheid**: Is de informatie in de ontslagbrief feitelijk juist en verifieerbaar?
3. **Samenhang/structuur**: Is de ontslagbrief logisch opgebouwd en goed gestructureerd?
4. **Relevantie**: Bevat de ontslagbrief alleen informatie die relevant is voor de ontvanger

Om de kwaliteit van de LLM-gegenereerde ontslagbrieven grondig te beoordelen, koos het UMC Utrecht voor een tweeledige aanpak waarbij zowel een detailanalyse (1) als holistische vergelijking (2) werd uitgevoerd. Vanuit het literatuuronderzoek bleek dit een passende manier om de subjectieve ervaring van kwaliteit voor de LLM-toepassing in het UMC Utrecht te evalueren.

Stap 1: Een gedetailleerde, handmatige evaluatie

Medische studenten controleerden nauwkeurig de door de LLM-gegenereerde ontslagbieven (uitkomsten). Er werd gecontroleerd of (i) alle belangrijke informatie in de brieven stond (relevantie) en (ii) geen onjuiste informatie was toegevoegd (accuraatheid). Hierbij werden de ontslagbrieven regel-voor-regel vergeleken met de bijbehorende patiëntdossiers. Bij twijfel konden de medische studenten de informatie uit de door de LLM-gegenereerde ontslagbrief aan de artsen van de afdeling voorleggen. Dit bleek een cruciale stap want artsen wezen ruim 30% van de informatie die door studenten werd aangemerkt als ‘hallucinatie’ (onjuiste informatie) aan als klinisch logische of correcte informatie. Dit laat zien dat de praktijkervaring van artsen een belangrijke factor is voor het ‘goed’ kunnen beoordelen van medische informatie. Voor een student foute informatie is voor een ervaren arts mogelijk een logische interpretatie van de informatie uit het patiëntendossier. Dit roept direct vragen op over het vermogen van medische studenten om de door de LLM-gegenereerde teksten te beoordelen en andere fouten of omissies niet te missen. De discrepantie tussen de beoordeling van medische studenten en ervaren artsen bij de vermeende hallucinaties suggereert dat het mogelijk is dat nuanceringen en klinisch relevante afwijkingen in de door de LLM-gegenereerde teksten door de studenten niet als problematisch zijn herkend, waar dit wel nodig was.

Stap 2: Een vergelijkende evaluatie op briefniveau
Voor de aspecten samenhang en structuur werd gekozen voor een directe vergelijking waarbij artsen twee versies van dezelfde ontslagbrief kregen te zien: één geschreven door een arts en één door GPT. Ze moesten blind (zonder te weten hoe de informatie in de ontslagbrief gegeneerd was) bepalen welk van de twee brieven de beste brief was. Het was een behoorlijk 'strenge' test voor GPT, omdat de door een arts opgestelde ontslagbrief een definitieve, gecontroleerde versie van een ontslagbrief was. Daarentegen was de door GPT-gegenereerde ontslagbrief een eerste, onbewerkt concept. Toch scoorde GPT verrassend goed al waren er wel verschillen tussen afdelingen zichtbaar. Dit laat zien dat zelfs iedere afdeling binnen een ziekenhuis een eigen schrijfstijl en verwachtingen voor een ontslagbrief hanteert.

Door zowel de details de ontslagbrief als het geheel te beoordelen in stap 1 en in stap 2 kreeg het team zicht op hoe goed de LLM-toepassing presteert. Tegelijkertijd maakte het ook inzichtelijk op welke aspecten menselijke betrokkenheid (op basis van kennis & ervaring) onmisbaar blijft voor het beoordelen en het gebruik van AI-gegenereerde medische teksten.

4.4 Bevindingen gerelateerd aan kwaliteitscriteria

De evaluatiestudie van het UMC Utrecht leverde belangrijke inzichten op over hoe verschillende kwaliteitscriteria presteren in een specifieke LLM-toepassing, waaronder enkele van de geselecteerde kwaliteitscriteria uit het TNO-onderzoek. Per kwaliteitscriterium relevant voor de specifieke LLM-toepassing van het UMC Utrecht zijn de belangrijkste bevindingen uit de evaluatiestudie opgehaald:

Accuraatheid: De GPT-gegenereerde brieven vertoonden hallucinaties, maar minder dan verwacht. Opvallend was dat hallucinaties vaak te maken hadden met causale verbanden die in het dossier niet expliciet werden gelegd (bijv. "deze patiënt krijgt medicatie X omdat zij kortademig is"). Het model genereerde in deze gevallen zelf klinische redeneringen die niet letterlijk in het dossier stonden, maar probeerde hiermee ontbrekende verbanden te overbruggen. Deze vorm van 'eigen invulling' door het model - het zelfstandig infereren van causale relaties tussen klinische observaties en interventies zonder expliciete documentatie hiervan in de brongegevens - vormt een specifiek type hallucinatie waarbij het model medische logica toepast die mogelijk correct zou kunnen zijn, maar niet wordt ondersteund door de beschikbare patiëntgegevens. Wat voldoende is in het kader van accuraatheid voor verdere implementatie is in dit praktijkvoorbeeld aan afdelingen en eindgebruikers.

Betrouwbaarheid: De consistentie varieerde per klinische afdeling, wat wijst op de invloed van domein specifieke factoren. Deze variatie roept vragen op over de trainingsdata van het model. Mogelijk presteert het model beter op afdelingen waarvan de medische documentatie sterker vertegenwoordigd was in de trainingsset, of waarvan het taalgebruik en de documentatiestandaarden meer overeenkomen met hetgeen het model tijdens training heeft geleerd. Dit suggereert dat betrouwbaarheid van AI-toepassingen voor de zorg is niet universeel en (deels) afhankelijk is van de trainingsdata en de mate waarin de data representatief is voor de specifieke context waarin de AI-toepassing wordt ingezet.

Output controle: De noodzaak van menselijke controle wordt ook in het praktijkvoorbeeld onderstreept. Daarnaast geldt dat artsen in de praktijk ontslagbrieven altijd na moeten

kijken, ongeacht de ontslagbrief handmatig of door AI gegeneerd wordt. Echter dient men scherp te zijn op deze aanname, waar automation bias mogelijk optreedt voor AI-gegenereerde teksten. Dit houdt in dat gebruikers van de AI-toepassing neigen overmatig te vertrouwen op de geautomatiseerde systemen en de uitkomsten minder (of zelf niet meer) controleren, vergeleken met de controle van handmatig gegenereerde teksten.

Een belangrijk aspect van de validatie in dit praktijkvoorbeeld was het blind beoordelen van de ontslagenbrieven waarbij artsen niet wisten of de teksten handmatige of door AI-gegenereerde teksten waren. Dit blinde beoordelingsproces test en biedt inzicht in de beoordelingscapaciteit van de artsen, zonder beïnvloeding van de artsen door voorkennis over de herkomst. In een praktijksituatie waarbij artsen wél weten dat ze AI-gegenereerde brieven controleren, is het mogelijk dat automation bias sterker optreedt. Verder onderzoek moet uitwijzen in welke mate dit plaats vindt en wat het effect hiervan is op het ‘goed’ beoordelen van door AI-gegenereerde ontslagbrieven.

Relevantie: De relevantie van de ontslagbrieven was in algemene zin goed. De structurering van de teksten was vergelijkbaar met structuur die wordt aangehouden door artsen, hoewel er variatie zichtbaar was tussen klinische afdelingen binnen het UMC Utrecht.

Gebruiksvriendelijkheid: De gebruiksvriendelijkheid van de LLM-toepassing werd in deze fase beperkt geëvalueerd. In de volgende fase waarin de applicatie (in een pilotsetting) in de praktijk in gebruik wordt genomen speelt dit aspect een belangrijkere rol.

Privacy: De LLM-toepassing in dit praktijkvoorbeeld maakt gebruik van een beveiligde verbinding met een Cloud-gebaseerde GPT-dienst maar waarbij de data binnen Nederland blijft. Dit is voor het UMC Utrecht goedgekeurd door de Data Security Officer van het UMC Utrecht en voor de evaluatiestudie getoetst. Hoewel niet bekend is of de LLM-toepassing op een Nederlandse server draait, is contractueel vastgelegd dat gegevens uitsluitend binnen de EU worden verwerkt en niet voor andere doeleinden worden aangewend.

Uit dit praktijkvoorbeeld blijkt dat accuraatheid, relevantie en menselijke controle als de belangrijkste kwaliteitscriteria werden beschouwd voor de specifieke LLM-toepassing. Dit is te verklaren met de aard van het gebruik: de ontslagbrieven (uitkomst van de LLM-toepassing) bevatten kritische medische informatie waarvoor juistheid essentieel is om gezondheidsimplicatie bij de patiënt te voorkom dus worden altijd door zorgprofessionals gecontroleerd voordat de ontslagbrieven definitief gemaakt worden. Het cruciaal is dat het gebruikssysteem op een manier wordt ingericht dat de ontslagbrief pas kan worden verstuurd aan de patiënt als de controle afdoende is uit gevoerd om te voorkomen dat gebruikers (de zorgprofessional) te snel en blind vertrouwen op de LLM. Het risico van de eerder beschreven automation bias kan op die manier worden tegengegaan. Deze pragmatische benadering weerspiegelt de realiteit van zorgprocessen en toont ten aanzien van administratieverlichting aan dat perfectie niet voor iedere toepassing of context noodzakelijk is, zolang er passende controlemechanismen zijn ingericht.

De bevindingen uit de evaluatiestudie van het UMC Utrecht zijn waardevol voor het traject naar bredere adoptie de LLM-toepassing. De resultaten van de evaluatiestudie worden voorgelegd aan de verschillende afdelingen binnen het UMC Utrecht. De afdelingen kunnen op basis van die bevindingen zelf beslissen of ze de LLM-toepassing willen implementeren in

hun werkprocessen. Dit markeert (potentieel) de overgang van Pilot A naar Pilot B van de AI-levenscyclus (zie Figuur 1) waarbij de LLM-toepassing getest zal worden met echte patiënten. Het UMC Utrecht erkent dat de uiteindelijke adoptie afhangt van specifieke behoeften en werkwijze per afdeling. Het is daarom van belang dat iedere afdeling zelfstandig kan oordelen of de balans tussen administratieve verlichting en benodigde menselijke controle voor de afdeling waardevol is.

4.5 Uitdagingen voor automatische evaluatie

De casus van het UMC Utrecht belicht een tweeledig doel achter de grondige evaluatieaanpak. Enerzijds streeft men naar directe verlichting van administratieve lasten voor zorgprofessionals door de introductie van LLM-technologie voor het genereren van ontslagbrieven. Anderzijds – en dit is een strategisch belang – werkt het UMC bewust toe naar een toekomst waarin technische validatie door geautomatiseerde evaluatie ondersteund kan worden.
Door het uitvoeren van een uitgebreide handmatige evaluatie bouwt het UMC Utrecht in feite een waardevolle gouden standaard: een dataset op van ontslagbrieven. De dataset zal als fundament dienen voor toekomstige automatische evaluatiemethoden en uiteindelijk een (interne) benchmark. Waar artsen en medische studenten op dit moment nog regel-voor-regel de kwaliteit moeten controleren, kan men op termijn steeds meer aspecten geautomatiseerd beoordelen door te vergelijken met de opgebouwde datasets. Deze aanpak is een voorbeeld van hoe zorgorganisaties in de technische validatie fase (Pilot A) kunnen investeren om later in de AI-levenscyclus, met name in de fase tijdens/na productie, efficiënter kunnen evalueren en monitoren. Echter is deze strategische route naar automatische evaluatie niet zonder uitdagingen. De volgende uitdagingen zijn in het praktijkvoorbeeld van het UMC Utrecht naar voren gekomen:

Arbeidsintensiviteit: De validatiemethode van de evaluatiestudie is zeer arbeidsintensief en niet schaalbaar. Dit is cruciaal voor een sector waarin personeelstekort reeds een significant uitdaging vormt. Het inzetten van medische studenten en artsen voor uitgebreide handmatige evaluaties is voor lange termijn geen houdbare aanpak wanneer er onvoldoende capaciteit beschikbaar is, of wanneer meerdere LLM-toepassingen of -versies van LLM-toepassingen geëvalueerd moeten worden.

Variëteit binnen zorgcontext: De significante verschillen tussen klinische afdelingen maakt generalisatie uitdagend, en te verwachten is dat de verschillen tussen sectoren binnen de zorg alleen maar groter zijn. Deze domeinvariatie compliceert de ontwikkeling van gestandaardiseerde, geautomatiseerde evaluatiemethoden die voor alle afdelingen en instellingen even waardevol zijn. Een belangrijke vraag is hoe deze variatie kan worden geadresseerd in verder ontwikkelingen van een benchmark. Een oplossingsrichting is het verzamelen van grotere en beter passende gouden standaarden datasets. Een andere vraag hierbij is waar deze variatie vandaan komt. Mogelijk komt dit voort uit de data waarop de LLM getraind is. Met de ontwikkeling van nieuwe LLM's zou dit verholpen kunnen worden.

Subjectiviteit van kwaliteitscriteria: Het praktijkvoorbeeld toont aan dat enkele essentiële kwaliteitscriteria, zoals samenhang, structuur en relevantie, inherent subjectief zijn. Ze vereisen contextueel begrip en klinisch beoordelingsvermogen dat moeilijk te vatten is in geautomatiseerde metrieken De discrepantie tussen studentbeoordelingen en artsenbeoordelingen ten aanzien van inconsistenties van de LLM vergeleken met de

originele tekst illustreert deze uitdaging: automatische systemen zouden mogelijk dezelfde ‘strenge’ benadering hanteren als studenten, zonder het klinische inzicht van artsen te kunnen gebruiken om context-specifieke interpretaties te herkennen en vice versa.

Correlatie met automatische metrieken: Een belangrijk inzicht uit het literatuuronderzoek van het UMC Utrecht is dat bestaande automatische evaluatiematen vaak slecht correleren met menselijke beoordelingen. Dit is een fundamenteel obstakel voor volledige automatisering en pleit voor hybride evaluatiemethoden, waarbij automatische metrieken worden geïmplementeerd en aangevuld met gerichte menselijke beoordelingen en is op dit moment de meest kansrijke benadering. Daarnaast zullen de discrepanties tussen de automatische evaluatiematen en menselijke beoordeling onderzocht moeten worden om de mogelijkheden voor automatisering te verbeteren.

Privacyoverwegingen: De kwestie van privacy vormt een significant uitdaging voor het delen van validatiedata ten behoeve van benchmarking en verdere automatisering. Patiënten geven namelijk niet expliciet toestemming voor het gebruik van hun gegevens in dergelijke benchmarks. Dit belemmert de ontwikkeling van gedeelde datasets als basis voor het trainen en valideren van robuuste, geautomatiseerde. Het gebruik van synthetisch gegenereerde patiëntdata zou een eerste stap kunnen zijn.

4.6 Vervolgstappen en lessons learned

Na de evaluatiestudie zijn twee van de drie betrokken afdelingen overgegaan tot het in gebruik nemen van de LLM-toepassing in een pilotsetting. Het UMC Utrecht werkt aan manieren om de validatie deels te automatiseren, met als doel een balans te vinden tussen grondige evaluatie en schaalbaarheid. Na deze technische validatie is het nu aan individuele afdelingen binnen het UMC Utrecht of ze de LLM-toepassingen ook willen testen, en mogelijk implementeren in hun werkprocessen.

Eén van de vraagstukken die in een volgende fase onderzocht moet worden is de kwaliteitsnorm die (binnen het UMC Utrecht) gehanteerd moet worden. Wanneer is de toepassing goed genoeg? De maatstaf voor acceptatie is niet perfectie, maar vergelijking met de huidige standaard (voor ontslagbrieven) terwijl de huidige standaard ook niet zonder fouten is, evenals de zorgverlener zelf. Welke foutmarge is acceptabel deze specifieke LLM-toepassing? Wanneer de foutmarge van de LLM-gegenereerde brieven vergelijkbaar of zelfs lager is dan van volledig handmatig opgestelde brieven, kan dit een sterk argument zijn voor adoptie, zelfs als het systeem niet perfect is. Dit perspectief verschuift de discussie van: ‘Is het perfect?’ naar ‘Is het beter dan wat we nu hebben?’, wat een realistischer kader biedt voor evaluatie en implementatiebeslissingen.

Belangrijke lessen uit deze casus:

- **Belang van multidisciplinaire evaluatie**: De discrepantie tussen student-beoordelingen en arts-beoordelingen onderschrijft het belang van verschillende perspectieven in het validatieproces.
- **Barrière voor adoptie**: De casus laat zien dat er veel onderzoek nodig is om een LLM-toepassing zorgvuldig in gebruik te nemen. Het UMC Utrecht heeft aanzienlijke middelen geïnvesteerd in het ontwikkelen van een validatiemethodiek, het opbouwen van een gouden standaard dataset, en het uitvoeren van arbeidsintensieve evaluaties.

- **Contextafhankelijkheid van kwaliteitscriteria**: De casus bevestigt dat de toepassing en weging van kwaliteitscriteria sterk afhankelijk is van de specifieke LLM-toepassing. Bij ontslagbrieven konden kleine afwijkingen worden geaccepteerd onder menselijke supervisie, terwijl bij andere toepassingen zoals automatische diagnosecoderingen het criterium accuraatheid veel zwaarder zou moeten wegen vanwege potentiële directe gezondheidsgevolgen. Een zinvolle evaluatie van LLM's in de zorg vereist daarom een context-specifieke benadering.
- **Behoefte aan collectieve benchmarks**: Ook uit het gesprek met de expert van het UMC Utrecht komt naar voren dat het waardevol is om collectieve benchmarks en gedeelde evaluatiekaders te ontwikkelen. Het evalueren van LLM's vraagt momenteel veel tijd en expertise, mede door de variatie tussen verschillende toepassingen, datastructuren en beoordelingscriteria. Door kennis en ervaringen te bundelen, zouden zorginstellingen wellicht kunnen leren van elkaars aanpak en middelen efficiënter kunnen inzetten. Het ontbreken van dergelijke referentiepunten maakt vergelijking van resultaten tussen instellingen uitdagend. Nader onderzoek zou kunnen uitwijzen in welke mate gezamenlijke benchmarks het validatieproces zouden kunnen versnellen en of dit de adoptie van betrouwbare AI-toepassingen in de zorg kan bevorderen.

Het praktijkvoorbeeld demonstreert de praktische toepassing van kwaliteitscriteria en brengt zowel de potentie als de uitdagingen van handmatige als automatische evaluatie van LLM-toepassingen in de zorg in beeld. Het is daarbij van belang rekening te houden met de toepassing- en contextafhankelijkheid en de weging (relevantie) van de kwaliteitscriteria.

5 Conclusie en aanbevelingen

5.1 Conclusie

Potentie en barrières van LLM's in de zorg

LLM's hebben aanzienlijke potentie om administratieve lasten in de zorgsector effectief te verminderen door zorgprofessionals te ondersteunen bij taken zoals documentatie, dossiervorming en overige administratieve werkzaamheden. Verschillende zorginstellingen in Nederland hebben al stappen gezet richting implementatie van dergelijke toepassingen, waarmee de behoefte aan betrouwbare en efficiënte evaluatiemethoden groter wordt.

Een belangrijke barrière voor grootschalige adoptie van AI-toepassingen en specifiek LLMs is het beperkte inzicht in de kwaliteit en de borging hiervan. Vanwege de complexiteit en soms onvoorspelbare gedrag (zoals hallucinaties en inconsistenties) van LLMs is het momenteel uitdagend om betrouwbare kwaliteitsbeoordelingen uit te voeren. Met de huidige evaluatiemethoden kost dit veel tijd. Dit kan een obstakel vormen voor verantwoorde opschaling, met name omdat veel zorginstellingen beperkte capaciteit hebben voor uitgebreide evaluaties.

Geselecteerde kwaliteitscriteria

Op basis van dit onderzoek is er een overzicht van de belangrijkste (relevante en bruikbare) kwaliteitscriteria voor het evalueren van LLM-toepassingen in de zorg opgesteld, specifiek voor het verlichten van administratieve lasten:

- Accuraatheid van gegenereerde inhoud
- Betrouwbaarheid (consistentie)
- Privacy, gegevensbescherming en aanvullende wetgeving
- Relevantie van uitkomst
- Menselijke controle (en geautomatiseerde output controle)
- Gebruiksvriendelijkheid van de applicatie
- Bias & discriminatie
- Transparantie
- Traceerbaarheid

De geselecteerde kwaliteitscriteria vormen de bouwstenen voor een gestructureerde en praktijkgerichte beoordeling van de technische kwaliteit van LLM's gedurende hun volledige levenscyclus: van ontwikkeling tot implementatie en monitoring.

Lessen uit de praktijkcasus UMC Utrecht

De praktijkcasus van het UMC Utrecht toont aan dat het valideren van een specifieke LLM-toepassing voor het genereren van ontslagbrieven een tijdrovend en kostbaar proces is. De onvermijdelijke inzet van zorgprofessionals voor gedetailleerde handmatige evaluaties is waardevol, maar in het kader van personeelstekorten en ervaren werkdruk in de zorg (voor de lange termijn) niet wenselijk én niet schaalbaar. De casus toont ook aan dat de toepassing en weging van kwaliteitscriteria sterk afhankelijk zijn van de specifieke context

waarin de LLM wordt ingezet. Hierdoor kan een criterium zoals consistentie of betrouwbaarheid voor de ene toepassing als meer relevant worden gezien dan voor een andere toepassing. Verder laat het zien dat de data die verzameld is gedurende de pilot mogelijk gebruikt kan worden voor het automatisch evalueren van de kwaliteitscriteria.

Uitdagingen voor automatische evaluatie

Automatische evaluatiemethoden bieden potentieel om het evaluatieproces efficiënter en minder arbeidsintensief te maken. Bestaande automatische methoden en benchmarks (zoals MedHELM) blijken echter beperkt bruikbaar door hun generieke karakter en daarmee slechte correlatie met de ervaren kwaliteit van gebruikers in de zorgpraktijk. Automatische evaluatie lijkt vooralsnog geschikt ter ondersteuning voor technische validatie zonder patiënten (technische pilotfase) en het monitoren van een geïmplementeerde LLM-toepassing in de praktijk. Echter hangt de effectiviteit van de benchmark sterk af van de kwaliteit van gouden standaard, de gekozen metrieken en aansluiting bij de zorg-specifieke context.

5.2 Aanbevelingen

Verkennen van automatische validatiemogelijkheden

Gezien het belang van het verminderen van de administratielast en de behoefte aan automatische validatie is het van waarde voor het zorgveld om de mogelijkheden hiervoor verder te onderzoeken. Met name vroege technische validatie en continue monitoring van geïmplementeerde toepassingen lijken te profiteren van het gebruik van benchmarks. De mate waarin een benchmark in staat is menselijke evaluatie te ondersteunen of zelfs te vervangen lijkt sterk af te hangen van de mate waarin de gebruikte gouden standaarden representatief zijn voor de beoogde toepassing. Toekomstig onderzoek zou zich kunnen richten op het bepalen van de bruikbaarheid van benchmarks in verschillende fasen van het implementatieproces en de noodzakelijke omvang en aard van de benodigde gouden standaarden. Op de lange termijn is het wenselijk om gezamenlijk te werken aan de ontwikkeling van (nationale) zorgbenchmarks die gebaseerd zijn op realistische, praktijkgerichte scenario's.

Dataverzameling en-deling voor benchmarkontwikkeling

Voor het realiseren van kwalitatieve benchmarks die in staat zijn menselijke evaluatie te ondersteunen is een aanzienlijke hoeveelheid kwalitatief hoogwaardige en representatieve data nodig. Momenteel verzamelen grote(re) zorginstellingen datasets in het kader van individuele validatiestudies, die waardevolle input kunnen leveren voor benchmarks die voor het hele veld zijn te gebruiken. Allereerst moet gevalideerd worden of een dergelijke benchmark in staat is inderdaad menselijke evaluatie te ondersteunen of zelfs vervangen. Vervolgens kan de individueel verzamelde data op een gecoördineerde en verantwoorde wijze toegankelijk gemaakt worden. Zo kunnen zorginstellingen gezamenlijk profiteren van de opgebouwde kennis en voorkomen dat elke instelling afzonderlijk gelijksoortige, kostbare evaluaties uitvoert.

Aanpak van juridische en privacy-gerelateerde barrières

De stap van individuele validatiestudies naar collectieve benchmarks vraagt echter om een gecoördineerde aanpak en een zorgvuldige afweging van juridische en privacy-gerelateerde barrières. Het delen van validatiedata buiten de oorspronkelijke studiecontext is vaak uitdagend vanwege privacywetgeving en het ontbreken van expliciete toestemmingen voor datadeling.

Om deze uitdagingen effectief aan te pakken zullen betrokken partijen - zoals zorginstellingen, beleidsmakers, technologieontwikkelaars en juridische experts –

gezamenlijk moeten (willen) verkennen welke maatregelen nodig zijn om veilige, verantwoorde en zinvolle datadeling mogelijk te maken. Naast juridische oplossingen vraagt dit ook om investeringen in de benodigde infrastructuur, mogelijk innovatieve oplossingen zoals 'data spaces' die gebruik maken van Privacy Enhancing Technologies (PET's) en passend beleid. PET's zijn een verzameling van technieken die ontwikkelaars in staat stellen data te benutten terwijl privacy gerespecteerd wordt. Een voorbeeld is synthetische data. Dit is een realistische maar kunstmatige kopie zonder echte persoonsgegevens te bevatten. Een vervolgonderzoek moet uitwijzen welke PET in staat is een werkzame oplossing te bieden.

Bevordering van samenwerking en kennisdeling

Daarnaast is het belangrijk om samenwerkingsverbanden en kennisdeling binnen de zorgsector actief te stimuleren. Door gezamenlijk validatiestudies te organiseren, waarbij data worden verzameld met het oog op benchmarkontwikkeling zal op termijn schaalbare evaluatiemethoden opgezet kunnen worden. Tegelijkertijd lijkt een hybride aanpak waarin menselijke en automatische evaluaties gecombineerd worden het meest kansrijk wat er voor pleit kostbare menselijke evaluaties gecoördineerder in te zetten om minder beslag te leggen op het schaarse zorgpersoneel.

Concrete vervolgstappen en rol VWS

Een eerstvolgende stap is doorontwikkelen en aanscherpen van de geselecteerde kwaliteitscriteria. Vervolgens het uitvoeren van een haalbaarheidsonderzoek naar de technische, juridische en organisatorische aspecten van zorgbenchmarks voor LLM's. Hierbij is het essentieel om zorgprofessionals en patiënten intensief te betrekken in zowel de ontwerpfase als de implementatie van LLM-toepassingen en de evaluatiemethoden zelf. Door gezamenlijk aan deze aanbevelingen te werken, ontstaat een duurzame en effectieve basis voor het verantwoord inzetten van LLM's in de zorg, wat ten goede komt aan de gehele sector door administratieve lasten te verminderen zonder compromissen op kwaliteit en veiligheid.

Het ministerie van Volksgezondheid, Welzijn en Sport zou vanuit haar bestaande rol, in haar netwerk en op basis van de kennis die binnen het ministerie beschikbaar is een faciliterende en aanjagende rol moeten spelen in de nodige vervolgstappen.

6 Bijlagen

Overzicht van bijlagen
Bijlage 1. Lijst mogelijk relevante en bruikbare kwaliteitscriteria
Bijlage 2. Memorandum
Bijlage 3. Digitale vragenlijst
Bijlage 4. Aanpak onderzoek
Bijlage 5. Aangevulde lijst mogelijk relevante en bruikbare kwaliteitscriteria
Bijlage 6. Lijst geselecteerde kwaliteitscriteria
Bijlage 7. Bronnenlijst

Ondertekening

TNO Healthy Living & Work Leiden, 28 maart 2025

ValidSigned by Daan Kloet
on 2025-03-28 15:11:50

Daan Kloet
Research Manager

ValidSigned by Roos Boereboom
on 2025-03-28 14:57:22

Roos Boereboom
Auteur

Health & Work

Sylviusweg 71
2333 BE Leiden
www.tno.nl

TNO innovation for life