AI-schrijfassistent voor heldere gemeentelijke brieven, gebaseerd op de OVER-schrijfwijzer

KAAN GÖGCAY, AMBER VAN HASSEL, AMIR JACOBS, TERENCE TERZOL, and BIBIËNE WÜST*,
Hogeschool van Amsterdam, Nederland

In dit project is onderzocht hoe gemeentelijke brieven van de OVER-gemeenten begrijpelijker kunnen worden gemaakt voor inwoners, met behulp van een AI-tool voor tekstversimpeling. Hiervoor is de LiNT-II-score, een maat voor leesbaarheid, gebruikt om de effectiviteit van de herschreven teksten te evalueren. Het project combineerde een decoder-only taalmodel met een naïve-RAG-architectuur, zodat het model relevante regels uit de OVER-schrijfwijzer kan toepassen tijdens het herschrijven van brieven. Door deze aanpak kan de AI-tool contextueel onderbouwde suggesties doen, terwijl de medewerker eindverantwoordelijk blijft voor de inhoud. De resultaten laten zien dat de AI-tool de leesbaarheid van brieven kan verbeteren ten opzichte van een baseline-model, maar dat de prestaties per briefcategorie variëren mede door een scheve verdeling van voorbeelden in de dataset. Daarnaast blijkt dat de responstijd van het model en het gebruikersbegrip van meldingen aandachtspunten vormen voor verdere optimalisatie. Vervolgonderzoek kan zich richten op het toepassen van alternatieve tekstvereenvoudigingstechnieken, zoals een advanced-RAG-implementatie, het vaststellen van aanvullende maatregelen om automation bias verder te beperken en het gebruik van aanvullende prestatiematen naast de LiNT-II-score om de prestaties van het model volledi ger te evalueren.

Additional Key Words and Phrases: OVER-gemeenten, OVER-schrijfwijzer, Large Language Model, naïve-RAG, baseline, LiNT-II, B1-taalniveau, vector-database, decoder-only, human-in-the-loop, gebruikerstest

*Alle auteurs hebben in gelijke mate bijgedragen aan dit onderzoek.

Authors' Contact Information: Kaan Gögcay, kaan.gogcay@hva.nl; Amber van Hassel, amber.van.hassel@hva.nl; Amir Jacobs, amir.jacobs@hva.nl; Terence Terzol, terence.terzol@hva.nl; Bibiëne Wüst, bibiene.wust.wust@hva.nl,
Hogeschool van Amsterdam, Amsterdam, Nederland.

1 Introductie

1.1 Context

In Nederland hebben overheden de taak om inwoners duidelijk te informeren over wetgeving, voorzieningen en procedures [43]. Schriftelijke communicatiemiddelen, zoals brieven, spelen hierbij een belangrijke rol, omdat deze vaak informatie bevatten over rechten en plichten (Bijlage A).

Op nationaal niveau is er toenemende aandacht voor begrijpelijke taal binnen publieke organisaties [34]. Dit is belangrijk, omdat ongeveer 2,5 miljoen mensen in Nederland moeite hebben met lezen en schrijven [16]. Het is essentieel dat iedere inwoner overheidscommunicatie goed kan begrijpen. Daarom gebruiken veel gemeenten een schrijfwijzer die bepaalt hoe zij teksten op B1-niveau opstellen. Zo heeft onder andere de Gemeente Amsterdam een schrijfwijzer ontwikkeld [35]. Ook de OVER-gemeenten Oostzaan en Wormerland maken gebruik van een schrijfwijzer, OVER-schrijfwijzer genoemd, om medewerkers te ondersteunen bij hun communicatie met inwoners [36]. Daarnaast worden online hulpmiddelen, zoals [60], gebruikt om woordkeuze te beoordelen op begrijpelijkheid (Bijlage A). Ten slotte verkennen de OVER-gemeenten de inzet van nieuwe technologieën, zoals kunstmatige intelligentie (AI), ter ondersteuning van haar communicatieve processen.

1.2 Probleem

Ondanks bestaande richtlijnen voor begrijpelijke communicatie, zoals gemeentelijke schrijfwijzers [35, 36], blijken gemeentelijke brieven in de praktijk vaak moeilijk te begrijpen voor een groot deel van de Nederlandse bevolking. Dat dit geen incidenteel probleem is, blijkt uit onderzoek naar Nederlandse publieke organisaties: een analyse van 240 teksten van 70 organisaties toont aan dat veel officiële communicatie te complex is, onder andere door lange zinnen en ingewikkeld woordgebruik [61]. Dit kan leiden tot misverstanden en problemen, zoals het niet lezen van essentiële informatie of het missen van kansen doordat mensen niet begrijpen wat er van hen wordt verwacht [34, 43].

Ook binnen de OVER-gemeenten doet dit probleem zich voor. Medewerkers beoordelen elkaars teksten op naleving van de regels van de OVER-schrijfwijzer, maar doordat de schrijfwijzer omvangrijk is en niet altijd consequent wordt toegepast, kunnen deze beoordelingen onbetrouwbaar zijn. Hierdoor ontvangen inwoners soms brieven die zij niet goed begrijpen, wat de kans op miscommunicatie vergroot. Daarnaast blijkt uit een interview (Bijlage A) dat het evalueren en herschrijven van teksten volgens de OVER-schrijfwijzer extra tijd kost.

Deze combinatie van een algemeen landelijk probleem en knelpunten binnen de OVER-gemeenten creëert de behoefte aan een ondersteunend systeem dat medewerkers helpt bij het vereenvoudigen of beoordelen van teksten, zodat brieven begrijpelijker worden en de communicatie met inwoners verbetert.

1.3 Bestaand werk over het probleem

Om de begrijpelijkheid van overheidscommunicatie te verbeteren, hebben verschillende gemeenten de afgelopen jaren maatregelen genomen. Voorbeelden hiervan zijn lokale schrijfwijzers [35, 36] en algemene richtlijnen voor duidelijke taal [33]. Daarnaast bestaan er op landelijk niveau initiatieven, zoals Direct Duidelijk [34]. Direct Duidelijk is een programma van de Rijksoverheid dat organisaties helpt hun communicatie toegankelijker te maken voor iedereen. Deze hulpmiddelen bieden praktische adviezen over woordkeuze, zinslengte en toon, maar de toepassing in de praktijk blijkt blijft beperkt en inconsistent, waardoor richtlijnen niet altijd structureel worden gevolgd.

Binnen het onderzoeksveld van Automatic Text Simplification (ATS) zijn al technologische ontwikkelingen gedaan. Taalmodellen zoals T5 en GPT worden internationaal gebruikt voor het vereenvoudigen van vooral Engelstalige teksten [54, 55]. Hoewel deze modellen laten zien dat AI in staat is teksten te analyseren en te versimpelen, worden zij binnen de OVER-gemeenten nog niet ingezet (Bijlage A). Overheidsorganisaties kunnen hun documenten vanwege privacy- en veiligheidsredenen niet zomaar aanbieden aan externe platforms [20]. Dit kan ertoe leiden dat privacygevoelige informatie uitlekt. Daarnaast zijn bestaande modellen niet specifiek getraind op de eisen en schrijfrichtlijnen die gelden voor Nederlandse gemeentelijke communicatie. Daarom is er behoefte aan een veilige AI-oplossing die kan helpen bij het herschrijven of beoordelen van brieven, waarbij bestaande gemeentelijke richtlijnen worden gevolgd en de menselijke controle behouden blijft (Bijlage A).

1.4 Wetenschappelijk gat in bestaand werk

Het wetenschappelijke gat ligt bij de beperkingen van automatische tekstvereenvoudiging voor Nederlandse gemeentelijke communicatie met juridische inhoud. Bestaand onderzoek naar Automatic Text Simplification (ATS) en Lexicale Simplificatie (LS) richt zich vooral op Engelstalige domeinen en houdt nauwelijks rekening met teksten met juridische gevolgen [39, 68].

Recent Nederlands onderzoek heeft op verschillende manieren gekeken naar tekstcomplexiteit en tekstvereenvoudiging. Met taalmodellen zoals "Geen makkie" en "LSBertje" is onderzocht welke taalkenmerken bijdragen aan de moeilijkheid van Nederlandse teksten [42], terwijl benchmarkonderzoek inzicht geeft in de complexiteit van gemeentelijke teksten [62]. Daarnaast laten monitorstudies zien hoe burgers overheidsteksten ervaren [28, 29] en beschrijven schrijfwijzers en richtlijnen hoe overheden begrijpelijker kunnen schrijven [33–36]. Deze onderzoeken richten zich echter vooral op algemene leesbaarheid en doelgroepgericht schrijven en besteden weinig aandacht aan juridische gemeentelijke brieven, waarin formuleringen directe rechtsgevolgen hebben. Hierdoor sluiten zij nog niet volledig aan bij de eisen van gemeentelijke communicatie, waarin juridische nauwkeurigheid essentieel is.

In dit onderzoek wordt daarom ingezoomd op twee tekortkomingen in de bestaande wetenschap:

(1) **Behoud van juridische betekenis en verantwoordelijkheid bij tekstvereenvoudiging**
Bestaand onderzoek naar ATS en LS richt zich vooral op leesbaarheid, zoals kortere zinnen en eenvoudigere woorden [42, 62], terwijl gemeentelijke brieven juridische bepalingen bevatten waarbij kleine herformuleringen tot andere interpretaties en rechtsgevolgen kunnen leiden. Het huidige ATS-onderzoek biedt geen methode om automatisch te waarborgen dat vereenvoudigde teksten juridisch en semantisch correct blijven [68]. Volledige automatisering is daarom problematisch, waardoor medewerkers eindverantwoordelijk moeten blijven. Hoewel human-centered AI dit uitgangspunt ondersteunt [46, 64], ontbreekt een concreet ontwerp voor een Nederlandse AI-schrijfassistent die deze menselijke eindverantwoordelijkheid behoudt.

(2) **Domein-specifieke evaluatie van juridische gemeentelijke teksten**
Gangbare evaluatiemethoden zoals SARI en n-gram-metrics meten vooral taalkundige vereenvoudiging [62, 68], maar geven geen inzicht in naleving van Nederlandse schrijfwijzers of het behoud van juridische juistheid [33, 35, 36]. Om deze tekortkoming te ondervangen, wordt in dit onderzoek naast een prestatiemaat een Human-in-the-Loop-benadering toegepast, waarbij gemeentelijke medewerkers suggesties beoordelen op juridische juistheid, naleving van de OVER-schrijfwijzer en praktische toepasbaarheid. Deze combinatie biedt een realistischer en verantwoordelijker evaluatiekader voor tekstvereenvoudiging in een juridische context.

Door deze tekortkomingen ontbreekt momenteel een domeinspecifiek evaluatiekader voor AI-systemen die gemeentelijke teksten vereenvoudigen, terwijl tegelijkertijd de oorspronkelijke betekenis en juridische juistheid geborgd moeten blijven.

1.5 Voorstel

Het voorstel is om een AI-gestuurde applicatie (AI-tool) te ontwikkelen die medewerkers van de OVER-gemeenten ondersteunt tijdens het schrijven van brieven aan de bewoners. Hierbij wordt een taalmodel gebruikt om bijvoorbeeld automatisch moeilijke woorden en lange zinnen te signaleren. Hierbij worden automatisch verbeteringssuggesties gegeven die direct aansluiten op de schrijfwijzer van de OVER-gemeenten.

Bij elke suggestie krijgt de medewerker een toelichting en verwijzing naar de regel in de schrijfwijzer waar de suggestie betrekking tot heeft, zodat niet alleen de tekst wordt verbeterd, maar medewerkers ook inzicht krijgen in het waarom van de voorgestelde aanpassingen. Er wordt onderzocht hoe de AI-tool kan worden geïntegreerd in de bestaande schrijfprocessen van medewerkers, zodat het hulpmiddel eenvoudig en laagdrempelig kan worden gebruikt. Medewerkers stellen eerst de volledige brief op, waarna de AI-tool wordt ingezet om de tekst te beoordelen.

Met behulp van deze AI-tool kunnen medewerkers praktisch en betrouwbaar geholpen worden bij het schrijven van begrijpelijke brieven aan inwoners. Hiermee wordt de kwaliteit van gemeentelijke brieven verhoogd en wordt de dienstverlening aan inwoners toegankelijker en duidelijker.

2 Achtergrond

2.1 Basiskennis

Het automatisch versimpelen van teksten is een taalverwerkingsprobleem waarvoor taalmodellen in de praktijk worden ingezet [38, 69]. Verschillende studies laten zien dat taalmodellen alternatieve woorden of zinnen kunnen aanbieden voor het herschrijven van complexe formuleringen [38, 69]. In [69] wordt bijvoorbeeld een model voorgesteld dat zinnen vereenvoudigt door een Transformer-architectuur te combineren met een externe parafrase-database (Simple PPDB). In plaats van simpelweg 'moeilijke' woorden te vervangen door 'makkelijkere' synoniemen, leert het model welke vereenvoudigingsregels het meest passend zijn en past deze gecontroleerd toe tijdens het genereren van de nieuwe zin. Taalmodellen kunnen dus eventueel gebruikt worden als ondersteuning voor het (her)schrijven van (begrijpelijke) gemeentelijke brieven.

2.1.1 State-of-the-art.

Er is veel onderzoek verricht naar tekst-versimpeling [26]. Traditionele bag-of-words-modellen negeerden woordvolgorde, waardoor context verloren ging. N-grammodellen verbeterden dit deels door korte reeksen mee te nemen, maar misten nog steeds de langere verbanden in een tekst. De interesse in sequentiemodellen nam in 2015 toe met de heropleving van Recurrent Neural Networks (RNN's) [26]. Hoewel RNN's meer context behouden dan N-grammodellen, kampen ze met hoge rekenkosten en moeite bij het verwerken van zeer lange teksten. De introductie van Transformer-modellen zorgde vervolgens voor een doorbraak binnen NLP [26]. Dankzij (self-)attention herkennen zij relaties over de gehele tekst en zijn ze efficiënter schaalbaar. Voor de OVER-gemeenten is dit cruciaal, zodat suggesties logisch aansluiten bij de volledige briefcontext. Transformers zijn daarom het meest geschikt voor de herschrijf-tool, bijvoorbeeld in een naïve-RAG implementatie [52].

2.2 Stakeholderanalyse

Op basis van de door de OVER-gemeenten gegeven introductie en de daaraan gekoppelde vragenrondes is een stakeholderanalyse uitgevoerd (Bijlage A). In Tabel 1 staat een overzicht van de stakeholders, hun belangen, ervaren problemen en modelvereisten binnen het huidige proces van het herschrijven en controleren van brieven om te beoordelen of deze voldoen aan de richtlijnen van de OVER-schrijfwijzer.

Table 1. Stakeholderanalyse - belangen, ervaren problemen en modelvereisten

Stakeholder	Belangen	Ervaren problemen	Modelvereisten
OVER-gemeenten	Een snelle tool die ervoor zorgt dat de werkdruk van de medewerkers minder belast wordt, waardoor het werk sneller afkomt.	Op dit moment kost het controleren en schrijven van deze brieven veel tijd. Ook hebben niet alle medewerkers dezelfde (specialistische) kennis over het B1-taalniveau.	Een snelle tool die de medewerkers van de gemeente helpt om de werkdruk te verlagen en werk sneller af te ronden.
Inwoners van de OVER-gemeenten	Begrijpelijke teksten waarmee duidelijk is wat de mededeling is. Daarnaast moeten de teksten ook inclusief geschreven zijn; niemand zou zich buitengesloten of gediscrimineerd moeten voelen.	Verwarring over de beschreven situatie en vervolgstappen, waardoor problemen ervaren worden bij het krijgen van hulp op bijvoorbeeld sociaal vlak en bij betalingen.	Privacygevoelige informatie mag niet worden gelekt en teksten moeten inclusief geschreven zijn. Daarnaast moet de brief een persoonlijke toon bevatten, als de context dat vraagt.
Medewerkers van OVER-gemeenten die de brieven schrijven/controleren	De tool moet medewerkers helpen om de brieven begrijpelijker te verwoorden met gebruik van de schrijfwijzer die gebaseerd is op B1-niveau. Hierbij moet rekening gehouden worden met juridische aspecten en het beschermen van privacygevoelige informatie.	Op dit moment kost het controleren en schrijven van deze brieven veel tijd. Ook hebben niet alle medewerkers dezelfde (specialistische) kennis over het B1-taalniveau.	De verbeterde teksten moeten gebaseerd zijn op de schrijfwijzer die de OVER-gemeenten aanbiedt. De verbeterde teksten moeten daarbij ook refereren naar de schrijfwijzer.
ICT-afdeling	De tool moet lokaal gerund worden, of op de eigen server. Hierbij moet voorkomen worden dat privacygevoelige gegevens gelekt worden.	Als data extern verwerkt wordt of als er onvoldoende controle is over wie de tool gebruikt, kunnen beveiligingsrisico's ontstaan.	Aangezien het model lokaal moet kunnen draaien op de bestaande infrastructuur van de gemeente, wordt geprobeerd het model zo licht mogelijk te ontwerpen.
Overheid	De tool moet voldoen aan de wetten en regelgeving die door de overheid zijn opgesteld met betrekking tot technische, ethische en juridische eisen.	Inwoners van gemeenten begrijpen niet altijd waar ze recht op hebben, doordat zij brieven niet begrijpen.	De brieven die verstuurd worden moeten op B1-niveau geschreven zijn, volgens de richtlijnen van de overheid (Bijlage A).

Figuur 1 toont daarnaast de positie van de stakeholders in de macht-belangdiagram. Medewerkers die brieven schrijven en controleren, evenals de OVER-gemeenten, hebben zowel veel belang als veel macht en spelen daarom een centrale rol in de ontwikkeling van de AI-tool. De ICT-afdeling beschikt over veel macht vanwege de technische randvoorwaarden, maar heeft minder direct inhoudelijk belang. Inwoners van de OVER-gemeenten en de overheid hebben juist veel belang bij duidelijke communicatie, maar weinig macht over het ontwerp van de tool.

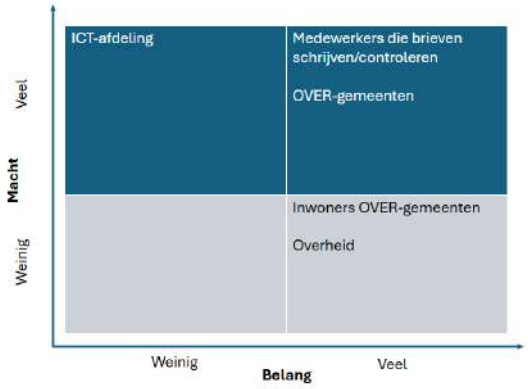


Fig. 1. Macht-belangdiagram

2.3 Huidige situatie

Bij het opstellen van een brief doorloopt de OVER-gemeenten een twee-staps controle. Een medewerker schrijft de brief in Microsoft Word eerst inhoudelijk correct en herschrijft deze vervolgens naar begrijpelijke taal met hulpmiddelen zoals de OVER-schrijfwijzer. Daarna controleren ten minste twee medewerkers de begrijpelijkheid van de brief en plaatsen opmerkingen in hetzelfde Word-document. Deze wordt vervolgens teruggestuurd naar de medewerker die de brief heeft opgesteld, zodat diegene de feedback kan verwerken. Na de feedback-verwerking ontvangt de inwoner de brief. Begrijpt de inwoner deze niet, dan kan de inwoner contact opnemen met de gemeente. Bij begrip volgt eventueel actie op basis van de inhoud van de brief. Dit proces is visueel weergegeven in Figuur 2.

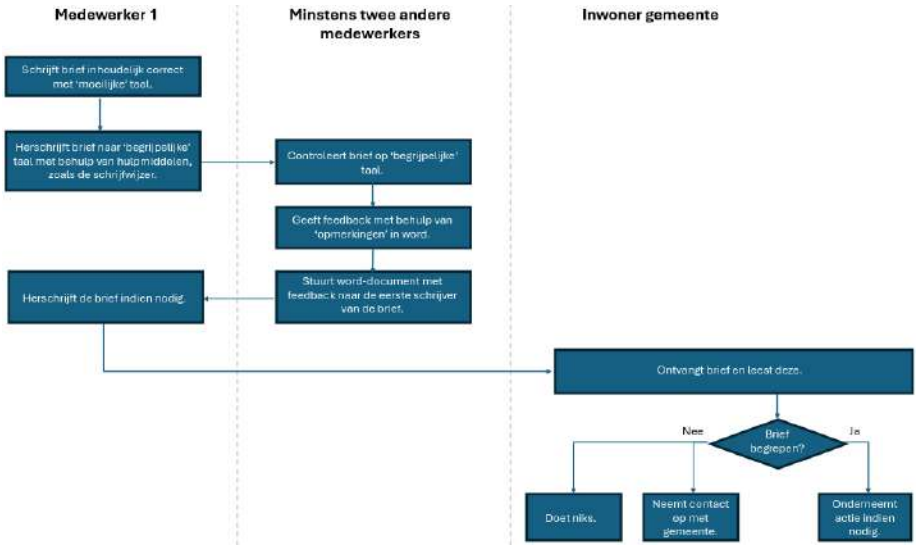


Fig. 2. Huidige situatie binnen de OVER-gemeenten

2.4 Geschiktheid van AI voor deze opdracht en bijbehorende technische, maatschappelijke en ethische overwegingen

De inzet van AI bij het herschrijven van gemeentelijke brieven volgens de OVER-schrijfwijzer vraagt om een zorgvuldige afweging van technische, ethische en maatschappelijke aspecten. Technisch gezien zijn bestaande ATS-modellen niet specifiek afgestemd op Nederlandse gemeentelijke communicatie, waardoor zij onnauwkeurige of juridisch onjuiste suggesties kunnen doen die niet aansluiten bij de OVER-schrijfwijzer. Tegelijkertijd zijn huidige AI-modellen in staat om teksten te analyseren en te vereenvoudigen, mits zij expliciet worden aangestuurd met deze richtlijnen en binnen een veilige, lokale omgeving worden toegepast. Vanwege privacy- en veiligheidseisen (AVG) mogen persoonsgegevens niet via externe servers worden verwerkt, wat betekent dat de tool lokaal moet draaien en dat opslag van tekst zoveel mogelijk moet worden beperkt of direct moet worden verwijderd [20].

Vanuit ethisch perspectief vormt automation bias een belangrijk risico omdat medewerkers onder tijdsdruk te veel kunnen vertrouwen op AI-suggesties en deze zonder voldoende kritische beoordeling overnemen [15]. Om dit te voorkomen moet het systeem werken volgens een human-in-the-loop-principe, waarbij medewerkers altijd eindverantwoordelijk blijven voor de inhoud van de brief en bewust zijn van de ondersteunende, niet-beslissende rol van AI [47] (Bijlage E). Transparantie over de beperkingen van het systeem is hierbij essentieel.

Maatschappelijk gezien kan AI bijdragen aan toegankelijkere gemeentelijke communicatie voor een brede doelgroep, waaronder laaggeletterden. Tegelijkertijd vraagt dit om training en bewustwording van medewerkers, zodat zij de beperkingen van het systeem kennen en hun eigen schrijfvaardigheid blijven ontwikkelen. De ontwikkeling en implementatie van de AI-tool vereisen daarom nauwe samenwerking met stakeholders om knelpunten tijdig te signaleren en aan te pakken. Binnen de geschetste technische, maatschappelijke en ethische overwegingen wordt AI in deze opdracht beschouwd als een geschikt ondersteunend hulpmiddel binnen het schrijfproces, mits het wordt toegepast in combinatie met de OVER-schrijfwijzer en menselijke beoordeling.

3 Requirements

Voorafgaand aan de ontwikkeling van de AI-tool zijn requirements opgesteld en geprioriteerd met behulp van een MoSCoW-analyse. Op deze manier zijn de belangrijkste eisen vastgesteld waaraan het uiteindelijke product moet voldoen (Tabel 2). De eisen die geen must have waren zijn opgenomen in Bijlage C. Hierbij is onderscheid gemaakt in verschillende categorieën: ethische, juridische, functionele, technische, organisatorische en duurzaamheids requirements. De in de tabel genoemde stakeholders geven niet de bron van de requirement aan, maar op wie de requirement van toepassing is of die direct door de requirement worden beïnvloed.

De organisatorische en duurzaamheidsvereisten zijn niet in dit hoofdstuk opgenomen maar uitsluitend in Bijlage C vermeld, omdat het regelen van organisatorische en duurzame zaken binnen de OVER-gemeenten buiten de scope van dit onderzoek valt. Dit onderzoek richt zich namelijk op het onderzoeken en ontwerpen van de AI-oplossing en niet op de interne organisatie-inrichting of implementatie binnen de OVER-gemeenten. Deze eisen zijn echter wel van belang en zullen als must-haves gelden wanneer de tool daadwerkelijk wordt geïmplementeerd binnen de OVER-gemeenten.

Table 2. Alle must-have requirements

RQ-nr	Categorie	Requirement	Stakeholder	Toelichting
RQ01	Ethisch	Gebruikers moeten weten hoe de AI-tool werkt, zonder dat AI-geletterdheid nodig is [23].	Medewerkers die brieven schrijven/controleren	Medewerkers: De medewerker moet in de AI-tool weten welke knoppen waarvoor dienen, zodat de tool correct gebruikt kan worden voor het herschrijven van brieven.
RQ02	Ethisch en functioneel	Een suggestie mag niet zonder toestemming van de medewerker worden aangepast in de brief (bijv. om bias en fouten van het LLM, zoals hallucinaties, te voorkomen), vandaar dat het systeem een human-in-the-loop moet hebben.	Medewerkers die brieven schrijven/controleren; OVER-gemeenten; Overheid	Medewerkers: De medewerker wil het herschrijven niet volledig aan de AI overlaten omdat hij/zij eindverantwoordelijk blijft voor de inhoud van de brief en de AI-tool is slechts ondersteunend is [23]. OVER-gemeenten en overheid: Willen dat inwoners correct worden geïnformeerd en dat risico op verkeerde verwoordingen wordt geminimaliseerd zodat wetgeving correct wordt vertaald.
RQ03	Ethisch en functioneel	Het gebruik van de AI-tool moet optioneel zijn voor de medewerker [23].	Medewerkers die brieven schrijven/controleren	Medewerkers: Medewerkers moeten de AI-tool kunnen aan- of uitzetten wanneer zij deze niet willen gebruiken. Medewerkers hebben namelijk tijdens een bijeenkomst gemeld, dat ze soms eerst zelf de brief willen schrijven, zonder meldingen van de AI-tool in beeld te krijgen (Bijlage A).

RQ-nr	Categorie	Requirement	Stakeholder	Toelichting
RQ04	Ethisch en functioneel	Om automation bias te voorkomen moet de AI-tool, de medewerker informeren dat de suggesties slechts ter ondersteuning dienen en dat de medewerker de eindbeslissing neemt [65].	Medewerkers die brieven schrijven/controleren	Medewerkers: Medewerkers die brieven schrijven of controleren nemen inhoudelijke en juridische beslissingen. Wanneer zij onvoldoende inzicht hebben in de beperkingen van de AI-tool, bestaat het risico dat AI-suggesties onkritisch worden overgenomen (automation bias). Dit kan leiden tot feitelijke onjuistheden of onjuiste formuleringen in brieven, omdat de AI-tool niet altijd de beste suggesties geeft.
RQ05	Juridisch en technisch	Persoonlijke informatie van burgers mag niet worden opgeslagen of gebruikt voor (verdere) ontwikkeling van het model, dus moet de gebruikte data voor ontwikkeling uit anonieme brieven bestaan [5, 20, 59].	Inwoners van OVER-gemeenten	Inwoners: De persoonlijke gegevens van inwoners mogen niet zonder toestemming worden verspreid.
RQ06	Juridisch en technisch	De AI-tool mag geen persoonsgegevens delen met externe partijen of servers [5, 57, 58] (Bijlage A).	Inwoners van OVER-gemeenten; OVER-gemeenten; ICT-afdeling; Overheid	Inwoners: Persoonlijke gegevens van inwoners mogen niet zonder toestemming worden verspreid. OVER-gemeenten: Zijn eindverantwoordelijk en mogen geen persoonlijke gegevens delen met externe partijen of servers. ICT-afdeling: De ICT-afdeling moet de tool onderhouden op de lokale server. Overheid: Moet controleren of deze requirement in de praktijk wordt nageleefd.
RQ07	Functioneel en technisch	De AI-tool moet beter presteren dan het baseline model.	OVER-gemeenten	OVER-gemeenten: OVER-gemeenten hebben als doel om schriftelijke communicatie begrijpelijk en toegankelijk te maken voor inwoners. De LiNT-II-score biedt een objectieve maat om leesbaarheid te beoordelen. Door te eisen dat herschreven brieven beter scoren dan de met het baseline-model herschreven versies, kan worden vastgesteld dat de AI-tool daadwerkelijk bijdraagt aan verbetering van de communicatie en toegevoegde waarde heeft ten opzichte van bestaande werkwijzen.
RQ08	Functioneel	Resultaten moeten betrouwbaar zijn voor alle verschillende briefcategorieën.	OVER-gemeenten	OVER-gemeenten: Hierdoor is de AI-tool bruikbaar voor alle brieven die geschreven worden binnen de OVER-gemeenten.

RQ-nr	Categorie	Requirement	Stakeholder	Toelichting
RQ09	Functioneel en juridisch	De ontwikkeling van de AI-tool (datasets, modelkeuze, training) en de prestaties moeten gedocumenteerd worden [23].	Medewerkers die brieven schrijven/controleren; Overheid	Medewerkers: Moet geïnformeerd worden over de prestaties van het model. Overheid: Wil inzicht in de documentatie om te controleren of het model voldoet aan wetgeving.
RQ10	Technisch	Bewonersinformatie wordt direct verwerkt en niet opgeslagen (Bijlage A).	Inwoners van de OVER-gemeenten; OVER-gemeenten	Inwoners: Gegevens van inwoners mogen niet opgeslagen worden zonder toestemming, voor de ontwikkeling van een AI-tool [20]. OVER-gemeenten: Mogen persoonlijke gegevens niet zomaar verwerken en opslaan [20].
RQ11	Technisch	De gebruikte datasets moeten gesplitst worden in een validatie- en testset van 50%, 50%.	ICT-afdeling	ICT-afdeling: Dit voorkomt data-leakage en zorgt voor correcte evaluatie.
RQ12	Technisch	De AI-tool moet in 2 seconden de brief controleren, fouten markeren en suggesties tonen.	Medewerkers die brieven schrijven/controleren	Medewerkers: Met deze tijdslimiet werkt de AI-tool aantoonbaar sneller dan een medewerker, zoals gewenst door leidinggevenden (Bijlage A). Zonder de AI-tool zou de medewerker handmatig de schrijfwijzer moeten raadplegen om te controleren of de brief aan de richtlijnen voldoet.

De requirements worden geëvalueerd aan de hand van vooraf vastgestelde acceptatiecriteria. Voor elke requirement uit Tabel 2 is in Tabel 3 de bijbehorende evaluatiemethode opgenomen (hoe deze requirement wordt beoordeeld) en het acceptatiecriteria dat aangeeft wanneer aan de requirement is voldaan.

Table 3. Acceptatiecriteria

RQ-nr	Evaluatiemethode	Acceptatiecriteria
RQ01	Interactief prototype	Gebruikerstesten uitvoeren waarbij wordt onderzocht of alle instructies duidelijk zijn (medewerkers weten welke knoppen waarvoor dienen) en of testers hierdoor het systeem correct kunnen gebruiken.
RQ02	Interactief prototype	Bij het ontwerpen van het AI-systeem moet de mens eindverantwoordelijk blijven voor de beslissing of een suggestie wordt goedgekeurd.
RQ03	Interactief prototype	Medewerkers hebben in het interactief prototype de mogelijkheid om het gebruik van de AI-tool vrijwillig aan of uit te schakelen.
RQ04	Interactief prototype	User Interface (UI) elementen tonen aan dat de AI-tool ondersteunend is en niet beslissend.

RQ-nr	Evaluatiemethode	Acceptatiecriteria
RQ05	Code en dataset controleren	Brieven die worden gebruikt voor het valideren en testen van het model moeten voor gebruik worden geanonimiseerd.
RQ06	Code controleren	Het opslaan en herschrijven van de brief met behulp van de AI-tool gebeurt lokaal.
RQ07	Evaluatie	In de evaluatie worden de LINT II-scores van de AI-tool en het baseline model tegenover elkaar gezet, zodat onderzocht kan worden of de AI-tool een lagere LiNT-II-score behaalt dan het baseline model.
RQ08	Evaluatie	De gemiddelde LiNT-II-score moet voor elke briefcategorie gelijk zijn.
RQ09	Documentatie	In de documentatie worden de ontwikkeling van het AI-systeem en de nauwkeurigheid van de suggesties weergegeven.
RQ10	Code controleren	In de code staat dat persoonlijke informatie verwijderd wordt direct na het opslaan van de brief.
RQ11	Code controleren	In de code staat een datasplitsing volgens een validatie- en test-set van 50% en 50%.
RQ12	Code testen	De code geeft binnen 2 seconden alle suggesties voor een volledig herschreven brief.

4 Methodologie

4.1 Concept

4.1.1 Waarde proposities.

Het voorgestelde systeem ondersteunt medewerkers van de OVER-gemeenten bij het herschrijven van brieven naar begrijpelijke taal volgens de regels van de OVER-schrijfwijzer. Door suggesties te geven besparen medewerkers tijd en hoeven zij minder taalkundige kennis te hebben om consistent teksten te schrijven volgens richtlijnen van de OVER-schrijfwijzer. Meer inwoners zouden hierdoor de brieven beter kunnen begrijpen, waardoor misverstanden worden verminderd. De OVER-gemeenten verbetert daarmee de kwaliteit van haar dienstverlening. Aangezien de tool lokaal draait, blijven privacy en veiligheidsvereisten gewaarborgd. De volledige waardeproposities staan in Bijlage B.

4.1.2 Flow diagram en gebruikte design patterns.

In Paragraaf 2.3 is de huidige werkwijze voor het opstellen van brieven door medewerkers van de gemeente beschreven. In dit onderzoek wordt een AI-tool toegevoegd die medewerkers ondersteunt bij het herschrijven van teksten naar de richtlijnen van de OVER-gemeenten (Figuur 3). De AI-tool wordt geïmplementeerd als een Word-extensie, zodat medewerkers direct binnen het programma waarmee zij brieven opstellen gebruik kunnen maken van de suggesties (Bijlage A). Het model beoordeelt of zinnen eenvoudig genoeg zijn of te complex, en genereert vervolgens suggesties voor het herschrijven van tekst met een verwijzing naar de bijbehorende regel in de schrijfwijzer. Deze suggesties worden aan de medewerker getoond, die deze naar eigen inzicht kan toepassen. Daarna verloopt het verdere proces op dezelfde manier als in de huidige situatie.

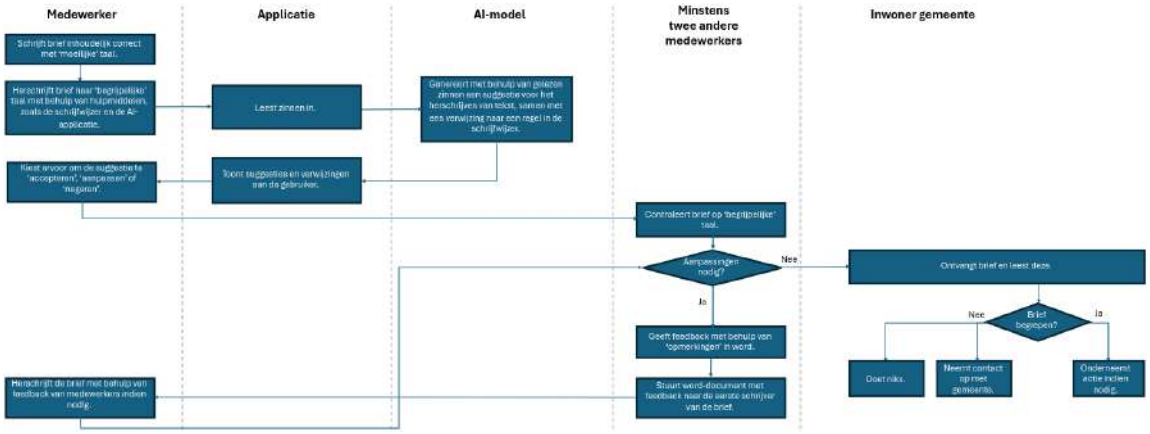


Fig. 3. Flow-diagram van de voorgestelde situatie.

Om inzicht te geven in de mate van automatisering binnen de toegevoegde stappen in het flowdiagram, is per stap van de AI-tool het bijbehorende Level of Automation (LoA) bepaald. Daarbij is gekeken naar de mate waarin de AI-tool zelfstandig handelt en in hoeverre de medewerker de inhoudelijke controle over de brief behoudt.

Vanwege de gestelde requirement RQ02, is volledige automatisering (LoA 6-10) niet wenselijk. De schrijfassistent is daarom bewust gepositioneerd op een laag tot beperkt middelmatig automatiseringsniveau (LoA 1–3). Levels 4 en 5 zijn niet van toepassing, omdat de tool meer dan één of twee mogelijke suggesties kan genereren waaruit de medewerker kan kiezen. Het hoogste automatiseringsniveau binnen het systeem is LoA 3. De AI-tool ondersteunt, maar neemt geen beslissingen over het herschrijven van de brief. Tabel 4 toont per stap welk automatiseringsniveau is toegepast en waarom.

Table 4. Level of Automation per stap in de AI-tool

Stap	Beschrijving	LoA	Rol van de AI-tool / Rol van de medewerker
1	Tekst schrijven (menselijk).	1	**Medewerker:** schrijft de inhoud zoals in de huidige situatie in Figuur 2, zonder gebruik van AI.
2	AI-tool analyseert zinnen.	2	**AI-tool:** markeert moeilijke woorden, lange zinnen, passieve vorm en toonafwijkingen, etc. op basis van de schrijfwijzer.
3	AI-tool geeft herschrijfsuggesties en legt uit waarom.	3	**AI-tool:** genereert één of meerdere herschrijfsuggesties en legt uit waarom iets moeilijk is.
4	Medewerker ontvangt suggestie(s) en kiest ervoor om deze te accepteren, aanpassen of negeren.	3	**AI-tool:** past nooit automatisch wijzigingen toe. elke verandering vereist expliciete actie. **Medewerker:** kiest tussen *Accepteren*, *Aanpassen* of *Negeren*.

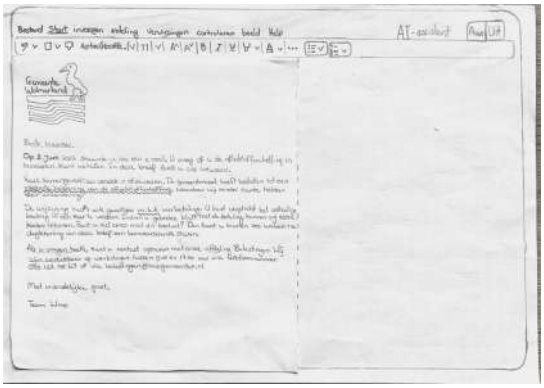
Tijdens het ontwerpproces zijn design patterns gebruikt om transparantie en controle binnen het systeem te vergroten:

- "Geef de controle terug aan de gebruiker wanneer de automatisering faalt" [40], door de medewerker de mogelijkheid te geven een suggesties te accepteren, aan te passen of te weigeren als deze de tekst niet herschreven is naar de OVER-schrijfwijzer-regels.
- "Leg uit voor begrip, niet voor volledigheid" [40], door met een verwijzing naar de schrijfwijzer toe te lichten waarom een tekst als ‘niet volgens richtlijnen van de OVER-schrijfwijzer’ wordt beschouwd.

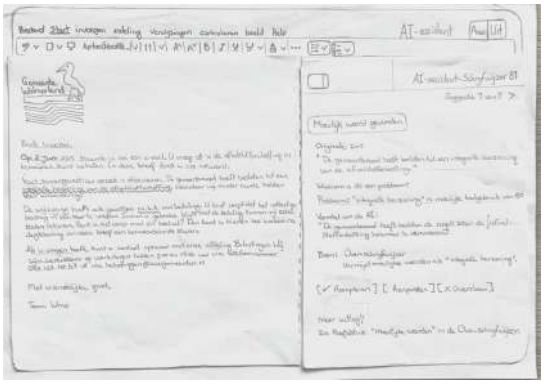
4.1.3 *Testopzet en evaluatiemethoden voor concept- en gebruikerstests.*

Er is een paper prototype ontwikkeld om medewerkers van de OVER-gemeenten de AI-tool visueel te laten beoordelen. Deze staat weergeven in Figuur 4. Daarnaast is een testplan opgesteld (Bijlage D), waarvan de vragen zijn voorgelegd aan medewerkers van de OVER-gemeenten. De vragen richtten zich op wenselijkheid, vertrouwen en uitlegbaarheid en werden beoordeeld vanuit het perspectief van de medewerker die de brief schrijft. Enkele vragen uit het testplan staan hieronder:

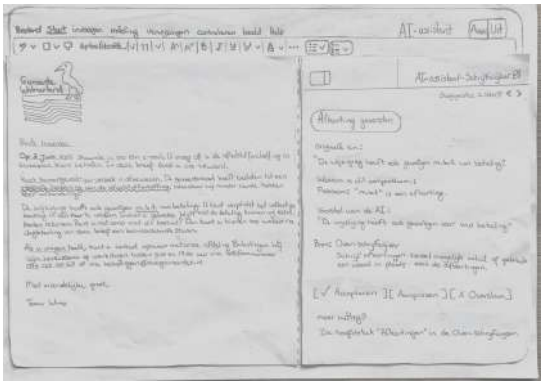
- Wenselijkheid: Zijn er andere stappen in het prototype wenselijk die er op dit moment nog niet instaan?
- Uitlegbaarheid: Denk je dat jij (en je collega's) de AI-tool snel kunt leren gebruiken?
- Vertrouwen: Zou je te veel beïnvloed worden door de suggesties die de AI-tool geeft, en zou je de AI-tool daardoor te veel vertrouwen?



(a) Paper prototype slide 1



(b) Paper prototype slide 2



(c) Paper prototype slide 3

Fig. 4. Paper prototype slides

Op basis van de verkregen feedback zijn verbeteringen aangebracht aan het prototype, dat vervolgens als Figma-prototype is uitgevoerd. Voor de eerste iteraties was een paper prototype handig, omdat dit snel experimenteren en itereren mogelijk maakte. Het tweede prototype in Figma maakte het mogelijk om verbeteringen concreet te visualiseren en de gebruikerservaring nauwkeuriger

te evalueren voordat verdere implementatie plaatsvindt. De onderbouwing voor de aanpassingen wordt toegelicht in Paragraaf 6.2. Dit Figma-prototype is opnieuw getest met een nieuw testplan en verschillende Communicatie en Multimedia Design-experts om de relevantie en gebruiksvriendelijkheid te beoordelen (Bijlage H). De AI-tool is beoordeeld vanuit het perspectief van een OVER-medewerker die brieven opstelt, aangevuld met hun expertise in het ontwikkelen en ontwerpen van digitale tools. Enkele vragen uit dit testplan staan hieronder:

- Relevantie: Zijn de meldingen/knoppen die het systeem geeft relevant en niet storend?
- Gebruiksvriendelijkheid: Is het voor u duidelijk wat er van u verwacht wordt wanneer een suggestie verschijnt (onderstreping + suggestie)?

De uitkomsten van deze test hebben geleid tot verdere verbeteringen aan het prototype, die worden onderbouwd in Paragraaf 6.3.

4.2 Data

Dit onderzoek maakt gebruik van drie typen data: de OVER-schrijfwijzer, externe woordenlijsten en bestaande brieven van de OVER-gemeenten. Samen vormen deze datasets de basis voor de ontwikkeling van de AI-tool, die gebruikmaakt van een tekstverwerkingsmodel om teksten te vereenvoudigen.

4.2.1 OVER-schrijfwijzer.

De OVER-schrijfwijzer bevat richtlijnen voor duidelijke gemeentelijke communicatie op B1-niveau, met nadruk op korte en concrete zinnen, één kernboodschap per alinea en een logische opbouw. Medewerkers schrijven vanuit het perspectief van de inwoner, vermijden jargon en formele taal, en hanteren een actieve, positieve en adviserende toon. Daarnaast beschrijft de schrijfwijzer regels voor inclusief taalgebruik, consequent gebruik van de u-vorm en vaste notaties voor onder andere data, bedragen en adressen. De schrijfwijzer vormt de basis van de AI-tool: alle suggesties moeten hiermee in lijn zijn. Een beperking is dat de schrijfwijzer weinig concrete versimpelde synoniemen bevat voor moeilijke woorden, waardoor aanvullende externe data nodig is voor lexicale vereenvoudiging.

4.2.2 Externe data.

Ter aanvulling zijn vijf externe woordenlijsten gevonden die ondersteuning bieden bij het herkennen en vereenvoudigen van moeilijke woorden. Hierbij zijn twee woordenlijsten tot samengevoegd tot één woordenlijst, omdat beide woordenlijsten versimpelde synoniemen bevatten. Een overzicht van deze woordenlijsten, inclusief hun kenmerken, voor- en nadelen, is opgenomen in Tabel 5. Deze data dienen als aanvulling op de schrijfwijzer.

Table 5. Overzicht van gevonden Nederlandse woordenlijsten met uitleg, voor- en nadelen.

Naam dataset / bron	**Uitleg**	**Aantal woorden**	**Datatype**	**Voordeel**	**Nadeel**
Moeilijk-woordenboek van de Gemeente Amsterdam [19]	Synoniemenlijst voor 'moeilijke' Nederlandse woorden.	747	String	Woorden zijn afgestemd op B1-niveau.	Geen zinnen opgenomen, waardoor geen zinsverbeteringen geleerd kunnen worden.
Helder juridisch woordenboek van de Gemeente Amsterdam [17]	Lijst met juridische termen en alternatieven op B1-niveau.	134	String	Vermijd juridisch jargon en stimuleert B1-niveau.	Geen zinnen opgenomen, waardoor geen zinsverbeteringen mogelijk.
Inclusieve woordenlijst van de Gemeente Amsterdam [18]	Lijst met termen en bijbehorende inclusieve alternatieven.	43	String	Bevordert inclusief taalgebruik.	Termen kunnen verouderen.
Synoniemen woordenlijst van B1-teksten [1], Dijk en Waard [7]	Lijst met versimpelde synoniemen.	17	String	Gericht op B1-niveau en ondersteuning van eenvoudiger taalgebruik.	Beperkte lijst en geen zinnen voor zinsverbetering.

De OVER-gemeenten hebben naast de OVER-schrijfwijzer bestaande brieven aangeleverd die de huidige communicatiesituatie representeren. Omdat deze nog niet zijn herschreven volgens de schrijfwijzer, worden ze niet gebruikt als trainingsdata. Wel worden ze ingezet als validatie- en testdata om de werking van de AI-tool in een realistische context te evalueren.

4.2.3 *Data-preparatie, splitsing en vector database.*

De woordenlijsten waren niet beschikbaar als kant-en-klare datasets en zijn via webscraping verzameld, wat resulteerde in vier gestructureerde datasets (Tabel 5). De gemeentelijke brieven zijn handmatig gegroepeerd op basis van bestandsnamen. Hierdoor ontstaan categorieën, zoals WMO en Participatiewet (Tabel 6). Vervolgens zijn de brieven met een stratified split [30] gesplitst. Door gebruik te maken van een stratified split is elke briefcategorie evenredig vertegenwoordigd, waardoor de briefcategorieën eerlijk verdeeld zijn in de datasets [30]. Vaak wordt bijvoorbeeld een 80% train-, 10% validatie- en 10% testset gebruikt, maar omdat in dit onderzoek geen gebruik wordt gemaakt van een trainset is ervoor gekozen om de brieven ook gelijk te verdelen verdeeld in 50% validatie- en 50% testset [8].

Table 6. Overzicht van de categorieën waarin de brieven verdeeld zijn.

Categorie	**Uitleg**	**Aantal brieven**
Schuldhulp	Brieven die betrekking hebben tot schuldhulp vanuit de OVER-gemeenten.	4
WMO	Brieven die betrekking hebben tot WMO en zorgvoorzieningen vanuit de OVER-gemeenten.	13
Participatiewet (PW)	Brieven die betrekking hebben tot PW en uitkeringszaken.	13
Participatiefonds (PF)	Brieven die betrekking hebben tot participatiefondsen.	4
Overig	Brieven die niet vallen binnen de bovenste vier categorieën.	2

Figuur 5 toont de verdeling van de briefcategorieën in de validatieset; brieven over de Participatiewet en de Wmo komen relatief vaak voor en worden daardoor vaker meegenomen in de evaluatie. Dit betekent dat de resultaten mogelijk niet volledig generaliseerbaar zijn naar alle briefcategorieën binnen de OVER-gemeenten.

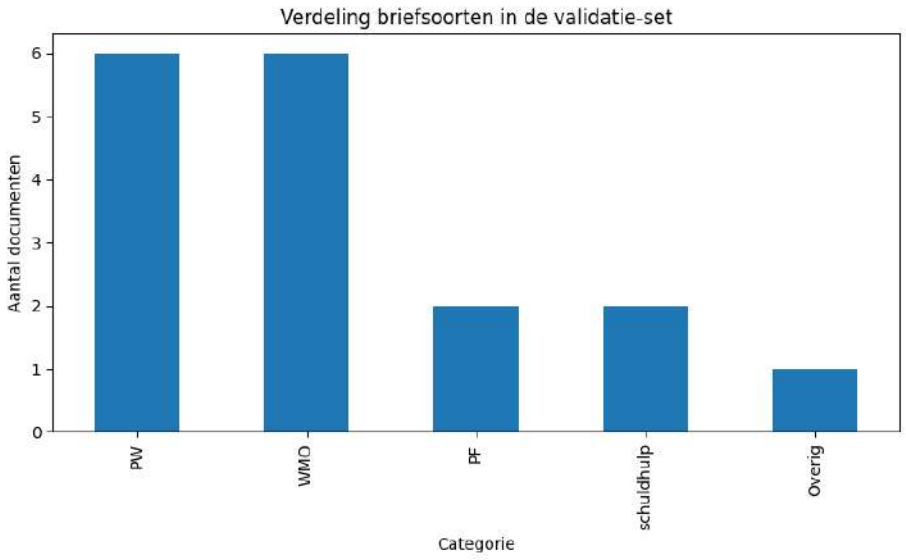


Fig. 5. Histogram van de briefcategorieën.

Om de schrijfwijzer als context beschikbaar te maken voor het taalmodel, wordt gebruikgemaakt van een vector database. Tekst wordt hierin opgeslagen als vectorrepresentaties, waardoor semantisch relevante regels kunnen worden opgehaald [13, 37, 53]. Overwogen opties van vector databases voor dit onderzoek zijn ChromaDB, LanceDB en Qdrant [3, 13, 14]. In de context van deze opdracht is het van belang dat het model lokaal kan draaien en aansluit bij de privacy- en AVG-eisen van de OVER-gemeenten, vandaar dat Qdrant minder geschikt is voor dit onderzoek (Tabel 7). Aangezien LanceDB meer filtermogelijkheden heeft dan ChromaDB is daarom gekozen voor LanceDB (Tabel 7).

Table 7. Vergelijking van vector databases

	ChromaDB [3]	**LanceDB [13]**	**Qdrant [14]**
Embedded/client-server	Hybride	Embedded	Client-server
Open-source	Ja (Apache 2.0)	Ja (Apache 2.0)	Ja Apache (2.0)
Metadata filtering	Filteren op basis van metadata en document inhoud.	Gebruikt SQL-strings en integreert met Pandas.	Geografische, tekst, en numerieke filters die in de index zijn verwerkt.

Voor opslag van de OVER-schrijfwijzer in de vector database is een chunking-strategie toegepast (Tabel 8) [13, 37, 53]. Eenvoudige methoden, zoals opsplitsen per zin of vaste lengte, zijn ongeschikt omdat richtlijnen vaak uit meerdere samenhangende zinnen en subsecties bestaan. Daarnaast is de semantische methode ongeschikt, omdat het zinnen die semantisch vergelijkbaar zijn dicht bij elkaar zet in een embedding en er hierdoor context verloren gaat. Daarom is gekozen voor een gespecialiseerde methode, “opsplitsen op structuur” (Tabel 8). Hierbij is de schrijfwijzer eerst handmatig omgezet naar het Markdown-formaat. Vervolgens is deze gebruikt om de OVER-schrijfwijzer in de vector database te plaatsen. Structureel samenhangende regels worden hierbij gezamenlijk opgeslagen, zodat voldoende context behouden blijft.

Table 8. Strategieën voor text chunking

	Strategie	**Voordelen**	**Nadelen**
Klassieke methoden	Opsplitsen op basis van een ingestelde lengte (bijvoorbeeld elke 100 karakters).	Eenvoudig te implementeren; voorspelbare chunk-grootte.	Contextverlies; kan zinnen of woorden middenin breken en negeert documentstructuur.
	Opsplitsen op zinnen.	Behoudt grammaticale structuur; geen afgebroken woorden.	Gebrek aan context en zinnen variëren in lengte.
	Opsplitsen op basis van een ingestelde lengte, met een overlap, meestal rond de 10/20%.	Vermindert contextverlies op de grenzen.	Redundantie in zoekresultaten (dubbele info); hogere opslag- en indexeringskosten.
Gespecialiseerde methoden	Opsplitsen op structuur (HTML, Markdown, of LaTeX)	Behoudt semantische hiërarchie (secties, regels).	Afhankelijk van bronkwaliteit, want slechte conversie leidt tot slechte chunks.
	Semantisch	Groepeert inhoudelijk gerelateerde zinnen; optimale contextbehoud.	Rekenintensief met onvoorspelbare grenzen. Gevoelig voor hyperparameters.

4.3 Model

4.3.1 Modelarchitectuur.

In de eerste iteratie is gekozen voor een naïeve-RAG-architectuur [50] (Figuren 6 en 7). Een RAG-architectuur is een benadering waarbij het genereren van output actief externe, relevante informatie ophaalt uit een kennisbron en deze context gebruikt om nauwkeurigere en beter onderbouwde antwoorden te produceren [52]. Hierbij wordt een vectordatabase gebruikt om aanvullende context aan het taalmodel mee te geven. Geavanceerdere RAG-varianten zijn in de eerste iteratie bewust buiten beschouwing gelaten om een eenvoudig en uitbreidbaar uitgangspunt te behouden. De vectordatabase van het model is gevuld met richtlijnen uit de OVER-schrijfwijzer. Relevante passages worden opgehaald en toegevoegd aan een prompt, die vervolgens wordt ingediend bij een taalmodel dat runt binnen Ollama (software om taalmodellen te runnen). Zo kan het taalmodel deze richtlijnen meenemen bij het herschrijven van teksten.

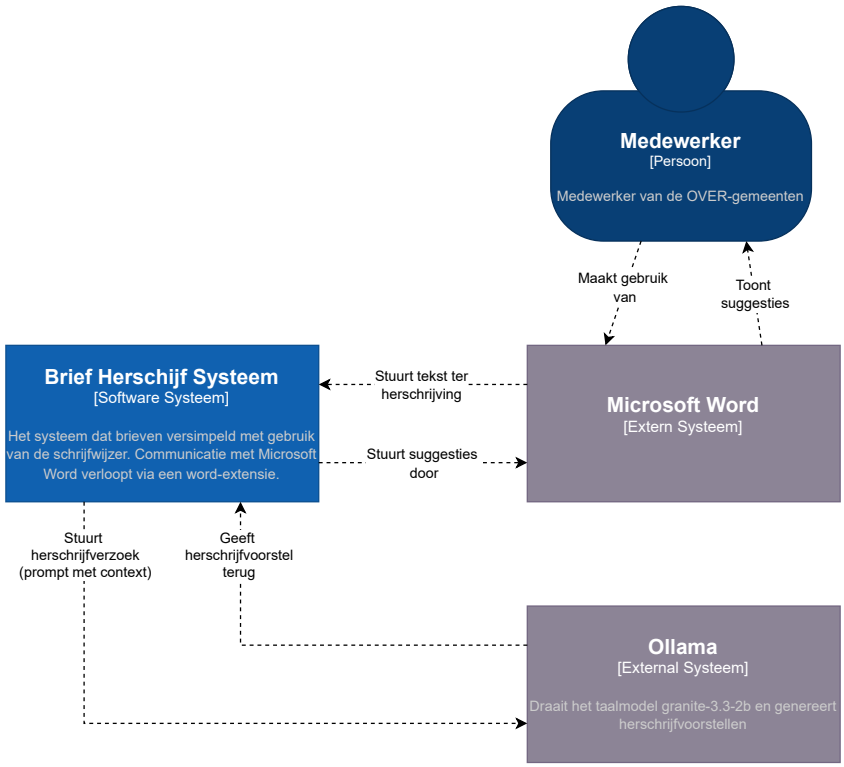


Fig. 6. Modelarchitectuur iteratie 1 (high-level)

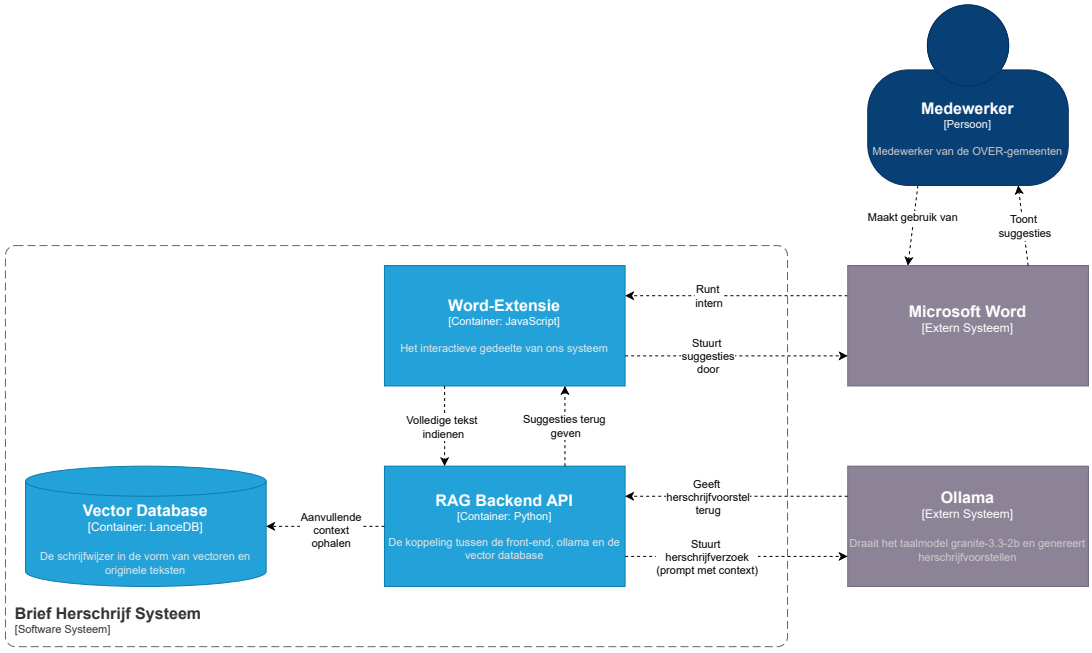


Fig. 7. Modelarchitectuur iteratie 1 (technischer)

4.3.2 Het gebruikte LLM.

Voorafgaand aan de selectie van het specifieke taalmodel is literatuuronderzoek uitgevoerd naar geschikte modelarchitecturen voor Nederlandstalige tekstvereenvoudiging in een gemeentelijke context. Daarbij is onderscheid gemaakt tussen encoder-only, decoder-only en encoder-decoder modellen, en is de inzet van een naïve-RAG-architectuur overwogen. Op basis van deze analyse is gekozen voor een decoder-only architectuur, waarbij de Leesplank Noot-modellen als meest geschikte kandidaten naar voren kwamen vanwege hun geschiktheid voor de Nederlandse taal en gemeentelijke communicatie. De volledige onderbouwing van deze globale modelkeuze, inclusief de afgevallen alternatieven, is opgenomen in Bijlage J.

Voor de naïve-RAG-architectuur is onderzocht welk LLM het meest geschikt is voor Nederlandstalige tekstvereenvoudiging in een gemeentelijke context. Het GEITje-model viel aanvankelijk op door zijn Nederlandstalige focus en uitgebreide training, maar is niet meer beschikbaar vanwege auteursrechtelijke bezwaren [56] en is daarom uitgesloten. Alternatief zijn de Leesplank Noot-modellen, ontwikkeld voor Nederlandse tekstvereenvoudiging richting B1-niveau en expliciet bedoeld voor publieke communicatie. Deze modellen voldoen aan eisen rondom transparantie, controleerbaarheid en veilige inzet binnen overheden. Een overzicht van de beschikbare Leesplank Noot-modellen is opgenomen in Tabel 9. Het bevat een set van drie Nederlandstalige LLM's, elk gefinetuned op een samengestelde Nederlandstalige dataset voor tekstvereenvoudiging naar B1-niveau. De modellen zijn ontwikkeld voor overheids- en publieke communicatie en sluiten aan bij de EU-AI-Act: transparant, herleidbaar, controleerbaar en veilig inzetbaar in gemeentelijke communicatie. In Tabel 9 worden de drie modellen weergegeven. Op basis van Tabel 9 wordt één van deze drie modellen geselecteerd voor inzet als LLM binnen de naïve-RAG-architectuur.

Table 9. Overzicht van verschillende LLM's met bron, aantal parameters, voor- en nadelen

LLM-model en bron	Voordelen	Nadelen	Aantal parameters
Granite-3.3-2b [10]	Biedt de hoogste kwaliteit volgens [6], levert consistente en betrouwbare output en blijft dankzij de compacte omvang efficiënt inzetbaar op gangbare hardware.	Het model is trager dan de alternatieven en vraagt meer rekenkracht dan het lichtste model, waardoor het minder geschikt is voor situaties waar snelheid de hoogste prioriteit heeft.	2 miljard
Llama-3.2-3b [11]	Combineert een goede outputkwaliteit met hogere snelheid dan Granite en profiteert van extra modelcapaciteit die het breed inzetbaar maakt.	De grotere modelgrootte zorgt voor hogere hardware-kosten, terwijl deze extra omvang geen duidelijke kwaliteitswinst oplevert ten opzichte van Granite.	3 miljard
EuroLLM-1.7b [9]	Werkt snel, is lichtgewicht en daarmee geschikt voor toepassingen met beperkte hardware.	De lagere kwaliteit en minder stabiele verwerking van complexe context maken het minder betrouwbaar voor taken waar nauwkeurigheid belangrijk is.	1.7 miljard

Voor de naïve-RAG-architectuur wordt gekozen voor het Granite-3.3-2B-model (Tabel 9). Volgens [6] behaalt dit model de hoogste kwaliteitsscore van de drie modellen bij het vereenvoudigen van gemeentelijke teksten naar B1-niveau. Het Granite-3.3-2B-model weet teksten beter te structureren en inhoudelijk nauwkeuriger weer te geven, waarbij het model qua omvang compact genoeg blijft om efficiënt te gebruiken. De andere modellen bieden voordelen, maar sluiten minder goed aan bij de doelstelling. Llama-3.2-3B is groter, maar levert ondanks het hogere aantal parameters waarschijnlijk geen kwaliteitsverbetering. Daardoor wegen de extra rekenkosten niet op tegen de beperkte meerwaarde [11]. EuroLLM-1.7B is het snelste model, wat aantrekkelijk kan zijn voor taken die snel uitgevoerd dienen te worden. Volgens [6] behaald EuroLLM-1.7B lagere scores, wat de betrouwbaarheid van het model kan verminderen [9]. Daarom is het Granite-3.3-2B-model de meest geschikte keuze voor implementatie in het naïve-RAG-model, vanwege de hoogste kwaliteit en stabiele verwerking van teksten.

4.3.3 Baseline model.

Als baseline om het ontwikkelde naïve-RAG-model mee te vergelijken is gekozen voor een LLM dat teksten herschrijft naar eenvoudiger Nederlands op basis van een vaste prompt. Deze baseline wordt gebruikt als referentiepunt voor latere iteraties, waarin aanvullende context (zoals de OVER-schrijfwijzer en externe woordenlijsten) worden toegevoegd. Voor het baseline-model is gebruikgemaakt van hetzelfde model als in de naïve-RAG-architectuur (Paragraaf 4.3.1). De eerder geprepareerde validatieset (brieven van de OVER-gemeenten) is ingeladen en per zin herschreven met behulp van een prompt. Deze prompt is bewust kort gehouden om de prestaties van een generiek LLM te meten, zonder invloed van aanvullende richtlijnen of regels:

Je bent een assistent die teksten herschrijft naar eenvoudiger Nederlands (B1-niveau).

Herschrijf de volgende tekst:

Prompt

Geef uitsluitend de herschreven tekst, zonder toelichting.

Door deze aanpak wordt inzicht verkregen in de mate waarin een LLM, zonder domeinspecifieke kennis of expliciete schrijfregels, in staat is om gemeentelijke teksten te vereenvoudigen. De baseline vormt daarmee het uitgangspunt voor vergelijking met vervolgmodellen waarin de OVER-schrijfwijzer en human-in-the-loop-principes worden geïntegreerd.

4.3.4 Evaluatiemetrics voor het model.

De prestaties van het taalmodel zullen geëvalueerd worden met verschillende kwaliteitscriteria. Eerst is bepaald welke prestatiemaat wordt gebruikt om iteraties te vergelijken. Voor dit onderzoek is een prestatiemaat (performance metric) nodig die specifiek inzicht geeft in de begrijpelijkheid van herschreven teksten. Het doel is om vast te stellen of gemeentelijke brieven na vereenvoudiging daadwerkelijk leesbaar zijn en voldoen aan de regels in de OVER-schrijfwijzer. Daarom worden in dit onderzoek meerdere prestatiematen overwogen, waaronder de tien meest gebruikte metrics voor het beoordelen van versimpelde teksten uit [49], aangevuld met de LiNT-II-score en FKBLEU, om de eenvoud van gemeentelijke brieven te beoordelen. Tabel 10 geeft een overzicht van de overwogen prestatiematen, inclusief een korte toelichting op hun relevantie voor dit onderzoek.

Table 10. Overzicht van overwogen prestatiematen

Prestatiemaat	**Beschrijving**
SARI	SARI is een veelgebruikte prestatiemaat voor tekstversimpeling die zich richt op de manier waarop een vereenvoudigde tekst afwijkt van het origineel [66]. Omdat SARI altijd een originele zin en een corresponderende vereenvoudigde versie nodig heeft om een score te berekenen, kan de metric niet worden toegepast op enkel de originele gemeentelijke brieven. Bovendien meet SARI vooral de kwaliteit van de vereenvoudiging zelf en niet de feitelijke leesbaarheid voor eindgebruikers, waardoor de metric minder geschikt is voor dit onderzoek.
FKGL	De Flesch-Kincaid Grade Level (FKGL) meet leesbaarheid en complexiteit op basis van woord- en zinslengte. Omdat deze metriek is ontwikkeld voor Engelstalige teksten [45], is de toepasbaarheid op Nederlandstalige teksten beperkt. Hierdoor is deze prestatiemaat minder geschikt voor dit onderzoek.
BLEU	BLEU is ontworpen voor het evalueren van machinevertalingen en meet n-gram-overlap met referentieteksten. Het richt zich op vloeiendheid en behoud van inhoud in plaats van eenvoud, wat van belang is voor dit onderzoek [49]. Vandaar dat deze prestatiemaat minder geschikt is voor dit onderzoek.
FKBLEU	FKBLEU combineert een maat voor parafrasekwaliteit (iBLEU) met de leesbaarheidsindex Flesch-Kincaid Grade Level (FKGL) [66]. De metriek probeert zowel betekenisbehoud als vereenvoudiging te meten. Omdat de FKGL is ontwikkeld voor Engelstalige teksten [45], is de toepasbaarheid van FKBLEU op Nederlandstalige teksten ook beperkt. Hierdoor is deze prestatiemaat minder geschikt voor dit onderzoek.
Precision	Met behulp van precision kan je meten hoe beknopt een tekst is en hoeveel overbodige woorden aanwezig zijn [21]. Deze metriek is daardoor vooral geschikt voor samenvattingstaken en minder geschikt voor het beoordelen van de eenvoud en leesbaarheid van gemeentelijke brieven. Vandaar dat deze niet gebruikt zal worden in dit onderzoek.
Recall	Met Recall kan gemeten worden in welke mate de inhoud van de referentietekst behouden blijft [21]. Het meet dus niet of tekst eenvoudiger of beter leesbaar is. Deze metriek is daardoor vooral geschikt voor het beoordelen van samenvattingstaken en minder geschikt voor dit onderzoek.
F1-Score	De F1-score is het harmonisch gemiddelde van precision en recall. Aangezien zowel precision als recall minder geschikt zijn voor het beoordelen van eenvoud en leesbaarheid, wordt de F1-score ook niet gebruikt voor dit onderzoek.
BERT-score	De BERT-score meet vooral of de betekenis behouden blijft tussen een referentietekst en een gegenereerde tekst [49]. Voor het berekenen van de BERT-score is daarom altijd een herschreven tekst nodig [67]. Omdat er geen standaardherschrijvingen van de originele brieven beschikbaar zijn, kan de BERT-score niet toegepast worden op de brieven zelf. Daarom is deze metriek minder geschikt voor dit onderzoek.
Rouge	Rouge meet hoeveel woorden of woordgroepen de gegenereerde tekst gemeen heeft met de originele tekst [21]. Voor samenvattingen is dit nuttig, maar voor tekstvereenvoudiging minder geschikt. Bij vereenvoudigen willen we dat zinnen makkelijker worden en de betekenis behouden blijft, zonder dat exacte woorden hetzelfde zijn. Daarom geeft Rouge niet goed weer of een tekst leesbaarder of begrijpelijker is en wordt het in dit onderzoek niet gebruikt.
FRES-score	De Flesch Reading Ease Score (FRES) is een leesbaarheidsmetriek die tekstcomplexiteit inschat op basis van zinslengte en het aantal lettergrepen per woord [32]. De metriek is ontwikkeld en afgestemd voor Engelstalige teksten en niet formeel gevalideerd voor het beoordelen van Nederlandstalige (juridisch-administratieve) communicatie. Om deze redenen is FRES niet opgenomen als evaluatiemetric voor dit onderzoek.
Accuracy	Accuracy geeft het aandeel correct geclassificeerde uitkomsten weer. Aangezien tekstvereenvoudiging een generatieve taak is zonder eenduidige 'juiste' uitkomst, is deze metriek niet geschikt voor het beoordelen van herschreven brieven.
LiNT-II	LiNT-II is een leesbaarheidsinstrument dat specifiek is ontwikkeld voor Nederlandstalige teksten [44]. De metric bepaalt de moeilijkheidsgraad van een tekst op basis van vier taalkundige kenmerken: woordfrequentie, syntactische afhankelijkheidslengte, het aantal inhoudswoorden per zin en het aandeel concrete zelfstandige naamwoorden [44]. Omdat LiNT-II is afgestemd op Nederlandstalige teksten en empirisch is gevalideerd met beoordelingen door docenten en scholieren [44, 51], is deze metric geschikt om te beoordelen in hoeverre teksten voldoen aan de richtlijnen.

Op basis van de overwegingen in Tabel 10 is LiNT-II gekozen als primaire prestatiemaat, omdat deze het meest geschikt is om de leesbaarheid van Nederlandstalige gemeentelijke teksten te beoordelen. De gemiddelde LiNT-II-score van de herschreven brieven in de validatie-set wordt berekend op een schaal van 0 tot 100. Op een aparte dataset is gekeken naar de samenhang tussen LiNT-II en het CEFR-leesniveau (Bijlage K). Voor B1-teksten lag 50% van de LiNT-II-scores tussen 36.2 en 50.1, wat een indicatieve richtlijn vormt voor gemeentelijke brieven. Het is echter belangrijk te vermelden dat CEFR-annotaties niet volledig betrouwbaar zijn. Sommige teksten in de aparte dataset kregen namelijk uiteenlopende niveaus toegewezen, soms met één of twee niveaus verschil [25], waardoor dit bereik slechts als indicatieve richtlijn kan worden beschouwd.

De LiNT-II-score wordt niet aan gebruikers van de AI-tool getoond, maar dient als intern instrument om de leesbaarheid objectief te kwantificeren. Het doel van de tool is medewerkers te ondersteunen bij het schrijven van duidelijke brieven, zonder hen te belasten met numerieke scores, die hun oordeel of schrijfgedrag zou kunnen beïnvloeden. Door de score achter de schermen te gebruiken, kan het model verbeterd worden zonder de gebruiker te overladen met technische details of complexiteit die voor praktische toepassing niet relevant zijn.

Naast LiNT-II wordt alleen het best presterende model beoordeeld op robuustheid, uitlegbaarheid, modelcomplexiteit en resource demand (Tabel 11). Het model wordt niet beoordeeld op schaalbaarheid, waarmee wordt bedoeld in hoeverre de AI-tool inzetbaar is bij andere gemeenten om ook daar brieven volgens de richtlijnen van de OVER-schrijfwijzer leesbaar te maken. Dit is in dit onderzoek niet mogelijk, omdat uitsluitend brieven van de OVER-gemeenten beschikbaar zijn en de AI-tool daardoor niet kan worden getest op brieven van andere gemeenten.

Table 11. Overzicht van aanvullende kwaliteitscriteria

Kwaliteitscriterium	**Evaluatiemethode**
Robuustheid	Er wordt onderzocht of de LiNT-II-score van de verbeterde brieven consistent is over verschillende briefcategorieën (Schuldhulp, WMO, Participatiewet (PW), Participatiefonds (PF) en Overig). Dit toetst of het model niet alleen voor één type tekst goed presteert (RQ08). De resultaten worden weergegeven als gemiddelde LiNT-II-scores per briefcategorie. Vervolgens wordt gecontroleerd of deze scores per categorie gelijk zijn.
Uitlegbaarheid	Geëvalueerd wordt of de UI-elementen duidelijk maken dat de tool ondersteunend is, fouten kan maken en dat de medewerker de eindbeslissing neemt (RQ01, RQ02, RQ04). Dit wordt getest met het prototype en met testpersonen.
Modelcomplexiteit	Aangezien de AI-tool geen persoonsgegevens mag delen met externe partijen of servers (RQ06), en hierdoor de tool lokaal gedraaid moet worden, wordt de gebruikte LLM-architectuur (Granite-3.3-2B in de naïeve RAG beschreven. Dit gebeurd in termen van het aantal parameters en de modelopbouw, om de 'zwaarte' van het model voor de gemeentelijke infrastructuur te beoordelen, aangezien het model daarom niet te computationeel belastend nag zijn. De complexiteit wordt daarom vergeleken met die van andere tekstversimpelingsmodellen.
Resource demand	Per zin wordt de verwerkingstijd gemeten. De AI-tool moet binnen een acceptabele tijd van ongeveer twee seconden een suggestie kunnen genereren, zodat de tool in de praktijk bruikbaar is zonder de werkstroom te vertragen (RQ12).

4.3.5 *Iteraties en systematische aanpak.*

Voor het ontwikkelen van het LLM-model werden meerdere iteraties uitgevoerd. Eerst werd een baseline-model ontwikkeld, gebaseerd op de naïve-RAG-architectuur (Paragraaf 4.3.3). Vervolgens werd een eerste iteratie uitgevoerd, waarbij een vector database werd toegevoegd voor het ophalen van onderdelen uit de OVER-schrijfwijzer (Figuur 7). In deze eerste iteratie werden alle zinnen aangepast, ook wanneer de aangepaste zin een slechtere LiNT-II-score opleverde dan de originele zin. Om dit te voorkomen, werd in de tweede iteratie een if-statement toegevoegd, zodat suggesties voor zinnen alleen werden weergegeven wanneer deze een lagere LiNT-II-score opleverden (Figuur 8). In een derde iteratie (Figuur 9) werd het model uit iteratie 2 gebruikt en onderzocht of het model beter presteerde wanneer de volgende woordenlijsten werden meegenomen bij het ontwikkelen van het model:

- Moeilijk-woordenboek van de Gemeente Amsterdam [19]
- Helder juridisch woordenboek van de Gemeente Amsterdam [17]
- Inclusieve woordenlijst van de Gemeente Amsterdam [18]
- Synoniemenwoordenlijsten van B1-teksten [1, 7]

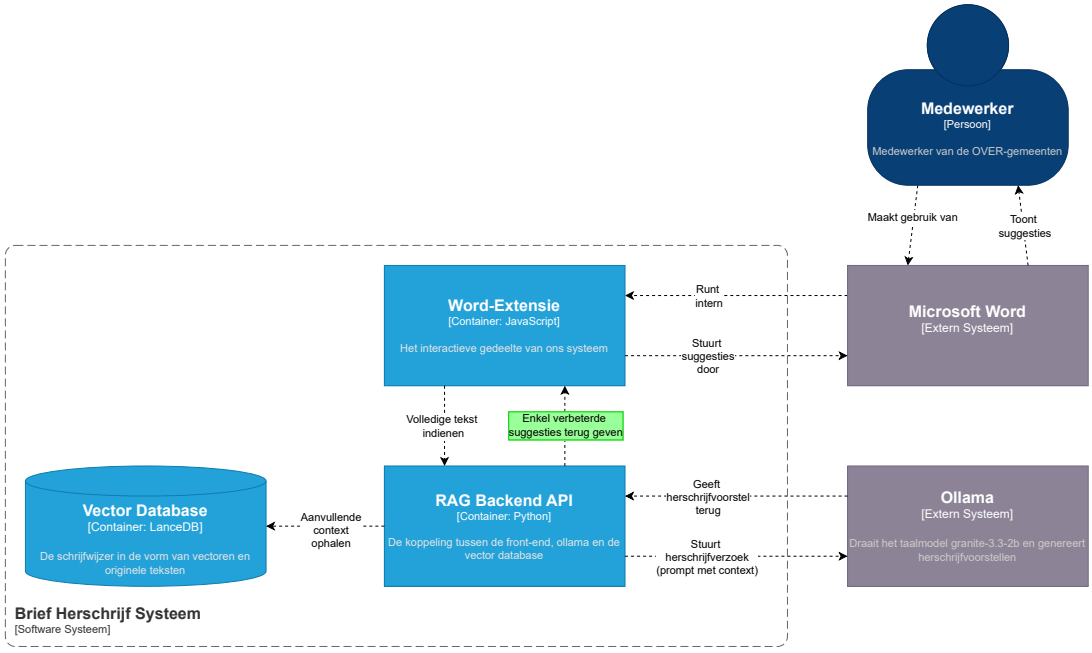


Fig. 8. Modelarchitectuur iteratie 2 (aanpassing in groen)

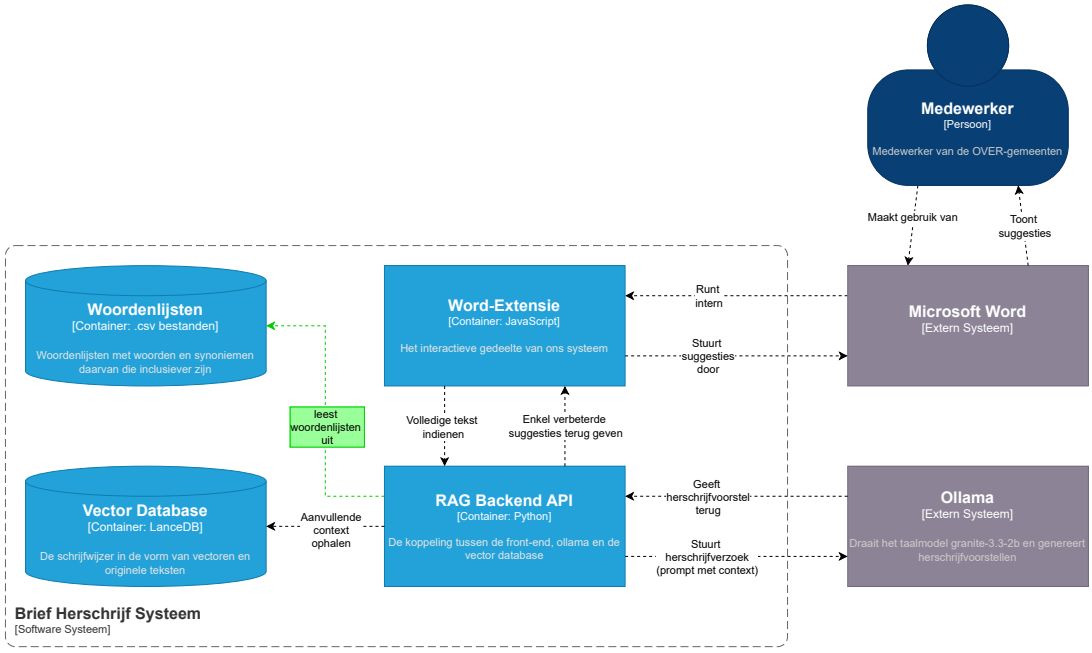


Fig. 9. Modelarchitectuur iteratie 3 (aanpassing in groen)

5 Resultaten

5.1 Resultaten van de modeltests

De prestaties van het LLM-model zijn geëvalueerd aan de hand van de LiNT-II-score, om te bepalen of herschreven brieven daadwerkelijk beter leesbaar zijn dan de originele versies. Voor deze evaluatie zijn meerdere iteraties van het model uitgevoerd, zoals beschreven in Paragraaf 4.3.5.

De zinnen uit de brieven in de validatieset zijn herschreven met alle ontwikkelde modellen: het baseline-model, het naïve-RAG-model (inclusief OVER-schrijfwijzer) met en zonder if-statement, en het model met if-statement aangevuld met de woordenlijsten. Voor elk model zijn de LiNT-II-scores berekend voor de originele brieven en de herschreven versies. Tabel 12 geeft een overzicht van de gemiddelde LiNT-II-scores en de gemiddelde verandering ten opzichte van de originele brieven.

Table 12. LiNT-II-scores

Model	**Gemiddelde LiNT-II-score**	**Gemiddelde verandering in LiNT-II-score t.o.v. originele brieven**
Geen model (originele brieven)	47.61	-
Baseline model	45.83	-1.78
Iteratie 1	44.80	-2.81
Iteratie 2	42.63	-4.98
Iteratie 3	43.26	-4.35

De originele brieven behalen een gemiddelde LiNT-II-score van 47.61. Na vereenvoudiging met het baseline-model daalt deze score naar 45.83 (−1.78), wat laat zien dat het model teksten vereenvoudigt, maar met een beperkt effect. Iteratie 1 behaalt een gemiddelde score van 44.80 (−2.81), wat een duidelijkere maar nog steeds beperkte verbetering vormt ten opzichte van het baseline-model. In iteratie 2 is een if-statement toegevoegd waardoor alleen suggesties worden getoond die de LiNT-II-score daadwerkelijk verlagen. Dit resulteert in een gemiddelde score van 42.63 (−4.98), wat aantoont dat het beperken tot effectieve herschrijvingen een significant positief effect heeft op de leesbaarheid. In iteratie 3 zijn woordenlijsten toegevoegd aan het model van iteratie 2, wat leidt tot een gemiddelde score van 43.26 (−4.35). Deze combinatie zorgt voor consistentere vereenvoudiging, al wijst de iets hogere score erop dat sommige suggesties minder effectief zijn dan in iteratie 2.

Voorbeeldzinnen van het baseline-model en de verschillende iteraties van het naïve-RAG-model zijn opgenomen in Tabellen 13, 14, 15 en 16. De voorbeelden van het baseline-model en iteratie 1 tonen aan dat zinnen minder formeel worden, maar nog fouten bevatten en soms geen correct Nederlands vormen, zoals het gebruik van het woord ‘noemd’. In iteratie 2 worden zinnen over het algemeen correcter en consistenter vereenvoudigd, maar veranderen sommige zinnen inhoudelijk of worden zij langer doordat extra informatie wordt toegevoegd. Hoewel het if-statement bepaalde fouten voorkomt, kan niet worden uitgesloten dat het model nieuwe informatie genereert. In iteratie 3 blijven de meeste zinnen vergelijkbaar met iteratie 2, maar sommige zinnen zijn grammaticaal minder correct of bevatten opnieuw extra context die niet in de originele tekst aanwezig was.

Table 13. Vergelijking originele en herschreven versies m.b.v. het baseline model

Originele zin	**Versimpelde zin**
Wij gaan daarna aan de slag om de ondersteuning voor u in te zetten.	We gaan daarna aan de slag om ondersteuning in te zetten voor u.
U heeft aangegeven geen voorkeur te hebben voor een zorgaanbieder.	Geen keuze gemaakt voor zorgaanbieder.
Wij gaan voor u op zoek naar een passende zorgaanbieder.	We zoeken voor u naar een geschikte zorgaanbieder.
Dit heet Zorg in Natura.	Zorg in Natura
U ontvangt hierover ons besluit met de belangrijkste informatie.	U ontvangt besluit met belangrijk info.
Tijdens het gesprek zijn enkele algemene en voorliggende voorzieningen benoemd.	Tijdens het gesprek werden enkele algemene voorzieningen genoemd.
Hiermee bedoelen wij voorzieningen, producten en diensten die voor alle inwoners van de gemeente beschikbaar zijn.	We bedoelen voorzieningen, producten en diensten die iedere burger in de gemeente heeft.
Hier is geen maatwerkvoorziening van de Wmo voor nodig.	Er is geen maatwerkvoorziening van de Wmo voor nodig.
Stichting WelzijnWonenPlus is een vrijwilligersorganisatie welke verschillende diensten biedt, onder andere tuinonderhoud, een sociaal maatje, de wijkbus/belbus en mantelzorgondersteuning.	Stichting WelzijnWonenPlus is een vrijwilligersorganisatie dat verschillende diensten aanbiedt, zoals tuinonderhoud en sociale contacten, de wijkbus/belbus, en mantelzorgondersteuning.
U kunt contact opnemen met WelzijnWonenPlus voor meer informatie.	Neem contact op met WelzijnWonenPlus voor meer details.

Table 14. Vergelijking originele en herschreven versies m.b.v. iteratie 1

Originele zin	**Versimpelde zin**
Wij gaan daarna aan de slag om de ondersteuning voor u in te zetten.	Wij beginnen met inzet van ondersteuning voor u.
U heeft aangegeven geen voorkeur te hebben voor een zorgaanbieder.	U heeft geen voorkeur voor een specifieke zorgaanbieder.
Wij gaan voor u op zoek naar een passende zorgaanbieder.	Wij zoeken een geschikte zorgaanbieder voor u.
Dit heet Zorg in Natura.	Dit heet Zorg in Natura.
U ontvangt hierover ons besluit met de belangrijkste informatie.	U ontvangt het besluit met de belangrijkste informatie.
Tijdens het gesprek zijn enkele algemene en voorliggende voorzieningen benoemd.	In het gesprek werden enkele algemene voorzieningen noemd.
Hiermee bedoelen wij voorzieningen, producten en diensten die voor alle inwoners van de gemeente beschikbaar zijn.	We bieden inhoud, producten en diensten die beschikbaar zijn voor iedere burger van onze gemeente.
Hier is geen maatwerkvoorziening van de Wmo voor nodig.	Je hebt geen specifieke maatwerkvoorziening nodig uit de Wmo.
Stichting WelzijnWonenPlus is een vrijwilligersorganisatie welke verschillende diensten biedt, onder andere tuinonderhoud, een sociaal maatje, de wijkbus/belbus en mantelzorgondersteuning.	Stichting WelzijnWonenPlus is een vrijwilligersorganisatie die verschillende diensten biedt, waaronder onder andere tuinonderhoud, een sociaal maatje, de wijkbus/belbus en mantelzorgondersteuning.
U kunt contact opnemen met WelzijnWonenPlus voor meer informatie.	U kunt contact opnemen met WelzijnWonenPlus voor meer informatie.

Table 15. Vergelijking originele en herschreven versies m.b.v. iteratie 2

Originele zin	**Versimpelde zin**
Wij gaan daarna aan de slag om de ondersteuning voor u in te zetten.	Wij beginnen met het inzetten van de ondersteuning voor u.
U heeft aangegeven geen voorkeur te hebben voor een zorgaanbieder.	U heeft aangegeven geen voorkeur te hebben voor een zorgaanbieder.
Wij gaan voor u op zoek naar een passende zorgaanbieder.	Wij zoeken voor u een geschikte zorgaanbieder. Ons onderzoek gaat uitdrukking geven aan de volgende stappen. We zullen contact opnemen met zorgaanbieders om passende opties te vinden. U krijgt van ons een overzicht van de beschikbare mogelijkheden.
Dit heet Zorg in Natura.	Dit heet Zorg in Natura.
U ontvangt hierover ons besluit met de belangrijkste informatie.	U ontvangt hierover ons besluit met de belangrijkste informatie.
Tijdens het gesprek zijn enkele algemene en voorliggende voorzieningen benoemd.	Tijdens het gesprek zijn enkele voorzieningen besproken.
Hiermee bedoelen wij voorzieningen, producten en diensten die voor alle inwoners van de gemeente beschikbaar zijn.	Hiermee bedoelen wij voorzieningen, producten en diensten die voor alle inwoners van de gemeente beschikbaar zijn.
Hier is geen maatwerkvoorziening van de Wmo voor nodig.	Hier is geen maatwerkvoorziening van de Wmo voor nodig.
Stichting WelzijnWonenPlus is een vrijwilligersorganisatie welke verschillende diensten biedt, onder andere tuinonderhoud, een sociaal maatje, de wijkbus/belbus en mantelzorgondersteuning.	Stichting WelzijnWonenPlus is een vrijwilligersorganisatie die verschillende diensten aanbiedt, waaronder tuinonderhoud, een sociaal maatje en de wijkbus/belbus. We bieden ook mantelzorgondersteuning.
U kunt contact opnemen met WelzijnWonenPlus voor meer informatie.	U kunt contact opnemen met WelzijnWonenPlus voor meer informatie.

Table 16. Vergelijking originele en herschreven versies m.b.v. iteratie 3

Originele zin	**Versimpelde zin**
Wij gaan daarna aan de slag om de ondersteuning voor u in te zetten.	Wij beginnen met de ondersteuning voor u in te zetten.
U heeft aangegeven geen voorkeur te hebben voor een zorgaanbieder.	U heeft geen voorkeur voor een specifieke zorgaanbieder. We zullen op zoek gaan naar een passende keuze in overleg met u.
Wij gaan voor u op zoek naar een passende zorgaanbieder.	Wij gaan voor u op zoek naar een passende zorgaanbieder.
Dit heet Zorg in Natura.	Dit heet Zorg in Natura.
U ontvangt hierover ons besluit met de belangrijkste informatie.	U ontvangt hierover ons besluit met de belangrijkste informatie.
Tijdens het gesprek zijn enkele algemene en voorliggende voorzieningen benoemd.	Tijdens het gesprek zijn enkele algemene en voorliggende voorzieningen benoemd.
Hiermee bedoelen wij voorzieningen, producten en diensten die voor alle inwoners van de gemeente beschikbaar zijn.	Hiermee bedoelen wij voorzieningen, producten en diensten die voor alle inwoners van de gemeente beschikbaar zijn.
Hier is geen maatwerkvoorziening van de Wmo voor nodig.	Hier is geen maatwerkvoorziening van de Wmo voor nodig.
Stichting WelzijnWonenPlus is een vrijwilligersorganisatie welke verschillende diensten biedt, onder andere tuinonderhoud, een sociaal maatje, de wijkbus/belbus en mantelzorgondersteuning.	Stichting WelzijnWonenPlus is een vrijwilligersorganisatie welke verschillende diensten biedt, onder andere tuinonderhoud, een sociaal maatje, de wijkbus/belbus en mantelzorgondersteuning.
U kunt contact opnemen met WelzijnWonenPlus voor meer informatie.	De vorige bewoner ontvangt veel post. U wilt hem uitschrijven vanwege dit. Neem contact op met WelzijnWonenPlus voor meer informatie.

5.2 Prestaties van het model in de context van het concept

Wanneer het huidige model in de OVER-gemeenten wordt toegepast, laten de resultaten zien dat het model potentie heeft om de leesbaarheid van brieven te verbeteren, maar dat er nog belangrijke aandachtspunten zijn. Met name iteratie 2 en iteratie 3 laten een duidelijke daling van de gemiddelde LiNT-II-score zien ten opzichte van de originele brieven (respectievelijk 4.98 en 4.35 punten), wat wijst op een significante verbetering in leesbaarheid. Tegelijkertijd tonen de voorbeeldzinnen aan dat deze verbeteringen soms gepaard gaan met inhoudelijke veranderingen, grammaticale fouten of het toevoegen van extra context die niet in de oorspronkelijke tekst aanwezig was. Hierdoor bestaat het risico dat de herschreven brief inhoudelijk afwijkt van het origineel of niet volledig voldoet aan de regels van de OVER-schrijfwijzer (Paragraaf 5.1).

De medewerker fungeert daarom nadrukkelijk als human-in-the-loop en blijft eindverantwoordelijk voor de uiteindelijke inhoud van de brief. Ondanks de verbeterde LiNT-II-scores is het essentieel dat de medewerker de suggesties van de AI-tool kritisch beoordeelt en controleert op correctheid, volledigheid en consistentie met de OVER-schrijfwijzer. In de huidige vorm kan het model ondersteuning bieden bij het vereenvoudigen van teksten, maar is het nog niet geschikt om zonder menselijke controle volledig betrouwbare herschrijvingen te leveren.

6 Evaluatie

6.1 Evaluatie van modelresultaten

In deze paragraaf worden de kwaliteitscriteria van het huidige best presterende model (iteratie 2) geëvalueerd, aangezien deze de laagste LiNT-II-score heeft behaald. Daarbij wordt ingegaan op de prestatiemaat, robuustheid, uitlegbaarheid, modelcomplexiteit en resource demand van het model (Tabel 17).

Table 17. Kwaliteitscriterium

Kwaliteitscriterium	Uitleg
Prestatiemaat	Uit de resultaten is gebleken dat het model uit iteratie 2 (gem. LiNT-II-score: 42.63) iets beter presteert dan het baseline model (gem. LiNT-II-score: 45.83), met een verschil van 3.2 punten. Hoewel het verschil in gemiddelde LiNT-II-score tussen het baseline-model (45.83) en het model uit iteratie 2 (42.63) relatief klein lijkt, is dit verschil betekenisvol wanneer het wordt geplaatst in de context van leesniveaus. Op basis van het eerder vastgestelde verband tussen LiNT-II-scores en CEFR-niveaus (Bijlage K) correspondeert een score van 42.63 met teksten die gemiddeld rond B1-niveau worden beoordeeld, terwijl hogere scores vaker samenhangen met hogere leesniveaus. Dit suggereert dat iteratie 2 de teksten gemiddeld naar een lager leesniveau verschuift dan het baseline-model. Tegelijkertijd moet worden benadrukt dat de koppeling tussen LiNT-II-scores en CEFR-niveaus geen harde grenzen kent. Uit eerder onderzoek blijkt dat LiNT-II-scores overlappen tussen leesniveaus en dat dezelfde tekst door verschillende beoordelaars op uiteenlopende CEFR-niveaus kan worden ingeschaald. Hierdoor kan niet worden geconcludeerd dat het model consequent teksten naar een lager leesniveau herschrijft, maar wel dat de kans op een eenvoudiger leesniveau toeneemt.
Robuustheid	In Tabel 18 worden de gemiddelde LiNT-II-scores van de originele en herschreven brieven weergegeven om te onderzoeken of deze consistent zijn over de verschillende briefcategorieën. Dit toetst of het model niet alleen voor één type brief goed presteert (RQ08). De resultaten worden weergegeven als gemiddelde LiNT-II-scores per briefcategorie en er wordt gecontroleerd of deze scores per categorie gelijk zijn. Uit de resultaten blijkt dat de gemiddelde LiNT-II-scores per briefcategorie van elkaar verschillen. Het model presteert dus niet consistent over de verschillende briefcategorieën.
Uitlegbaarheid	Uit de testen van de prototypes met de testpersonen (Bijlage G en I) en de daaruit aangepaste onderdelen in de nieuwe prototypen weergegeven in Paragrafen 6.2 en 6.3, kan geconcludeerd worden dat de UI-elementen duidelijk zijn en de tool als ondersteunend zal worden ervaren en niet als beslissend. Hierbij is ervoor gezorgd dat de medewerker altijd de eindverantwoordelijke is en dat dit ook duidelijk wordt gemaakt aan de medewerker door gebruik van een melding in het scherm als de tool aangezet wordt.
Modelcomplexiteit	De gebruikte LLM-architectuur in iteratie 2 is Granite-3.3-2B. In Paragraaf 4.3.2 werd aangegeven dat dit model 2 miljard parameters bevat. Dit zijn, vergeleken met bijvoorbeeld RoBERTa die 300 miljoen parameters bevat, veel parameters [26] voor een model. Daarentegen worden bij het Taalloket ook modellen benoemd die boven de 200 miljard parameters bevatten. In vergelijking met deze modellen bevat het model uit iteratie 2 weinig parameters. Met de informatie die beschikbaar is gesteld (Bijlage A) kan nog niet worden vastgesteld of het model binnen de OVER-gemeenten lokaal kan draaien zonder het aanschaffen van nieuwe hardware.

Resource demand Per zin wordt de verwerkingstijd gemeten. De AI-tool dient in staat te zijn om binnen ongeveer twee seconden een suggestie te genereren (RQ12). Met het huidige model blijkt dat het genereren van een suggestie voor één zin gemiddeld ongeveer 1,5 seconde duurt, waardoor het systeem deze tijdslimiet bij meerdere zinnen per brief overschrijdt. Hiermee kan worden geconcludeerd dat deze requirement in de huidige vorm nog niet wordt behaald.

Table 18. Gemiddelde LiNT-II-scores per briefcategorie

Groep	Gem. LiNT-II-score originele brieven	Gem. LiNT-II-score herschreven brieven (iteratie 2)	Gem. verschil LiNT-II-scores
Participatiefonds (PF)	46.00	42.71	3.29
Participatiewet (PW)	45.49	41.42	4.07
Schuldhulp	57.17	48.6	8.57
WMO	47.27	42.2	5.07
Overig	43.85	41.62	2.23

6.2 Onderbouwing van wijzigingen tussen concepten na eerste testronde

Na de eerste test met het paper prototype is de ontvangen feedback van de OVER-gemeenten medewerkers en studenten (Bijlage G) verwerkt in een verbeterd paper prototype. De belangrijkste aanpassingen en toevoegingen worden hieronder toegelicht aan de hand van relevante design patterns (Tabel 19).

Table 19. Overzicht van aanpassingen en toevoegingen na testronde 1

Aanpassing / toevoeging	**Uitleg en figuurnummer**
Aanpassing 1: Terminologie	De oorspronkelijke term 'AI-assistent' werd als technisch ervaren. Deze is vervangen voor 'Schrijfwijzer assistent', zodat de betekenis voor de medewerkers duidelijker is. Deze aanpassing is gebaseerd op de design pattern 'Voeg context toe vanuit menselijke bronnen' [40], omdat de naam van de AI-tool nu direct verwijst naar een bestaande en vertrouwde schrijfwijzer binnen de organisatie, waardoor gebruikers beter begrijpen waarvoor de tool bedoeld is. (Figuur 10)
Aanpassing 2: Toegevoegde knop 'Nieuwe suggestie'	Om gebruikers meer keuze en controle te geven is een extra knop 'nieuwe suggestie' toegevoegd. Hiermee kunnen zij alternatieve formuleringen opvragen wanneer de eerste niet bevalt. Dit verhoogt de gebruiksvriendelijkheid en voorkomt dat medewerkers aan één voorstel vastzitten. De aanpassing volgt het design pattern 'Laat gebruikers de automatisering begeleiden' [40], omdat gebruikers de suggesties niet alleen kunnen aanpassen, maar ook meerdere opties kunnen opvragen en daaruit kunnen kiezen. (Figuur 11)
Aanpassing 3: Bronvermelding eenmaal tonen	De bronverwijzing naar de schrijfwijzer werd twee keer weergegeven en wordt nu nog maar één keer weergegeven. Hoewel deze verwijzingen als nuttig en leerzaam werden ervaren, zorgen dubbele vermelding voor onnodige herhaling. Door de bron slechts eenmaal te tonen, blijft de informatie aanwezig zonder dat de interface onoverzichtelijk wordt. Deze aanpassing is gebaseerd op de design pattern 'Leg uit voor begrip, niet voor volledigheid' [40], omdat gebruikers voldoende informatie krijgen om het systeem te begrijpen zonder te worden belast met herhalende of overbodige informatie. (Figuur 12)
Toevoeging 1: Foutgerichte navigatie	Door op een gemarkeerde fout te klikken, wordt de gebruiker direct doorgestuurd naar de bijbehorende uitleg en verbetersuggestie. Deze aanpassing is gebaseerd op de design pattern 'Ga verder dan uitleg op het moment zelf' [40], omdat de gebruiker direct relevante informatie te zien krijgt, zonder zelf te zoeken naar de juiste toelichting of mogelijke verbeteringen.
Toevoeging 2: Melding eigen verantwoordelijkheid	Bij het inschakelen van de tool wordt een melding weergegeven met de tekst 'U bent zelf verantwoordelijk voor de inhoud van deze brief'. Deze melding is toegevoegd om medewerkers bewust te maken van hun eigen rol en verantwoordelijkheid bij het gebruik van de AI-tool. Deze toevoeging is gebaseerd op de design pattern 'Stel de juiste verwachtingen' [40], omdat hiermee duidelijk wordt gemaakt dat de AI slechts een hulpmiddel is en de medewerker eindverantwoordelijk blijft. (Figuur 13)
Toevoeging 3: Extra pagina als er geen suggesties zijn gevonden	Wanneer de AI-tool geen moeilijke woorden of zinnen vindt, wordt een pagina getoond met de melding dat er geen verbetersuggesties zijn. Daarbij wordt aangegeven dat de brief voldoet aan de B1-richtlijnen, maar dat de medewerker de brief nog wel kritisch dient na te lezen. Dit is belangrijk omdat de medewerker verantwoordelijk blijft voor de inhoud. Deze toevoeging is gebaseerd op het design pattern 'Stel de juiste verwachtingen' [40], omdat duidelijk wordt gemaakt dat de AI-tool geen verbeteringen heeft gevonden, maar dat de medewerker de brief zelf nog moet controleren. Dit voorkomt dat een lege extensie (doordat er geen suggesties worden weergegeven) wordt gezien als teken dat de tekst perfect is. (Figuur 14)

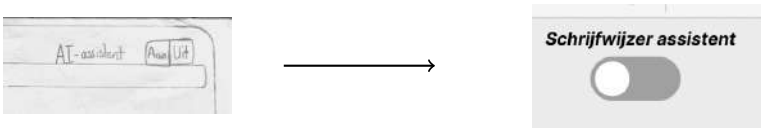


Fig. 10. Aanpassing 1: Terminologie

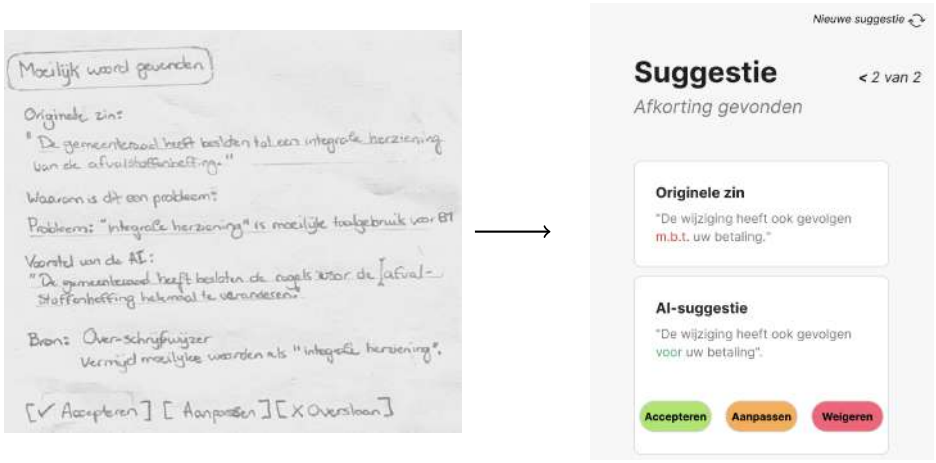


Fig. 11. Aanpassing 2: Toegevoegde knop 'Nieuwe suggestie'

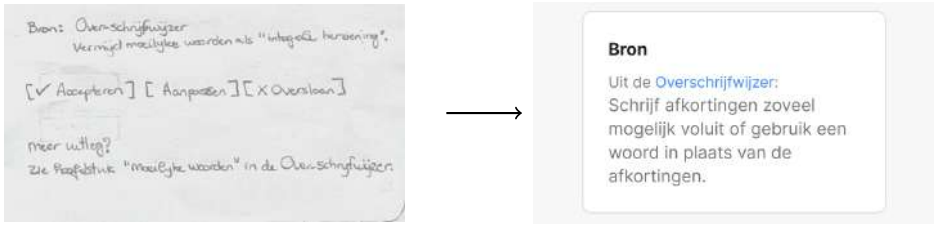


Fig. 12. Aanpassing 3: Bronvermelding eenmaal tonen

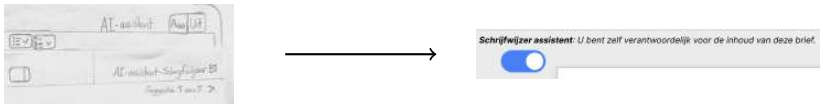


Fig. 13. Toevoeging 2: Melding eigen verantwoordelijkheid

Fig. 14. Toevoeging 3: Extra pagina als er geen suggesties zijn gevonden

6.2.1 Aanbevelingen die niet zijn meegenomen in het verbeterde concept.

De volgende aanbevelingen die tijdens het testen door de testpersonen zijn genoemd, zijn niet doorgevoerd in het verbeterde prototype:

- **Alle fouten met verbetersuggesties onder elkaar tonen in plaats van navigeren met pijlen.**
  Deze suggestie is niet doorgevoerd, omdat de meeste testpersonen aangaven het juist prettig te vinden dat niet alle fouten tegelijk zichtbaar zijn. Hierdoor worden gebruikers gestimuleerd per fout stil te staan bij de uitleg en de voorgestelde verbetering, wat het leerproces bevordert. Dit sluit beter aan bij het doel van de tool om gebruikers te helpen hun schrijfvaardigheid te verbeteren door inzicht te geven in concrete verbeterpunten.
- **Een knop ‘alles accepteren’ toevoegen.**
  Hoewel sommige testpersonen deze functie als praktisch beschouwden, is ervoor gekozen om deze niet te implementeren. Een dergelijke knop verhoogt het risico dat medewerkers suggesties automatisch overnemen zonder deze kritisch te beoordelen. Dit kan leiden tot onjuist gebruik van de AI-tool en een te groot vertrouwen in de AI, wat niet wenselijk is gezien de rol van de medewerker als eindverantwoordelijke.
- **Extra ondersteuning voor nieuwe gebruikers, bijvoorbeeld in de vorm van onboarding.**
  Eén testpersoon wilde bij de eerste keer gebruiken van de tool meer uitleg. Deze aanbeveling wordt als waardevol beschouwd, maar valt buiten de scope van dit project. Het toevoegen van onboarding-functionaliteit wordt daarom gezien als een mogelijke toekomstige verbetering.

6.3 Onderbouwing van wijzigingen tussen concepten na tweede testronde

Na de tweede test met het prototype is de ontvangen feedback van de CMD-experts (Bijlage I) verwerkt in een verbeterd prototype. De belangrijkste aanpassingen en toevoegingen worden hieronder toegelicht aan de hand van relevante design patterns (Tabel 20).

Table 20. Overzicht van aanpassingen en toevoegingen na testronde 2

Aanpassing / toevoeging	**Uitleg en figuurnummer**
Aanpassing 1: Vergroten melding eigen verantwoordelijkheid	Na de eerste testronde is een melding toegevoegd aan de tool met de tekst 'U bent zelf verantwoordelijk voor de inhoud van deze brief' (Paragraaf 6.2). Uit de tweede testronde bleek dat deze melding niet altijd duidelijk genoeg werd opgemerkt. Op basis van deze feedback is de melding groter weergegeven in het systeem. Deze aanpassing is gebaseerd op het design pattern 'Stel de juiste verwachtingen' [40], omdat hiermee duidelijk wordt gemaakt dat de AI een hulpmiddel is en de medewerker eindverantwoordelijk blijft. (Figuur 15)
Aanpassing 2: Kleurgebruik aanpassen	Uit de gebruikerstesten bleek dat visuele signalen, zoals kleurgebruik en onderstrepingen, effectief zijn om snel te zien waar aanpassingen in de tekst worden voorgesteld. Tegelijkertijd gaven testers aan dat het grote aantal kleuren en het sterke contrast bij eerste gebruik overweldigend kan zijn. Op basis van deze feedback zijn de volgende aanpassingen doorgevoerd: • Highlights lichter weergeven: visuele signalen zijn aangepast om minder overweldigend te zijn. • Alleen daadwerkelijk gewijzigde tekst ten opzichte van de suggestie in het word document markeren: – Eén woord highlighten wanneer slechts één woord is aangepast. – De gehele zin highlighten wanneer de hele zin is gewijzigd. • Knoppen 'Accepteren', 'Aanpassen' en 'Weigeren' neutraal weergeven: de felle kleuren (groen, oranje en rood) zijn vervangen door grijze knoppen, conform de standaardstijl in Word, zodat de knoppen niet onnodig de aandacht trokken. Deze aanpassingen zijn gebaseerd op het design pattern 'Leg uit voor begrip, niet voor volledigheid' [40]. De aanpassing helpt gebruikers om visuele aanwijzingen en uitleg snel en intuïtief te begrijpen, terwijl onnodige details en complexiteit achterwege worden gelaten. (Figuren 16, 17, 18)
Aanpassing 3: Verwoording eindmelding aanpassen	De eindmelding die verschijnt nadat alle suggesties zijn verwerkt, waarin expliciet wordt benadrukt dat de gebruiker zelf verantwoordelijk blijft voor de inhoud van de brief, werd als waardevol ervaren. Tegelijkertijd werd opgemerkt dat de bewering dat de brief 'op B1-niveau is geschreven' niet met volledige zekerheid kan worden gedaan. Omdat deze claim inderdaad te stellig was, is de verwoording ervan aangepast. Deze aanpassing is gebaseerd op de design pattern 'Stel de juiste verwachtingen' [40], omdat hiermee duidelijk wordt gemaakt wat de AI wel en niet kan, waardoor gebruikers een realistisch beeld krijgen van de rol van de tool en zich bewust blijven van hun eigen verantwoordelijkheid. (Figuur 19)
Aanpassing 4: Suggestie opnieuw-knop verplaatsen	Het lijkt alsof de gehele AI-tool wordt vernieuwd bij het klikken op de 'suggestie opnieuw'-knop, daarom is besloten de knop te verplaatsen. De knop zelf sluit aan bij het design pattern 'geef controle terug aan de gebruiker wanneer de automatisering faalt'. Het verplaatsen van de knop sluit aan bij het design pattern 'Stel de juiste verwachtingen', omdat voor medewerkers nu duidelijk is dat de knop alleen een nieuwe suggestie genereert en niet de hele AI-tool ververst [40]. (Figuur 20)

Aanpassing / toevoeging	Uitleg en figuurnummer
Toevoeging 1: Melding eigen verantwoordelijkheid herhalen	Na de eerste testronde is een melding toegevoegd aan de tool met de tekst ‘U bent zelf verantwoordelijk voor de inhoud van deze brief’ (Paragraaf 6.2). Uit de tweede testronde bleek dat deze melding niet altijd voldoende opviel. Daarnaast gaf een testpersoon aan dat de melding vaker in beeld mocht komen, bijvoorbeeld wanneer een medewerker meerdere keren (bijvoorbeeld drie keer) achter elkaar op de knop ‘Accepteren’ klikt, zodat zij vaker worden herinnerd aan hun eigen verantwoordelijkheid. Deze toevoeging is gebaseerd op het design pattern ’Wees verantwoordelijk voor fouten’ [40], omdat het medewerkers bewust maakt van hun verantwoordelijkheid voor eventuele fouten in de brief die kunnen ontstaan wanneer (alle) suggesties zonder kritisch te beoordelen worden geaccepteerd.
Toevoeging 2: Knop ’terug’ toevoegen.	Een ander punt dat in de tweede testronde werd genoemd, was dat gebruikers eerder geaccepteerde of geweigerde suggesties niet opnieuw konden bekijken. Daarom is een ‘terug’-knop toegevoegd. Deze aanpassing is gebaseerd op het design pattern Laat gebruikers de automatisering begeleiden [40]. De knop stelt gebruikers in staat hun eerdere keuzes te controleren en behoudt zo de controle over het proces. (Figuur 21)

Fig. 15. Aanpassing 1: Vergroten melding eigen verantwoordelijkheid

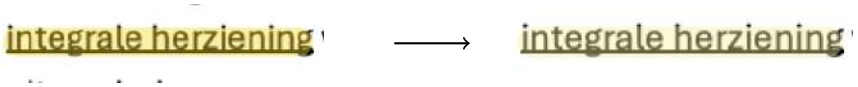


Fig. 16. Aanpassing 2.1: Kleurgebruik aanpassen

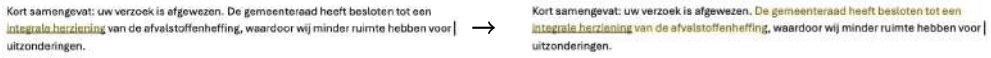


Fig. 17. Aanpassing 2.2: Kleurgebruik aanpassen

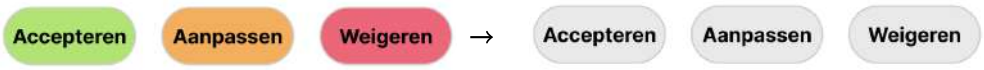


Fig. 18. Aanpassing 2.3: Kleurgebruik aanpassen

Fig. 19. Aanpassing 3: Verwoording eindmelding aanpassen

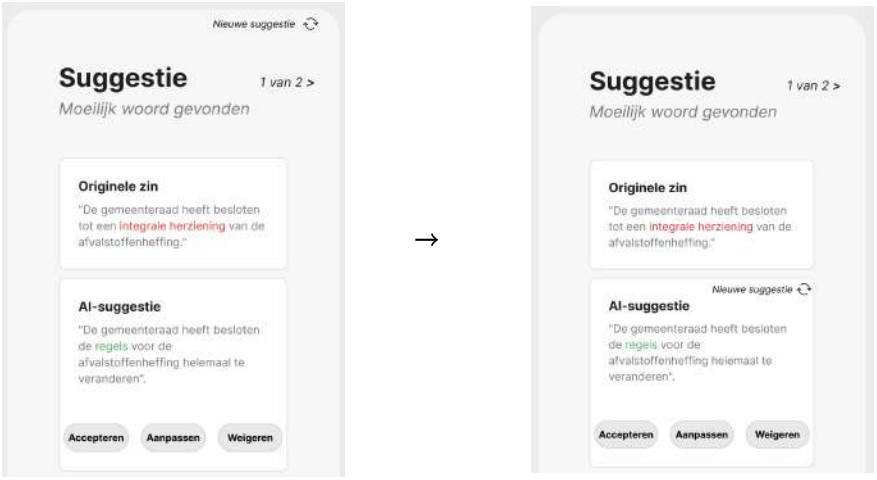


Fig. 20. Aanpassing 4: Suggestie opnieuw-knop verplaatsen

Fig. 21. Toevoeging 2: Knop 'terug' toevoegen

6.3.1 *Aanbeveling die niet is meegenomen in het verbeterde concept.*

De relevantie van de suggesties werd overwegend positief beoordeeld. Vooral de inhoudelijke uitleg en de koppeling aan de OVER-schrijfwijzer werden als leerzaam ervaren. Wel werd voorgesteld om de bronverwijzing prominenter te plaatsen, zodat gebruikers eerst begrijpen waarom een suggestie wordt gedaan voordat zij een keuze maken. Uit de eerste gebruikerstest met medewerkers van de OVER-gemeenten (Bijlage G) bleek echter dat zij de bronverwijzing liever eenmalig onderaan in de AI-tool willen hebben staan. Daarom is deze aanbeveling niet meegenomen bij het aanpassen van het prototype. Daarnaast werd door een testpersoon voorgesteld om de 'aanpassen'-knop te verwijderen. Deze suggestie is niet meegenomen, omdat andere testpersonen juist aangaven dat zij de knop nuttig vinden.

6.4 Feedbackloop en doorontwikkeling van de AI-tool

Om de AI-tool na implementatie verder te verbeteren, wordt een feedbackloop ingericht die inzicht geeft in het gebruik en effectiviteit van de AI-tool. Hierbij wordt onderscheid gemaakt tussen evolving en adaptive verbetermechanismen. De evolving component richt zich op periodieke hertraining van het taalmodel met nieuwe, externe en algemeen beschikbare taaldata, zoals geactualiseerde B1-richtlijnen of schrijfwijzers. Deze verbetering is niet gebaseerd op gebruikersfeedback en valt daarmee buiten de directe feedbackloop, maar draagt wel bij aan de structurele kwaliteit en actualiteit van het model.

De adaptive component vormt de kern van de feedbackloop en is gebaseerd op het gebruik van de AI-tool in de praktijk. Medewerkers kunnen per suggestie kiezen om deze te accepteren, aan te passen, te negeren of een nieuwe suggestie aan te vragen. Deze interacties leveren impliciete gebruikersfeedback op, waarmee onder andere acceptatiepercentages, verschillen per zinstype en overlap tussen oorspronkelijke en uiteindelijke formuleringen kunnen worden geanalyseerd. Deze inzichten worden gebruikt om de formuleringen en kwaliteit van de gegenereerde suggesties verder te optimaliseren. Omdat gemeentelijke brieven persoonsgegevens kunnen bevatten, wordt privacybescherming expliciet meegenomen in de feedbackloop door middel van automatische anonimisering, bijvoorbeeld met Named Entity Recognition (NER), en het uitsluiten van vaste tekststructuren en invulvelden van analyse [12].

6.5 Volledig geïntegreerde evaluatie van het model binnen het concept

Het prototype is getest door medewerkers van de OVER-gemeenten, studenten en Communicatie en Multimedia Design-experts. De meeste testpersonen gaven aan dat zij de tool als prettig en ondersteunend zouden ervaren in de praktijk. In het prototype was daarbij zichtbaar dat de AI-tool uitleg kon geven met een verwijzing naar de OVER-schrijfwijzer. Met het huidige model is dit echter nog niet mogelijk. Naast het feit dat deze functionaliteit nog niet beschikbaar is, presteert het model uit iteratie 2 iets beter dan het baseline model (Paragraaf 6.1), dat eveneens onvoldoende in staat is om originele brieven te herschrijven volgens richtlijnen van de OVER-gemeenten. Daarnaast maakt de gekozen architectuur het momenteel niet mogelijk om brieven snel genoeg te herschrijven voor praktische inzet.

Wanneer het model zich wel consistent kan houden aan de regels van de OVER-schrijfwijzer, wordt het systeem gezien als een waardevol hulpmiddel dat het werk kan vereenvoudigen en versnellen. Tegelijkertijd wijzen testpersonen op het risico dat suggesties, met name onder tijdsdruk, te snel worden overgenomen zonder voldoende controle. Dit kan leiden tot het versturen van brieven met foutieve informatie en daarmee tot ongewenste of risicovolle situaties voor inwoners. Hoewel de AI-tool tijdens gebruik expliciet meldt dat medewerkers eindverantwoordelijk blijven voor de inhoud, is het noodzakelijk dit ook bij implementatie duidelijk te communiceren. Daarbij moet worden benadrukt dat de AI-tool statistische fouten kan maken en daarom nooit volledig te vertrouwen is.

7 Discussie

In dit hoofdstuk worden de discussiepunten besproken:

- Uit de DEDA-analyse (Bijlage F), blijkt dat de inzet van een AI-tool binnen gemeentelijke context risico's met zich meebrengt. Een belangrijk aandachtspunt is automation bias: medewerkers kunnen geneigd zijn om AI-suggesties te accepteren zonder deze kritisch te beoordelen. Dit is risicovol bij gemeentelijke brieven, omdat fouten gevolgen voor de inwoners kunnen hebben. Daarom zijn naar aanleiding van de feedback van de tweede testronde meldingen toegevoegd om gebruikers te wijzen op hun eindverantwoordelijkheid. In vervolgonderzoek kan worden onderzocht welke aanvullende maatregelen nodig zijn om automation bias verder te beperken.
- Het model (iteratie 2) presteert niet voor alle briefcategorieën even goed. Dit is te verklaren door de beperkte en scheef verdeelde dataset van de OVER-gemeenten, waarin sommige briefcategorieën oververtegenwoordigd zijn (Figuur 5). Hierdoor kan de AI-tool op dit moment nog niet als generaliseerbaar binnen de OVER-gemeenten worden beschouwd.
- Daarnaast is er nagedacht over de inrichting van een mogelijke feedbackloop, zoals beschreven in Paragraaf 6.4. In deze paragraaf wordt ingegaan op een toekomstscenario waarin het model leert van gebruikersbeslissingen, bijvoorbeeld door het accepteren of negeren van suggesties. De concrete implementatie van een dergelijke feedbackloop zijn echter nog niet uitgewerkt. Vervolgonderzoek is nodig om te bepalen hoe gebruikersfeedback veilig en effectief kan worden ingezet, zonder dat foutieve goedkeuringen worden versterkt of privacygevoelige informatie wordt meegenomen bij het verbeteren van het model.
- Hoewel de tool ondersteuning biedt bij het herschrijven van brieven, blijft een deel van de inwoners onvoldoende bereikt, namelijk de mensen die een lager taalniveau bezitten dan het B1-niveau [2]. Het toepassen van B1-niveau is een eis vanuit de opdrachtgever, maar lost het bredere probleem van begrijpelijkheid in gemeentelijke teksten hiermee niet volledig op. Dit roept de vraag op of aanvullende communicatievormen nodig zijn.
- Het model presteert op dit moment nog onvoldoende wat betreft responstijd. Het genereren van een zin-suggestie duurt gemiddeld ongeveer 1,5 seconde, terwijl de beoogde functionaliteit was dat het model binnen ongeveer twee seconden een volledige brief zou kunnen herschrijven (RQ12). Deze beperkte snelheid heeft invloed op de gebruikservaring en vormt daarmee een aandachtspunt voor verdere optimalisatie in vervolgonderzoek.
- Hoewel de Lint-II-score van de herschreven brieven is verbeterd, kan deze mogelijk verder worden verlaagd door aanvullende of alternatieve methoden toe te passen. In dit onderzoek is gebruikgemaakt van één specifieke aanpak met naïve-RAG, terwijl andere methoden mogelijk tot verdere verbetering kunnen leiden. In vervolgonderzoek kan worden onderzocht welke methoden kunnen bijdragen aan de leesbaarheid van gemeentelijke brieven.
- In dit onderzoek is de werking van het model voornamelijk beoordeeld aan de hand van de LiNT-II-score. Hoewel deze maat inzicht geeft in de leesbaarheid van teksten, biedt zij slechts een beperkt perspectief op de kwaliteit van de herschreven brieven. Andere prestatiematen zijn in dit onderzoek niet meegenomen bij het evalueren van het model. In vervolgonderzoek kan worden onderzocht welke aanvullende prestatiematen nodig zijn om het model vollediger en evenwichtiger te beoordelen.

8 Conclusie

8.1 Terugkoppeling op de requirements

Op ethisch vlak zijn meerdere requirements succesvol ingevuld. De medewerker bleef eindverantwoordelijk voor de inhoud van de brief (RQ02) en het gebruik van de AI-tool was volledig optioneel (RQ03). Daarnaast werd expliciet gecommuniceerd dat de gegenereerde suggesties uitsluitend ter ondersteuning dienen en dat de eindbeslissing bij de medewerker ligt, wat bijdraagt aan het voorkomen van automation bias (RQ04). Een belangrijk aandachtspunt blijft echter de begrijpelijkheid van de werking van de AI-tool voor gebruikers zonder AI-kennis (RQ01), aangezien niet alle testers de meldingen direct begrepen.

Ook op juridisch vlak zijn belangrijke waarborgen gerealiseerd. Tijdens de ontwikkeling is uitsluitend gebruikgemaakt van geanonimiseerde brieven (RQ05) en het taalmodel werd lokaal uitgevoerd, waardoor geen persoonsgegevens met externe partijen werden gedeeld (RQ06). Tegelijkertijd is er technisch geen expliciete verwerking ingebouwd om persoonsgegevens te verwijderen tijdens runtime (RQ10), wat in toekomstige toepassingen een risico kan vormen.

Functioneel presteerde de AI-tool beter dan het baseline model (RQ07), wat wijst op toegevoegde waarde. De betrouwbaarheid van de resultaten over verschillende briefcategorieën heen kon echter niet overtuigend worden aangetoond (RQ08), mede door een beperkte en scheef verdeelde dataset. Technisch gezien bleek de responstijd een beperkende factor omdat het systeem niet voldeed aan de beoogde snelheid voor het verwerken van volledige brieven (RQ12), ondanks dat de dataset correct is gesplitst en de ontwikkeling en prestaties zijn gedocumenteerd (RQ09 en RQ11).

8.2 Toekomstig werk en aanbevelingen

Vanwege tijdsbeperkingen konden niet alle aspecten van dit onderzoek volledig worden uitgewerkt. De volgende onderwerpen worden daarom aangedragen voor vervolgonderzoek:

- Onderzoeken of het ontwerp geïmplementeerd kan worden binnen een word-extensie.
- Onderzoeken of het mogelijk is om bij elke suggestie de bijbehorende regel uit de OVER-schrijfwijzer te tonen.
- Verder onderzoeken welke methoden, naast de huidige aanpak, kunnen bijdragen aan verbeterde leesbaarheid, waaronder alternatieve tekstvereenvoudigingstechnieken zoals een advanced-RAG-implementatie.
- In een nieuwe testronde vaststellen welke aanvullende maatregelen nodig zijn om automation bias verder te beperken, naast de geïmplementeerde meldingen.
- Uitbreiden en beter balanceren van de dataset, met een gelijkere verdeling van brieven per categorie, om sterkere uitspraken te kunnen doen over de generaliseerbaarheid van het model.
- De mogelijke feedbackloop, waarbij het model leert van gebruikersbeslissingen, explicieter uitwerken en onderzoeken hoe deze veilig en privacybewust kan worden geïmplementeerd.
- Optimaliseren van de responstijd van het model, zodat het beter aansluit bij de beoogde gebruikssituatie in de gemeentelijke praktijk.
- Het gebruik van aanvullende prestatiematen naast de LiNT-II-score, om de prestaties van het model vollediger te evalueren.
- Ons model uit iteratie 2, zonder woordenlijsten zoals de inclusiviteits-lijst, behaalde een LiNT-II-score die 0.63 lager lag dan die van iteratie 3. Mogelijk zouden zij de voorkeur geven aan een inclusief model, ook als dit gepaard gaat met een iets hogere LiNT-II-score. In vervolgonderzoek kan onderzocht worden of gemeenten het marginale verschil in LiNT-score belangrijker vinden dan het toevoegen van de woordenlijsten.
- Ontwikkelen en toevoegen van een onboarding-functionaliteit, zodat gebruikers meer uitleg krijgen over de werking en mogelijkheden van de tool.

8.3 Ethische verantwoording: impact van de AI-oplossing op individu en maatschappij

In de volgende paragrafen wordt de impact van de AI-oplossing besproken op zowel individueel niveau, voor medewerkers van de OVER-gemeenten, als op maatschappelijk niveau, voor de burgers die deze gemeenten bedienen.

8.3.1 Impact van de AI-oplossing op individu (medewerker van de OVER-gemeenten).

De AI-tool is in staat voorstellen te genereren die aansluiten bij de OVER-schrijfwijzer en kan daarmee medewerkers van OVER-gemeenten ondersteunen bij het herschrijven van brieven volgens de vastgestelde richtlijnen. Voor medewerkers is het daarbij belangrijk dat deze voorstellen worden voorzien van een toelichting, bijvoorbeeld in de vorm van een onderliggende reden en een expliciete verwijzing naar de relevante schrijfwijzerregel. In de huidige vorm biedt het model deze onderbouwing nog niet, waardoor medewerkers geen houvast hebben bij het beoordelen van de suggesties. Dit kan de effectiviteit van de tool verminderen en ertoe leiden dat extra tijd nodig is voor controle en correctie.

Deze beperkingen hebben ook ethische implicaties voor de inzet van het model. Hoewel de AI-tool kan bijdragen aan efficiëntere en meer consistente toepassing van B1-richtlijnen, bestaat het risico dat medewerkers voorstellen onder tijdsdruk te snel overnemen zonder grondige beoordeling. Om dit risico te beperken is het noodzakelijk dat de AI-tool uitsluitend een ondersteunende, adviserende rol vervult en dat de medewerker eindverantwoordelijk blijft voor de inhoud van de brief. Door deze verantwoordelijkheid expliciet te verzorgen via een human-in-the-loop-benadering, kan de AI-oplossing onder duidelijke voorwaarden op ethisch verantwoorde wijze worden ingezet, mits medewerkers zich bewust blijven van de beperkingen van het systeem en actief betrokken blijven bij de eindbeoordeling.

8.3.2 Impact van de AI-oplossing op maatschappij (burgers van de OVER-gemeenten).

Medewerkers herschrijven brieven voor burgers in Word, waarbij de AI-tool lokaal wordt uitgevoerd. Dit betekent dat gegevens van burgers de gemeente niet verlaten en dus niet extern worden gedeeld. Op deze manier wordt voorkomen dat privacygevoelige informatie van burgers wordt verspreid naar derden.

Afhankelijk van de effectiviteit van de suggesties kunnen inwoners beter begrijpelijke brieven ontvangen. Tegelijkertijd bestaat het risico dat brieven juist minder duidelijk worden, bijvoorbeeld door automation bias of door het overnemen van onjuiste suggesties. Het is daarom belangrijk dat een medewerker de brief altijd zorgvuldig controleert op inhoudelijke en (juridische) correctheid.

References

[1] [n. d.]. *B1-teksten in de praktijk: voorbeelden en tips | B1-teksten.* https://b1teksten.nl/artikel/voorbeelden-van-b1-teksten

[2] [n. d.]. *Begrijpelijke communicatie | CommunicatieRijk.* https://www.communicatierijk.nl/vakkennis/uitgangspunten-en-organisatie/inclusieve-communicatie/begrijpelijke-communicatie

[3] Chroma 2025. *Chroma-Core/Chroma.* Chroma. https://github.com/chroma-core/chroma

[4] [n. d.]. *EDIA.* https://www.edia.nl

[5] [n. d.]. *Focus op AI bij de rijksoverheid | Algemene Rekenkamer.* https://www.rekenkamer.nl/documenten/2024/10/16/focus-op-ai-bij-de-rijksoverheid

[6] 2025. *nluwv/Leesplank_Noot.* https://github.com/nluwv/Leesplank_Noot original-date: 2024-10-14T07:40:41Z.

[7] [n. d.]. *Taalniveau B1.* https://site.dijkenwaard.nl/huisstijl/heldere-taal/direct-duidelijk-dijk-en-waard/taalniveau-b1

[8] [n. d.]. *Train Test Validation Split: How To & Best Practices [2024].* https://www.v7labs.com/blog/train-validation-test-set

[9] 2024. *UWV/leesplank-noot-eurollm-1.7b · Hugging Face.* https://huggingface.co/UWV/leesplank-noot-eurollm-1.7b

[10] 2025. *UWV/leesplank-noot-granite-3.3-2b · Hugging Face.* https://huggingface.co/UWV/leesplank-noot-granite-3.3-2b

[11] 2025. *UWV/leesplank-noot-llama-3.2-3b · Hugging Face.* https://huggingface.co/UWV/leesplank-noot-llama-3.2-3b

[12] 2023. *What Is Named Entity Recognition? | IBM.* https://www.ibm.com/think/topics/named-entity-recognition

[13] 2025. lancedb/lancedb. https://github.com/lancedb/lancedb original-date: 2023-02-28T01:15:17Z.

[14] 2025. qdrant/qdrant. https://github.com/qdrant/qdrant original-date: 2020-05-30T21:37:01Z.

[15] Moustafa Abdelwanis, Hamdan Khalaf Alarafati, Maram Muhanad Saleh Tammam, and Mecit Can Emre Simsekler. 2024. Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. 5, 4 (2024), 460–469. doi:10.1016/j.jnlssr.2024.06.001

[16] Algemene Rekenkamer. 2016. Volwassenen met moeite met taal of rekenen: kloof tussen probleem en aanpak via rijksbeleid. https://www.rekenkamer.nl/publicaties/persberichten/2016/04/20/volwassenen-met-moeite-met-taal-of-rekenen-kloof-tussen-probleem-en-aanpak-via-rijksbeleid.

[17] Amsterdam. [n. d.]. *Helder juridisch woordenboek.* https://www.amsterdam.nl/schrijfwijzer/heldere-taal-basis-onze-huisstijl/helder-juridisch-schrijven/helder-juridisch-woordenboek/ Last Modified: 2025-12-02; Publisher: Gemeente Amsterdam.

[18] Amsterdam. [n. d.]. *Inclusieve woordenlijst.* https://www.amsterdam.nl/schrijfwijzer/inclusieve-taal-richtlijnen-tips/inclusieve-woordenlijst/ Last Modified: 2025-12-02; Publisher: Gemeente Amsterdam.

[19] Amsterdam. [n. d.]. *Moeilijke-woordenboek (inclusief niet te gebruiken ambtelijke taal).* https://www.amsterdam.nl/schrijfwijzer/moeilijke-woorden/ Last Modified: 2025-12-07; Publisher: Gemeente Amsterdam.

[20] Autoriteit Persoonsgegevens. 2025. *AVG-Randvoorwaarden voor generatieve AI.* Technical Report. Autoriteit Persoonsgegevens. Accessed: 2025-11-18.

[21] Marcello Barbella and Genoveffa Tortora. 2022. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317* (2022).

[22] Marvin van Bekkum and Frederik Zuiderveen Borgesius. 2025. De spanning tussen het non-discriminatierecht en het gegevensbeschermingsrecht: heeft de AVG een nieuwe uitzondering nodig om discriminatie door kunstmatige intelligentie tegen te gaan? arXiv:2509.08836 [cs] doi:10.48550/arXiv.2509.08836

[23] Betabit. 2018. ETHISCHE RICHTSNOEREN voor BETROUWBARE KI Deskundigengroep op hoog niveau inzake kunstmatige intelligentie. Report. https://www.betabit.nl/media/4614/ethicsguidelinesfortrustworthyai-nl.pdf

[24] Punyakeerthi BL. 2024. *Understanding Encoder-Only Models: Simplifying Text with Single-Direction Power.* https://medium.com/@punya8147_26846/understanding-encoder-only-models-simplifying-text-with-single-direction-power-0a0c4f0a7846

[25] Mark Breuker. 2022. Cefr labelling and assessment services. In *European Language Grid: A Language Technology Platform for Multilingual Europe.* Springer International Publishing Cham, 277–282.

[26] François Chollet. 2023. *Deep Learning with Python.* Chollet. https://deeplearningwithpython.io/ Online boek, geraadpleegd op 2025-11-21.

[27] Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An Exploration into Neural Text Simplification Models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille, France, 2020-05), Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 5588–5594. https://aclanthology.org/2020.lrec-1.686/

[28] Mischa Corsius, Henk Pander Maat, Els van der Pool, and Wouter Sluis-Thiescheffer. 2023. Monitor Begrijpelijkheid overheidsteksten 2022. Report. https://taalunie.org/publicaties/218/monitor-begrijpelijkheid-overheidsteksten-2022

[29] Mischa Corsius, Els Van Der Pool, and Wouter Sluis-Thiescheffer. 2024. Begrijp jij het? Mixed-method monitor van overheidsteksten. 45, 1 (2024), 40–65. doi:10.5117/TVT2023.3.003.CORS

[30] Analyttica Datalab. 2021. *What is meant by 'Stratified Split'?* https://medium.com/@analyttica/what-is-meant-by-stratified-split-289a8a986a90

[31] European Union Agency for Fundamental Rights. 2025. Artikel 21 – Non-discriminatie. `EU-Handvest van de grondrechten`. https://fra.europa.eu/nl/eu-charter/article/21-non-discriminatie Geraadpleegd op 24 september 2025.

[32] Rudolf Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221–233.

[33] GebruikerCentraal. [n. d.]. Toolkit Taal. https://toolkittaal.gebruikercentraal.nl/richtlijnen/.

[34] GebruikerCentraal. [n. d.]. Van Direct Duidelijk naar duidelijke overheidscommunicatie. https://www.gebruikercentraal.nl/meedoen/netwerk-direct-duidelijk/duidelijke-overheidscommunicatie/ Accessed: November 13, 2025.

[35] Gemeente Amsterdam. [n. d.]. Schrijfwijzer v4. https://openresearch.amsterdam/nl/page/94061/schrijfwijzer-v4.pdf.

[36] Gemeente Oostzaan. [n. d.]. Schrijfwijzer. https://dlo.mijnhva.nl/d2l/le/content/691830/viewContent/2743225/View.

[37] Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent Chunking Mechanisms for Long-Text Machine Reading Comprehension. doi:10.48550/arXiv.2005.08056 arXiv:2005.08056 [cs].

[38] Sian Gooding and Manuel Tragut. 2022. One Size Does Not Fit All: The Case for Personalised Word Complexity Models. arXiv:2205.02564 [cs] doi:10.48550/arXiv.2205.02564

[39] Sian Gooding and Manuel Trägut. 2022. One Size Does Not Fit All: The Case for Personalised Word Complexity Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, USA, 353–365. https://aclanthology.org/2022.findings-naacl.27/

[40] Google PAIR. 2022. People + AI Guidebook: Patterns. https://pair.withgoogle.com/guidebook/patterns Accessed: 2025-12-02.

[41] Grondwet voor het Koninkrijk der Nederlanden. 2023. Artikel 1: Gelijke behandeling en discriminatieverbod. `DeNederlandseGrondwet.nl`. https://www.denederlandsegrondwet.nl/id/vgrnb2er8avw/artikel_1_gelijke_behandeling_en Geraadpleegd op 6 december 2025.

[42] Eliza Hobo, Charlotte Pouw, and Lisa Beinborn. 2023. "Geen makkie": Interpretable Classification and Simplification of Dutch Text Complexity. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 503–517. https://aclanthology.org/2023.bea-1.42/

[43] Kayleigh Hoogenboom. 2022. Duidelijke Overheidscommunicatie voor gemeenten. *Hoogenboom* (2022).

[44] Jenia Kim and Henk Pander Maat. 2025. *LiNT-II: readability assessment for Dutch*. https://github.com/vanboefer/lint_ii Python package.

[45] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Technical Report Research Branch Report 8–75. Naval Technical Training Command, Millington, TN, USA.

[46] Akshay Kore. 2022. *Designing Human-Centric AI Experiences: Applied UX Design for Artificial Intelligence*. Apress.

[47] Akshay Kore. 2022. *Designing Human-Centric AI Experiences: Applied UX Design for Artificial Intelligence*. Apress. doi:10.1007/978-1-4842-8088-1

[48] Anthony Lamelas. 2026. Evaluating Small Decoder-Only Language Models for Grammar Correction and Text Simplification. arXiv:2601.03874 [cs] doi:10.48550/arXiv.2601.03874

[49] Victoria Muñoz-García, Paloma Moreda, and Manuel Palomar. 2025. Exploring Evaluation Methods on Text Simplification: A Systematic Review. In *2025 7th International Conference on Natural Language Processing (ICNLP)*. IEEE, 277–281.

[50] Gustavo H. Paetzold and Lucia Specia. 2017. A Survey on Lexical Simplification. 60 (2017), 549–593. doi:10.1613/jair.5526

[51] Henk Pander Maat, Suzanne Kleijn, and Servaas Frissen. [n. d.]. LiNT: een leesbaarheidsformule en een leesbaarheids instrument. *Tijdschrift voor Taalbeheersing* 45, 1 ([n. d.]), 2–39.

[52] Patronus AI. 2023. RAG Evaluation Metrics: Best Practices for Evaluating RAG Systems. https://www.patronus.ai/llm-testing/rag-evaluation-metrics. Accessed: 2025-11-28.

[53] Renyi Qu and Ruixuan Tu. [n. d.]. Is Semantic Chunking Worth the Computational Cost? ([n. d.]).

[54] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2022. Improving Language Understanding by Generative Pre-Training. (2022).

[55] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs] doi:10.48550/arXiv.1910.10683

[56] Edwin Rijgersberg. 2025. *Het einde van GEITje 1*. https://goingdutch.ai/nl/posts/geitje-takedown/ Section: posts.

[57] SecureDataService GmbH. 2023. Artikel 25 EU-AVG: Gegevensbescherming door ontwerp en door standaardinstellingen. `Privacy-Regulation.eu`. https://www.privacy-regulation.eu/nl/artikel-25-gegevensbescherming-door-ontwerp-en-door-standaardinstellingen-EU-AVG.htm Laatste wijziging volgens site: 04-04-2023..

[58] SecureDataService GmbH. 2023. Artikel 32 EU-AVG: Beveiliging van de verwerking. `Privacy-Regulation.eu`. https://www.privacy-regulation.eu/nl/artikel-32-beveiliging-van-de-verwerking-EU-AVG.htm Laatste wijziging volgens site: 04-04-2023. Geraadpleegd op <vul-hier-je-consultatiedatum-in>.

[59] SecureDataService GmbH. 2023. Artikel 5 EU-AVG: Beginselen inzake verwerking van persoonsgegevens. `Privacy-Regulation.eu`. https://www.privacy-regulation.eu/nl/artikel-5-beginselen-inzake-verwerking-van-persoonsgegevens-EU-AVG.htm Laatste wijziging volgens site: 04-04-2023. Geraadpleegd op 6-12-2025.

[60] KLINKENDE TAAL. [n. d.]. Is het B1? https://ishetb1.nl/. Accessed: 2025-12-10.

[61] Taalunie. 2022. Monitor Begrijpelijkheid overheidsteksten 2022. https://taalunie.org/publicaties/218/monitor-begrijpelijkheid-overheidsteksten-2022.

[62] Daniel Vlantis, Iva Gornishka, and Shuai Wang. 2024. Benchmarking the Simplification of Dutch Municipal Text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. European Language Resources Association (ELRA), Torino, Italy, 2217–2226. https://aclanthology.org/2024.lrec-main.199/

[63] Tong Wang, Ping Chen, Kevin Amaral, and Jipeng Qiang. 2016. An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification. arXiv:1609.03663 [cs] doi:10.48550/arXiv.1609.03663

[64] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A Survey of Human-in-the-loop for Machine Learning. 135 (2022), 364–381. arXiv:2108.00941 [cs] doi:10.1016/j.future.2022.05.014

[65] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A Survey of Human-in-the-loop for Machine Learning. 135 (2022), 364–381. arXiv:2108.00941 [cs] doi:10.1016/j.future.2022.05.014

[66] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics (TACL)* 4 (2016), 401–415. doi:10.1162/tacl_a_00107

[67] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[68] Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating Transformer and Paraphrase Rules for Sentence Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3164–3173. doi:10.18653/v1/D18-1355

[69] Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating Transformer and Paraphrase Rules for Sentence Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3164–3173. doi:10.18653/v1/D18-1355

A Informatie vanuit OVER-gemeenten

A.1 Kennismaking OVER-gemeenten (13 nov)

Op 13 november bezochten drie medewerkers van de WMO-afdeling van de OVER-gemeenten de Hogeschool van Amsterdam om het project over tekstversimpeling toe te lichten waar studenten van de Master Applied AI aan gaan werken. De OVER-gemeenten is de ambtelijke organisatie voor de Gemeente Oostzaan en Wormerland. Het doel van dit project is medewerkers te ondersteunen bij het opstellen van brieven op B1-niveau.

De bijeenkomst startte met een presentatie van Annamaria van Teeseling, waarna de studenten de gelegenheid kregen om vragen te stellen aan de drie aanwezige medewerkers. De belangrijkste notities naar aanleiding van deze kennismaking zijn hieronder vastgelegd.

A.1.1 Introductie van Annamaria van Teeseling voer het project.

Introductie medewerkers OVER-gemeenten

- Annamaria van Teeseling werkt voor de ambtelijke organisatie OVER-gemeenten (voor Gemeente Oostzaan en Wormerland) en is teamleider van de afdeling WMO.
- Franny van der Bijl is WMO-consulent bij OVER-gemeenten. Zij houdt zich onder andere bezig met het herschrijven van brieven volgens de interne schrijfwijzer.
- Michael is projectleider bij de ICT-afdeling. Hij bewaakt de privacyaspecten binnen het gebruik van AI en is betrokken bij het onderzoeken van mogelijkheden voor veilige inzet van AI binnen de gemeente.

Wat doet de OVER-gemeenten?

De OVER-gemeenten is verantwoordelijk voor taken zoals openbare orde, zorg, infrastructuur en sociale voorzieningen. De besluitvorming ligt bij het college van B&W en de gemeenteraad, en de uitvoering bij de ambtelijke organisatie. Voorbeelden van werkzaamheden zijn:

- Ze zorgen voor huizenbouw
- Ze geven vorm aan de leefomgeving
- Ze lichten de raad in, helpen en steunen deze
- Ze schrijven beleid
- Ze behandelen WMO-aanvragen
- Ze vangen Oekraïners op
- Ze geven vergunningen af
- Ze voeren veel gesprekken met inwoners en versturen brieven

Bij brieven is het belangrijk dat de juridische inhoud begrijpelijk blijft, zonder dat deze te eenvoudig of kinderachtig overkomt.

Werken met AI in een gemeente, de uitdagingen

Herschrijven van teksten met AI is op dit moment niet toegestaan vanwege de bescherming van persoonsgegevens en andere vertrouwelijke informatie. Belangrijke aandachtspunten zijn:

- Privacy: bescherming van persoonsgegevens (AVG).
- Bedrijfsgevoelige informatie: risico op datalekken.
- Transparantie: AI-beslissingen moeten uitlegbaar zijn.
- Ethiek: gelijke behandeling en inclusie (bijvoorbeeld voorkomen van situaties zoals de toeslagenaffaire).

Hoe werkt de afdeling WMO?

De afdeling WMO voert de Wet maatschappelijke ondersteuning uit (hulp bij zelfredzaamheid). Taken zijn onder meer:

- Beoordelen van indicaties
- Bieden van ondersteuning
- Samenwerken met zorgaanbieders
- Zorgen voor begrijpelijke en toegankelijke communicatie met inwoners

Communicatie op B1-niveau

- Waarom B1? Teksten op B1-niveau zijn begrijpelijk voor ongeveer 80% van de inwoners (bron: CBS, volgens medewerkers).
- Onze ambitie: eenduidige en duidelijke communicatie.
- Hulpmiddel: interne schrijfwijzer.

De gemeente wil dat alle communicatie op dezelfde manier wordt opgesteld, zodat inwoners ervaren dat zij met één organisatie te maken hebben. Ter illustratie van de taalniveaus:

- A1: eenvoudig basisniveau (bijvoorbeeld taalgebruik voor op vakantie).
- A2: niveau dat vereist is voor inburgering.
- B1: helder en begrijpelijk voor een brede doelgroep.
- Etc.

Schrijfwijzer & standaard WMO-brieven

- De schrijfwijzer is opgesteld om te zorgen dat de communicatie consistent wordt.
- De standaardbrieven (van de WMO) moeten nog getoetst worden aan deze schrijfwijzer.

Opdrachten voor de studenten

- De gemeente vraagt de studenten om een slimme AI-oplossing te ontwikkelen:
  - Die medewerkers eenvoudig ondersteunt bij het schrijven van brieven volgens de schrijfwijzer,
  - Gebruikmakend van meegeleverde voorbeeldbrieven,
  - Rekening houdend met gemeentelijke kaders en privacyrichtlijnen.

A.1.2 Vragen en antwoorden met betrekking tot het onderzoek.

Stakeholders

- **Vraag:** Wat valt er onder 'OVER'-gemeenten?
- **Antwoord:** OVER-gemeenten is de ambtelijke organisatie voor de Gemeente Oostzaan en Wormerland.

Inhoud van brieven & misverstanden van burgers

- **Vraag:** Welke onderwerpen bevatten de brieven?
- **Antwoord:** De brieven gaan onder andere over schulden, uitkeringen, werk en inkomen, hulp bij het huishouden, en informatie voor Oekraïense vluchtelingen.
- **Vraag:** Welke problemen ontstaan door miscommunicatie? (Met miscommunicatie wordt bedoeld dat burgers inhoud van de brief niet begrijpen.)
- **Antwoord:**
  - Het gaat om een groot aantal brieven; binnen de WMO-afdeling alleen al ongeveer 250.
  - Daarnaast is schrijven een vak. Consulenten moeten naast hun reguliere werkzaamheden ook nog goed kunnen schrijven. De teksten moeten niet alleen juridisch dekkend zijn volgens wettelijke verplichtingen, maar ook begrijpelijk.
  - Daarnaast is de situatie van elke inwoner persoonlijk, dus daar moeten brieven op worden aangepast. (Een kanttekening bij het antwoord is dat de vragen zijn beantwoord door mensen van het team WMO met kennis in het sociaal domein.)
  - Wetgeving verplicht de gemeente tot het gebruik van begrijpelijke taal.
  - Uit data van het CBS blijkt dat veel mensen moeite hebben met het begrijpen van complexe teksten.
  - Laaggeletterden melden zich niet bij de gemeente, waardoor het probleem niet, waardoor het probleem niet direct zichtbaar is.
  - Er is een experiment gestart waarbij inwoners stickers op brieven kunnen plakken bij tekst die voor hen onduidelijk is. Dit staat nog in de kinderschoenen.
- **Vraag:** Welke doelgroepen hebben de meeste moeite met het begrijpen van brieven?
- **Antwoord:** Burgers met schulden, een uitkering en Oekraïners.

Niveau van de brieven

- **Vraag:** Kunnen brieven ook op andere niveaus dan B1 worden geschreven?
- **Antwoord:** Ja. Het taalniveau kan worden aangepast, bijvoorbeeld naar A1 als dit beter past bij de situatie van de inwoner. Soms kan het niveau juist hoger worden gezet, zoals B2.
- **Vraag:** Heeft de gemeente voorkeur voor A1 of B1, als A1 ook technisch mogelijk is?
- **Antwoord:** B1.
- **Vraag:** Welke tools worden op dit moment gebruikt?
- **Antwoord:** Medewerkers werken voornamelijk in Word, Outlook en een eigen klantsysteem genaamd Zorgnet. Word wordt het meest gebruikt.
- **Vraag:** Maken medewerkers nu gebruik van AI-tools, zoals ChatGPT?
- **Antwoord:** Nee. Het herschrijven van brieven met AI is niet toegestaan vanwege bescherming van persoonlijke gegevens.
- **Vraag:** Zijn er al conclusies uit de lopende pilot met de schrijfwijzer?
- **Antwoord:** De eerste pilotbrieven zijn ontvangen, maar over de resultaten waren ze nog niet tevreden. Franny heeft de brieven zelf nog moeten herschrijven.
- **Vraag:** Wie bepaalt nu of een tekst begrijpelijk is voor iedereen en hoe wordt de schrijfwijzer toegepast in de huidige situatie?

- **Antwoord:**
  - Medewerkers bepalen zelf of een tekst begrijpelijk is voor iedereen. Zij hebben een training gehad en toetsen elkaars stukken door brieven onderling na te lezen. Volledige garantie dat een tekst op B1-niveau is, bestaat echter niet.
  - De teamleider gaf aan dat de voorkeur uitgaat naar minder toetsing. Zij wil liever vertrouwen op de deskundigheid van de medewerkers, zonder dat onderlinge controle altijd nodig is.
- **Vraag:** Waarom is Copilot nog niet genoeg?
- **Antwoord:** Vanwege de privacy, servers, etc.
- **Vraag:** Worden er al al tools gebruikt voor het begrijpelijker maken van brieven?
- **Antwoord:** Ja. Tools zoals ishetb1.nl, maar eigenlijk werkt dit niet ideaal en werken ze met hun eigen kennis, kennis van een training en de schrijfwijzer.
- **Vraag:** Moeten medewerkers nog rekening houden met andere wetgeving dan de AVG, etc.?
- **Antwoord:** Nee.

Context waarom de gemeente een hulpmiddel wil

- **Notities:**
  - Medewerkers willen niet continu moeten kijken naar de schrijfwijzer. (Deze is namelijk is gigantisch en niet uit het hoofd te leren)
  - Er is geen tijd bekend over hoelang medewerkers gemiddeld bezig zijn met het schrijven van een brief. Medewerkers van de OVER-gemeenten hebben in principe geen tijdslimiet voor het schrijven van brieven. Sommige gemeenten hebben dat wel.
  - De gemeente wil voorkomen dat het proces geautomatiseerd wordt. Medewerkers moeten de controle behouden over de tekst.

Data en beschikbaarheid van brieven

- **Vraag:** Is er gebruikersdata over hoe inwoners brieven ervaren?
- **Antwoord:** Nee, er zijn geen gegevens bekend over wat mensen als onduidelijk zien.
- **Vraag:** Kunnen er brieven van andere afdelingen worden gedeeld?
- **Antwoord:** Daar gaan we achteraan.
- **Notities:**
  - De brieven die we hebben gekregen zijn standaard teksten. Tussen rode haken moet dus nog tekst geplaatst worden. Dit zijn een soort templates. Binnen de hele gemeente moeten alle afdelingen brieven schrijven op B1 niveau.
  - De brieven die we hebben gekregen zijn standaard teksten. Tussen rode haken moet dus nog tekst geplaatst worden. Dit zijn een soort templates. Binnen de hele gemeente moeten alle afdelingen brieven schrijven op B1 niveau.
  - In de beschikkingen zijn de juridische stukjes hetzelfde in elke brief. De gedeeltes over beschikkingen zijn variabel en worden per burger en per werknemer anders geschreven.

Modelkeuze (AI-technisch stuk

- **Vraag:** Wordt het aangeraden om een model te trainen vanaf scratch?
- **Antwoord:** Nee. Dit heeft te maken met de grootte van de modellen. Ze zijn te groot om zelf te trainen. Het wordt dus aangeraden om een voorgetraind model te finetunen op eigen data. Je moet dan wel verantwoorden waarom je een bepaalde architectuur kiest. Het mag dus wel. Finetunen hoeft ook niet. Onthoud dat een NLP echt niet alleen taalmodellen zijn.
- **Notities:**
  - Het zou een nice to have zijn als het herleidbaar is waar gegenereerde content op gebasseerd is en hoe het model werkt.
  - Als tekst vereenvoudigd wordt, moet deze niet ineens 4 pagina's lang worden. Ook moet belangrijke informatie niet verdwijnen en moet het voor inwoners wel duidelijk zijn dat ze bijvoorbeeld in bezwaar kunnen gaan.
  - De tool moet door alle afdelingen gebruikt kunnen worden.

Interface

- **Vraag:** Heeft de gemeente voorkeur voor een bepaald soort interface?
- **Antwoord:** Ze werken nu voornamelijk met Firefox en sinds kort ook weer Edge (tijdelijk niet door security breach in AI-functies). Ze hebben geen algemene voorkeur voor een ideale situatie.
- **Vraag:** Willen medewerkers tijdens het schrijven al suggesties zien van bijvoorbeeld synoniemen, of pas later bij het scannen?
- **Antwoord:**
  - Ze hebben nu al een tool die dit doet. Mensen vinden het juist makkelijk dat er tussendoor al voorstellen komen (aldus Michael).
  - Michael beschreef hoe hij zelf met een tool heeft gewerkt die exact klonk als wat overleaf doet en vond het een zeer fijne tool.
  - Er moeten niet te veel meldingen tegelijkertijd in beeld komen. Niet te veel streepjes en kleurtjes tegelijkertijd (aldus Franny).
  - Het liefst zien ze dit dus al tijdens het schrijven en niet pas na het inscannen.
  - De tool moet niet te druk worden (aldus Franny).
  - Het liefst willen ze ook iets van de tool leren (aldus Franny). Ze ziet dus liever suggesties in plaats van dat automatisch een hele brief wordt herschreven.

Model lokaal draaien/technische eisen

- **Vraag:** Moet de tool lokaal kunnen draaien?
- **Antwoord:** Het liefst wel, zodat je ook met persoonlijke gegevens kan werken.
- **Notities:**
  - Het zou fijn zijn als de tool offline zou kunnen runnen.
  - Het is interessant en relevant om duurzaam te werk te gaan zover het kan, probeer hier rekening mee te houden.

Contact & beschikbaarheid

- **Vraag:** Kunnen we jullie bereiken voor vragen tijdens het project?
- **Antwoord:** Hiervoor zijn afgesproken data.

Overige vragen

- **Vraag:** Is er momenteel een feedbackloop?
- **Antwoord:** Nee, nogmaals deels door schaamte. We proberen een inclusiebeleid op te zetten. We willen meer nadruk leggen op dat iedereen de brieven begrijpt.
- **Notities:**
  - Er zijn andere gemeenten zie apps gebruiken (volgens Pim). Gemeente OVER gebruikt dat niet. Ze worden getraind en dan losgelaten op de werkvloer.

A.2 Vragen en antwoorden OVER-gemeenten (21 nov)

Om meer inzicht te krijgen in de huidige situatie en context van de OVER-gemeenten konden alle projectgroepen tot 21 november vragen indienen. De docenten selecteerden vervolgens de vragen om doublures te voorkomen. Medewerkers van de OVER-gemeenten hebben deze vragen beantwoord. De antwoorden zijn hieronder te vinden.

Q1: Projectomvang – Hoe belangrijk is het om schrijvers een assistent te bieden die hen ondersteunt tijdens het schrijfproces?

Het bieden van een assistent kan voor sommige collega's helpend zijn, omdat er direct feedback wordt gegeven. Wij hopen/verwachten dat collega's hier snel van kunnen leren. Wij kunnen ons voorstellen dat dit niet voor alle collega's prettig werkt en dat zij hier niet allemaal gebruik van zullen maken.

Wij vinden daarom een oplossing die achteraf het stuk verbetert en aanpast conform de schrijfwijzer ook heel prettig. Mogelijk met optie voor functie aan- of uitzetten naar wens van de gebruiker?

Q2: Projectomvang – Hoe belangrijk is het om te beoordelen of brieven voldoen aan de vereiste leesbaarheidsnormen?

Het is heel belangrijk om te beoordelen of brieven voldoen aan de vereiste leesbaarheidsnormen. Gemeentelijke brieven gaan vaak over zaken die direct invloed hebben op het leven van inwoners, zoals belastingen, vergunningen, ondersteuning of regelingen. Als de tekst te ingewikkeld is, kunnen mensen de informatie verkeerd begrijpen of belangrijke stappen missen. Dat kan leiden tot onzekerheid, fouten of zelfs problemen.

Door te controleren of een brief op B1-niveau is geschreven, weet je zeker dat de meeste inwoners de tekst goed kunnen volgen. Dit maakt de communicatie eerlijker, duidelijker en toegankelijker voor iedereen. Bovendien scheelt het tijd en vragen voor de gemeente, omdat inwoners sneller begrijpen wat er van hen wordt verwacht.

Q3: B1-niveau & Begrijpelijke Taal – Worden er momenteel Engelse woorden in brieven gebruikt, of uitsluitend Nederlandse?

Uitsluitend Nederlands.

Q4: Huidig Schrijfproces en Workflow – Hoe verloopt de volledige procedure van het schrijven van een brief (van aanvraag tot verzending)?

Dat kan verschillen, veel afdelingen werken met standaardbrieven, waarbij veel al voorgeschreven is en de medewerker alleen kleine aanpassingen hoeft te doen die van toepassing zijn op de situatie van de inwoner.

Echter zijn er ook teams, denk aan juridische zaken, die brieven of stukken moeten schrijven die direct op de vraag aansluiten; hierbij moet de medewerker zelf een stuk schrijven.

Standaardbrieven worden vaak op voorhand door een medewerker van de afdeling gemaakt. Ook is het per team verschillend of brieven getoetst worden; daar hebben we als organisatie geen eenduidige werkwijze in.

Q5: Huidig Schrijfproces en Workflow – Welke rollen zijn betrokken bij het schrijfproces? (bijv. consulent, juridisch medewerker, kwaliteitsmedewerker, communicatie)

Alle rollen, door de hele organisatie. Ook beleid, vergunningen etc.

Q6: Huidig Schrijfproces en Workflow – Wat zijn de huidige pijnpunten of vertragingen in het proces? (bijv. juridisch taalgebruik, inconsistentie, veel correctierondes)

Niet op een eenduidige manier communiceren, onbekend met de werkwijze, veel tijd om brieven van anderen te toetsen/controleren.

Q7: Kwaliteitscontrole & Werkinstructies – Welke specifieke onderdelen van de schrijfwijzer zijn de grootste knelpunten in het huidige gebruik?

Het is een groot document, waarvan wij denken dat deze in de praktijk bij weinig mensen bekend is. Hierdoor is het gevaar dat dit een papieren werkelijkheid is en dat de medewerkers allemaal op eigen wijze door blijven communiceren.

Wij zoeken een tool waarmee we op een meer eenduidige manier, conform afspraken, kunnen communiceren.

De schrijfwijzer geeft ons, naast de schrijfstijl, ook veel detailrichtlijnen voor een eenduidige werkwijze, zoals de manier van een telefoonnummer noteren en het gebruiken van opsommingstekens. Om dit allemaal handmatig te wijzigen bij de huidige brieven zal ons enorm veel tijd kosten. De AI-tool moet ervoor zorgen dat dit voor elke brief (en dus elke afdeling) gelijk is.

Q8: AI-gebruik, Behoeften & Beslissingen – Wat is volgens jullie de ideale werkwijze voor het schrijven/evalueren van brieven met AI?

Het liefst hebben we een tool die het schrijven makkelijk maakt voor de schrijver (tijdwinst) en die de brieven zodanig toetst dat dit niet door een (tweede) collega hoeft te gebeuren.

We zoeken een tool voor zowel het herschrijven van huidige brieven, als een tool die ondersteunend is bij het schrijven van nieuwe brieven. Een tool die ons hierbij ondersteunt qua schrijfwijze (taalgebruik, zinsopbouw) en die het ons ook "leert", zodat wij zien waar we vaak "de fout" in gaan qua schrijfstijl.

Opmerking van Maikel: Mee eens, als de grootheid van het document in kleine stappen kan worden gerefereerd zal dat helpen. Ter illustratie: in de tooling voor het digitaal anonimiseren wordt er aan het einde van het document een grondslag gegeven voor de anonimisering. Zie Q6. Bijvoorbeeld: "Gegevens geanonimiseerd o.b.v. grondslag Financiële gegevens B.6.1".

Q9: AI-gebruik, Behoeften & Beslissingen – Is er binnen de gemeente of elders al onderzoek gedaan naar AI-systemen om teksten begrijpelijker te maken?

Binnen de pilot M365 Copilot zijn er zeker mogelijkheden om teksten te redigeren naar begrijpelijkere leesniveaus. Dit biedt zeker standaard al aardige mogelijkheden, maar dat is niet gekoppeld aan de schrijfwijzer. Juist die combinatie zou de tool extra zinvol maken.

A.3 Vragen en antwoorden Gemeente Oostzaan (6 jan)

Om meer inzicht te krijgen in de huidige situatie en context van de OVER-gemeenten konden alle projectgroepen tot 6 januari vragen indienen. De docenten selecteerden vervolgens de vragen om doublures te voorkomen. Medewerkers van de OVER-gemeenten hebben deze vragen beantwoord. De antwoorden zijn hieronder te vinden.

Q1: Vinden jullie het goed als deze woordenlijsten worden toegepast in de toepassing voor het herschrijven van de gemeentelijke brieven?

- **Juridisch jargon:** https://www.amsterdam.nl/schrijfwijzer/heldere-taal-basis-onze-huisstijl/helder-juridisch-schrijven
- **Synoniemen woordenlijst:** https://b1teksten.nl/artikel/voorbeelden-van-b1-teksten en https://site.dijkenwaard.nl/huisstijl/heldere-taal/direct-duidelijk-dijken-waard/taalniveau-b1
- **Moeilijke woorden:** https://www.amsterdam.nl/schrijfwijzer/moeilijke-woorden
- **Inclusieve lijst:** https://www.amsterdam.nl/schrijfwijzer/inclusieve-taal-richtlijnen-tips/inclusieve-woordenlijst/

Bedoelen jullie hiermee dat het jargon uit deze woordenlijsten wordt goedgekeurd en dus niet als suggestie wordt aangegeven om te verbeteren naar B1? Als dat het geval is, dan zou ik zeggen: Nee liever niet.
Deze woordenlijsten zijn erg uitgebreid en een hoop kan zeker aangepast worden naar B1.
Het is fijner als wij zelf de mogelijkheid hebben om in de AI-tool woorden toe te voegen die hij moet "negeren". Dit kan namelijk ook per afdeling verschillen ivm vaktermen. Als we al deze woorden zouden toevoegen en accepteren dan blijft er naar mijn idee te veel moeilijke taal in staan.
Mocht je deze websites echter willen gebruiken om om te zetten in B1/synoniemen, dan hebben we het liefste dat je gebruik maakt van het advies van de stichting lezen en schrijven, zij adviseren de volgende websites:

- Eenvoudige taal
- Wil je weten of een woord op B1-niveau is? Gebruik de website https://www.ishetb1.nl
- Zoek eenvoudige woorden op taalniveau A2 en B1 via https://www.zoekeenvoudigewoorden.nl

Q2: Welke specificaties heeft de server waarop het model kan runnen? (Dan weten wij hoe zwaar het model kan worden)

- OVG heeft geen pre-designed AI-server specificaties, maar:
  - Het moet een Windows Server 2019 of hoger zijn
  - Deze moet virtueel gehost kunnen worden
  - Deze moet voorzien kunnen worden van policies ten behoeve van security
  - Deze moet M365 Copilot kunnen draaien
- OVG heeft aangegeven dat een Word-plugin en/of in combinatie met een Copilot Agent de voorkeur heeft boven een dedicated server
- Ik denk dat het zinvoller is om eerst een model neer te zetten, waarna aan OVG wordt aangegeven wat de specificaties voor een server zouden moeten zijn om dat model efficiënt te laten draaien (dus omgekeerd)
- Daarnaast is het niet gezegd dat OVG per definitie akkoord gaat met een dedicated server
  - Een AI-server moet een zinvolle en aantoonbare kosten-batenanalyse hebben (kosten niet ten laste van het ICT-budget)
  - Een voorstel voor een AI-server zal worden beoordeeld op de criteria van BIO2, Common Ground (VNG) en gemeentelijke architectuurprincipes en security-overwegingen, waarbij de CISO een kritieke rol heeft

Q3: Welke fouten in een tekst vindt u zo ernstig dat deze altijd moeten leiden tot een lage beoordeling volgens de schrijfwijzer? (Denk bijvoorbeeld aan te lange zinnen, een onjuiste lay-out of onnodig moeilijke formuleringen.

- Te lange zinnen
- Geen of weinig alinea's
- Geen kernkopzinnen
- Onnodig moeilijke formulering
- Spel- en grammaticafouten

Q4: Wat moet de schrijfcoach strenger beoordelen: Zinnen die al duidelijk en eenvoudig zijn, maar ten onrechte als 'te complex' worden aangemerkt of zinnen die eigenlijk ingewikkeld zijn, maar door de schrijfcoach als 'eenvoudig genoeg' worden gezien?

Zinnen die eigenlijk ingewikkeld zijn, maar door de schrijfcoach als eenvoudig genoeg worden gezien.

Q5: Wanneer zou u vertrouwen hebben in het oordeel van een AI-schrijfcoach die beoordeelt of het voldoet aan de schrijfwijzer?

Als ik tips krijg waar ik mij in herken, en die mij helpen mijzelf en mijn teksten te verbeteren. Als zowel ik als de AI-schrijfcoach van elkaar kunnen leren, om zo veel voorkomende "fouten" te voorkomen van beide kanten.

Q6: De AI-schrijfcoach geeft een cijfer als beoordeling. Wat maakt een tekst volgens u een 10 en wat een 6?

Een dikke 10 als alles volgens de schrijfwijzer is (duidelijke tekst, kernkopzinnen, alle details volgens de richtlijnen, geen spel/grammatica fouten). Een 6 wanneer de brief geen spel/grammatica fouten heeft, maar nog wel ingewikkeldere taal gebruikt.

Q7: Welke juridische termen mogen niet worden herschreven?

Daar kan ik zo geen antwoord op geven, er zou misschien een waarschuwing kunnen komen dat deze tekst niet voldoet aan B1, om … reden en dat het aan de schrijver is om in te schatten of dit wel of niet onvermijdelijk is.

Handig zou zijn als we zelf de optie hebben om deze termen te kunnen toevoegen. BV bij Word kan je ook woorden toevoegen aan de woordenlijst of er voor kiezen een woord te negeren.

Q8: Hoe vaak komt het voor dat er data van de bewoner lopend/per ongeluk in de brief terecht komt?

Dit zijn incidenten, vaak wordt vanuit een systeem gewerkt en niet vanuit standaard brieven, wij proberen dit te voorkomen. Wel staan standaard de NAW gegevens in de brieven die gemaakt worden vanuit de systemen waar mee we werken. De brief wordt automatisch gegenereerd vanuit het systeem in Word. Hier passen we dan de inhoudt aan van de brief.

Q9: Hoe belangrijk is de herkomst van de dataset waarop getraind is?

Ik begrijp de vraag niet helemaal, voor de volledigheid geef ik onze kaders aan:

- OVG werkt met M365 Copilot en niet met andere AI's zoals ChatGPT, Perplexity, enz.
- De LLM van M365 Copilot mag gebruikt worden, maar de data mag **niet** gebruikt worden door de LLM.
- De herkomst van de dataset moet volledig transparant en dus duidelijk zijn.
- Als er data naar een dataset wordt gestuurd, is er zeer waarschijnlijk een verwerkersovereenkomst nodig en mogelijk zelfs een DPIA.

B Waardeproposities per stakeholder

Stakeholder 1: OVER-gemeenten

- **Conceptnaam:**
  AI-applicatie om brieven begrijpelijker te maken.
- **Door gebruik te maken van:**
  Een lokaal draaiend AI-systeem dat opties geeft om brieven te vereenvoudigen.
- **Om:**
  De werkdruk van medewerkers te verlagen en het schrijfproces van brieven te versnellen.
- **Wij kunnen helpen:**
  Door medewerkers te ondersteunen bij het herschrijven en controleren van brieven.
- **Met een betere manier om:**
  Brieven sneller begrijpelijk te maken volgens de OVER-gemeenten schrijfwijzer.
- **Zodat:**
  Medewerkers minder tijd kwijt zijn aan het begrijpelijker maken van tekst (volgens de schrijfwijzer) en meer tijd overhouden voor andere werkzaamheden.
- **Met:**
  Behoud van gemeentelijke kwaliteitsstandaarden, zonder afhankelijk te zijn van externe (onveilige) AI-platforms.

Stakeholder 2: Inwoners van de OVER-gemeenten

- **Conceptnaam:**
  AI-applicatie om brieven begrijpelijker te maken.
- **Door gebruik te maken van:**
  AI-ondersteunde taalcontrole die medewerkers suggesties geeft voor het vereenvoudigen en inclusief schrijven van tekst.
- **Om:**
  Inwoners duidelijkheid te bieden over hun rechten en acties die ze eventueel moeten ondernemen.
- **Wij kunnen helpen:**
  Door met behulp van de AI-applicatie medewerkers te ondersteunen bij het begrijpelijker, toegankelijker en inclusiever schrijven van brieven voor inwoners.
- **Met een betere manier om:**
  Misverstanden bij inwoners te voorkomen bij belangrijke onderwerpen.
- **Zodat:**
  (Alle) inwoners begrijpen wat ze moeten weten of doen.
- **Zonder:**
  Moeite of stress.

Stakeholder 3: Medewerkers die brieven schrijven of controleren

- **Conceptnaam:**
  AI-applicatie om brieven begrijpelijker te maken.
- **Door gebruik te maken van:**
  AI-ondersteunde taalcontrole die medewerkers suggesties geeft voor het vereenvoudigen en inclusief schrijven van tekst.
- **Wij kunnen helpen:**
  Door verbeteringen voor te stellen die het B1-taalniveau waarborgen.
- **Met een betere manier om:**
  Sneller teksten te controleren, herschrijven en verduidelijken.
- **Zodat:**
  Medewerkers hun werk efficiënter kunnen uitvoeren.
- **Met:**
  Automatische verwijzingen naar de bijbehorende regel uit de schrijfwijzer, zodat medewerkers precies weten waarop de suggestie is gebaseerd.

Stakeholder 4: ICT-afdeling

- **Conceptnaam:**
  AI-applicatie om brieven begrijpelijker te maken.
- **Door gebruik te maken van:**
  Een lichtgewicht, lokaal te hosten AI-applicatie.
- **Om:**
  De gemeente een veilige, controleerbare oplossing bieden.
- **Wij kunnen helpen:**
  Door een systeem te leveren dat past binnen bestaande infrastructuur.
- **Met een betere manier om:**
  AI in te zetten zonder risico op datalekken of afhankelijkheid van derde partijen.
- **Zodat:**
  De ICT-afdeling in controle blijft van de gegevens en beveiliging ervan.
- **Met:**
  Lokale dataverwerking.

Stakeholder 5: Overheid (wet- en regelgevers)

- **Conceptnaam:**
  AI-applicatie om brieven begrijpelijker te maken.
- **Door gebruik te maken van:**
  AI-ondersteunde taalcontrole die medewerkers suggesties geeft voor het vereenvoudigen en inclusief schrijven van tekst.
- **Om:**
  Overheidscommunicatie begrijpelijker en toegankelijker te maken.
- **Wij kunnen helpen:**
  Door gemeenten te ondersteunen bij het naleven van richtlijnen voor begrijpelijke communicatie naar inwoners.
- **Met een betere manier om:**
  Burgers hun rechten te laten begrijpen of acties die ze moeten ondernemen.
- **Zodat:**
  Overheidscommunicatie transparant blijft en voldoet aan nationale richtlijnen, zoals Direct Duidelijk en de OVER-gemeente schrijfwijzer.
- **Met:**
  Naleving van taal- en privacyrichtlijnen.

C Requirements

C.1 MoSCoW-analyse

De MoSCoW-analyse is uitgevoerd op alle ethische, juridische, organisatorische, functionele en technische vereisten. M staat voor *must have*, S staat voor *should have*, C staat voor *could have* en W staat voor *won't have*. De nummers bij een requirement (RQ) staan voor het type requirement. ER = ethische requirement, JR = juridische requirement, OR = organisatorische requirement, FR = functionele requirement, TR = technische requirement en DR = duurzaamheids requirement.

C.2 Ethische requirements

Table 21. Ethische requirements

RQ-nr	Requirement	MoSCoW	Stakeholder
ER05	De AI-tool moet worden beveiligd tegen kwetsbaarheden (bijv. hacks of technische storingen) [20].	S	ICT-afdeling; Overheid; Inwoners van Gemeente Oostzaan
ER06	De AI-tool moet een uitwijkplan hebben voor situaties met problemen zoals hacks of technische storingen [20].	S	ICT-afdeling; Medewerkers die brieven schrijven/controleren
ER07	De stakeholder moet worden geïnformeerd dat een AI-tool is gebruikt bij het schrijven van de brief [20].	C	Inwoners van Gemeente Oostzaan
ER08	Na installatie van de AI-tool moet regelmatig feedback worden gevraagd aan belanghebbenden [20].	W	Medewerkers die brieven schrijven/controleren; Inwoners van Gemeente Oostzaan
ER09	De gevolgen van het systeem op sociaal gebied moeten onderzocht worden [20].	W	Inwoners van Gemeente Oostzaan; Gemeente Oostzaan

C.3 Organisatorische requirements

Table 22. Organisatorische requirements

RQ-nr	Requirement	MoSCoW	Stakeholder
OR01	De Gemeente Oostzaan moet het gebruik van AI verwerken in hun beleid [64].	W	Gemeente Oostzaan
OR02	De Gemeente moet periodieke evaluaties inplannen om te controleren of brieven correct worden geschreven [64].	W	Gemeente Oostzaan
OR03	De medewerkers moeten een training krijgen over hoe de AI-tool werkt voor het herschrijven van brieven [64].	W	Medewerkers die brieven schrijven/controleren
OR04	De gemeente moet een beleid opstellen waarin wordt vastgelegd wie de eindverantwoordelijke is voor beslissingen die met behulp van de AI-tool tot stand komen [64].	W	Gemeente Oostzaan
OR05	De gemeente moet duidelijk maken aan de medewerkers dat zij niet volledig moeten vertrouwen op de tekstsuggesties van de AI-tool [64].	W	Gemeente Oostzaan; Medewerkers die brieven schrijven/controleren
OR06	De gemeente moet een feedbackmechanisme implementeren waarmee medewerkers suggesties over de AI-tool kunnen indienen [64].	W	Gemeente Oostzaan; Medewerkers die brieven schrijven/controleren
OR07	De gemeente moet een risicoanalyse opstellen waarin wordt beschreven wat de risico's zijn van de implementatie van de AI-tool bij het herschrijven van brieven [64].	W	Gemeente Oostzaan; Medewerkers die brieven schrijven/controleren

C.4 Juridische requirements

Table 23. Juridische requirements

RQ-nr	Requirement	MoSCoW	Stakeholder
JR03	De AI-tool mag alleen worden gebruikt voor het herschrijven van brieven voor inwoners van de gemeente en niet voor andere doeleinden [20, 57, 59].	S	Overheden/toezichthouders
JR04	Er moet aangetoond kunnen worden waar de trainingsdata vandaan komt en dat er toestemming is voor gebruik hiervan [20, 22], .	S	Overheden/toezichthouders
JR05	Het model moet worden getest op een representatieve set brieven om mogelijke fouten te identificeren. Hierbij moet specifiek worden gecontroleerd op bias (o.a. geslacht) zodat geen enkele groep wordt benadeeld en discriminatie wordt voorkomen [31, 41].	S	Overheden/toezichthouders

C.5 Functionele requirements

Table 24. Functionele requirements

RQ-nr	Requirement	MoSCoW	Stakeholder
FR06	Resultaten moeten betrouwbaar zijn voor alle verschillende briefcategorieën.	S	OVER-gemeenten
FR07	De ontwikkeling van de AI-tool (datasets, modelkeuze, training) en de prestaties moeten gedocumenteerd worden [23].	Overheid; Medewerkers die brieven schrijven/controleren	Overheid: Wil inzicht in de documentatie om te controleren of het model voldoet aan wetgeving. Medewerkers die brieven schrijven/controleren: Moet geïnformeerd worden over de prestaties van het model.
FR08	De AI-tool biedt mogelijkheid om door een mens gecontroleerd te worden [64].	C	ICT-afdeling

C.6 Technische requirements

Table 25. Technische requirements

RQ-nr	Requirement	MoSCoW	Stakeholder
TR07	De AI-tool moet een representatieve dataset gebruiken [22].	S	ICT-afdeling; Overheden/toezichthouders
TR08	Het model moet bias-mitigatie technieken gebruiken, zodat het om kan gaan met incorrect beoordeelde resultaten [22].	S	ICT-afdeling; Overheden/toezichthouders
TR09	Het systeem moet beveiligd zijn tegen hacks door middel van netwerksegmentatie.	W	ICT-afdeling; Overheden/toezichthouders

C.7 Duurzaamheids requirements

Table 26. Duurzaamheids requirements

RQ-nr	Requirement	MoSCoW	Stakeholder
DR01	De AI-tool wordt zo ontwikkeld dat het energie- en datagebruik minimaal is door inzet van efficiënte modellen en infrastructuur, waarbij deze keuzes aantoonbaar bijdragen aan een zo laag mogelijke milieu-impact.	W	Overheden/toezichthouders; ICT-afdeling

D Testplan: maandag 1 december 2025

D.1 Introductie

Wij studeren met zijn vijven de Master Applied Artificial Intelligence hier op de Hogeschool van Amsterdam. Vanuit jullie, ook wel de OVER-Gemeenten, hebben wij de opdracht gekregen om een AI-tool te maken dat de gemeente kan helpen om begrijpelijker Nederlands te schrijven. In dit geval betekend begrijpelijker Nederlands ook wel het B1-niveau Nederlands, aangezien ruim 80% van de Nederlanders B1-niveau goed kan begrijpen.
Tijdens de vorige bijeenkomst gaven jullie aan dat er een schrijfwijzer is ontwikkeld (de OVER-schrijfwijzer), zodat medewerkers handvatten hebben om brieven begrijpelijker te schrijven. Daarbij werd vermeld dat deze schrijfwijzer vrij uitgebreid is en daardoor in de praktijk nog niet altijd goed wordt toegepast. Daarom is het idee ontstaan om medewerkers te ondersteunen met een AI-tool die is gekoppeld aan de OVER-schrijfwijzer, zodat teksten eenvoudiger en consistenter kunnen worden herschreven.
Voor dit onderzoek willen we graag jullie expertise gebruiken om te onderzoeken of het design van ons idee naar jullie wens is. We willen jullie graag ons prototype laten zien en zullen jullie gaan vragen om feedback te geven vanuit een werknemersperspectief.

D.2 Ons idee in het kort voor dit project

In ons project willen we een AI-model ontwikkelen dat de medewerkers van OVER-gemeenten ondersteunt bij het schrijven van brieven. De medewerker krijgt tijdens het typen suggesties gericht op taalniveau en zinsopbouw, om ervoor te zorgen dat de geschreven teksten versimpeld worden.
De AI-tool gaat suggesties geven op basis van wat hij geen B1-niveau Nederlands vindt, of wat hij bijvoorbeeld te complexe zinnen vindt volgens de OVER-schrijfwijzer. Bij deze suggestie zal dan ook doorverwezen worden naar de OVER-schrijfwijzer. Er is een knop aanwezig op het scherm, die het toelaat voor de medewerker om de AI-tool aan of uit te zetten. Hierdoor zal de medewerker niet gestoord worden tijdens het schrijven, tenzij hij of zij dat natuurlijk wel wil.
De extensie zal, als de AI-tool ingeschakeld is, rechts in beeld komen te staan. Hierbij zal per gevonden te lange zin of een te moeilijk woord, één pagina gebruikt worden. Bij 20 gevonden fouten in één document, zullen dus 20 doorklikbare pagina's beschikbaar zijn in de extensie. Zo blijven alle suggesties los van elkaar zichtbaar en kan er ook later teruggekomen worden op een suggestie.

D.3 Doel en taakverdeling

Het doel voor ons is dat wij meer te weten komen over de wenselijkheid en de uitlegbaarheid/ het vertrouwen in deze AI-tool. We willen testen of dit het gewenste prototype is, aangezien we hebben vernomen dat de medewerkers het fijn zouden vinden als de tool niet telkens in beeld springt, een overvloed aan kleurtjes heeft en over het algemeen afleidend is. We willen graag observeren wat jij van de AI-tool vindt en of je eventuele verbeterpunten ziet die zorgen voor een prettige werkervaring met deze AI-tool. Tijdens het testen wordt de rolverdeling aangehouden zoals weergegeven in de onderstaande tabel.

Naam	Taak
Amber	Aantekeningen maken
Amir	Aantekeningen maken
Kaan	Introductie geven
Terence	Het prototype uitleggen en ondersteunen tijdens de test
Bibiëne	Afwezig

Table 27. Taakverdeling tijdens de test

D.3.1 Uitleg door Terence tijdens de test.

Tijdens de test legt Terence uit hoe het prototype werkt. Hierbij gaat hij in op de volgende punten:

- Wat het prototype doet
- Het gebruik van de AI-knop (aan- en uitzetten)
- De extensie
- Verwijzingen per pagina binnen de extensie
- Verwijzingen naar de OVER-schrijfwijzer.

Vervolgens laat Terence de medewerker naar het prototype kijken en test hij of zij de aan- en uitknop van de AI-tool kan vinden. Wanneer deze knop wordt geactiveerd, verschijnt de extensie in beeld. Daarna laat Terence de testpersoon de AI-tool verder bekijken en wordt geobserveerd welke vragen of onduidelijkheden er ontstaan. Alle bevindingen worden vastgelegd en gedocumenteerd (Bijlage G).

D.4 Vragen aan de medewerkers van de OVER-gemeenten

D.4.1 Desirability (wenselijkheid).

- Voldoet de optie om de AI-tool aan en uit te zetten, aan jullie wensen?
- Zou je als medewerker prettig om kunnen gaan met de AI-tool? *(Wat zouden jullie eventueel anders willen?)*
- Is het wenselijk dat de AI-tool aangeeft dat er een suggestie is voor een bepaalde zin of een bepaald woord?
  - Zijn alle stappen wenselijk, overbodig of zijn er extra stappen nodig?
- Denk je dat je de AI-tool snel kunt leren gebruiken?
- Zijn er andere stappen in het prototype wenselijk die er op dit moment nog niet instaan?

D.4.2 Explainability & Trust (uitlegbaarheid & vertrouwen).

- Zijn de signalen van de AI-tool nuttig en voldoende uitgelegd om jouw werk te ondersteunen?
- Zou je deze AI-tool vertrouwen?
- Geeft de AI-tool informatie die jou helpt om beslissingen te nemen, of juist te veel/te weinig?
- Zou je te veel beïnvloed worden door de suggesties die de AI-tool geeft, dus zou je de AI-tool te veel vertrouwen?
- Vind je deze AI-tool vertrouwelijk in de gemeentelijke context?

E AI-breakdown & Error Flows

Tijdens het ontwerpproces is een AI-breakdown opgesteld en zijn error-flows gemaakt. De AI-breakdown en error-flows geven mogelijke fouten weer en de impact van deze fouten binnen de AI-tool. Daarnaast worden het percentage fouten weergegeven die wij nog acceptabel vinden. Ook worden mogelijke oplossingen voor deze fouten weergegeven.

In	Taak	Uit	Mogelijke fouten (FP / FN, hallucinaties, meaning drift, inconsistentie, geen output)	Impact van fouten low (recommended) medium (consequential) high (life-critical)	Percentage van fouten waarbij het nog waardevol blijft	Mogelijke oplossingen
Tekst in brief van de gemeente	Detecteren van moeilijke woorden	Markeringen bij woorden die als 'moeilijk' geclassificeerd worden.	**FP**: woorden markeren als `moeilijk' die al B1-niveau zijn (context verkeerd, eigennamen/afkortingen). **FN**: moeilijke woorden missen (samenstellingen, vervoegingen, jargon/beleidswoorden, spelfouten).	**Medium**: FP-markeringen kunnen leiden tot onnodige aanpassingen en vertraging in het schrijfproces. **High**: FN kan ertoe leiden dat kernproblemen onopgemerkt blijven, waardoor B1-niveau niet wordt gehaald.	20% omdat de medewerker markeringen kan negeren.	Medewerker kan suggesties ten alle tijde afkeuren/wegklikken (HITL) Blacklist/whitelist (jargonlijst), frequentie-/B1-woordenlijst en (optioneel) confidence-drempel toevoegen. Toon kort waarom een woord als moeilijk is gemarkeerd
Tekst in brief van de gemeente	Detecteren van zinscomplexiteit (te lang, veel bijzinnen, actief of passief geschreven)	Markering van complexe zinnen.	**FP**: zinnen markeren als complex puur door lengte, terwijl ze goed leesbaar zijn (bv. opsommingen). **FN:** complexe zinnen missen (bijzinnen, lijdende vorm, juridische constructies). **Span-fout:** verkeerde zinsdelen markeren (niet de echte oorzaak van complexiteit).	**Medium**: FP kan leiden tot onnodige herschrijfacties en vermindert vertrouwen in de tool. **High:** FN kan ertoe leiden dat complexe zinnen blijven staan, waardoor de brief niet op B1-niveau komt. **High:** Span-fouten kunnen tot verwarring leiden (medewerker begrijpt de fout niet en corrigeert verkeerd).	20% omdat markeringen en suggesties door de medewerker genegeerd kunnen worden.	Duidelijk vermelden waarom een zin complex is (bv. lijdende vorm, aantal bijzinnen, lengte). Markeer (indien mogelijk) alleen het relevante zinsdeel en laat medewerker de markering corrigeren/afkeuren (HITL).
Tekst in brief van de gemeente	Checken van toon en inclusiviteit (o.a. vermijden van afstandelijke of stigmatiserende termen)	Feedback op toon en suggesties voor alternatieve formuleringen.	**FP**: neutrale woorden onterecht als `niet-inclusief/stigmatiserend' markeren (context/bedoeling verkeerd). **FN:** subtiele niet-inclusieve formuleringen missen. **Generatieve fout**: Alternatief voorstel dat de betekenis/tone-of-voice ongewenst verschuift (te hard/te informeel)	**Low**: FP is meestal herkenbaar en kan door medewerkers gecorrigeerd worden. **Medium**: generatieve fouten of FN kan leiden tot minder passende formuleringen of gemiste verbeterkansen, wat kwaliteit en begrijpelijkheid verlaagt.	10% omdat medewerkers suggesties mogen negeren.	Randvoorwaarden toevoegen (vaste lijst + contextregel(s) en toon kort de reden van markering. Laat medewerker altijd kiezen tussen meerdere alternatieven of afkeuren (HITL).
Tekst in brief van de gemeente	Controleren of tekst voldoet aan specifieke regels van de OVER-schrijfwijzer.	Verwijzing naar relevante regel/passage uit de schrijfwijzer waarom zinnen/woorden aangepast dienen te worden.	**FP**: verwijzen naar een regel die niet van toepassing is (irrelevante retrieval). **FN:** geen passende regel tonen terwijl er wel een overtreding is (retrieval faalt) of te algemene verwijzing geven. **Hallucinatie**:AI noemt een regel/bron die niet in de schrijfwijzer staat. **Geen output/timeout:** geen verwijzing beschikbaar door fout/timeout.	**Low**: FP is vervelend maar kan genegeerd worden. **Medium**: FN/te algemene verwijzing kan ertoe leiden dat noodzakelijke aanpassingen niet worden gedaan. **High**: hallucinaties kunnen leiden tot onterechte regels en daarmee tot verkeerde beslissingen of onjuiste onderbouwing.	20% omdat de medewerker de verwijzing kan negeren en zelf kan controleren in de schrijfwijzer.	Toon altijd een snippet/citaat van de relevante passage + link naar sectie (verkleint hallucinatie-impact). Als er geen regel gevonden wordt: expliciete fallback-zin ("Geen passende regel gevonden in de schrijfwijzer"). Retry bij timeout + loggen/monitoren van retrieval-fouten.
Tekst in brief van de gemeente	Meedere suggesties genereren voor herschrijven volgens B1-niveau	Knop 'Nieuwe suggestie' die nieuwe zins- en woordsuggesties per zin/woord voorstellen.	**Meaning drift**: betekenis verandert (juridische implicaties/voorwaarden/nuances verdwijnen). **Hallucinatie**: AI introduceert info die niet in de teskt staat. **Omissie**: essentiële details (bedragen, data, verplichtingen) verdwijnen. **Geen output/timeout**: geen herschrijfsuggestie terwijl dit wel nodig is. **Inconsistentie:** verschillende output per run (Nieuwe suggestie geeft tegenstrijdige adviezen).	**High**: betekenisverandering/omissie/additie kan juridische of inhoudelijke strekking aanpassen (risico op verkeerde besluitvorming of communicatie). **High**: hallucinaties kunnen verwarring veroorzaken en het schrijfproces verstoren. **High**: geen output/timeout kan ertoe leiden dat noodzakelijke vereenvoudiging uitblijft (B1 niet gehaald). **High**: tegenstrijdige output vergroot de kans op een (juridisch risicovolle) verkeerde keuze en maakt resultaten minder betrouwbaar/repliceerbaar.	25% omdat de medewerker de suggestie of gehele tekst kan negeren.	Medewerker kan suggesties te allen tijde afkeuren (HITL). Prompt-constraints: behoud betekenis, behoud specifieke gegevens/getallen, geen nieuwe feiten. Toon verschillen (highlight wijzigingen) en bied 2–3 alternatieven.

Fig. 22. AI-breakdown

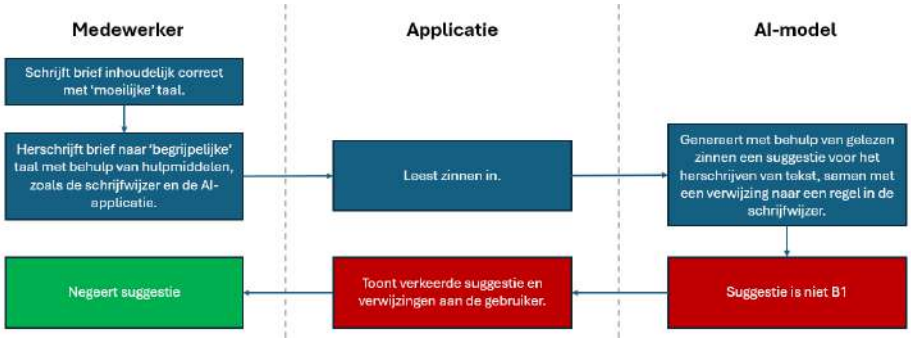


Fig. 23. Gegenereerde suggestie is niet B1

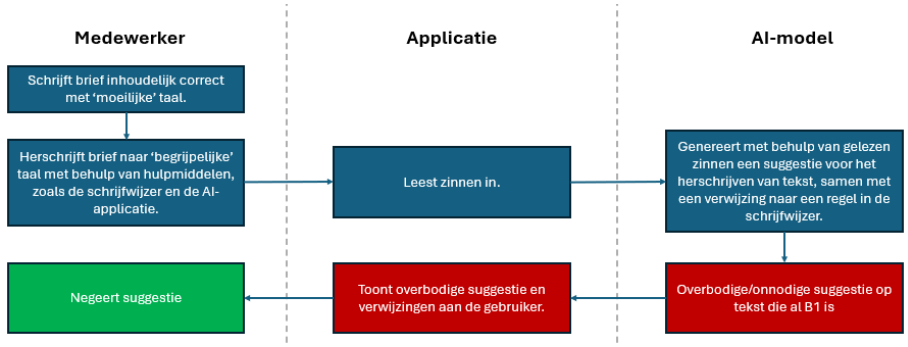


Fig. 24. De gegenereerde suggestie is overbodig of onnodig bij de huidige tekst

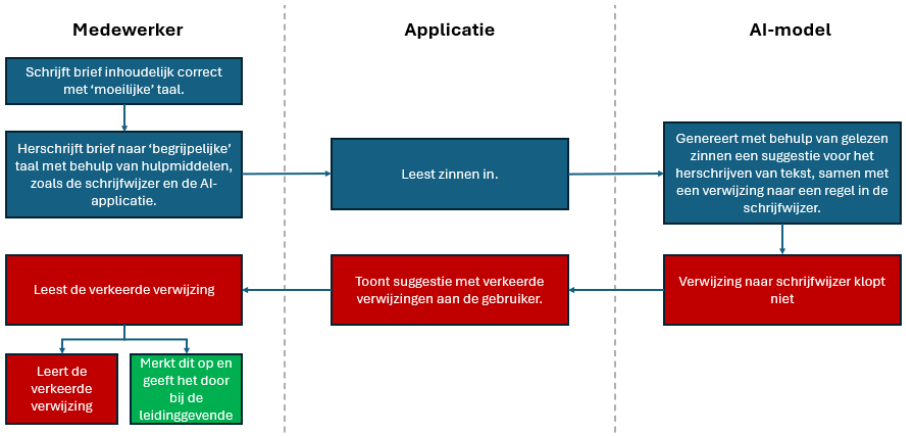


Fig. 25. De gegenereerde suggestie bevat een verwijzing die niet klopt

F DEDA-analyse

Tijdens de literatuurles op 4 december is de DEDA-tool gebruikt om te reflecteren op de ethische vraagstukken die spelen bij de ontwikkeling van de AI-tool voor de OVER-gemeente. Deze tool bestaat uit vier fasen waarin verschillende vragen worden beantwoord.

F.1 Fase één

(1) **Projectnaam en datum:**
AI als schrijfassistent voor heldere gemeentelijke brieven, gebaseerd op B1-niveau Nederlands, 4 december 2025.

(2) **Projectteam:**
Team 2.

(3) **Wat houdt het project in en wat is het doel?:**
Gemeentelijke brieven van de OVER-gemeenten worden herschreven volgens de schrijfwijzer, met als doel deze begrijpelijk te maken op B1-niveau.

(4) **Wat zijn de eventuele maatregelen of acties op basis van de uitkomsten van dit project?:**
Wanneer het project succesvol verloopt, worden brieven eenvoudiger geschreven. Dit draagt bij aan betere begrijpelijkheid voor burgers en vergroot de kans dat de inhoud daadwerkelijk wordt begrepen.

(5) **Wat voor data gebruikt u? Geef een korte beschrijving van de inhoud hiervan.:**
De volgende databronnen worden gebruikt:
- Brieven van de gemeente Oostzaan;
- De schrijfwijzer van de OVER-gemeenten;
- Een inclusiviteitswoordenlijst;
- Een lijst met moeilijke woorden (Gemeente Amsterdam).

(6) **Wie zijn de stakeholders van dit project en op wie/wat heeft het impact?**
Dit zijn de stakeholders van het project:
- De ICT-afdeling;
- Medewerkers die brieven schrijven en controleren;
- De OVER-gemeente;
- Inwoners van de OVER-gemeente;
- De overheid.

Het project heeft impact op alle stakeholders, maar vooral op de medewerkers van de OVER-gemeenten zullen te maken krijgen met de AI-tool. De burgers zullen hierdoor hopelijk de brieven beter begrijpen (Paragraaf 2.2).

(7) **Wat zijn de nagestreefde voordelen van dit project?**
- Consistenter en begrijpelijker schrijven van gemeentelijke brieven, waardoor er minder misverstanden en problemen ontstaan zoals beschreven in Paragraaf 1.2.

(8) **Zijn er (mogelijke) problemen met het project?**
- Het beoordelen van B1-niveau is deels subjectief;
- Onjuiste suggesties kunnen leiden tot betekenisverlies of verkeerd woordgebruik;
- Privacygevoelige gegevens mogen de gemeente niet verlaten en mogen niet worden gebruikt voor het trainen van het model.

(9) **Welke van de onderstaande onderwerpen (algoritmen, bron, anonimiseren etc.) zijn van toepassing op uw project? Beslis welke er eventueel tijdens de workshop kunnen worden overgeslagen.**
De volgende vragen worden tijdens de workshop overgeslagen:
- Vraag 41
- Vraag 42

F.2 Fase twee

Bij de start van de analyse is vastgesteld dat de volgende waarden belangrijk zijn voor de OVER-gemeente bij de ontwikkeling en inzet van de AI-tool:

- **Privacy:** Privacy is essentieel, omdat persoonsgegevens van inwoners niet buiten de organisatie mogen worden verwerkt of gedeeld.
- **Menselijke eindverantwoordelijkheid:** Medewerkers van de gemeente vinden het belangrijk dat de mens eindverantwoordelijk blijft voor de output van de AI-tool (Bijlage A).
- **Gebruiksvriendelijkheid:** De AI-tool moet zo worden ontworpen dat deze daadwerkelijk consistent in de praktijk gebruikt kan worden, in tegenstelling tot de huidige schrijfwijzer die als lastig in gebruik wordt ervaren.
- **Inclusiviteit:** In de schrijfwijzer staat dat teksten bedoeld zijn voor alle inwoners en dat het belangrijk is dat iedereen zich betrokken voelt. Daarom moet de AI-tool inclusief taalgebruik ondersteunen en bevorderen.

F.3 Fase drie

In deze fase worden vragen beantwoord met betrekking tot datagerelateerde overwegingen en algemene overwegingen.

F.3.1 Datagerelateerde overwegingen.

(10) Gebruikt u een algoritme in uw project?
- Ja

(11) Hoe gaat u om met false positives en false negatives?
- De medewerker is eindverantwoordelijk en keurt suggesties zelf.

(12) Is er iemand in het team die kan uitleggen hoe het gebruikte algoritme werkt? Is het noodzakelijk dat iemand kan uitleggen wat het doet?
- Nee, niet noodzakelijk dat iemand het exact kan uitleggen, globaal wel als het gaat om de werking van de applicatie.

(13) Gebruikt u de uitkomsten van het model als leidend of aanvullend in uw beslissingsmodel?
- Nee, alleen aanvullend.

(14) Is er menselijke controle op fouten die het algoritme kan maken? Hoeveel ruimte heeft een mens om van het systeem af te wijken wanneer nodig?
- Ja, medewerkers moeten de suggesties goed of afkeuren. De AI-tool vervangt niet meteen woorden die als 'foutief' worden gezien.

(15) Wat is de bron van de data?
- Gemeente Amsterdam
- OVER-schrijfwijzer
- Interne brieven van de OVER-gemeenten

(16) Heeft u de kwaliteit van de dataset(s) gecontroleerd?
- Nee de brieven zijn niet gecontroleerd, omdat deze enkel voor validatie en testen wordt gebruikt. Als we de kwaliteit zouden controleren is er sprake van data leakage. Wel hebben medewerkers van de gemeente deze voor ons geselecteerd.
  De woordenlijsten zijn doorgenomen en lijken gebruikt te kunnen worden, maar dit moet nog nagevraagd worden bij de OVER-gemeenten.

(17) Hebben de data een houdbaarheidsdatum?
- Nee, maar bijhouden is noodzakelijk, aangezien taal in de loop der jaren verandert.

(18) Verzamelt u de juiste informatie voor uw doel?
- Grotendeels, mogelijk aanvullende bronnen nodig.

(19) Is het nodig om de dataset(s) te anonimiseren, pseudonimiseren of te generaliseren?
- Alleen bij gebruik van echte brieven, wat in dit geval niet zo is. In de praktijk dus wel.

(20) In het geval van anonimiseren: is gecheckt of de data echt niet meer herleidbaar is?
- Dit is waarschijnlijk in de toekomst mogelijk, wanneer de tool geïmplementeerd wordt. Voor dit onderzoek is het in deze fase nog niet van toepassing.

(21) In het geval van pseudonimiseren: wie heeft de sleutel om de pseudonimisering terug te draaien?
- Dit is waarschijnlijk in de toekomst mogelijk, wanneer de tool geïmplementeerd wordt. Voor dit onderzoek is het in deze fase nog niet van toepassing.

(22) Hoe worden de uitkomsten van het project weergegeven? Worden deze gevisualiseerd?
- In een Word-extensie, zoals in het prototype is weergegeven.

(23) Wat zou een andere manier van visualiseren zijn?
- Dat de suggesties onder elkaar worden getoond is voorgesteld door testpersonen, maar omdat dit mogelijk leidt tot minder menselijke controle en een te groot vertrouwen in de tool (automation bias), passen we dit niet toe.

(24) Wie heeft toegang tot de dataset(s)?
- Iedereen heeft toegang tot de woordenlijsten, want deze zijn publiekelijk;
- De mensen die brieven schrijven, docenten van de master applied AI en studenten van deze master hebben toegang tot enkele brieven van de OVER-gemeente.
- Wanneer de applicatie echt gebruikt gaat worden, moet genoteerd worden hoe de data wordt opgeslagen en gebruik.

(25) Hoe wordt de toegang gemonitord?
- Wanneer de applicatie echt gebruikt gaat worden, moet later wel genoteerd worden hoe de data wordt opgeslagen en gebruik.

(26) Zijn de resultaten geschikt om te worden hergebruikt? Zo ja, onder welke voorwaarden?
- Er zijn nog geen resultaten, maar de manier waarop de data gebruikt zal worden, wordt vastgelegd zodat hier in de toekomst rekening mee kan worden gehouden.

(27) Zijn (delen) van de (input)data geschikt om te worden hergebruikt? Zo ja, onder welke voorwaarden?
- Ja want deze worden nu lokaal opgeslagen en de gebruikte instellingen ook. Echter moet de data in de toekomst misschien geüpdatet worden, vanwege bijvoorbeeld verandering in de samenleving over wanneer worden worden gezien als inclusief

F.3.2 Algemene overwegingen.

(28) Bestaan er binnen de organisatie beleid of richtlijnen die van toepassing zijn op dit project? Zo ja, welke?
- Nee, maar ze moeten zich houden aan algemene wetten en regelgeving, zoals artikel 50 en de AI-act.

(29) Wie is/zijn eindverantwoordelijk voor het project?
- De mensen die besluiten het systeem te implementeren (medewerkers van de gemeente).

(30) Zijn de verantwoordelijkheden van die persoon/personen helder?
- Nee, want die kent het systeem nog niet.

(31) Is dit project geschikt voor samenwerking met (commerciële) partners? Zo ja, welke partijen zouden dat kunnen zijn?
- Ja, niet met commerciële partijen, maar wel met andere gemeenten.

(32) Wat is de communicatiestrategie voor dit project? (Zowel voor de positieve als negatieve impact hiervan). In het geval van samenwerkingspartners: is deze strategie met hen afgestemd?
- Ja, er is een communicatiestrategie voor dit project. Er zijn momenten afgesproken waarop vragen gesteld kunnen worden, zowel digitaal als in persoon. Daarnaast is op 1 december 2025 een prototype getest met medewerkers van de gemeente.

(33) Zijn er communicatiestrategieën voor het geval er iets mis gaat?
- Dat weten we niet, aangezien docenten als tussenpartij fungeren. Indien nodig kunnen wij zelf altijd nog contact zoeken.
- Mogelijk moet aan de gemeente worden doorgegeven dat zij een communicatiestrategie opstellen voor het geval er iets misgaat, zodat er een duidelijk plan is voor hoe te handelen in zo'n situatie.

(34) Wie is er verantwoordelijk voor deze strategieën?
- De docenten van de opleiding applied AI.
- De communicatieafdeling en de mensen die het systeem maken (een ontwikkelteam).

(35) Bestaat het risico op publieke verontwaardiging, nu of in de toekomst?
- Ja, aangezien burgers zich kunnen afvragen of de inhoud van de brief wel correct is en of hun persoonlijke gegevens worden gebruikt in een AI-systeem.

(36) Hoe transparant bent u over uw project naar burgers toe?
- Tijdens het ontwikkelen van het project worden de burgers nog niet geïnformeerd over het project.

(37) Hoe worden burgers actief betrokken?
- Aan verschillende burgers worden vragen gesteld over wat zij ervan vinden dat de gemeente AI gebruikt bij het herschrijven van brieven en over hoe transparant de gemeente hierover moet zijn.
- Ook worden medewerkers van de gemeente betrokken bij het ontwerpproces.

(38) Kunnen burgers bezwaar maken tegen uitkomsten van het project?
- Ja, burgers kunnen altijd klachten indienen. Zij worden geïnformeerd over de ontwikkeling van een AI-systeem dat brieven herschrijft. Daarnaast kunnen zij een klacht indienen als zij een brief niet begrijpen of het niet eens zijn met de inhoud.

(39) Is een opt-out mogelijk voor burgers? Zo ja, op welk moment kunnen burgers ervoor kiezen om deelname te beëindigen?
- Nee, deze is er niet. Wel wordt er tijdens het ontwikkelen van het systeem rekening gehouden met de privacyrechten van burgers, etc.

(40) Welke wetten, voorschriften en/of richtlijnen zijn van toepassing op uw project?
- Artikel 50, AI-act, etc. Zie requirements in het verslag.

(41) Heeft u de privacyfunctionaris en/of functionaris gegevensbescherming betrokken bij het project?
- Deze vraag slaan we over.

(42) Heeft u een DPIA (Data Protection Impact Assessment) gehanteerd?
- Deze vraag slaan we over.

(43) Wat is uw onderbuikgevoel over dit project? Heeft u zorgen? Er zijn verschillende onderbuikgevoelens gedeeld door Bibiëne Wüst, Kaan Göcay en Amir Jacobs.
- Bibiëne Wüst is benieuwd of het model zich kan aanpassen naar het format van de brieven.
- Amir Jacobs heeft zorgen over hoe het B1-niveu beoordeeld kan worden, aangezien dit subjectief lijkt om te beoordelen.

- Kaan Göcay maakt zich zorgen over het project, omdat ongeveer 20 procent van de Nederlanders B1-niveau niet kan lezen en deze groep dus niet wordt bereikt.

(44) Bestaat het gevaar dat bepaalde mensen of groepen gediscrimineerd zouden kunnen worden door uw project?
- Ja, mensen die onder B1-niveau lezen, worden nog steeds niet bereikt. Dat ligt niet aan de uitvoering van het project, maar aan de eis die gesteld wordt door de overheid dat teksten op B1-niveau moeten zijn.

(45) Zijn alle verschillende groepen burgers vertegenwoordigd in de dataset(s)? Wie missen er of zijn niet zichtbaar?
- Ja, verschillende groepen burgers zijn vertegenwoordigd omdat er is gewerkt met een inclusiviteitslijst. Deze lijst zorgt ervoor dat rekening wordt gehouden met passende benamingen, zoals neutrale aanduidingen voor ouderen en het gebruik van “kinderen” in plaats van “jongens/meisjes”.

(46) Zit er een feedback loop in het model die negatieve consequenties kan hebben?
- Ja, er kan een feedbackloop ontstaan wanneer een medewerker een suggestie accepteert die in feite onjuist is. Als het model deze acceptaties gebruikt om verder te leren, kan het verkeerde patronen gaan overnemen. Dit risico wordt groter wanneer medewerkers te veel vertrouwen op de gegenereerde tekst en minder kritisch controleren. Daarom is het belangrijk dat datasets en modeluitvoer regelmatig worden beoordeeld en bijgewerkt door mensen om fouten en vertekening te voorkomen.

(47) Function creep: kunt u zich een toekomstig scenario voorstellen waarin de uitkomsten van dit project voor een ander doeleinde gebruikt worden?
- Ja, de uitkomsten van dit project kunnen ook gebruikt worden voor het versimpelen van brieven voor andere gemeenten.

(48) Veranderen uw antwoorden als u de mogelijke langetermijneffecten in acht neemt? Waarom?
- Ja, mijn antwoord verandert wanneer ik rekening houd met mogelijke langetermijneffecten. Taal verandert namelijk in de loop van de tijd. Het model zal daarom regelmatig moeten worden hertraind met bijgewerkte woordenlijsten en voorbeelden om relevant en bruikbaar te blijven.

(49) Wanneer wordt dit project geëvalueerd?
- Het project wordt geëvalueerd zodra de ontwikkelaars de vooraf vastgestelde doelen samen met de gemeente hebben behaald. Daarna kan het systeem worden ingevoerd bij de OVER-gemeenten, zodat medewerkers ermee kunnen testen. Tijdens de ontwikkeling wordt de AI-tool bovendien tussentijds beoordeeld aan de hand van verschillende kwaliteitscriteria.

F.4 Fase vier

In deze fase wordt een conclusie gevormd aan de hand van antwoorden op een aantal vragen.

(1) Zijn de organisatiewaarden en persoonlijke waarden voldoende gewaarborgd?
- **Privacy**: Er wordt uitsluitend getraind met openbare, niet-privacygevoelige data. Brieven met privacygevoelige informatie worden alleen intern gebruikt. Vandaar dat de waarde privacy voldoende is gewaarborgd.
- **Gebruiksvriendelijkheid**: Het systeem moet praktisch inzetbaar zijn en niet omslachtig in gebruik. Door het model te ontwikkelen als een Word-extensie wordt de gebruiksvriendelijkheid vergroot. Dit wordt geëvalueerd op basis van feedback van testgebruikers op het verbeterde prototype. Op deze manier kan deze waarde voldoende worden gewaarborgd.
- **Menselijke eindverantwoordelijkheid**: In het ontwerp van de AI-tool is vastgelegd dat gemeentemedewerkers eindverantwoordelijk blijven voor de inhoud van hun brieven. De testpersonen hebben op 1 december 2025 bevestigd dat deze waarde in het prototype voldoende wordt gewaarborgd.
- **Inclusiviteit**: De brieven moeten inclusief worden geschreven volgens de OVER-schrijfwijzer. Door een inclusieve woordenlijst van de Gemeente Amsterdam in het model te implementeren, wordt aan deze waarde voldaan.

(2) Wat zijn de belangrijkste ethische knelpunten?
    - De belangrijkste ethische knelpunten zijn dat privacygevoelige informatie mogelijk wordt gebruikt bij het hertrainen van het model, wat risico's oplevert voor gegevensbescherming. Daarnaast is er het risico dat een deel van de bevolking, ongeveer 20 procent van de Nederlanders, ook bij teksten op B1-niveau nog steeds niet goed wordt bereikt. Tot slot bestaat de kans dat gebruikers te veel vertrouwen op het systeem en suggesties klakkeloos accepteren, waardoor fouten onopgemerkt blijven en het model op basis van onjuiste input wordt hertraind.
(3) Wat zijn nieuwe en verrassende inzichten?
    - Er is nog onvoldoende nagedacht over een mogelijke feedback-loop.
(4) Onder welke voorwaarden willen we wel of niet doorgaan met dit project?
    - Wanneer medewerkers het systeem te veel vertrouwen, waardoor de inhoudelijke kwaliteit van brieven achteruitgaat.
(5) Bekijk alle actiepunten die hieronder worden weergegeven en schrijf voor elk punt de actiehouder op.
    - De actiepunten worden verder uitgewerkt tijdens het schrijven van het verslag, met name in het hoofdstuk 'Discussie'.

G Gebruikerstesten op 1 december

G.1 Testronde 1: Uitgevoerd door Franny van der Bijl, WMO-consulent

G.1.1 Algemene bevindingen en verbeterpunten.

De eerste indruk van de testpersoon is dat zij het leuk vond dat we een paper prototype gemaakt hadden, ondanks dat we een ICT-opleiding doen. De testpersoon heeft zelf de knop 'AI-assistent' gevonden en heeft deze zelf aangeklikt. Hierbij heeft teamlid Terence verteld dat het onderstreepte deel in de tekst waarschijnlijk een highlight wordt in het geel in de tweede iteratie.

De optie 'aanpassen' in de AI-tool vond de testpersoon onduidelijk omdat je rechts in de extensie op aanpassen klikt en vervolgens een zin in het document kan aanpassen. De testpersoon gaf aan dat ze het idee had dat niet alle collega's zullen snappen hoe dit eruit ziet. Hierbij begreep ze wel hoe ze door kon gaan naar de volgende fout door op de pijl te klikken.

G.1.2 Vragen van de testpersoon.

- Als je op accepteren drukt, gaat hij dan automatisch door naar de volgende te verbeteren punt in het document? Je wilt namelijk door naar de volgende suggestie als je op accepteren of overslaan klikt.
- Kan het model ook leren? (Antwoord Amber: Ja dat is wel het idee, dat als medewerkers altijd op niet accepteren klikken dat het model dan gaat leren dat dit niet meer verbeterd hoeft te worden.)

G.1.3 Desirability (wenselijkheid).

- **Voldoet de optie om de AI-tool aan en uit te zetten, aan jullie wensen?**
  Ze vinden het fijn dat deze knop aan- en uitgezet kan worden. Ze geeft aan dat het goed is dat je de tool aan kan zetten als je denkt dat de brief goed is.
- **Zou je als medewerker prettig om kunnen gaan met de AI-tool?**
  De testpersoon vindt het fijn dat onderstreept wordt waar de fout is en dat er wordt uitgelegd waarom. Op deze manier leren de medewerkers zelf ook van hun fouten.
- **Is er te veel tekst in beeld?**
  De testpersoon vindt het fijn dat de bron te zien is, maar geeft ook aan dat dit dubbelop is omdat er twee keer naar de schrijfwijzer verwezen wordt.
- **Is het wenselijk dat de AI-tool aangeeft dat er een suggestie is voor een bepaalde zin of een bepaald woord?**
  De testpersoon geeft aan dat je wilt naar de volgende als je op accepteren of overslaan klikt.
- **Denk je dat je de AI-tool snel kunt leren gebruiken?**
  De testpersoon zelf wel en denkt andere medewerkers ook. Er wordt aangeraden niet te veel kleur te gebruiken (wel duidelijk en overzichtelijk).
- **Zijn er andere stappen in het prototype wenselijk die er op dit moment nog niet instaan?**
  Een knop 'alles accepteren' zou in sommige gevallen wel handig zijn. Soms zouden ze er door de tijd te veel op vertrouwen en op alles accepteren willen drukken en daarna de brief na willen lezen om vervolgens te controleren of deze goed werkt. Ook geeft de testpersoon aan dat je een melding krijgt dat de brief op B1-niveau is, als er niks verbeterd hoeft te worden.

G.1.4 Explainability & Trust.

- **Zijn de signalen van de AI-tool nuttig en voldoende uitgelegd om jouw werk te ondersteunen?**
  Ja, misschien de bron dus dubbelop.
- **Zou je deze AI-tool vertrouwen?**
  Ja, maar eerst nog een aantal keer gebruiken en dan zien of het werkt. Op alles accepteren en dan nalezen zelf.
- **Geeft de AI-tool informatie die jou helpt om beslissingen te nemen, of juist te veel/te weinig?**
  Alleen de bron één keer noemen en verder is het fijn dat de tool aan- en uitgezet kan worden.
- **Zou je te veel beïnvloed worden door de suggesties die de AI-tool geeft?**
  Dat weet de testpersoon niet, soms wel en soms niet. Ze zouden wel nog een vakterm moeten gebruiken, waardoor een assistent kan aangeven dat dit moeilijk is, maar deze overslaan.
- **Vind je deze AI-tool vertrouwelijk in de gemeentelijke context?**
  Ja

G.2 Testronde 2: Uitgevoerd door Esmee en Pim (team 5)

G.2.1 Algemene bevindingen en verbeterpunten.

De testpersonen hebben zelf de knop 'AI-assistent' gevonden en hebben deze zelf aangeklikt. Hierbij heeft teamlid Terence verteld dat het onderstreepte deel in de tekst waarschijnlijk een highlight wordt in het geel in de tweede iteratie.
Esmee geeft aan dat we rechtsboven het woord 'suggestie' kunnen vervangen voor 'fouten'. Ook Pim vindt, net zoals de persoon in de eerste testronde, dat twee keer de bron noemen overbodig is. Zowel Esmee als Pim geven aan deze bronnen te combineren tot één.

G.2.2 Desirability (wenselijkheid).

- **Voldoet de optie om de AI-tool aan en uit te zetten, aan jullie wensen?**
  Ja, dat is juist fijn.
- **Zou je als medewerker prettig om kunnen gaan met de AI-tool?**
  Ja
- **Is het wenselijk dat de AI-tool aangeeft dat er een suggestie is voor een bepaalde zin of woord?**
  Ja die zijn ook wel fijn. Misschien kan er nog een knop toegevoegd worden voor 'volgende suggestie'.
- **Zijn er andere stappen in het prototype wenselijk?**
  'suggestie' vervangen voor 'fouten' rechtsboven.

G.2.3 Explainability & Trust.

- **Zou je deze AI-tool vertrouwen?**
  Opzich niet.
- **Geeft de AI-tool voldoende informatie?**
  Ja, er wordt voldoende informatie getoond om een suggestie te geven die onderbouwd is.
- **Zou je te veel beïnvloed worden door de suggesties?**
  Nee, want doordat er stap voor stap door de suggesties wordt gelopen leer je er ook van.
- **Zou je een knop 'alles accepteren' willen?**
  Nee want nu moet iemand er nog over nadenken.
- **Vind je deze AI-tool vertrouwelijk in de gemeentelijke context?**
  Over het algemeen vinden ze hem wel nice.

G.3 Testronde 3: Uitgevoerd door Annemaria van Teeseling, teamleider WMO

G.3.1 Algemene bevindingen en verbeterpunten.

De testpersoon heeft zelf de knop 'AI-assistent' gevonden en heeft deze zelf aangeklikt. Hierbij heeft teamlid Terence verteld dat het onderstreepte deel in de tekst waarschijnlijk een highlight wordt in het geel in de tweede iteratie. De testpersoon kon de knop vinden, waarmee je naar de volgende fout kon gaan. Het prototype viel in de smaak.
De testpersoon gaf aan dat het beter is om het woord 'AI-assistent' te vervangen door 'schrijfwijzer assistent'. Ook zou de testpersoon een optie toegevoegd willen hebben om alle suggesties te accepteren. Maar ze moeten wel echt zelf na blijven denken. Blijkt ook uit iets wat nu loopt met 'copilot'.

G.3.2 Vragen van de testpersoon.

- Wat is het verschil tussen aanpassen en accepteren?
- Wat gebeurt er als ze hem aan het begin aanzet en aan het eind?
- Wanneer kan je hem aanzetten?

G.3.3 Desirability (wenselijkheid).

- **Voldoet de optie om de AI-tool aan en uit te zetten?**
  Ja dat vindt ze heel erg fijn! Ze vindt het echt heel fijn dat deze tool ook achteraf aangezet kan worden, zodat ze eerst losjes kunnen schrijven. Ze vindt het ook heel leuk dat je fouten gemarkeerd kan krijgen, zonder de suggesties al te zien, zodat je jezelf kan uitdagen.
- **Zijn alle stappen wenselijk?**
  Misschien alle suggesties onder elkaar, want dan kun je alle suggesties bekijken voordat je op alles accepteren klikt.
- **Denk je dat je de AI-tool snel kunt leren gebruiken?**
  Ja (duidelijke ja), ze vindt het echt fijn dat hij aan en uit kan en dat er uitgelegd wordt waarom er een suggestie gedaan wordt. Het aanpassen vindt ze heel leuk dat dat ook rechts kan. Het verschil tussen aanpassen en accepteren moet je net even weten, maar als je het weet vindt ze het geweldig.
- **Andere wenselijke stappen?**
  Een knop voor 'andere suggestie' zou wel heel mooi zijn.

G.3.4 Explainability & Trust.

- **Zou je te veel beïnvloed worden?**
  Nee zelf niet. Het is wel goed zoals het nu is.
- **Vind je dit vertrouwelijk in gemeentelijke context?**
  Ze kennen tolkie nog niet. Soms zeggen gemeenten dat ze al ver zijn, maar in de praktijk ben je al snel ver. Als je werkt met 'copilot'. Siso vinden hun heel risicogevoelig. Misschien heeft het ook wel te maken met geld, maar dat weten ze nu niet.

G.4 Testronde 4: Uitgevoerd door Savannah Werner en Rainesh Rewat

G.4.1 Algemene bevindingen en verbeterpunten.

De testpersonen hebben zelf de knop 'AI-assistent' gevonden en hebben deze zelf aangeklikt. Hierbij heeft teamlid Terence verteld dat het onderstreepte deel in de tekst waarschijnlijk een highlight wordt in het geel in de tweede iteratie. Savannah vond de inhoud erg duidelijk en de tool mooi.

Rainesh vindt dat de oorspronkelijke zin niet opnieuw getoond hoeft te worden rechts. Savannah wil dat er een onboarding komt voor als je nog nooit met de AI-tool hebt gewerkt.

G.4.2 Vragen van de testpersonen.

- Werkt de tool live?
- Hoe gaat het model om met structuur van de tekst, zoals kopjes?

G.5 Testronde 5: Uitgevoerd door Elsbeth Francisca, beleidsadviseur van de gemeente

G.5.1 Algemene bevindingen en verbeterpunten.

De testpersoon heeft zelf de knop ‘AI-assistent’ gevonden en heeft deze zelf aangeklikt. Hierbij heeft teamlid Terence verteld dat het onderstreepte deel in de tekst waarschijnlijk een highlight wordt in het geel in de tweede iteratie. Savannah vond de inhoud erg duidelijk en de tool mooi. Als het model zou leren, vind ze wel een risico als de brief in de database beland. Ze twijfelt of de term AI-gebruikt moet worden, misschien iets als ‘schrijfwijzer assistent’. Bijvoorbeeld voor collega’s die ouder zijn.

G.5.2 Desirability (wenselijkheid).

- **Voldoet de optie om de AI-tool aan en uit te zetten?**
  Ja, want sommige zullen de tool misschien wel niet willen gebruiken, maar het is goed om de autonomie aan de schrijver te laten. Daarnaast kon ze de knop snel vinden.
- **Is het wenselijk dat de AI-tool aangeeft dat er een suggestie is voor een zin of woord?**
  Ja. Het is wel goed dat je als schrijver nog leert en je verantwoordelijk voelt voor de inhoud. Als je in de modus van gemak zou zitten is niet goed natuurlijk. Het is goed dat een suggestie wordt gegeven en niet meteen wordt toegepast.
- **Denk je dat je de AI-tool snel kunt leren gebruiken?**
  Ja en collega’s op leeftijd ook wel. Het is heel gebruiksvriendelijk.
- **Andere stappen wenselijk?**
  Een melding met ‘je bent zelf verantwoordelijk voor de inhoud van deze brief’ had een andere groep en dat was erg goed. Ze zou de knop voor de volgende fout niet vinden, maar zou op het onderstreepte woord zelf klikken. Een kleur gebruiken als highlight of een strook eronder die niet zo ‘in your face is’.

G.5.3 Explainability & Trust.

- **Geeft de AI-tool informatie die helpt bij beslissingen?**
  Ik zou misschien wel te veel vertrouwen op de suggesties.
- **Zou je te veel beïnvloed worden?**
  Met een knop ‘Alles accepteren’ wel. Ik zou de bal bij de schrijver laten liggen.

H Testplan: maandag 12 januari 2026

Op 12 januari gaan we ons idee testen met 3 Communicatie en Multimedia Design-experts. Hieronder staat het testplan wat we deze dag gaan uitvoeren.

H.1 Introductie

Wij studeren met zijn vijven de Master Applied Artificial Intelligence hier op de Hogeschool van Amsterdam. Vanuit de OVER-gemeenten hebben wij de opdracht gekregen om een AI-tool te maken dat de gemeente kan helpen om begrijpelijker Nederlands te schrijven. In dit geval betekend begrijpelijker Nederlands ook wel het B1-niveau Nederlands, aangezien ruim 80% van de Nederlanders B1-niveau goed kan begrijpen. Nu hebben zij daar een OVER-schrijfwijzer voor. Een tool met richtlijnen/regels over hoe een gemeentelijke brief begrijpelijk en op B1-niveau geschreven moet worden.

- We gaan ervan uit dat de medewerkers die de brieven schrijven, minimaal B1 niveau Nederlands bezitten. Dit komt omdat de medewerker altijd de eindverantwoordelijke is om te bepalen of een suggestie wel of niet doorkomt en de tekst daadwerkelijk wordt aangepast.

Voor dit onderzoek willen we graag jullie mening gebruiken om te onderzoeken of het design van ons idee naar jullie wens is. We willen jullie graag ons prototype laten zien en zullen jullie gaan vragen om feedback te geven vanuit een werknemers perspectief.

H.2 Ons idee in het kort voor dit project

In ons project willen we een AI-model ontwikkelen dat de medewerkers van de OVER-gemeenten assisteert bij het schrijven van brieven. Als de AI-tool ingeschakeld is krijgt de medewerker na het typen van een onjuiste zin, een gemarkeerde lijn onder de foutieve zin te zien. Als de medewerker hierop klikt, krijgt hij of zij suggesties te zien die gericht zijn op taalniveau en zinsopbouw, om ervoor te zorgen dat de geschreven tekst versimpeld kan worden.
De AI-tool gaat suggesties geven op basis van wat hij geen B1-niveau Nederlands vindt, of wat hij bijvoorbeeld te complexe zinnen vindt volgens de OVER-schrijfwijzer. Bij deze suggestie zal dan ook doorverwezen worden naar de OVER-schrijfwijzer. Er is een knop aanwezig op het scherm, die het toelaat voor de medewerker om de AI-tool aan of uit te zetten. Hierdoor zal de medewerker niet gestoord worden tijdens het schrijven, tenzij hij of zij dat natuurlijk wel wil.
De extensie zal, als de AI-tool ingeschakeld is, rechts in beeld komen te staan. Hierbij zal per gevonden te lange zin of een te moeilijk woord, één pagina gebruikt worden. Bij 20 gevonden fouten in één document, zullen dus 20 doorklikbare pagina's beschikbaar zijn in de extensie. Zo blijven alle suggesties los van elkaar zichtbaar en kan er ook later teruggekomen worden op een suggestie.

H.3 Doel en taakverdeling

Het doel voor ons is dat wij meer te weten komen over de relevantie en de gebruiksvriendelijkheid van deze AI-tool. We willen testen of dit het gewenste prototype is, aangezien we hebben vernomen dat de medewerkers het fijn zouden vinden als de tool niet telkens in beeld springt, een overvloed aan kleurtjes heeft en over het algemeen afleidend is. We willen graag observeren wat jij van de AI-tool vindt en of je eventuele verbeterpunten ziet die zorgen voor een prettige werkervaring met deze AI-tool. Tijdens het testen wordt de rolverdeling aangehouden zoals weergegeven in Tabel 28.

Naam	Taak
Amber	Aantekeningen maken + vragen stellen
Amir	Aanwezig
Kaan	Aantekeningen maken
Terence	Het prototype uitleggen en ondersteunen tijdens de test
Bibiëne	Vragen stellen

Table 28. Taakverdeling tijdens de test

H.3.1 Uitleg door Terence tijdens de test.

Tijdens de test legt Terence uit dat de experts het prototype moeten beoordelen vanuit het perspectief van de medewerkers van de OVER-gemeenten die de brieven schrijven, aangevuld met hun expertise in het ontwikkelen en ontwerpen van digitale tools.

Vervolgens legt Terence uit hoe het prototype werkt. Hij gaat daarbij in op de volgende punten:

- wat het prototype doet,
- het gebruik van de AI-knop (aan- en uitzetten),
- de extensie,
- verwijzingen per pagina binnen de extensie,
- verwijzingen naar de OVER-schrijfwijzer.

Daarna laat Terence de medewerker het prototype bekijken en test hij of deze de aan- en uitknop van de AI-tool kan vinden. Wanneer deze knop wordt geactiveerd, verschijnt de extensie in beeld. Vervolgens onderzoekt Terence hoe de testpersoon de AI-tool gebruikt en observeert hij welke vragen of onduidelijkheden er ontstaan. Alle bevindingen worden zorgvuldig vastgelegd en gedocumenteerd (Bijlage I).

H.4 Vragen aan de medewerkers van de OVER-gemeenten

H.4.1 Relevance (relevantie).

- Zijn de signalen van het systeem nuttig om jouw werk te ondersteunen?
- Zijn de meldingen/knoppen die het systeem geeft relevant en niet storend?
- Is de knop ‘aanpassen’ relevant of vind u deze overbodig?
- Past deze manier van feedback geven (losse suggesties per zin/woord) bij jouw huidige manier van werken?
- Hoe relevant vindt u de knoppen Accepteren, Aanpassen en Weigeren bij het verwerken van een suggestie?
  - Mist u hier een optie, of is er juist een knop overbodig?

H.4.2 Usability (gebruiksvriendelijkheid).

- Is het voor u duidelijk wat er van u verwacht wordt wanneer een suggestie verschijnt (onderstreping + suggestiekaart)?
- Vindt u de opbouw van de extensie aan de rechterkant in Word (originele zin, AI-suggestie, knoppen) logisch en overzichtelijk? Wat zou volgens u verbeterd kunnen worden?
- Helpt dit systeem u om gemakkelijk een beslissing te nemen?
- Voelt het gebruik van de AI-tool ondersteunend, zonder dat u het gevoel hebt dat de tool uw schrijfkeuzes overneemt?
- Is er iets in het systeem dat verwarrend of onduidelijk aanvoelt?

I Gebruikerstesten op 12 januari

I.1 Testronde 1: Uitgevoerd door César Van Hardeveld

I.1.1 Algemene bevindingen en verbeterpunten.

De aan- uitknop staat nu links, dat kan niet in word. Deze aan-uit knop kon hij wel vinden. Hij ziet meteen een tekst met dat hij zelf verantwoordelijk is. Dat vind César goed om te zien. Hij ziet dat er een soort analyse is gemaakt, geel gehighlighte tekst en onderstreepte tekst. Ook ziet hij een suggestie. Hij ziet duidelijk om welk woord de nieuwe suggestie gaat. Hij vind de suggestie wel goed, dus klikt op accepteren. Nu ziet hij dat de suggestie aangepast is en dat er een nieuwe deel in het document wordt gehighlight. Hij ziet dat deze zin verbetert kan worden met een nieuwe suggestie. Hij ziet ook een bron, waar hij op kan klikken en dan doorverwezen wordt naar de schrijfwijzer. Hij vind de nieuwe suggestie ook goed. Nu heeft hij alles geaccepteerd en alles is op B1-niveau volgens de tekst die verschijnt, maar hij moet de brief zelf nog kritisch nalezen.

I.1.2 Relevance (relevantie).

- **Zijn de signalen van het systeem nuttig om jouw werk te ondersteunen?**
  Ja, wel zou ik zeggen dat de tekst ‘u bent zelf verantwoordelijk voor de inhoud van deze brief’ best wel belangrijker mag zijn. De medewerkers krijgen vast een goede demo waarbij dit wordt uitgelegd. Dit zorgt ervoor dat iemand niet op accepteren blijft klikken. De optie/knop voor een nieuwe suggestie vind hij wel goed. Daarnaast zou ik ervoor zorgen dat je terug kan gaan naar suggesties die eerder aangepast zijn, bijvoorbeeld met een ’undo’-knop. Het gebruik van kleur vind ik wel heel erg nice. Met rood zie je meteen duidelijk om welk woord het gaat.
- **Zijn de meldingen/knoppen die het systeem geeft relevant en niet storend?**
  Opzich wel. Het is ook fijn om te zien waarom een zin niet wordt goedgekeurd (alsin niet op B1-niveau is), zoals nu het geval is. Het is ook goed dat je een regel ziet in de schrijfwijzer, om te leren. Het stoort wel in het prototype dat bij een volgende suggestie de indeling aan de rechterkant verspringt, maar dat is puur een klein opmaakfoutje in Figma.
- **Is de knop ‘aanpassen’ relevant of vind u deze overbodig?**
  Nee, ik vind hem niet overbodig. Op eerste opzicht niet helemaal duidelijk wat het precies doet. Ik zou deze optie er wel inhouden om de gebruiker autonoom te houden over de aanpassingen.
- **Past deze manier van feedback geven (losse suggesties per zin/woord) bij jouw huidige manier van werken?**
  Ja sowieso. Misschien is het wel een handig idee om ook nog de hele zin te highlighten welke zin aangepast gaat worden en alleen het woord wat niet schrijfwijzervriendleijk is wordt aangepast. Dit zou ik doen bij enkel het woord wat niet voldoet aan de schrijfwijzer, maar dat dan de hele zin wel nog gehighlight wordt. Dit kan ook anders met kleuren worden aangegeven, dus alles bijvoorbeeld geel en een deel rood.
- **Hoe relevant vindt u de knoppen Accepteren, Aanpassen en Weigeren bij het verwerken van een suggestie? Mist u hier een optie, of is er juist een knop overbodig?**
  Nee ik vind niks overbodig. Het is goed om alle drie de opties te houden.

I.1.3 Usability (gebruiksvriendelijkheid).

- **Is het voor u duidelijk wat er van u verwacht wordt wanneer een suggestie verschijnt (onderstreping + suggestie)?**
  Ja.
- **Vindt u de opbouw van de extensie aan de rechterkant in Word (originele zin, AI-suggestie, knoppen) logisch en overzichtelijk? Wat zou volgens u verbeterd kunnen worden?**
  Euhm het ontwerp wat er nu staat kan denk ik niet exact nagemaakt worden in Word denk ik. Het ziet er heel cool uit, maar ik ben bang dat niet alles volledig nagemaakt kan worden, want je hebt bij het implementeren van zo'n Word-extensie best wel last van limitaties van Word, maar opzich maakt dat niet heel erg uit.
- **Helpt dit systeem u om gemakkelijk een beslissing te nemen?**
  Ja, 100%.
- **Voelt het gebruik van de AI-tool ondersteunend, zonder dat u het gevoel hebt dat de tool uw schrijfkeuzes overneemt?**
  Ja.
- **Is er iets in het systeem dat verwarrend of onduidelijk aanvoelt?**
  Euhm nee ik zou zeggen van niet. Misschien dat de aanpassen knop nu nog wat onduidelijk aanvoelt. Dat kan komen omdat de werking van deze knop nog niet volledig is uitgewerkt in het prototype. Ik ben er wel benieuwd naar.

I.2 Testronde 2: Uitgevoerd door Isabel Erven

I.2.1 Algemene bevindingen en verbeterpunten.

Oke ik ben dus een schrijfpersoon bij de gemeente en ik zie een toggle button. Hier klik ik op. Ik zie veel gebeuren. Er is iets geel geworden en ik weet niet waarom. (Waarneming: Isabel kijkt moeilijk). Ik zie dat de tekst 'de integrale herziening' fout is, omdat het rood gekleurd is. De AI-suggestie zegt regels. Ja omdat alleen 'regels' gekleurd is dacht ik dat alleen 'regels' geweizigd is, maar de hele zin is gewijzigd. De knop 'aanpassen' doet helaas nu niks. Ik denk dat ik dan zelf de zin of suggestiezin mag aanpassen en dat je hem daarna mag accepteren. Als ik op accepteren klik ga ik automatisch naar de volgende suggestie. Opzich prima. Ik vind 'de brief is op B1-niveau een grote belofte'.

I.2.2 Relevance (relevantie).

- **Zijn de signalen van het systeem nuttig om jouw werk te ondersteunen?**
  Je bedoelt de highlights? Misschien als ik de brief zelf had geschreven en ik had tekst onderstreept, kan het model verwarrend werken, aangezien het model ook teksten onderstreept. Echter heeft de streep die tevoorschijn komt door het model wel een andere kleur. Als medewerker hierover eerst uitleg hebben gehad is het wel duidelijk denk ik.
- **Zijn de meldingen/knoppen die het systeem geeft relevant en niet storend?**
  Er wordt veel getoond de allereerste keer dat je dit ziet is de tool daardoor wel overweldigend. Qua highlights zit het niet dwars, het is duidelijk wat voor suggesties het geeft. Dit komt vooral doordat er veel kleurtjes gebruikt zijn. Deze pakken wel erg je aandacht, maar de kleurtjes zijn zeker niet verkeerd. Kunnen misschien iets lichter. Verder gaat mijn aandacht nu direct naar de knoppen accepteren, aanpassen weigeren door het kleurgebruik. Dit is misschien niet nodig.Waarom krijg ik gelijk suggesties voor de hele brief? Misschien wil ik wel feedback op maar één alinea. En stel je model geeft een slechte suggestie of stel er is eigenlijk niks fout aan de zin, maar hij geeft wel een suggestie. Wordt er dan alsnog een suggestie getoond? Dit kunnen jullie nu enkel aantonen met LiNT-II scores, maar hoe ga je aan iemand uitleggen hoe de tool dit bepaald als ik een medewerker ben die niet weet wat een LiNT-II-score is?
- **Is de knop 'aanpassen' relevant of vind u deze overbodig?**
  Weet ik niet echt. Want ik ben geen medewerker, maar ik zou zelf een tekst ook kunnen aanpassen in word. Ik zou vooral deze knop (net als de knoppen aanpassen en weigeren minder prominent maken).
- **Past deze manier van feedback geven (losse suggesties per zin/woord) bij jouw huidige manier van werken?**
  Het ligt eraan. Is het niet te veel in beeld als je ook feedback 'opmerkingen' van een collega hebt? Verder is het erg passend bij de huidige manier van werken.
- **Hoe relevant vindt u de knoppen Accepteren, Aanpassen en Weigeren bij het verwerken van een suggestie? Mist u hier een optie, of is er juist een knop overbodig?**
  Ik vraag me af waarom deze zo prominent gekleurd zijn. Als deze wat lager geplaatst zou worden, dat ze eerst te zien krijgen waarvan ze leren, bijvoorbeeld de bron. De knopen accepteren en weigeren zou ik sowieso wel laten staan.

I.2.3 Usability (gebruiksvriendelijkheid).

- **Is het voor u duidelijk wat er van u verwacht wordt wanneer een suggestie verschijnt (onderstreping + suggestie)?**
  Ja, ik vind van wel. Ik zou het wel ook zeker nog met iemand testen die geen master AI doet. (De testpersoon heeft namelijk de opleiding CMD afgerond, maar is ook de Master Applied AI aan het volgen.)
- **Vindt u de opbouw van de extensie aan de rechterkant in Word (originele zin, AI-suggestie, knoppen) logisch en overzichtelijk? Wat zou volgens u verbeterd kunnen worden?**
  Sommige knoppen zijn te groot en de bron zou ik wat eerder plaatsen in plaats van onderaan. De kleuren van de knoppen vind ik ook te overheersend. Ik zou bij de hele zin duidelijker visueel maken dat de hele zin is gewijzigd.
- **Helpt dit systeem u om gemakkelijk een beslissing te nemen?**
  Nou nee, omdat ik niet snel zag dat de hele zin anders was. Zou ik het wel snel zien dan wel.

- **Voelt het gebruik van de AI-tool ondersteunend, zonder dat u het gevoel hebt dat de tool uw schrijfkeuzes overneemt?**
  Ja, ik zou de bron erboven zetten en dan vind ik het ondersteunend.
- **Is er iets in het systeem dat verwarrend of onduidelijk aanvoelt?**
  Ja, vanwege de dingen die ik net heb gezegd.

I.3 Testronde 3: Uitgevoerd door Julius Meeuwisse

I.3.1 Algemene bevindingen en verbeterpunten.

'Oke, ik zit in Word'. 'Ehm, ik heb net dit document zelf getypt en ik zie een nieuwe knop 'schrijfwijzer assistent". 'Verder lijkt het gewoon normaal. Ik kan niet scrollen maar de bedoeling is denk ik dat op 'schrijfwijzer assistent". Hij klikt op schrijfwijzer assistent. Hij kan het indrukken en dan gaat het aan. Hij krijgt rechts mooi een overzicht, zegt hij zelf. 'Moeilijk woord gevonden, integrale herziening (herhaalt wat er staat) etc. Oke vet'. Hij geeft aan dat je kan accepteren, weigeren en aanpassen. 'Aanpassen werkt niet 'work in progress". Hij klikt op weigeren en ziet dat de tekst is aangepast. Hij probeert op nieuwe suggestie te klikken, maar dit werkt nog niet. Hij vindt het wel fijn dat dat kan. 'Goede disclaimer dat je de brief zelf nog dient na te lezen'. Als ik perongeluk op weigeren klik, gaat dan de suggestie weg? Ja -> Misschien een fallback toevoegen

I.3.2 Relevance (relevantie).

- **Zijn de signalen van het systeem nuttig om jouw werk te ondersteunen?**
  Wat bedoel je met signalen? Hij vindt het een vaag woord. (Ik heb het nu uitgelegd). Hij zegt kwa uiterlijk dat dat gele in de brief verwarrend kan zijn. Want je kan ook geel markeren in Word. Je kan misschien iets anders proberen wat niet in Word zit. Je kan misschien op het woord hooveren zodat daar dan al op accepteren, aanpassen of weigeren kan klikken. Hij vindt het goed dat je door alle suggesties heen kan en dat je snel op accepteren kan klikken. Ik zou de knop 'Suggestie opnieuw' ergens anders plaatsen, want het lijkt nu alsof je de hele pagina refresht. Ik zou zelf bij AI suggestie de opnieuw knop plaatsen.
- **Zijn de meldingen/knoppen die het systeem geeft relevant en niet storend?**
  Ik vind het niet storend omdat je de tool aan en uit kan zetten. Als ik een oudere vrouw was zou ik het ook fijn vinden werken. Wat ik raar vind, is dat de knop links zit en dat de extensie rechts opent. Een uitklap pijl zou bijvoorbeeld misschien mooier en beter zijn voor de flow.
- **Is de knop 'aanpassen' relevant of vind u deze overbodig?**
  Ik weet niet wat die doet. (ik heb het uitgelegd). Ik vind het overbodig omdat je in de tekst zelf ook al zou kunnen aanpassen als je dat wilt. Accepteren en Weigeren is genoeg.
- **Past deze manier van feedback geven (losse suggesties per zin/woord) bij jouw huidige manier van werken?**
  Jazeker, want dan is het behapbaar. Ook kan je aan het einde van de workflow even rustig door alles heen. Het is duidelijk dat je alles rustig moet checken. Wat nog zou kunnen helpen dat als je op accepteren klikt dat er een prompt komt met 'Weet je het zeker". Hierdoor neemt de flow af, maar kan men niet meer zo snel achter elkaar op accepteren klikken.
- **Hoe relevant vindt u de knoppen Accepteren, Aanpassen en Weigeren bij het verwerken van een suggestie? Mist u hier een optie, of is er juist een knop overbodig?**
  Ik weet niet wat die doet. (ik heb het uitgelegd). Ik vind het overbodig omdat je in de tekst zelf ook al zou kunnen aanpassen als je dat wilt. Accepteren en Weigeren is genoeg. Er hoeft geen nieuwe knop bij van mij.

I.3.3 Usability (gebruiksvriendelijkheid).

- **Is het voor u duidelijk wat er van u verwacht wordt wanneer een suggestie verschijnt (onderstreping + suggestie)?**
  Ja dat is duidelijk.
- **Vindt u de opbouw van de extensie aan de rechterkant in Word (originele zin, AI-suggestie, knoppen) logisch en overzichtelijk? Wat zou volgens u verbeterd kunnen worden?**
  Ik vind hem heel mooi, ik weet niet of een ander uiterlijk dan Word überhaupt kan in een extensie. De knop zou hij dus verplaatsen van 'schrijfwijzer assistent'.
- **Helpt dit systeem u om gemakkelijk een beslissing te nemen?**
  Ja, ik kan gewoon op accepteren klikken of op weigeren.

- **Voelt het gebruik van de AI-tool ondersteunend, zonder dat u het gevoel hebt dat de tool uw schrijfkeuzes overneemt?**
  Ja want ik kan gewoon weigeren wanneer ik wil. Er is genoeg keus en ik zou het ook uit kunnen zetten als ik weet dat ik wel op B1-niveau kan schrijven.
- **Is er iets in het systeem dat verwarrend of onduidelijk aanvoelt?**
  De knop aanpassen was eerst verwarrend en onduidelijk. Ook de refresh knop lijkt alsof de hele pagina opnieuw geladen wordt, terwijl het dient als refresh voor de suggestie. Stel de mail is meerdere pagina's lang, hoe kan ik dan weten waar ik ben kwa suggesties? Er staat nergens pagina 1 van de iets.

J Globale modelkeuze

Tekstversimpeling kan worden gezien als een specifieke vorm van vertalen: van een complexe brontekst naar een eenvoudiger variant met behoud van kernbetekenis. Goede tekstversimpeling combineert begrip en vereenvoudiging [27]. Voor dit project moeten de taalmodellen de inhoud begrijpen en herformuleren volgens de regels van de OVER-schrijfwijzer. Om dit te ondersteunen zijn twee keuzes gemaakt: het gebruik van naïve-rag en het gebruik van een decoder-only model.

J.1 Keuze voor Naïve-RAG

De architectuur waar het uiteindelijk gekozen model zich in zal bevinden is de naïve-RAG-architectuur. Hiervoor is gekozen omdat er voor het herschrijven van de gemeentelijke brieven gebruik gemaakt moet worden van de OVER-schrijfwijzer. Deze OVER-schrijfwijzer, indien er aanpassingen nodig zijn, kan op elk moment worden geüpdatet binnen de vector database. Ook kan er aanvullende informatie worden toegevoegd aan deze database, waaronder bijvoorbeeld synoniemenlijsten.
De belangrijkste rol dat deze naïve-RAG-architectuur dient is het ondersteunen van de tekstgeneratieproces door middel van extra context, waaronder bijvoorbeeld regels uit de OVER-schrijfwijzer. Dit is nuttig voor tekst versimpeling, omdat het de consistentie verhoogt en de kans op onjuiste toevoegingen (hallucinaties) verkleint [53].

J.2 Modelsoorten

Bij de selectie van het taalmodel is eerst onderscheid gemaakt tussen verschillende architecturen: encoder-only, decoder-only en encoder-decoder. Hoewel deze indeling inzicht geeft in de globale werking van modellen, is de architectuur alleen niet voldoende om de geschiktheid voor deze taak te bepalen. Daarom is bij de verdere selectie gekeken naar aanvullende aspecten, zoals het kunnen volgen van de OVER-schrijfwijzer, ondersteuning voor de Nederlandse taal en de algehele betrouwbaarheid van het model.

J.2.1 Encoder-only.

Encoder-only modellen, zoals BERT, zijn sterk in tekstbegrip en analyse. Ze zijn geschikt voor taken zoals het inschatten van complexiteit of het vergelijken van betekenis. Omdat encoder-only modellen geen mechanisme hebben om zelfstandig nieuwe tekst te genereren, zijn ze niet bruikbaar als primair model voor het herschrijven van teksten [24]. Ze vallen daarom af voor de generatietaak.

J.2.2 Decoder-only modellen.

Decoder-only modellen zijn van oorsprong ontworpen voor tekstgeneratie. Waar oudere varianten soms minder accuraat waren, zijn moderne modellen verbeterd in het volgen van specifieke instructies (instruction following) [48]. Dit maakt ze geschikt om tekststijlen, zoals die van de OVER-schrijfwijzer, toe te passen. In combinatie met de juiste prompts en context kunnen ze worden ingezet voor tekstversimpeling.

J.2.3 Encoder-decoder modellen.

Encoder-decoder modellen worden traditioneel veel gebruikt voor 'sequence-to-sequence' taken zoals vertalen en samenvatten. De encoder verwerkt/transformeert de input en de decoder genereert de output, wat vaak goede resultaten oplevert op simplificatiebenchmarks [63]. Daarmee vormen zij een logische en valide keuze voor deze taak, al vergt het toepassen van specifieke stijlinstructies soms meer finetuning dan bij decoder-only modellen.

J.3 Modelkeuze binnen de Naïve-RAG

Voor dit onderzoek is uiteindelijk gekozen voor een decoder-only model. De doorslaggevende factor was de flexibiliteit waarmee deze modellen complexe instructies, zoals de specifieke regels van de OVER-schrijfwijzer, kunnen toepassen op de Nederlandse taal. Encoder-decoder modellen bieden hier minder flexibiliteit in. Verder zijn de risico's van decoder-only modellen, zoals het verzinnen van informatie in deze opzet verminderd door de RAG-architectuur. Hierbij gebruikt het model aanvullende informatie waarmee de herschrijf suggestie gegeven wordt [6].
Van de vele beschikbare modellen is uiteindelijk gekozen een model uit de Leesplank Noot-modellen te kiezen [6]. Aangezien deze modellen specifiek ontwikkeld zijn voor het schrijven naar B1-niveau, en voor publieke overheidscommunicatie. Dit past goed binnen de context van gemeentelijke teksten versimpelen.

K Samenhang tussen CEFR-leesniveaus en LiNT-II?

Deze bijlage laat zien hoe is onderzocht of LiNT-II-scores samenhangen met CEFR-leesniveaus in een geannoteerde dataset. Er wordt beschreven welke data is geselecteerd, welke analyses zijn uitgevoerd en hoe de resultaten zijn geïnterpreteerd. Dit dient als onderbouwing van de resultaten met betrekking tot LiNT-II-scores in het rapport.

K.1 Dataset

De dataset is een getransformeerde versie van data verzameld en geannoteerd door EDIA, een Amsterdams bedrijf gespecialiseerd in AI voor educatie [4]. De dataset bevat korte teksten die elk door drie taalexperts zijn gelabeld op CEFR-niveaus (A1 tot en met C2). Deze teksten zijn geen directe representatie van gemeentelijke brieven. De dataset bevatte de volgende kolommen:

- title
- lang
- source_name
- format
- category
- cefr_level
- licence
- text

De kolommen lang, source_name, format, category en license bevatten elk slechts één waarde, respectievelijk 'nl', 'elg-cefr-nl', 'document-level', 'reference' en 'CC BY-NC-SA 4.0'. Omdat deze waarden voor dit onderzoek naar de samenhang tussen LiNT-II-scores en CEFR-niveaus niet van belang waren, zijn deze kolommen verwijderd uit de dataset. De kolom title bevat meestal de naam van de tekst uit de originele bron, maar deze was voor de analyse ook niet relevant en is daarom ook verwijderd.
Verder bevat de dataset geen ontbrekende waarden (NaN). Wel komen er in de teksten nog kleine 'fouten' voor, zoals ongewenste regeleinden weergegeven in markdown-format. Om de analyses betrouwbaar te maken, zijn de teksten daarom opgeschoond.

K.2 Onderzoeksvariabelen

- **CEFR-niveaus:** De categorieën variëren van A1, A1+ tot C2+. Deze geven het leesniveau van de tekst aan zoals ingeschat door de experts.
- **LiNT-II:** LiNT-II is een leesbaarheidsinstrument dat specifiek is ontwikkeld voor Nederlandstalige teksten. De metric bepaalt de moeilijkheidsgraad van een tekst op basis van vier taalkundige kenmerken: woordfrequentie, syntactische afhankelijkheidslengte, het aantal inhoudswoorden per zin en het aandeel concrete zelfstandige naamwoorden. Omdat LiNT-II is afgestemd op Nederlandstalige teksten en empirisch is gevalideerd met beoordelingen door docenten en scholieren, is deze metric geschikt om te bepalen of teksten dichter bij de richtlijnen voldoen.

K.3 Data voorbereiding en analyse

Omdat sommige teksten door meerdere experts hetzelfde CEFR-label kregen, ontstonden er dubbele rijen in de dataset. Als deze niet verwijderd zouden worden, zouden sommige teksten te zwaar meewegen bij het berekenen van statistieken zoals gemiddelden en spreiding. Door alleen unieke combinaties van tekst en label te behouden, telt elke tekst even zwaar mee en zijn de resultaten betrouwbaarder. Daarom zijn de dubbele combinaties uit de dataset verwijderd.
Per tekst is de LiNT-II-score berekend. Vervolgens zijn de volgende descriptieve statistieken per CEFR-niveau verzameld:

- Gemiddelde LiNT-II-score
- Minimum LiNT-II-score
- Maximum LiNT-II-score
- Standaarddeviatie
- Eerste kwartiel (Q1)
- Derde kwartiel (Q3)
- Aantal teksten per niveau

De resultaten zijn gevisualiseerd in Figuur 26 en Figuur 27.

LiNT-II Scores per CEFR-niveau

CEFR-niveau	Gemiddelde LiNT-II	Min LiNT-II	Max LiNT-II	SD LiNT-II	Q1 LiNT-II	Mediaan LiNT-II	Q3 LiNT-II	Aantal teksten
A1	18.58	8.91	32.16	5.69	15.04	18.39	20.56	27.00
A1+	27.46	11.24	54.88	11.00	18.36	25.63	34.69	52.00
A2	35.02	14.45	62.09	10.51	27.49	35.51	42.52	177.00
A2+	37.83	11.24	65.35	9.76	30.90	38.73	44.37	247.00
B1	43.20	14.45	82.30	10.11	36.18	43.27	50.07	589.00
B1+	46.32	14.24	88.54	10.51	39.88	45.73	53.09	578.00
B2	49.87	22.52	100.00	10.87	42.56	49.29	56.11	543.00
B2+	51.39	22.15	100.00	11.99	43.59	51.03	57.71	267.00
C1	54.62	21.80	100.00	14.24	46.36	54.06	60.92	193.00
C1+	59.26	26.66	100.00	14.93	49.34	58.71	68.79	98.00
C2	65.22	27.03	100.00	18.54	53.61	61.68	76.12	69.00
C2+	63.81	40.11	100.00	18.53	49.54	62.99	75.26	13.00

Fig. 26. Statistieken LiNT-II-scores per CEFR-level

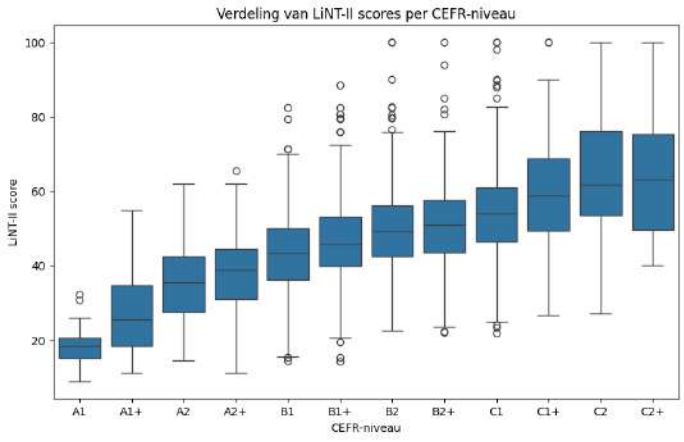


Fig. 27. Boxplot LiNT-II-scores per CEFR-level

Zoals te zien is in Figuren 26 en 27, is er aanzienlijke spreiding in LiNT-II-scores binnen elk CEFR-niveau. Hierdoor is het lastig om een directe koppeling te maken tussen een specifieke LiNT-II-score en een CEFR-niveau. Wanneer gekeken wordt naar de medianen en gemiddelde scores, is echter een oplopende trend zichtbaar bij hogere CEFR-niveaus, wat suggereert dat LiNT-II-scores indicatief kunnen zijn voor leesniveau. Tegelijkertijd zijn de teksten niet altijd consistent gelabeld. Sommige teksten kregen verschillende CEFR-leesniveaus toegewezen, soms zelfs met één of twee niveaus verschil [25]. Hierdoor blijft het lastig om een CEFR-niveau nauwkeurig aan een tekst te koppelen, en daarmee ook om een LiNT-II-score betrouwbaar aan een leesniveau te relateren.