AI Safety Rapport

Inzicht, controle en veiligheid bij gebruik van AI met Attendi

---

Rapport Q1 2026

attendi

Samenvatting

In dit AI Safety Rapport wordt uiteengezet hoe wij bij Attendi kunstmatige intelligentie inzetten op een veilige en verantwoorde manier binnen de zorg. We beschrijven hoe de Attendi app werkt bij het opnemen van gesprekken, het genereren van AI-samenvattingen en het herschrijven van ingesproken tekst via Schrijfhulp.

Voor elke functionaliteit brengen we de belangrijkste risico's in kaart zoals toestemming bij opnames, hallucinaties, misinterpretaties, automation bias en onbedoelde activatie. Vervolgens lichten we toe welke maatregelen in het product en de processen zijn ingebouwd om deze risico's te mitigeren.

Daarnaast gaat het rapport in op het volledige normenkader (AVG, NEN 7510, ISO 27001 en EU AI Act) en beschrijven we hoe Attendi gegevens verwerkt, minimaliseert en beveiligt. We geven inzicht in bewaartermijnen per type betrokkene en de technische en organisatorische beveiligingsmaatregelen die worden toegepast binnen de cloud infrastructuur van Microsoft Azure.

We leggen uit hoe Attendi gebruikmaakt van live monitoring, menselijke annotatie op schaal en een geautomatiseerde evaluatiepipeline om de kwaliteit, volledigheid en gegrondheid van AI-output continu te bewaken.

Inhoudsopgave

1. Inleiding

Bij Attendi werken we vanuit één missie: meer tijd voor zorg met aandacht. Waar zorgprofessionals en cliënten echt de tijd krijgen om elkaar goed te leren kennen ontstaat kwalitatieve zorg. Technologie kan dit verstoren, maar het kan daar ook bij helpen.

Bij Attendi ontwikkelen we oplossingen waarmee we tijd en aandacht proberen terug te geven door technologie op de achtergrond vooral ondersteunend te maken. De Attendi App is ontwikkeld met dat doel voor ogen. De app luistert mee met zorggesprekken en genereert automatisch een eerste versie van een samenvatting, die de professional vervolgens controleert en aanvult.

Daarnaast kan in de Attendi App (en in het ECD) gebruik worden gemaakt van “Schrijfhulp”. Schrijfhulp geeft zorgprofessionals de mogelijkheid om op basis van commando’s een ingesproken tekst door AI te laten herschrijven tot bijvoorbeeld een SOEP rapportage.

De inzet van AI ondersteunt hierbij de zorgprofessional op verschillende manieren bij het opstellen van verslaglegging zonder daarbij storend te zijn in de dynamiek met de cliënt. Tegelijkertijd vraagt de inzet van AI in de zorg om zorgvuldige toepassing.

AI in de zorg biedt kansen, maar brengt ook nieuwe verantwoordelijkheden met zich mee. Waar traditionele software een vaste, voorspelbare uitkomst kent, 0 of 1, goed of fout, heeft de inzet van AI in de praktijk altijd een foutmarge.

Een door AI geschreven samenvatting kan bijvoorbeeld onjuistheden bevatten of nuances missen.

Als leverancier vraagt dit dat je je op verschillende manieren inzet om deze toepassingen in de praktijk zo veilig mogelijk te laten verlopen. Met dit rapport maken we die risico's inzichtelijk en beschrijven we hoe we deze mitigeren.

2. Normenkader & Compliance

De verwerking van persoonsgegevens binnen de Attendi App vindt plaats in een zorgcontext en omvat het opnemen van gesprekken, het verwerken van transcripties en het genereren van AI-ondersteunde samenvattingen.

Dit brengt specifieke privacy- en veiligheidsverplichtingen met zich mee. Daarom werkt Attendi vanuit een helder normenkader. Het normenkader rust op vier pijlers:

- AVG – Algemene verordening gegevensbescherming
- NEN 7510 - Informatiebeveiliging in de zorg (ISMS)
- ISO 27001 - Informatiebeveiligingsmanagementsysteem (ISMS)
- EU AI Act - Verantwoorde en uitlegbare AI

De Algemene Verordening Gegevensbescherming (AVG)

De AVG vormt de kern van alle dataverwerking binnen Attendi. Attendi borgt de naleving van de AVG door:

- het uitsluitend verwerken van persoonsgegevens met een rechtmatige grondslag;
- het minimaliseren van gegevens (dataminimalisatie);
- het toepassen van privacy-by-design en privacy-by-default in de software;

- het uitvoeren van Data Protection Impact Assessments (DPIA's) bij nieuwe functionaliteiten met privacyrisico's;
- het inrichten van bewaartermijnen en toegangsbeperkingen;
- het aanbieden van volledige transparantie richting zorgorganisaties over datastromen.

In de DPIA is per verwerkingsstap vastgesteld welke persoonsgegevens worden verwerkt, wat het doel is en welke risico's daarbij ontstaan, inclusief de genomen mitigerende maatregelen.

NEN 7510 (informatiebeveiliging in de Zorg)

NEN 7510 is de Nederlandse norm voor informatiebeveiliging in de zorgsector. Attendi werkt volgens en is gecertificeerd voor de NEN-7510 norm, waarmee we garanderen dat:

- vertrouwelijkheid, integriteit en beschikbaarheid van gezondheidsinformatie gewaarborgd zijn;
- technische maatregelen zoals encryptie, logging en toegangsbeheer volgens zorgstandaard worden toegepast;
- risico's rond dataverwerking worden geïdentificeerd en periodiek herbeoordeeld;
- beveiligingsincidenten volgens een vast incidentresponsproces worden behandeld.

ISO 27001

Naast de zorgspecifieke norm NEN 7510 hanteert Attendi ook de internationale norm ISO 27001. Deze norm beschrijft eisen aan een volledig Information Security Management System (ISMS). ISO 27001 garandeert onder andere:

- een systematische aanpak om risico's te identificeren, evalueren en mitigeren;
- aantoonbare beveiligingsmaatregelen, zoals SSO/MFA, encryptie, versiebeheer en change management;
- periodieke security-audits en onafhankelijke certificering;
- rollen en verantwoordelijkheden binnen het securitybeheer;
- structurele monitoring, logging en documentatie.

Attendi is ISO 27001-gecertificeerd, wat betekent dat deze processen extern zijn getoetst en jaarlijks worden herbeoordeeld.

EU AI Act - Kaders voor veilige en uitlegbare AI

De Attendi App maakt gebruik van AI-functionaliteit voor transcriptie en samenvatting. De functionaliteit wordt ontworpen en geëvalueerd in lijn met relevante principes uit de EU AI Act en is voorbereid op de vereisten voor hoog-risico AI-systemen, voor zover van toepassing.

De volgende maatregelen zijn momenteel ingericht en operationeel, en sluiten aan bij kernvereisten uit de EU AI Act:

- Menselijke controle (human oversight): Zorgprofessionals beoordelen, corrigeren en accorderen AI-gegenereerde samenvattingen voordat deze in het dossier worden opgeslagen.
- Transparantie: AI-gegenereerde samenvattingen bevatten zichtbare disclaimers waaruit blijkt dat AI is ingezet.
- Geen geautomatiseerde besluitvorming: De AI ondersteunt uitsluitend; besluiten worden altijd door een mens genomen.
- Dataminimalisatie en bewaarbeperking: Audio-opnamen en transcripties worden binnen 24 uur verwijderd en niet gebruikt als trainingsdata zonder expliciete toestemming.
- Risicobeperking (inhoudelijk): Risico's met betrekking tot hallucinaties, misinterpretatie van output en het chilling effect op gebruikersgedrag zijn geïdentificeerd, geanalyseerd en vastgelegd in een risico- en maatregelentabel, inclusief bijbehorende mitigerende maatregelen.

Vooruitblik

In toekomstige AI Safety rapporten zal worden gerapporteerd over:

- de voortgang van bovenstaande implementaties;
- de effectiviteit van genomen mitigerende maatregelen;
- eventuele nieuwe risico's of aanpassingen naar aanleiding van wetgeving, gebruik of technische ontwikkelingen.

Betrokkenen en bewaartermijn

Attendi verwerkt persoonsgegevens uitsluitend voor het ondersteunen van zorgprofessionals bij verslaglegging. Hierbij hanteren we vaste principes: dataminimalisatie, transparantie, rechtmatigheid en controle door de gebruiker. Audio en transcripties worden uitsluitend tijdelijk verwerkt voor het genereren van de samenvatting en worden binnen 24 uur verwijderd. De uiteindelijke samenvatting wordt pas onderdeel van het zorgdossier nadat de zorgprofessional deze heeft gecontroleerd en goedgekeurd.

De Attendi App ondersteunt zorgprofessionals bij het efficiënt en accuraat vastleggen van rapportages via spraak. De app bevat functies voor spraakgestuurd rapporteren, het opnemen van gesprekken en het automatisch samenvatten via templates (standaard of eigen).

De oplossing van Attendi wordt gehost binnen Microsoft Azure datacenters in de Europese Unie. De verwerking van persoonsgegevens vindt uitsluitend plaats op Europese servers, waarop de AVG en overige Europese privacywet- en regelgeving volledig van toepassing zijn. Hierbij wordt gebruikgemaakt van Azure Data Zone Standard (Europa), waardoor opslag, verwerking en replicatie uitsluitend binnen Europese Azure-regio's plaatsvinden.

De functionaliteiten worden geleverd door Attendi Technology B.V., die optreedt als verwerker. De verwerkingsverantwoordelijkheid ligt bij de zorgorganisatie die Attendi inzet.

Categorie betrokkenen	Categorie persoonsgegevens	Persoonsgegevens	Type (AVG)	Bron /herkomst	Bewaartermijn
Cliënten	Identificatiegegevens	Voor- en achternaam, cliëntnummer, ECD-ID	Gewoon / identificatie	Ingevoerd door zorgprofessional / ECD	Conform bewaartermijn zorgdossier (20 jaar, Wgbo)
	Contactgegevens (indien genoemd)	Telefoonnummer, adres, e-mail	Gewoon	Inspraak / notities	Alleen zolang notitie in dossier bewaard blijft
	Gezondheidsgegevens	Medische toestand, behandeling, medicijngebruik, observaties, triage-info	Bijzonder (art. 9 AVG)	Audio-opname, transcript, samenvatting	Conform bewaartermijn zorgdossier
	Audio- en transcriptgegevens	Stemgeluid, uitspraken tijdens zorggesprek	Bijzonder (biometrisch mogelijk)	Opname via Attendi App	Audio max. 24 uur, transcript max. 24 uur, samenvatting conform dossier
	Metadata opname	Tijdstip, duur gesprek, device-ID	Gewoon / technisch	Attendi App backend	Max. 1–2 jaar (logging/technisch beheer)
	Toestemmingsgegevens (indien opgeslagen)	Digitale check	Gewoon (juridisch)	App (pop-up of opname)	Zolang opname/notitie wordt bewaard
Zorg-professionals (gebruikers)	Accountgegevens	Naam, e-mail, gebruikers-ID, rol	Gewoon / identificatie	Attendi (SSO/IdP)	Zolang account actief is
	Werkgerelateerde data	Functie, afdeling, toegangsrechten	Gewoon	HR / accountbeheer	Zolang account actief is
	Audio- en transcriptgegevens	Gesproken tekst professional tijdens opname	Gewoon (kan bijzondere data bevatten in combinatie met cliëntgegevens)	Attendi App	Audio/transcript max. 24 uur, samenvatting conform dossier
	Logging / gebruiksdata	IP-adres, inlogtijd, acties (start opname, goedkeuring)	Gewoon / technisch	Attendi App (audit trail)	1–2 jaar (afhankelijk van klantafspraken)
Overige betrokkenen (indirect)	Incidenteel genoemde derden	Namen van familieleden, mantelzorgers etc	Gewoon	Audio-opname / transcript	Zelfde bewaartermijn als opname/notitie
Attendi IT- & supportmedewerkers	Support- & beheersdata	Incidenteel toegang tot metadata, nooit tot inhoud	Gewoon	Supportcases / beheersystemen	Alleen zolang supportcase loopt

3. Bescherming van data & infrastructuur

Attendi verwerkt persoonsgegevens binnen een gecontroleerde en beveiligde Cloud omgeving. De Attendi App draait volledig binnen Microsoft Azure, waarbij gebruik wordt gemaakt van de beveiligingsmechanismen die standaard beschikbaar zijn binnen deze omgeving. De infrastructuur is opgebouwd met meerdere beveiligingslagen die gezamenlijk zorgen voor bescherming tegen ongeautoriseerde toegang en datalekken.

Logische scheiding en toegangsbeheer

Alle klantomgevingen worden logisch van elkaar gescheiden zodat gegevens nooit tussen organisaties kunnen worden uitgewisseld. Toegang tot systemen en data wordt strikt geregeld via role-based access control (RBAC). Alleen medewerkers met een aantoonbaar functionele noodzaak krijgen toegang tot specifieke onderdelen van het platform. Dit geldt zowel voor ontwikkelomgevingen als voor productiecomponenten. Toegang wordt continu gemonitord en volgens vaste procedures vastgelegd.

Netwerkbeveiliging en dataversleuteling

De applicatie draait binnen afgeschermde netwerksegmenten die worden beschermd met firewalls, netwerkrestricties en aanvullende Azure-beveiligingslagen. Alle gegevens worden standaard versleuteld. Tijdens transport wordt TLS 1.2 of hoger gebruikt. Voor opslag past Azure end-to-end AES-256 encryptie toe. Hierdoor zijn zowel audio, transcripties, samenvattingen als metadata beschermd tegen ongeautoriseerde inzage, ook wanneer fysieke opslag systemen gecompromitteerd zouden raken.

SSO, MFA en audit logging

Authenticatie verloopt via Single Sign-On (SSO) op basis van de identiteitsprovider van de zorgorganisatie. Wanneer de identiteitsprovider Multi-Factor Authenticatie ondersteunt, wordt deze automatisch afgedwongen binnen het loginproces. Daarnaast wordt alle relevante activiteit gelogd, waaronder inlogpogingen, systeemacties en gebruikershandelingen. Deze logging wordt centraal beheerd via Azure Application Insights en Log Analytics, waardoor monitoring, auditability en incidentanalyse zijn geborgd.

Incidentresponsproces in hoofdlijnen

Alle beveiligingsmaatregelen maken onderdeel uit van het Information Security Management System (ISMS) van Attendi. Daardoor worden risico's periodiek geëvalueerd, maatregelen geactualiseerd en beveiligingsincidenten volgens vaste procedures behandeld. In combinatie met jaarlijkse audits, pentests en certificering volgens NEN 7510 en ISO 27001 wordt structureel geborgd dat de infrastructuur en dataverwerking blijven voldoen aan de eisen voor veilige en betrouwbare inzet van AI in de zorg.

Dataminimalisatie en toegangsrechten

Attendi verwerkt uitsluitend gegevens die noodzakelijk zijn voor het functioneren van de Attendi App. Audio en transcripties worden alleen tijdelijk bewaard voor verwerking en daarna automatisch verwijderd. Alleen de gecontroleerde samenvatting blijft beschikbaar voor de zorgprofessional volgens de geldende bewaartermijnen van het zorgdossier.

Toegang tot gegevens is strikt geregeld via role-based access control en interne autorisatieprocedures. Alleen medewerkers met een aantoonbare functionele noodzaak krijgen toegang, waarbij de data van iedere zorgorganisatie logisch van elkaar gescheiden is. Loggegevens worden volgens ISMS-normen bewaard om toezicht, auditing en incidentanalyse mogelijk te maken, zonder onnodige opslag van gevoelige informatie.

4. Attendi App prestaties & risico's

Het gebruik van de Attendi App brengt risico's met zich mee die actief moeten worden gemitigeerd. Hieronder beschrijven we per functionaliteit de belangrijkste risico's, evenals de maatregelen die Attendi heeft ingericht om zorgprofessionals te ondersteunen bij veilig en verantwoord gebruik van de app.

Per risico lichten we toe welke mitigatiestrategie in het product is ingebouwd en hoe we zorgprofessionals helpen om zich bewust te zijn van het betreffende risico.

Risico	**Mitigatie Strategie**
1.1 Opname van gesprek zonder toestemming gesprekspartner	• In App E-Learning • Consent Flow in product
1.2 Zorgprofessional is onbewust dat samenvatting door AI is gemaakt	• AI-Disclaimer • In App E-Learning
1.3 Zorgprofessional keurt samenvatting goed zonder naar de inhoud te kijken.	• In App E-Learning • AI-Disclaimer
1.4 Samenvattingen met fouten worden wel inclusief fouten doorgezet naar andere systemen.	• AI-Disclaimer
1.5 Samenvatting bevat misinterpretaties of hallucinaties.	• AI Evaluaties & Product verbeteringen • In-App feedback op samenvattingen
1.6 Automatische kwaliteitsmonitoring van samenvattingen	• AI-gebaseerde kwaliteitsmonitoring • Classificatie van zinnen op risiconiveau • In-App feedback op samenvattingen

1.1 Opname van het gesprek zonder toestemming van gesprekspartner

Het opnemen van een gesprek met ten minste twee deelnemers is alleen mogelijk wanneer alle deelnemers expliciet akkoord zijn met de opname. Het risico dat hiermee wordt gemitigeerd, is dat een gesprek wordt opgenomen zonder dat iedere deelnemer daar uitdrukkelijk toestemming voor heeft gegeven. We mitigeren dit risico op twee manieren.

In App E-Learning

In de App E-Learning is binnen de leerroute een stap opgenomen met de titel "Het opnemen van een gesprek". In deze stap bekijken zorgprofessionals een video waarin wordt uitgelegd hoe het opnemen van een gesprek werkt, inclusief het bijbehorende risico dat alle deelnemers expliciet akkoord moeten zijn met de opname. Daarnaast lichten we toe hoe deze toestemming in de Attendi App wordt vastgelegd en toegepast. Via het Attendi-dashboard kan een projectleider of beheerder vervolgens inzicht krijgen in hoeveel zorgprofessionals deze competentie hebben behaald, nadat zij de opdracht succesvol hebben afgerond.

Consent flow in product

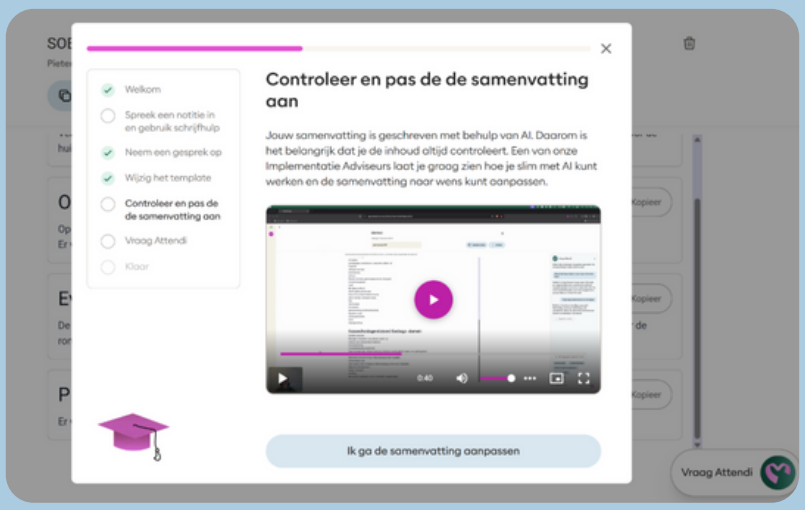
Wanneer binnen de Attendi App een gesprek wordt opgenomen, dient de zorgprofessional expliciet toestemming (consent) te geven. Alleen na het geven van deze toestemming kan de opname worden gestart. Door deze consent-flow in het product te integreren, mitigeren we het risico dat gesprekken worden opgenomen zonder toestemming van alle gesprekspartners.

1.2 Zorgprofessional is onbewust dat samenvatting door AI is gemaakt

Het is mogelijk dat een zorgprofessional onbewust een door AI gegenereerde samenvatting gebruikt in zijn of haar werk. Het risico dat we willen mitigeren, is dat de zorgprofessional niet doorheeft dat de samenvatting door AI is opgesteld. Dit risico pakken we aan via twee maatregelen.

In App E-learning

Een van de onderdelen van de in App E-Learning is de uitleg over het gebruik van AI-samenvattingen in hun werk. In deze E-Learning module leggen we de risico's van het gebruik van AI-samenvattingen uit en wordt het belang van menselijke controle benadrukt.



AI-Disclaimer

Elke AI-samenvatting bevat een standaarddisclaimer die duidelijk maakt dat:

- de samenvatting door AI is gegenereerd;
- de zorgprofessional verplicht is de inhoud te controleren voordat deze gebruikt kan worden.

Dit voorkomt misinterpretatie en ondersteunt transparantie richting collega's en cliënten. Deze disclaimer is zichtbaar bij de samenvatting in de Attendi App.

1.3 Zorgprofessional keurt samenvatting goed zonder naar de inhoud te kijken.

De gegenereerde samenvatting wordt zonder adequate en kritische blik goedgekeurd door de zorgprofessional. Dit risico pakken we als volgt aan:

In-App E-learning

In de In-App E-learning leggen we zorgprofessionals uit waarom het belangrijk is om AI-gegenereerde samenvattingen kritisch te beoordelen. Door deze bewustwording te vergroten, mitigeren we het risico dat samenvattingen worden gebruikt zonder inhoudelijke controle.

AI-Disclaimer: Expliciet boven de samenvatting

Wanneer een eerste versie van de samenvatting door Attendi zonder wijzigingen wordt gekopieerd, dan vermeldt Attendi dat altijd boven de samenvatting. Attendi plaatst daar dan de tekst:

"Deze samenvatting is gemaakt met AI in de Attendi App en is niet handmatig bewerkt. Lees deze zorgvuldig door."

Er bestaat een risico dat de gegenereerde tekst op basis van audio van zorgprofessionals en cliënten misinterpretaties of hallucinaties bevat, waardoor niet-onderbouwde informatie kan ontstaan.

1.4 Samenvattingen met fouten worden inclusief fouten doorgezet naar andere systemen

Het is mogelijk dat zorgprofessionals de gegenereerde samenvattingen direct gebruiken zonder deze goed te controleren op fouten. Hierdoor kunnen samenvattingen met fouten terecht komen in andere systemen.

Net als bij 1.3 passen we de AI disclaimer toe boven de samenvatting die zonder wijzigingen wordt doorgezet.

1.5 Samenvatting bevat misinterpretaties of hallucinaties.

Er bestaat een risico dat de gegeneerde tekst op basis van de audio van zorgprofessionals en cliënten misinterpretaties of hallucinaties niet gegronde informatie bevat.

In App Feedback

Zorgprofessionals kunnen binnen de Attendi App direct feedback geven. Dit kan in de App in zijn geheel of op specifieke uitkomsten of templates. Tijdens de Attendi App pilot periode is er feedback van zorgprofessionals binnen gekomen. Op basis daarvan zijn aanpassingen gedaan aan de prompts. Hierdoor ervaren zorgprofessionals in zeer beperkte mate hallucinaties.

Naast de feedback van zorgprofessionals is er ook automatische monitoring van de kwaliteit van de samenvattingen. Hier gaan we dieper op in, in de volgende paragraaf.

1.6 Automatische monitoring van de kwaliteit van samenvattingen

Gebruikersfeedback geeft ons een goed beeld van anekdotische ervaringen en is onmisbaar in productverbetering, maar om de kwaliteit van samenvattingen op schaal te kunnen waarborgen terwijl we de privacy van cliënten garanderen, automatiseren we ook de monitoring met AI. Hierbij gebruiken we Large Language Models (LLM's) die we de opdracht geven om op gedetailleerd niveau te kijken naar de volledigheid en gegrondheid van samenvattingen.

De volledigheid in deze context meten we door de kernpunten uit het transcript te halen en het LLM te vragen hoeveel van deze kernpunten daadwerkelijk terugkomen in de samenvatting.
De gegrondheid meten we door iedere zin in de samenvatting te classificeren als (i) laag risico: expliciet genoemd of eenvoudig af te leiden uit het transcript (ii) mild risico: er is een interpretatie geweest van wat is gezegd of (iii) verhoogd risico: er is een interpretatie die mogelijk tegenstrijdig is met het transcript of niet gegrond is.

Omdat we op het niveau van individuele zinnen en kernpunten beoordelen, kunnen we gericht zoeken naar gebreken in het systeem. Bovendien zijn hierdoor de metingen ook goed te interpreteren. Dit samen maakt het mogelijk om met gesimuleerde gesprekken incrementele stappen te zetten richting nog betere samenvattingen in de Attendi App.

Het gebruik van een LLM voor het automatiseren van kwaliteitsmonitoring komt echter met een extra uitdaging: het oordeel van het monitorende LLM moet namelijk zoveel mogelijk overeenkomen met dat van de zorgprofessional.

Naast echte gesprekken werken we ook met gesimuleerde gesprekken: zorgvuldig samengestelde of synthetische casussen waarmee we specifieke situaties en fouttypes kunnen nabootsen, zonder echte cliëntgegevens te hoeven gebruiken. Omdat we op het niveau van individuele zinnen en kernpunten beoordelen, kunnen we in zowel echte als gesimuleerde gesprekken gericht zoeken naar gebreken in het systeem. Bovendien zijn de metingen hierdoor ook goed te interpreteren. Dit samen maakt het mogelijk om incrementeel, scenario voor scenario, toe te werken naar nog betere samenvattingen in de Attendi App.

In toekomstige AI Safety Rapporten zullen we verder ingaan op volledigheid en gegrondheid door te kijken naar de eerste grootschalige resultaten van deze metrics en de correlatie tussen het menselijk oordeel versus dat van een LLM.

5. Schrijfhulp prestaties & risico's

Schrijfhulp ondersteunt zorgprofessionals bij het herschrijven van ingesproken tekst naar gestructureerde verslaglegging. Deze functionaliteit biedt belangrijke voordelen, maar brengt ook specifieke risico's met zich mee. De risico's bevinden zich op twee niveaus:

1. Het herkennen van de intentie (het moment waarop de gebruiker een herschrijfopdracht geeft)
2. De uitvoering van de transformatie (het herschrijven door een LLM)

Het herkennen van de intentie (het moment waarop de gebruiker een herschrijfopdracht geeft)

De intentiedetector van Schrijfhulp bepaalt of een zorgprofessional daadwerkelijk een opdracht geeft om tekst te laten herschrijven. Hierbij kunnen twee typen fouten optreden:

- False Positives (FP's): De detector activeert Schrijfhulp terwijl de gebruiker géén transformatie heeft bedoeld. Dit vormt een veiligheidsrisico, omdat hierdoor onbedoeld wijzigingen kunnen worden aangebracht in zorginhoudelijke tekst.
- False Negatives (FN's): Een geldige opdracht wordt niet herkend. Dit heeft vooral impact op de gebruikservaring, omdat Schrijfhulp dan niet wordt uitgevoerd wanneer dat wel gewenst is. FN's leveren geen direct veiligheidsrisico op.

Omdat onbedoelde activatie (FP's) kan leiden tot foutieve of ongewenste verslaglegging, ligt de focus van risicobeheersing op het minimaliseren van FP's.

Risico bij intentiedetectie	Mitigerende maatregelen
False Positives (FP's): Schrijfhulp wordt geactiveerd zonder dat de gebruiker een transformatie bedoelt. Dit kan leiden tot onbedoelde wijzigingen in zorginhoudelijke tekst.	- Eigen intern intentiemodel, getraind op domeinspecifieke zorgdata - Kalibratie gericht op een zeer lage FP-rate - Twee-stapsproces: intentiedetectie en aanvullende sanity check door LLM - LLM weigert transformatie bij ontbrekende of onduidelijke intentie
False Negatives (FN's): een geldige opdracht wordt niet herkend. Dit beïnvloedt vooral de gebruikservaring, maar vormt geen veiligheidsrisico.	- We accepteren bewust een beperkt aantal FN's om te voorkomen dat het systeem automatisch een herschrijfopdracht uitvoert zonder dat de gebruiker dit echt bedoelt (onveilige auto-triggering). - Doorlopende optimalisatie van intentiemodel en trainingsdata

Mitigerende maatregelen bij intentiedetectie

Attendi heeft verschillende maatregelen geïmplementeerd om de intentiedetectie van Schrijfhulp zo veilig en betrouwbaar mogelijk te maken.

Eigen intentiedetectiemodel

Attendi ontwikkelt en beheert een intern intentiemodel dat specifiek is getraind op domeingerichte datasets uit de zorgpraktijk. Hierdoor sluit de herkenning nauw aan op de manier waarop zorgprofessionals hun opdrachten formuleren.

Kalibratie gericht op minimale False Positives

Tijdens de afstelling wordt expliciet gestuurd op een zeer lage FP-rate. Een beperkt aantal False Negatives wordt daarbij geaccepteerd, omdat dit de veiligheid vergroot en voorkomt dat Schrijfhulp onbedoeld wordt geactiveerd.

Gelaagde systeemarchitectuur (twee-stapsproces)
Schrijfhulp werkt met een dubbele controle:

- Stap 1: Detectie van intentie.
- Stap 2: Een aanvullende ‘sanity check’ door het LLM, dat de opdracht weigert wanneer de input geen geldige intentie bevat.

Bewezen veiligheid in productie
Dankzij deze gelaagde aanpak zijn in de productieomgeving tot nu toe vrijwel geen False Positives geconstateerd, wat aangeeft dat de strategie effectief is in het voorkomen van onbedoelde transformaties.

De uitvoering van de transformatie (het herschrijven door een LLM)

Wanneer Schrijfhulp een ingesproken tekst herschrijft naar een gestructureerde rapportage, maakt de functionaliteit gebruik van een LLM (Large Language Model). Dit proces brengt een aantal risico's met zich mee die inherent zijn aan generatieve AI-modellen. Deze risico's kunnen, indien niet goed beheerst, leiden tot onjuiste of misleidende samenvattingen.

Risico bij transformatie (LLM)	**Mitigerende maatregelen**
Hallucinatie of ongewenste toevoegingen. Het model voegt informatie toe of wijkt af van de input.	- Strikt ontworpen prompts met harde grenzen - Verbod op toevoegingen, diagnostiek en conclusies - Iteratieve validatie met productievoorbeelden
Misinterpretatie van spraak door spraak-naar-tekst variatie	- Domeinspecifieke spraak-naar-tekst tuning - Verplichte handmatige controle door de gebruiker
Automation bias (oververtrouwen op AI-output)	- Duidelijke AI-disclaimers in de output - In-app e-learning over zorgvuldig gebruik
Structuurfouten of incorrect gebruik van verslagleggingsformats (bijv. SOEP)	- Templatevalidatie en hardcoded structuurregels - Monitoring van veelvoorkomende structuurafwijkingen
Regressie of nieuwe foutpatronen na prompt- of modelupdates	- Uitgebreide regressietests bij elke update (tests waarin we controleren of eerder goed werkende functionaliteit na een wijziging nog steeds correct functioneert, zodat we voorkomen dat bestaande kwaliteiten per ongeluk verslechteren) - Validatie met realistische productievoorbeelden - Snelle prompt- of modelaanpassingen binnen de safety-cycle

Om bovenstaande risico's te beperken, heeft Attendi een uitgebreid geheel aan safeguards ingericht die de kwaliteit, veiligheid en betrouwbaarheid van Schrijfhulp bewaken.

Strikt ontworpen prompts met harde beperkingen

Alle transformaties worden uitgevoerd binnen strikt afgebakende instructies die exact definiëren wat het LLM wel en niet mag doen.

Het model mag expliciet géén:

- diagnoses stellen,
- conclusies trekken of inferenties maken,
- informatie toevoegen die niet door de gebruiker is uitgesproken.

Deze instructies worden continu geoptimaliseerd op basis van productievoorbeelden.

Iteratieve validatie en regressietests

Bij iedere wijziging aan het model of de prompts worden uitgebreide regressietests uitgevoerd. Nieuwe foutpatronen worden zo vroeg mogelijk gedetecteerd en gecorrigeerd.

Snel aanpasbare instructies binnen de safety-cycle

Wanneer nieuwe risico's in de praktijk zichtbaar worden, kunnen prompt instructies onmiddellijk worden aangepast, waardoor fouten snel kunnen worden gemitigeerd.

Monitoring op schaal van Schrijfhulp

Attendi past meerdere lagen van monitoring toe om de juistheid en veiligheid van Schrijfhulp in de praktijk te waarborgen.

1. Live Monitoring

In de productieomgeving worden (waar toegestaan) gebruiksstatistieken en intentieverdeling gevolgd. Daarnaast wordt gecontroleerd op afwijkende patronen of onverwachte verschuivingen in transformaties. Modelupdates worden vóór uitrol getest met grootschalige regressie- en foutanalyses om regressies te voorkomen.

2. Menselijke evaluatie op schaal

Attendi maakt gebruik van een intern annotatiesysteem waarmee productievoorbeelden handmatig worden beoordeeld. Dit helpt om foutpatronen snel te herkennen en om verbeteringen te valideren voordat deze worden doorgevoerd.

3. Geautomatiseerde evaluatie (in ontwikkeling)

Attendi ontwikkelt een geautomatiseerde evaluatiepipeline die outputs beoordeelt op juistheid, trouw aan de input, hallucinatie en naleving van transformatierichtlijnen.

- Fase 1: Offline evaluatie naast productie om prestaties te benchmarken.
- Fase 2: Live beoordeling, waarmee verdachte outputs automatisch kunnen worden gemarkeerd of geblokkeerd.

Het human-annotation systeem is volledig operationeel. De geautomatiseerde evaluatie wordt de komende 1–2 maanden gefaseerd uitgerold.

6. Conclusie

Met dit AI Safety Rapport maken wij inzichtelijk op welke wijze wij bij Attendi AI op een veilige, transparante en controleerbare manier inzetten binnen de zorgpraktijk. We beschrijven zowel de technische als de organisatorische maatregelen die bijdragen aan verantwoorde toepassing van AI, inclusief duidelijke kaders voor gegevensverwerking, beveiliging, menselijke controle en continue kwaliteitsborging. De risico's die samenhangen met het gebruik van AI, zoals bij het luisteren, samenvatten en het gebruik van Schrijfhulp, hebben wij volledig in kaart gebracht en voorzien van concrete mitigerende maatregelen.

Bij de ontwikkeling van AI hanteren wij een stapsgewijze en zorgvuldige aanpak: we onderzoeken en testen nieuwe functionaliteit eerst uitgebreid, simuleren diverse scenario's en voeren gerichte verbeteringen door voordat we deze technologie op grotere schaal inzetten.

Pas wanneer de kwaliteit en veiligheid aantoonbaar zijn geborgd, brengen wij een AI model of functionaliteit in productie. Vanaf dat moment start de structurele monitoring, waarbij we samen met zorgprofessionals uit de praktijk inzichten verzamelen om de AI modellen en processen verder te verfijnen.

Om deze continue cyclus van leren en verbeteren te ondersteunen, zullen wij periodiek een bijgewerkte versie van dit AI Safety Rapport publiceren. In het vervolgrapport blikken we terug op het gebruik in de praktijk, de prestaties van de AI en de trends in risico's en mitigaties.

In elke nieuwe editie delen we de nieuwste inzichten, verbeteringen, modelupdates en veiligheidsbevindingen.

Op deze manier blijft de inzet van AI binnen de zorg niet alleen effectief, maar vooral ook veilig, verantwoord en transparant.

Berend Jutte
Mede-oprichter
Verantwoordelijk voor productontwikkeling

attendi