

# KMU Projekt “SmartDialogue Engine” (SDE) - Schlussbericht

Vorhaben: KI-Modell zur Optimierung des Dialogverhaltens digitaler Telefonassistenten (SDE - SmartDialogue Engine)

Teilvorhaben: Erforschung eines KI-Modells zur Optimierung des Dialogverhaltens am Beispiel eines bestehenden medizinischen Telefonassistenten

Förderkennzeichen: 16SV8846

Projektkoordinator: Aaron GmbH, Berlin

Zuwendungsempfänger: Aaron GmbH, Berlin

DFKI - Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken/Berlin

Autoren: Dr. Patrick Jähnichen, Aaron GmbH

Erstellungsdatum: 30.09.2024



Bundesministerium  
für Bildung  
und Forschung



# Teil I: Kurzdarstellung nach NKBF 98 8.2 I

## Aufgabenstellung

Das Ziel des vorliegenden Projekts war es, ein KI-Verfahren (die "Smart Dialogue Engine", SDE) zu erforschen, das das Kommunikationsvermögen telefon-basierter Sprachassistenten im Gesundheitsbereich besser an die Bedürfnisse und Signale von Anrufer:innen sowie an den jeweiligen Kontext anpasst. Dadurch sollte die Akzeptanz gesteigert, das Praxispersonal weiter entlastet und die Patient:innenzufriedenheit durch einen natürlichen Dialogfluss verbessert werden. Für andere Anbieter von Voice-Lösungen sollten zudem innovative Metriken zur Optimierung des Kommunikationserfolgs bereitgestellt werden, bis hin zur Messung der Konversationsgüte in Echtzeit.

## Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Projekt "SmartDialogue Engine" wurde als Kooperationsprojekt zwischen der Aaron GmbH Berlin und dem Deutschen Forschungszentrum für künstliche Intelligenz (DFKI) durch das deutsche Bundesministerium für Bildung und Forschung im Rahmen des Förderprogramms "KMU-innovativ" unter dem Förderkennzeichen 16SV8846 und im Förderbereich "Forschungsbereich Mensch-Technik-Interaktion für den demographischen Wandel" gefördert. Die Betreuung seitens des Projektträgers wurde durch die VDI/VDE sichergestellt. Die Projektlaufzeit betrug 24 Monate (01.03.2022-29.02.2024) wobei das Projekt einmalig kostenneutral um sechs Monate bis zum 31.08.2024 verlängert wurde.

## Planung und Ablauf des Vorhabens

Der Lösungsansatz zur dynamisch-adaptiven Optimierung des Dialogverhaltens von Sprachassistenzsystemen basierte auf drei Phasen und legte besonderen Fokus auf die zwei zu Beginn herausgestellten Bereiche: Metakommunikative Konventionen und Kontextinformation. Der Fokus von Aaron lag hierbei auf der technischen Umsetzung der Anforderungen in ein KI-Modell zur Dialogoptimierung.

Aaron brachte seine Expertise wie folgt ein:

- Datenschutzkonforme Bereitstellung von Trainingsdaten für die Erforschung und Entwicklung eines ML-Modells zur laufenden Vorhersage des Dialogerfolgs.
- Kenntnisse über Bedürfnisse und Anforderungen der Endnutzer sowie Netzwerk und Zugang zu interessierten Anwendern für Nutzerbefragungen.
- Experimentelle Entwicklung eines ML-Modells, zur laufenden Vorhersage des Dialogerfolgs sowie abgeleitetes Modell zur Identifikation analoger Teildialogverläufe mit gehäuften Kommunikationsproblemen. Dieses wurde anhand einer API-Schnittstelle und eines Web User Interface (UI) um die Möglichkeit zur Optimierung von Dialogen ohne eigenes Vorwissen zugänglich gemacht.

Es sei darauf hingewiesen, dass über die ursprünglich projektierten Ziele hinaus eine Verbesserung der Spracherkennung evaluiert und umgesetzt wurde, in erster Linie um die Umsetzbarkeit des Projektes sicherzustellen. Weiterhin führte die Verfügbarmachung leistungsstarker generativer Textmodelle nach Projektstart zu einer Neubewertung einzelner Arbeitspakete und ermöglichte stattdessen weitergehende Innovationsansätze. Details zu diesen Änderungen finden sich im Abschnitt "Herausforderungen und Anpassungen". Darüber hinaus wurden die notwendigen Anpassungen bereits in den eingereichten Zwischenberichten angekündigt und inhaltlich motiviert.

## Kooperationen

Neben der Aaron GmbH und dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) waren keine weiteren Kooperationspartner am Projekt beteiligt.

Zwischen der Aaron GmbH und dem DFKI wurde eine Rollenverteilung wie folgt vereinbart:

- Aaron GmbH:
  - Projektleitung und -koordination: Die Aaron GmbH war verantwortlich für die Leitung und Koordination des Projekts.
  - Entwicklungs-, Integrations- und Verwertungsleistungen: Die Aaron GmbH übernahm die technische Entwicklung, Integration der neuen Lösungen in das bestehende System und die Verwertung der Projektergebnisse.
- DFKI:
  - Wissenschaftliche Begleitung und Beratung: Das DFKI unterstützte das Projekt durch wissenschaftliche Beratung und Begleitung.
  - Erstellung der Erfolgsmetrik: Das DFKI entwickelte die Metrik zur Messung des Kommunikationserfolgs ("dialogue smoothness").
  - Stress-Framework und Modellentwicklung: Das DFKI war verantwortlich für die Entwicklung von Modellen zur Emotionserkennung und Sentimentanalyse sowie für das Stress-Framework zur Bewertung der Dialogqualität.

Im Rahmen des Projekts wurden mehrere Formen der Zusammenarbeit organisiert:

- Workshops: Es wurden drei ganztägige Workshops abgehalten: ein Kickoff-Workshop sowie zwei Treffen zur engeren Zielabstimmung und -justierung.
- Regelmäßige Treffen: Es fanden im zweiwöchigen Takt Treffen des Projektteams (alle direkt projektbeteiligten Personen) statt. Zusätzlich gab es im vierwöchigen Takt Treffen mit erweitertem Teilnahmekreis, einschließlich weiterer Stakeholder der Aaron GmbH (CTO, Produktverantwortliche, Platform-Team und Produktentwickler), zur Fortschritts- und Zielkontrolle.

Die enge und gute Zusammenarbeit zwischen der Aaron GmbH und dem DFKI führte zu mehreren wichtigen Ergebnissen:

- Produktverbesserungen: Es konnten vielversprechende Produktverbesserungen identifiziert und vorbereitet werden. Deren Auswirkungen auf die Wahrnehmung des Dialogverlaufs wurden gemessen und analysiert.
- Inklusive Ansprache: Die Zusammenarbeit zeigte, dass Dialoganpassungen in Inhalt und Form geeignet sind, eine inklusive Ansprache aller möglichen Gesprächspartner zu erreichen.
- Prompt-Design: Durch die enge Kooperation wurden schnelle und vielversprechende Fortschritte im Prompt-Design erzielt. Prompts sind natürlichsprachliche Anweisungen an LLMs zum Dialogverlauf, zu Anpassungen des Dialogs oder zur dynamischen Dialoggenerierung.
- Ergebnispublikation: Gemeinsame Publikation (Soliman et al., 2024) von Zwischenergebnissen für eine optimierte Nutzendenansprache in Zusammenarbeit mit DFKI.

# Teil II: Eingehende Darstellung

## Einordnung

Die Aaron GmbH entwickelt und vertreibt einen KI-basierten automatischen Telefonassistenten, der Ärzt:innen und medizinische Fachangestellte (MFAs) bei der Entgegennahme und effizienten Bearbeitung von Patient:innen-Anrufen unterstützt. Dabei nimmt der Assistent die Telefongespräche entgegen, kategorisiert sie nach Anliegen und sammelt in der Konversation per Spracheingabe die für die Praxis zur Bearbeitung notwendigen Informationen.

## Ausgangslage

### Stand der Technik und vergleichbare Lösungen

Der Stand der Technik beschäftigt sich schon länger mit den Themen KI und Chatbots, insbesondere im Marketing und der Kundenbetreuung. Die rasante Entwicklung von Sprachassistenten wie Amazon Alexa, Google Home oder Apple Siri zeigt, welche Bedeutung diese Entwicklungen für Marketing, Kundenbeziehungen und Kommunikation generell haben. Folgende Tabelle gibt einen Überblick über Produkte und Systeme, von denen sich das Vorhaben abgrenzt:

Name	Beschreibung	Abgrenzung
Software-Lösungen für Entwickler/Agenturen wie rasa.ai oder Google Dialogflow	Freie Dialogführung mit spontaner und intuitiver Interaktion. Komplexe oder gar unvollständige Eingaben werden verstanden, und bei Unklarheit fragen die modernsten Systeme bereits eigenständig nach. Die Berücksichtigung von Kontextwissen, von persönlichen Vorlieben und Bedürfnissen der Nutzer ermöglicht eine proaktive Unterstützung mit Empfehlungen und Vorschlägen.	Integration in direkt nutzbares Produkt ist nicht möglich. Die Anwendung ist nur für Entwickler ausgelegt und nicht für KMUs. Die Formulierungen im Dialog werden selbst gewählt. Die Funktionsweise basiert rein auf der Textebene. Hierdurch wird die nonverbale Meta Metakommunikation vollständig ignoriert.
Sprachassistenten-Systeme für Endanwender wie Alexa/Google Home/Siri	Intelligente Lautsprecher, gekoppelt an sprachgesteuerte, cloud-basierte, intelligente persönliche Assistenten wie	Integration in bestehende Telefonanlage schwierig bis unmöglich. Formulierungen im Dialog müssen selbst gewählt werden. Kein Fokus

	Alexa. Über Alexa Skills (bzw. Google Actions) können neue Funktionen über Drittanbieter integriert werden.	auf KMU. Kaum Unterstützung meta-sprachlicher Signale. Keine Erfolgsmetriken für das Dialogverhalten.
Echtzeit-Konversationsanalyse und -Coaching wie i2x.ai	Angeboten werden KI-Methoden, die das Verhalten von Servicepersonal in der Kundenkommunikation analysieren und vergleichen und darauf aufbauend Verbesserungsvorschläge machen ("Silent Advisor").	Bisher keine Anwendung auf Sprachassistenten, sondern nur auf menschliche Interaktionen. Fokus auf Wortlaut und teilweise Sprachemotion, aber noch kein besonderes Augenmerk auf metakommunikative Signale, da Menschen diese i.d.R. ohnehin besser beherrschen.

In Abhandlungen, wie (Suendermann et al., 2010) geht es um die automatisierte Gesprächsgestaltung bei Telefonanrufen. Der Fokus der Abhandlung liegt allerdings auf der statischen Optimierung der Entscheidungsfindung bei Sprachassistenten. (Suhm & Peterson, 2002) entwickelten eine Methodik zur datengesteuerten Bewertung von telefonischen Sprachassistenzsystemen, um die Benutzerfreundlichkeit und Kosteneffizienz zu steigern. Der Fokus der Veröffentlichung liegt auf großen Call-Centern und liefert eher allgemeine, statische Optimierungshinweise, die nicht anwenderbezogen sind und nicht von KI-basierten Telefonassistenten genutzt werden können. Erst jüngere Forschungsarbeiten beschäftigen sich mit der Bedeutung von Metakommunikation und zeigen, dass diese in heutigen Sprachassistenzsystemen nur rudimentär unterstützt wird (Ito, 2020).

Metakommunikative Signale seitens des Nutzers werden weitgehend ignoriert. Auch so Grundlegendes wie die Signale zum Sprecherwechsel (turn taking) oder Reparaturen (d.h. spontane Selbst-Korrekturen) folgen Konventionen und werden durch spezifische metakommunikative Signale verhandelt. Schlägt dies fehl, fallen sich die Kommunikationspartner ins Wort oder missverstehen sich, was bei Sprachassistenzsystemen oft zum Zusammenbruch der Kommunikation oder der Ablehnung des Systems führt.

Neben den metakommunikativen Aspekten bilden heutige Sprachassistenzsysteme die Einbindung von Kontextinformationen nur unzureichend ab. Insbesondere sind sie in ihren Möglichkeiten heute sehr eingeschränkt, was die nachträgliche Konfiguration von Gestaltungswünschen der Betreiber des Sprachassistenzsystems (z.B. System-Persona, Wortwahl, Dialogverhalten) anbetrifft. Diese Parameter sind heute oft "hartkodiert" und nicht adaptiv, d.h. sie passen sich nicht dynamisch an Kontextinformation an. Die größten technischen Verbesserungen der letzten Jahrzehnte an Sprachassistenten bestanden darin, die Variabilität der Nutzereingaben (z.B. bzgl. Akustik, Sprechereigenschaften, Artikulation, etc.) als Störgrößen zu modellieren. Ein Großteil der Kontextinformation sowie Signale zur Metakommunikation gehen dabei aber verloren und stehen in späteren Verarbeitungsstufen

nicht mehr zur Verfügung. Außerdem erhöht sich die Menge freier Parameter exponentiell. Aktuell gibt es für die dargestellte Problemstellung in der Gesamtheit noch keinen Lösungsansatz. Der im Rahmen dieses Projektes verfolgte Ansatz birgt relevante technische, linguistische, sozialwissenschaftliche und datenschutz-spezifische Herausforderungen und hebt sich in mehreren Punkten vom Stand der Technik und vergleichbaren Lösungen ab:

- Erforschung und Generalisierung von Erfolgsmetriken zur (Meta-)Kommunikation sowie der dafür benötigten Eingangsgrößen, dadurch für Dritte anwendbar.
- Erstmals Fokus auf bedürfnisgerechtes und adaptives Dialogverhalten von Sprachassistenten durch Berücksichtigung von Signalen der zwischenmenschlichen Metakommunikation, insbesondere Parasprache (z.B. Sprechpausen, Intonation, Betonung, etc.) sowie betreiber- und nutzerspezifischer Kontextinformation.
- Datenpunkte aus verschiedenen Kanälen (Anruf/Mail/Chat/Web) werden vergleichbar gemacht, um eine umfassende Repräsentation von Kontextinformation zu synthetisieren.
- Echtzeitoptimierung basierend auf aktuellem Nutzerkontext.

Zu Beginn des Projekts "SmartDialogue Engine" bestand die vorhandene Lösung aus einem modular aufgebauten System, das folgende Komponenten umfasste:

- Spracherkennungsmodul (STT, Speech-To-Text): Dieses Modul wurde über eine API-Anbindung von einem externen Anbieter bereitgestellt.
- Dialogmodul (FlowEngine): Eine Eigenentwicklung der Aaron GmbH, die als Zustandsraum-Lösung den Dialogfluss auf Grundlage der vorliegenden Nutzerkonfiguration definierte.
- Sprachsynthesemodul (TTS, Text-To-Speech): Auch dieses Modul wurde über eine API-Anbindung von einem externen Anbieter bereitgestellt.

Die vorhandene Lösung hatte mehrere Limitierungen:

- Keine Dialogoptimierung: Es fand keine automatische Anpassung des Dialogs statt; alle Anpassungen mussten manuell vorgenommen werden.
- Unterbrechungen: Das System erlaubte es den Endnutzenden nicht, Sprachausgaben verbal zu unterbrechen. Stattdessen war eine Heuristik notwendig, um das Ende der Spracheingabe des Endnutzenden zu detektieren. Eine Möglichkeit, die Maschinenausgabe verbal zu unterbrechen, bestünde darin, eine kontinuierliche Spracherkennung zu gewährleisten und bei detektierten Eingaben eine Reaktion zu veranlassen.
- Fehlende Erfolgsmetrik: Es gab keine etablierte Erfolgsmetrik zur Überprüfung der Ergebnisse, was die Bewertung und Optimierung des Dialogverlaufs erschwerte.

## Stand der Forschung

Bei gängigen Sprachassistentensystemen, wie z.B. Voicebots und KI-basierten Telefonassistenten erfolgen Optimierungen nicht durch die Betreiber, sondern nur durch Softwareentwickler, auch weil Zugang und Wissen typischerweise nicht bei KMU-Anwendern gegeben ist. Im Bereich der automatischen Echtzeit-Optimierung basierend auf Anrufkontext und betreiberspezifischen Anforderungen wurde bisher wenig bis keine Forschung betrieben.

Zum Zeitpunkt des Projektbeginns lag der Hauptfokus des Projektes darin, quasi-statische bzw. dynamische meta-kommunikative Signale vom Endnutzenden zu identifizieren. Ziel war es, sowohl den Inhalt als auch die Form der resultierenden Sprachausgabe so anzupassen, dass diese einen positiven Einfluss auf den erwarteten Anruferfolg haben. Die notwendigen Analysen sollten auf Grundlage der bereits vorhandenen Produktivdaten unter Einhaltung aller datenschutzrechtlichen Anforderungen erfolgen. Diese Anforderungen sind im Gesundheitsbereich besonders hoch und gehen über die üblichen Anforderungen hinaus.

## Vergleich mit bestehenden Lösungen

Aaron sind drei direkte Mitbewerber bekannt (Vitas.ai, Praxisconciierge und Susi & James). Von diesen grenzt sich Aaron über mehrere Aspekte ab:

- Plug & Play: Aaron bietet eine Telefon-Integration, die eine unbegrenzte Anzahl paralleler Anrufe zulässt, ohne Leitungen zu blockieren.
- Medizinischer Fokus: Aaron fokussiert sich seit 3 Jahren auf die Gesundheitsbranche und hat das Produkt im direkten Austausch mit relevanten Vereinigungen, Ärzten und Fachpersonal entwickelt.
- Netzwerkeffekte: durch die vergleichsweise hohe Marktdurchdringung kann Aaron auf eine höhere Anzahl an Trainingsdaten zurückgreifen, mit denen die Präzision des ML-Modells sich laufend verbessert.
- Forschungsfokus ML & Linguistik: Aaron beschäftigte zum Zeitpunkt des Antrags 6 Ingenieure mit relevantem Forschungshintergrund und hatte bereits mehrere Forschungsprojekte mit Partnern wie der Humboldt Universität Berlin, der Charité Berlin und dem Goethe-Institut durchgeführt.

## Meilensteine und Zeitplan

Das Projekt war in drei Hauptphasen unterteilt, die jeweils spezifische Meilensteine und Arbeitspakete (AP) umfassten. Um einen möglichst umfassenden Überblick über die Umsetzung zu geben, wird an dieser Stelle zuerst ein Überblick über die im Projektantrag geplanten Meilensteine gegeben. Eine Zusammenfassung der APs in den drei bereits erwähnten Phasen des Projektes, gemeinsam mit den darin erreichten Ergebnissen, schließt sich an. Es sei darauf hingewiesen, dass an dieser Stelle nur über APs des von der Aaron GmbH durchgeführten Teilvorhabens berichtet wird.

### Meilensteine (Zeitplanung bei Antragstellung)

#### MS1 - Metriken (August 2022)

Aaron hat die durch das DFKI erarbeiteten Merkmale und Metriken zum Dialogerfolg algorithmisch implementiert. Darauf aufbauend hat Aaron ein erstes mathematisches Modell zur

Vorhersage des Dialogerfolgs topologisch definiert, auf den annotierten Trainingsdaten trainiert und auf einem Testdatensatz validiert.

#### MS2 - Abschluss Phase 1 (Februar 2023)

Aaron hat die Metriken und Modelle zur Vorhersage des Dialogerfolgs unter Berücksichtigung metakommunikativer Signale erforscht und erstellt. Die Erfolgsmetrik kann über eine Web-App auf die Produktionsdaten angewendet und visualisiert werden. Das erste Benchmarking der optimierten Dialoge wurde durchgeführt. Ein Datenschutzmodell für Dritte wurde erstellt und geprüft. Wie zuvor ist die datenschutzkonforme Datennutzung weiterhin als Herausforderung zu berücksichtigen.

#### MS3 - Abschluss Phase 2 (Mai 2023)

Aaron hat basierend auf Vorschlägen und Konzeptionierung des DFKI ein ML- und NLP-basiertes Modell, das Parameter zur Definition des Dialogverhaltens auf quasi-statische, kundenspezifische Kontextinformation hin optimiert, algorithmisch implementiert und exemplarisch trainiert. Über API und Web-UI können kundenindividuelle Dialogverbesserungsvorschläge generiert und angewendet werden.

Beim Training von ML-Modellen muss stets mit einer Verzerrung (bias) und Datenunterrepräsentation (sparseness) in den Trainingsdaten gerechnet werden. Dem kann durch manuelle Annotation, Überprüfung und Gewichtung sowie behutsame Selektion der Modelltopologien und Trainingsparameter vorgebeugt werden.

#### MS4 - Modellerweiterung und Demonstrator (September 2023)

Aaron hat ein erweitertes Modell mit Heuristiken zur Dialoganpassung auf Basis metakommunikativer Signale und einen erweiterten Trainingsdatensatz erstellt, einen Detektor für systemverursachte Dialogirritationen implementiert und letzteren in das Web-UI integriert. Weiterhin hat Aaron einen am DFKI entwickelten ersten Demonstrator für die SDE an die API und das Web-UI angebunden und testfähig gemacht. Im Hinblick auf die Nutzererfahrung sollte die Systemakzeptanz gerade bei kanalübergreifender Kontextmodellierung und zunehmend intelligentem Systemverhalten laufend beobachtet werden. Möglichen Datenschutzbedenken und KI-bezogenen Ängsten ist bereits durch umsichtiges System- und Studiendesign vorzubeugen. Sollte die Systemakzeptanz stark hinter den Erwartungen zurückbleiben, sind vertrauensbildende Maßnahmen und mögliche Modellrestriktionen vorzusehen.

#### MS5 - Echtzeitimplementierung (Februar 2024)

Aaron hat die Echtzeit-SDE sowie ein nutzervalidiertes UI zur Konfiguration der Engine implementiert, im Produktivbetrieb evaluiert und die Implementierung iteriert. Außerdem wurde das Modell zur Dialogerfolgsvorhersage echtzeitfähig gemacht und über eine API für Dritte zugänglich gemacht.

Das Design der API/UI ermöglicht es Kunden, die autonome Echtzeit-Optimierung des nutzer-adaptiven Dialogverhaltens individuell einzustellen. Es erfolgte ein statistischer Vergleich

der Erfolgsquoten über die Modelle mit verschiedenen Kontext-Parametern. Eine Schnittstelle für den Echtzeit-Einsatz der erforschten Dialogerfolgsmetrik in Sprachassistenten Dritter wurde erforscht und validiert.

Die hohe zu erwartende Komplexität des resultierenden Modells bedeutet auch eine hohe Komplexität in der Konfiguration durch Kunden. Sollte sich dies als Hindernis herausstellen, sind Heuristiken und sinnvolle Beschränkungen anzuwenden.

Der Erfolg des Einsatzes der Metrik für Drittsysteme hängt von der Generalisierbarkeit der Modellannahmen bzw. der Generalisierungsfähigkeit des Modells selbst ab. Der aktuelle Erfolg von Technologieunternehmen im Bereich der affektiv-akustischen Sprachverarbeitung zeigt aber, dass bereits eine Minimallösung basierend auf wenigen generischen akustischen Merkmalen bedeutenden Mehrwert liefern könnte.

Arbeitsplan		Projektjahr 1												Projektjahr 2											
		2022	2022	2022	2022	2022	2022	2022	2022	2022	2022	2022	2022	2023	2023	2023	2023	2023	2023	2023	2023	2023	2023		
#	Arbeitspakete	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24
1	Literaturrecherche, Datenanalyse & Ableitung von Metriken des Kommunikationserfolgs ("Dialog Smoothness")	M1	M2	M3	M4	M5																			
2	Co-Creation mit künftigen Nutzern zur Anforderungvalidierung und Optimierung des ELSI-Profiles		M2	M3	M4	M5																			
3	ML-Modelle zur Vorhersage des Dialogerfolgs ("Dialog Smoothness")			M3	M4	M5																			
4	API & User-Interface, Benchmark				M4	M5	M6	M7	M8	M9	M10	M11	M12												
5	User-Interface				M4	M5	M6	M7	M8	M9	M10	M11	M12												
6	Benchmark				M4	M5	M6	M7	M8	M9	M10	M11	M12												
7	Datenmodell & Datenschutz																								
8	Erforschung eines ML-Modells zur Optimierung des Dialogverhaltens basierend auf quasi-statischen, kundenspezifischen Kontextinformationen							M7	M8	M9	M10	M11	M12	M13	M14	M15									
9	Anpassung API & UI, Benchmark																								
10	Anpassung API																								
11	Anpassung User Interface und Entwicklung eines Konfigurationsassistenten																								
12	Benchmark																								
13	Dialoganpassung an metakommunikative Signale des Anrufers																								
14	Priorisierung und Auswahl von metakommunikativen Signalen und Erstellung eines annotierten Trainingsdatensatzes																								
15	Optimierung der Dialoganpassungsfunktionalität auf Basis des Trainingsdatensatzes																								
16	Verbesserung metakommunikativer Signale des Systems																								
17	Entwicklung eines Detektors für Dialogbrüche																								
18	Dialogbrüche																								
19	Dialog-Integration																								
20	Echtzeit-Dialogoptimierung: Engine Design & Research																								
21	Modellierung Research																								
22	Engine Design und Demonstrator																								
23	Implementierung & Iterative Weiterentwicklung Echtzeit-Modell																								
24	Implementierung Echtzeit-Modell																								
25	Evaluation & Iterative Weiterentwicklung Echtzeit-Modell																								
26	Anpassung API/UI der SDE für Echtzeit-Modell																								
27	Dialogoptimierung & Benchmarking Echtzeit-Modell																								
28	Nutzerstudien und Design der API/ des UI für Echtzeit-Dialogoptimierung																								
29	Umsetzung API/UI der SDE für Echtzeit-Dialogoptimierung																								
30	Benchmark des Echtzeit-Modells																								
31	Entwicklung externe Anbindungsmöglichkeit für Echtzeit-Dialogerfolgsmetrik																								
32	Implementierung einer externen Anbindungsmöglichkeit für die Echtzeiterfolgsmetrik																								
33	Anbindungsmöglichkeit für die Dokumentation User-Test der Anbindungsmöglichkeit mit Partnern																								
Gesamt		76																							
M51	Datengrundlage und Modell																								
M52	Phase 1																								
M53	Phase 2																								
M54	Echtzeitmodell																								
M55	Phase 3																								

Abbildung: geplante Meilensteine und Arbeitspakete

## Umsetzung

Im Folgenden geben wir einen Überblick über die Bearbeitung der verschiedenen Arbeitspakete und setzen Planung und Umsetzung zueinander in Beziehung. Dabei ist zu berichten, dass das Projekt auf Grund von Verzögerungen bei der Stellenbesetzung erst mit ca. Sechsmonatiger Verspätung begonnen werden konnte. Dies wurde bereits über entsprechende Zwischenberichte kommuniziert, in Folge wurde das Projekt kostenneutral um sechs Monate verlängert.

Die folgenden Übersichten stellen die projektierten Arbeitspakete dar und verweisen auf die im Projekt entstandenen Arbeiten, die die angestrebten Ergebnisse abbilden. Details zu den entstandenen Arbeiten können dem Abschnitt "Detaillierte Darstellung" entnommen werden.

### Phase 1 - Dialoganalyse durch Metriken des Kommunikationserfolgs

Arbeitspaket	Ergebnisse der Arbeitspakete
AP 1: Literaturrecherche, Datenanalyse & Ableitung von Metriken des Kommunikationserfolgs ("Dialog Smoothness")	Anforderungsbestimmung, Datenanonymisierung
AP 2: Co-Creation mit künftigen Nutzern zur Anforderungvalidierung und Optimierung des ELSI-Profiles	Anforderungsbestimmung, Patientenumfrage, Datenanonymisierung, Transkriptionsverbesserung
AP 3: ML-Modelle zur Vorhersage des Dialogerfolgs ("Dialog Smoothness")	Dialogue Smoothness Metrik
AP 4.1: API	API
AP 4.2: User-Interface	Engine Demonstrator
AP 4.3: User Benchmarking	Ersetzt durch Stress FrameworkAusstehend, siehe "Herausforderungen und Anpassungen"

### Phase 2 - Kundenspezifische Dialogoptimierung durch ML-Modell und NLP-Methoden inkl. Implementierung Web-UI

Arbeitspaket	Ergebnisse der Arbeitspakete
AP 6: Erforschung eines ML-Modells zur Optimierung des Dialogverhaltens basierend	Einfluss Willkommensnachricht auf Anruferfolg, Vorschlag einer verbesserten

auf quasi-statischen, kundenspezifischen Kontextinformationen	Willkommensnachricht
AP 7.1: Anpassung API	Vorschlag einer verbesserten Willkommensnachricht
AP 7.2: User Interface und Entwicklung eines Konfigurationsassistenten	Vorschlag einer verbesserten Willkommensnachricht
AP 7.3: User Benchmarking	Ersetzt durch Stress FrameworkAusstehend, siehe "Herausforderungen und Anpassungen"

Phase 3 - Nutzerspezifische Echtzeit-Dialogoptimierung durch weiterentwickeltes ML-Modell

Arbeitspaket	Ergebnisse der Arbeitspakete
AP 8.1: Priorisierung und Auswahl von metakommunikativen Signalen und Erstellung eines annotierten Trainingsdatensatzes	Detektion meta-kommunikativer Signale
AP 8.2: Optimierung der Dialoganpassungsfunktionalität auf Basis des Trainingsdatensatzes	Anpassung der Nutzendenansprache
AP 9.1: Entwicklung eines Detektors für Dialogirritationen	Dialog Smoothness Metrik
AP 9.2: Erweiterung des Web-UI um die Anzeige von Dialog-Irritationen	Engine Demonstrator
AP 10.2: Engine Design und Demonstrator	Engine Demonstrator
AP 11.1: Implementierung Echtzeit-Modell	Anpassung der Nutzendenansprache
AP 11.2: Evaluation & iterative Weiterentwicklung Echtzeit-Modell	Stress Framework
AP 12.1: Nutzerstudien und Design der API/ des UI für Echtzeit-Dialogoptimierung	Anpassung der Nutzendenansprache
AP 12.2: Umsetzung API/UI der SDE für Echtzeit-Dialogoptimierung	API
AP 12.3: User Benchmarking	Stress Framework
AP 13.1: Design und Implementierung einer externen Anbindungsmöglichkeit für die Echtzeiterfolgsmetrik	API

## Zusammenfassung der erreichten Ergebnisse

### Hauptziele und Ergebnisse

Im Rahmen des Projekts "SmartDialogue Engine" wurden die Hauptziele überwiegend erreicht und spezifische Ergebnisse erzielt:

- Erstellung einer Erfolgsmetrik ("dialogue smoothness"): Eine Metrik zur Messung des Kommunikationserfolgs wurde entwickelt, die den Verlauf und den Erfolg von Dialogen bewertet.
- Erweiterung des bestehenden Produkts:
  - Assistent für textbasierte Konfiguration: Ein Assistent wurde entwickelt, der es Kunden ermöglicht, Dialoge textbasiert zu konfigurieren.
  - Erkennung von meta-kommunikativen Signalen: Ein System zur Erkennung von meta-kommunikativen Signalen wurde integriert, um den Dialogverlauf anzupassen.
  - Adaption der Nutzeransprache: Die Nutzeransprache wird nun auf Grundlage der erkannten Signale automatisch angepasst.
- Weiterentwicklung/Produktevolution:
  - Dynamische Dialoggenerierung mit LLM-Unterstützung: Die Möglichkeit, vollständig dynamische Dialoge mit Unterstützung von Large Language Models (LLMs) zu generieren, wurde erforscht.
  - Anpassung des Dialogumfangs: Die Anpassung des Dialogumfangs durch Konfigurationsänderungen wurde ermöglicht.
  - Mehrere Anliegen in einer Konversation: Es wurde die Möglichkeit geschaffen, mehrere Anliegen während einer Konversation zu behandeln, was im ursprünglichen Setup nicht möglich war.

### Technische und wissenschaftliche Fortschritte

#### Überblick

- Verbesserung der Transkriptionsqualität: Die Transkriptionsqualität wurde erheblich verbessert, von ca. 21% Word Error Rate (WER) zu 8% WER<sup>1</sup>, basierend auf dem Whisper-Modell und nach umfangreicher Evaluation verfügbarer Optionen.
- Integration von Emotionsbestimmung: Ein System zur Emotionsbestimmung auf dem Audiosignal wurde integriert, das die Nutzeransprache in Echtzeit automatisch anpasst.
- Prototypische Entwicklung eines Konversationsagenten: Ein Prototyp eines Konversationsagenten wurde entwickelt, der Dialoge automatisch führt, basierend auf

---

<sup>1</sup> Es wurde zum Zwecke dieser Messung ein Evaluationsdatensatz erstellt, siehe "Transkriptionsverbesserungen" im Abschnitt Detaillierte Darstellung.

definierten Datenpunkten, die innerhalb eines Konversationsschrittes gesammelt werden sollen.

## Konstruktionen

Im Rahmen des Projekts "SmartDialogue Engine" wurden mehrere spezifische technische Konstruktionen entwickelt, um die Qualität und Anpassungsfähigkeit der Dialogsysteme zu verbessern:

- Selbstbetriebene STT-Lösung Whisper: Eine Middleware, die Echtzeit-Transkriptionen ermöglicht und die Qualität der Audiotranskriptionen verbessert.
- Modell zur Vorhersage des Einflusses von Willkommens-Nachrichten: Ein Modell, das den Einfluss von fest konfigurierten Willkommens-Nachrichten (Start des Dialogs) auf den Anruferfolg vorhersagt und Nutzern verbesserte Willkommens-Nachrichten vorschlägt.
- Emotionserkennung und Sentimenterkennung: Ein System zur Erkennung von Emotionen im Audiosignal und Sentimenten im transkribierten Text.
- Vollständig dynamische Dialoggenerierung: Ein System, das die Dialoge auf Grundlage der Konfiguration der Dialogschritte dynamisch generiert. LLMs werden zur Informationsextraktion und Textgenerierung eingesetzt.

## Verfahren

Im Rahmen des Projekts wurden mehrere neue Verfahren oder Methoden entwickelt oder optimiert:

- Evaluationsverfahren zur Messung der Transkriptionsqualität: Ein Verfahren zur Bewertung der Qualität verschiedener Transkriptionsmodelle, um die beste Lösung für die Audiotranskription zu identifizieren.
- Verbesserungsvorschläge für Willkommens-Nachrichten: Ein Verfahren zur Analyse der Einflussfaktoren von Tokens (Wortbestandteile) in Willkommens-Nachrichten. Negative Einflüsse werden identifiziert und genutzt, um verbesserte Varianten der Nachrichten zu erstellen.
- Definition der Metrik "dialogue smoothness": Eine Metrik zur Messung der Erfolgswahrscheinlichkeit im aktuellen Dialog nach jedem Konversationsschritt. Ein Modell wurde entwickelt, um diese Metrik iterativ zu messen und den Dialog entsprechend anzupassen.

## Schutzrechte

Während des Projekts wurden keine Patente, Urheberrechte oder anderen Schutzrechte angemeldet oder erworben.

## Erfolgsmetriken und Evaluation

Die Ergebnisse des Projekts wurden anhand mehrerer Erfolgsmetriken evaluiert:

- Hauptmetriken:

- Anruferfolg: Diese Metrik misst, ob der Anrufende erfolgreich sein Anliegen geschildert und die minimal notwendigen Informationen übermittelt hat, die es der Praxis ermöglichen, das Anliegen zu bearbeiten.
- Dialogue Smoothness: Diese Metrik variiert während der Konversation und ist mit dem zum aktuellen Zeitpunkt vorhergesagten Anruferfolg auf Grundlage des bisherigen Gesprächs identisch.
- Weitere Metriken:
  - Dialogirritation: Diese Metrik misst die Wahrscheinlichkeit, dass der Anrufende in der nächsten Interaktion auflegt oder das Gespräch abbricht.
  - Stressfaktoren: Diese Metrik wurde während der Evaluation der Auswirkung von Dialoganpassungen auf den Stress von Probanden während der Interaktion gemessen.

## Herausforderungen und Anpassungen

Während des Projekts ergaben sich außerdem weitere Herausforderungen bzw. Anpassungserfordernisse:

- *Alters- und Geschlechtsbestimmung*: Ursprünglich war geplant, eine Dialoganpassung unter anderem auf Grundlage des vorhergesagten ungefähren Alters und Geschlechts der Anrufenden vorzunehmen. Es stellte sich jedoch heraus, dass eine Altersbestimmung allein auf Grundlage des Audiosignals keine zufriedenstellenden Ergebnisse lieferte. Die Bestimmung des Geschlechts war mit einer Vielzahl von Unwägbarkeiten verbunden, unter anderem spielte der Unterschied zwischen Geschlecht und Gender sowie die aus technischer Sicht unklare Definition von Gender eine große Rolle.
- *Kommerzielle Verfügbarkeit von Large Language Models*: Etwa sechs Monate nach Projektstart wurden mehrere (teils OpenSource, teils ausschließlich kommerziell verfügbare) außerordentlich leistungsstarke generative Textmodelle veröffentlicht (OpenAI, 1). Dies ermöglichte eine signifikante Verbesserung der Audio-Transkriptionsqualität (durch das frei verfügbare Whisper-Modell) sowie die Anpassung der Projektarbeitspakete. Im Speziellen wurden die Modellentwicklung und das Modelltraining in den APs 6, 8 und 9 obsolet, da nicht zu erwarten war, durch Training selbst entworfener Modelle mit den zur Verfügung stehenden Daten eine vergleichbare oder bessere Performanz zu erzielen.
- *User benchmarking*: Durch strukturelle Änderungen im Unternehmen konnten die entwickelten Produktverbesserungen bislang noch nicht direkt an der Nutzerbasis evaluiert werden, sondern nur im Rahmen von zielgerichteten Evaluationen mit Testnutzern. Dies führte zu einer Verzögerung bei der Validierung der Projektergebnisse und erforderte zusätzliche Anpassungen in der Projektplanung.

# Detaillierte Arbeitsergebnisse

Wie bereits im Überblick geschildert, teilte sich das Projekt grob in drei Phasen auf. Wir berichten im Folgenden über die erreichten Ergebnisse der einzelnen Phasen und weisen jeweils auf relevante Arbeitspakete wie oben beschrieben hin.

Hinweis: Zur Wahrung schützenswerter Interessen des Zuwendungsempfängers werden die Inhalte dieses Abschnitts (Seiten 17-38) als vertraulich klassifiziert.

## Phase 1 - Analyse des Dialogerfolgs

### Anforderungsbestimmung

Relevante APs: AP1, AP2

Zur Validierung der antizipierten Anforderungen durch zukünftige Nutzer des Systems wurden ausgewählte Praxisteams durch die Aaron GmbH interviewt sowie über Fragebögen ebenfalls Patient:innen Feedback eingeholt. Zusätzlich zu diesen Maßnahmen wurde weiterhin Support-Feedback ausgewertet, sodass ein umfassendes Gesamtbild erstellt werden konnte, welches sowohl funktionale als auch ELSI-Kriterien umfasst.

Zusammenfassend ist davon zu berichten, dass die Produkthanforderungen sich grundsätzlich zwischen den untersuchten Nutzergruppen unterscheiden:

- Für Patienten
  - ist ein als natürlich empfundener Dialog die Voraussetzung, um diesen nicht vorzeitig abzubrechen (“auflegen”)
  - Sollte das System das Anliegen klar erkennen, die für die Verarbeitung notwendigen Informationen abfragen und eine erfolgreiche Verarbeitung zurück melden
- Für Praxismitarbeitende
  - Ist eine hohe Transkriptionsqualität eine entscheidende Anforderung, da nur so eine Zeitersparnis entsteht (bei geringer Qualität müssen die Audioaufnahmen abgehört werden)
  - Müssen die gesammelten Informationen ausreichend sein, um ein Anliegen bearbeiten zu können (da sonst wiederum zeitraubende Rückfragen an Patient:innen notwendig werden)

Konkret wurden in sechs Arztpraxen, in denen Aaron genutzt wurde, folgende Aspekte untersucht:

- Demografische Fragen
  - Geschlechtszuweisungen, Altersgruppen (in Generationen) und genaue Berufsgruppen wurden eruiert. Die Auswertungen dazu sind unten bei den Ergebnissen zu finden (Abb. 1.1).
- ELSI

- Ethische, rechtliche und soziale Fragen (Parker et al., 2018) im Zusammenhang mit der Nutzung des digitalen Sprachassistenten wurden erfasst.
- Einfluss von Persönlichkeitseigenschaften
  - Dabei sollten befragte Praxismitarbeiter:innen beantworten, inwiefern sie den Einfluss von Persönlichkeitseigenschaften von anrufenden Nutzern (Patient:innen) auf den Dialogverlauf einschätzen.
  - Folgende Persönlichkeitseigenschaften nach McCrae und Costa (1997) (Standardmodell in der Persönlichkeitsforschung) wurden hierbei in Betracht gezogen:
    - Offenheit für neue Erfahrungen (Openness to new Experience)
    - Gewissenhaftigkeit (Conscientiousness)
    - Verträglichkeit (Agreeableness)
    - Extraversion
    - Neurotizismus (Neuroticism)
- Einfluss von Emotionen
  - Ebenfalls sollten Praxismitarbeiter:innen beantworten, wie sie den Einfluss von Basisemotionen (Standard bei Emotionsforschung) von anrufenden Nutzern (Patient:innen) auf den Dialogverlauf einschätzen.
  - Basisemotionen nach Ekman und Friesen (1971) sind:
    - Angst
    - Trauer
    - Freude
    - Wut
    - Ekel
    - Überraschung

## Ergebnisse

- Demografische Daten

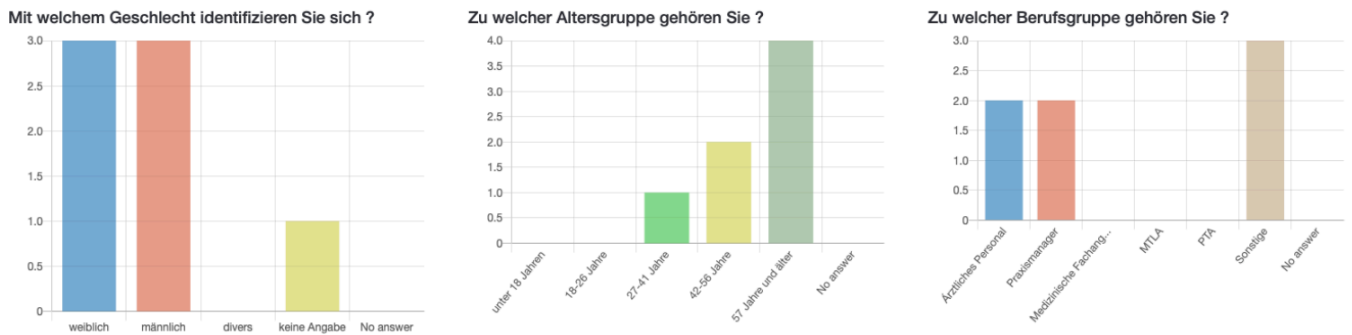


Abbildung 1.1 Demografische Daten

- Es liegt eine balancierte Stichprobe in Bezug auf die Geschlechter- und Berufsgruppe vor. Bezüglich der Altersgruppe liegen besonders viele Beantwortungen von über 56-jährigen Praxismitarbeiter:innen vor. Von Interesse wäre hierbei der Vergleich mit der Altersverteilung der Gesamtpopulation von Praxismitarbeiter:innen im deutschsprachigen Raum.
- ELSI
  - Ethische Fragen
    - Durch den Einsatz von Digitalassistenten entstehen bei den befragten Praxismitarbeiter:innen überwiegend keine ethischen Fragen
    - Der Digitalassistent wird u.a. als erweiterter Anrufbeantworter und als probates Mittel beschrieben
  - Rechtliche Fragen
    - Bei den Befragten Praxismitarbeiter:innen entstehen rechtliche Fragen zum Datenschutz, den Aussagen des Sprachassistenten und der Datenspeicherung
  - Soziale Fragen
    - Soziale Fragen ergeben sich bei Praxismitarbeiter:innen bzgl. der Personalressourcen und der guten Kommunikation gegenüber Patienten.
    - Es ergeben sich Fragen dazu, wer die Anfragen bearbeitet und hinterfragt, ob alles richtig aufgenommen wurde.
    - Es werden auch Fragen dazu geäußert, ob durch den Einsatz des Digitalassistenten ein Personalabbau in weiterer Zukunft möglich wäre

Aus den gesammelten Antworten ergaben sich keine neuen Erkenntnisse zu zusätzlichen funktionellen Anforderungen. Das Operieren im Gesundheitsbereich macht die Beachtung der europäischen Datenschutzgrundverordnung bei der Speicherung und Verarbeitung medizinisch relevanter persönlicher Informationen zu einer Grundvoraussetzung. Auch legt die Aaron GmbH seit längerer Zeit in der Außenkommunikation einen Schwerpunkt auf soziale Fragen wie z.B. verbesserte Work-Life-Balance für Mitarbeitende im Gesundheitswesen, größeren Fokus und mehr

Zeit im persönlichen Umgang mit Patienten und die daraus folgende verbesserte grundsätzliche medizinisch-soziale Versorgung von Patient:innen im Praxisbetrieb.

- Persönlichkeitseigenschaften

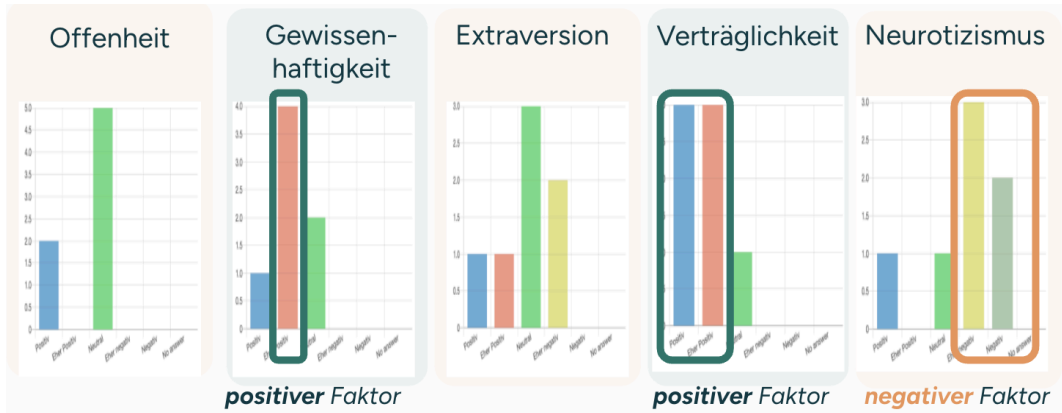


Abbildung 1.2 Ergebnisse Persönlichkeitseigenschaften

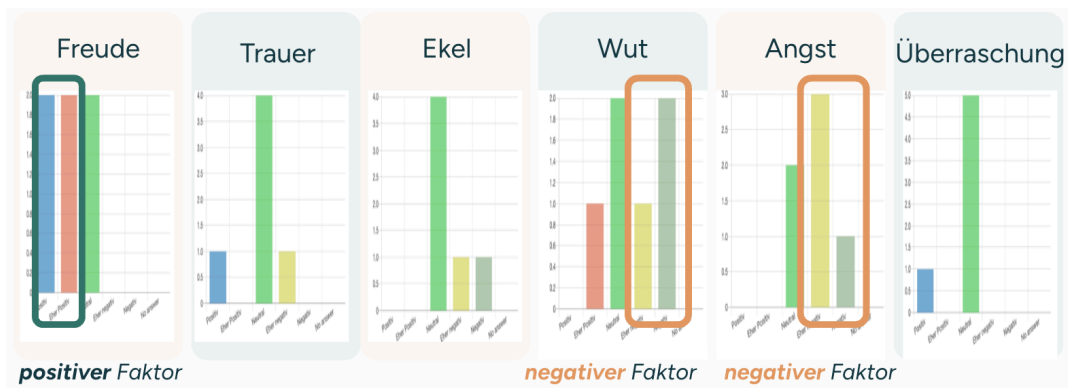


Abbildung 1.3 Ergebnisse Emotionen

## Transkriptionsverbesserung

Relevante APs: AP2, AP8, AP9, AP10

Um robuste Rückschlüsse auf die Auswirkungen von Patient:innen-Eingaben auf den Dialogverlauf bzw. Anruferfolg ziehen zu können, war es zunächst nötig, die grundlegende Qualität der Audio-Transkription zu verbessern. Diese bildet als textuelle Darstellung des Gesagten die Ausgangssituation für jegliche darauf aufbauende Datenverarbeitung. Eine unzureichende Transkriptionsqualität hätte alle weiteren Entwicklungen, Hypothesen und Experimente des Forschungsprojektes auf Grund des spezifischen Fokus auf meta-kommunikative Signale in der empfangenen Kommunikation und der Schwierigkeit, diese zu messen und zu nutzen in Frage gestellt oder möglicherweise in Gänze verhindert. Um diesem Umstand Abhilfe zu schaffen (und auch die durch Praxispersonal und durch die technische Kundenbetreuung der Aaron GmbH geäußerte Anforderung einer verbesserten Transkriptionsqualität zu befriedigen) wurde eine Evaluation verfügbarer kommerzieller und OpenSource-basierter Lösungen vorgenommen.

Dem Projektziel kam entgegen, dass die Firma OpenAI im Oktober 2023 ihr neu entwickeltes Transkriptionsmodell "Whisper" verfügbar machte (Radford et al., 2023), dessen Leistungsfähigkeit in der erwähnten Evaluation überprüft werden konnte.

### Evaluationsdaten

Um eine aussagekräftige Evaluation zu ermöglichen, war es zunächst notwendig, einen Goldstandard für Audioaufnahmen und deren korrekte Transkription zu erzeugen. Hierzu wurde aus den verfügbaren Produktionsdaten der Aaron GmbH ein Evaluationsdatensatz bestehend aus 600 Audioaufnahmen aus sechs verschiedenen Stati des Dialogflusses ausgewählt. In einem manuellen Prozess wurde diese Aufnahmen durch einen Muttersprachler transkribiert und außerdem sichergestellt, dass die Aufnahmen keine datenschutzrechtlich relevanten Informationen enthalten (siehe Abschnitt "Datenanonymisierung").

Darüber hinaus wurde auf einem frei verfügbaren Datensatz evaluiert, für den sowohl Audioaufnahmen als auch dazugehörige Transkriptionen vorliegen. Der Inhalt besteht vorrangig aus offiziellen Dokumenten des Europäischen Parlaments sowie der deutschen Wikipedia, aus denen einzelne Sätze eingesprochen wurden.

### Evaluationsprozess

Für die eigentliche Evaluation wurde ein einfaches webbasiertes Tool entwickelt, das die Transkription über API-Anbindung an die verschiedenen Anbieter für die beschriebenen Audio-Beispiele erzeugt und relevante Metriken berechnet. Hierbei seien  $s_1$  und  $s_2$  zwei in Frage stehende Wörter.  $S$ ,  $D$  und  $I$  seien die Anzahl von Ersetzungen, Entfernung bzw. Erweiterungen, um das eine Wort in das andere zu transformieren. Ferner sei

$$H(s_1, s_2) = |s_1| - (S(s_1, s_2) - D(s_1, s_2))$$

Die relevanten Metriken (wie oben beschrieben) errechnen sich dann wie folgt:

$$\text{WER}(s_1, s_2) = \frac{S(s_1, s_2) + D(s_1, s_2) + I(s_1, s_2)}{H(s_1, s_2) + S(s_1, s_2) + D(s_1, s_2)}$$

$$\text{MER}(s_1, s_2) = \frac{S(s_1, s_2) + D(s_1, s_2) + I(s_1, s_2)}{H(s_1, s_2) + S(s_1, s_2) + D(s_1, s_2) + I(s_1, s_2)}$$

$$\text{WIP}(s_1, s_2) = \frac{H(s_1, s_2)}{|s_1|} \frac{H(s_1, s_2)}{|s_2|}$$

## Datenanonymisierung

Relevante APs: AP1, AP2

Um die vorliegenden Produktionsdaten der Aaron GmbH für eine datenschutzkonforme Analyse, Verarbeitung und Grundlage für weitergehende Modelltrainings bzw. -verbesserungen nutzbar zu machen, war es notwendig, die vorliegenden Daten einem Anonymisierungsprozess zu unterziehen um die Persönlichkeitsrechte der Patient:innen zu schützen und den hohen Anforderungen der Aaron GmbH an den Schutz persönlicher Informationen sowie Gesundheitsdaten zu genügen.

Die Grundlage für eine Anonymisierung stellten die vorliegenden Audiotranskripte der Patient:innen dar.

In diesen wurden weiter alle potentiellen persönlichen Informationen identifiziert und durch artifizielle Daten ersetzt.

Zu den detektierten Information gehörten:

- Namen
- Orte
- Daten (Wochentage, Geburtsdaten)
- Telefonnummern
- Postleitzahlen

Name und Orte wurden durch zufällig generierte Werte ersetzt<sup>2</sup>, Daten und Telefonnummern wurden durch zufällig generierte Zahlenfolgen anonymisiert. Für Postleitzahlen wurde - um eine potentielle geographische Lokalisierung auch in Zukunft zu ermöglichen - ausschließlich die erste Ziffer weiter verwendet.

Alle weiteren Datenpunkte, die zu einer potentiellen Identifikation eine:r Patient:in führen könnten (wie z.B. die Arztpraxis) wurden aus dem Datensatz gelöscht.

---

<sup>2</sup> Zum Einsatz kam die Python Bibliothek Faker, die künstlich generierte Daten in verschiedenen Sprachen anbietet, <https://faker.readthedocs.io/en/master/>

## Dialog Smoothness Metrik

Relevante APs: AP1, AP3, AP6, AP9

Als zentraler Baustein bei der Bewertung der Güte des Dialogflusses wurde bereits bei Projektplanung das zu diesem Zeitpunkt noch auszuführende Maß der "Dialogue Smoothness" aufgenommen (als Maß dafür, wie flüssig eine Unterhaltung abläuft). Um diese Vorannahme weiter auszuführen, wurde durch den Projektpartner DFKI auf Grundlage von durch die Aaron GmbH zur Verfügung gestellten anonymisierten Produktionsdaten ein Modell entwickelt, das verschiedene statistische Eigenschaften der Konversation verwendet, um eine Wahrscheinlichkeit zu ermitteln, ob der betreffende Dialog erfolgreich zu Ende geführt werden kann (also einen als "Erfolg" definierten Endzustand erreicht).

Das initial verwendete Modell konnte mit den zur Verfügung gestellten Daten auf einen **F1 Wert von 88%** trainiert werden.

Es stellte damit eine gute Grundlage dar, um komplexere Modelle darauf aufbauen zu können. Das wurde vor Allem dadurch notwendig, dass das Modell Konversationen nur in Retrospektive betrachten kann, da für dieses Modell statistische Eigenschaften als Features verwendet wurden, die erst nach Beendigung des Dialogs zur Verfügung stehen.

Um eine kontinuierliche Bewertung des Gesprächsflusses zu ermöglichen, wurden die projektierten Maße für Dialogue Smoothness sowie für die Messung von Dialogirritationen zusammengeführt.

Im Ergebnis liegt nun ein Modell vor, das die Wahrscheinlichkeit eines Gesprächsabbruchs in der nächsten Patient:innen-Maschine-Interaktion modelliert und auf vor-trainierten Embeddings basiert. Dieses wurde weiter mit den vorliegenden Daten einem fine-tuning Prozess unterworfen.

Die Vorhersage der Abbruchwahrscheinlichkeit basiert im Anschluss nur auf Teilen des davor stattgefundenen Gesamtgesprächs, was zum einen zu einem einfacher generalisierbaren Modell führt und zum Anderen eine konstante Laufzeit des Algorithmus garantiert (ein Bezug auf den Gesamtkontext hätte zöge eine linear verlaufende Verlängerung der Bearbeitungszeit nach sich, die sich in größer werdenden Wartezeiten für Patient:innen während der Konversation niederschlagen würde).

Die Zielausgabe besteht aus der Information, ob das Gespräch weitergeführt wird bzw. erfolgreich zu Ende geht oder aber abbricht.

Das resultierende Ergebnis der beschriebenen Lösung erreicht einen um etwa 10% verringerten F1 Wert im Vergleich zur oben beschriebenen Baseline des initialen Modells. Da dieses allerdings jederzeit Zugriff auf die gesamte Konversation hat, während die beschriebene Lösung sich auf die aktuelle und vorherige Interaktionen beschränkt, wird der Leistungsabfall als akzeptabel angesehen<sup>3</sup>.

---

<sup>3</sup> In einer Ausbaustufe könnte durch die Berücksichtigung aller bisherigen Interaktionen vermutlich eine höhere Vorhersagegenauigkeit erreicht werden, der Einfluss durch auf diese Weise zu erwartende erhöhte Wartezeiten im Dialogfluss sollte dabei evaluiert werden.

## API

Relevante APs: AP4, AP7, AP12

Die einzelnen Datenverarbeitungsschritte (Einlesen, Verarbeiten, Anonymisierung, Analyse) sowie die entwickelten Metriken wurden im Rahmen einer REST API intern zur Verfügung gestellt. Die API wurde mit Hilfe der Bibliothek FastAPI<sup>4</sup> erstellt und ist innerhalb der vom Anbieter AWS zur Verfügung gestellten Unternehmenscloud zugänglich. Sowohl die Funktionalitäten aus Phase 1 als auch für die Phasen 2 und 3 wurden in einer gemeinsamen API zur Verfügung gestellt. Zusätzlich kann für einzelne Endpunkte (z.B. Bewertung der Dialog Smoothness) Zugriff für externe Parteien gewährt werden.

## Engine Demonstrator

Relevante APs: AP4, AP9, AP10

Aufbauend auf den durch die API bereitgestellten Funktionalitäten wurde eine separate Web-Applikation entwickelt, über die diese grafisch dargestellt werden können. In Phase 1 wurde die post-hoc Analyse bestehender (anonymisierter) Konversationsdaten priorisiert, wobei pro Interaktion die Dialogue Smoothness berechnet und dargestellt werden. Dieser Engine Demonstrator wurde für die Phasen 2 und 3 um die dort hinzugefügten Funktionalitäten erweitert. Im Speziellen wurde der Möglichkeit einer post-hoc Betrachtung bereits abgeschlossener Konversationen ein Simulator für vollständig dynamisch generierte Dialoge hinzugefügt. Dieser ermöglicht es dem Nutzenden, solche dynamisch erzeugten Gespräche mittels Sprach- oder auch Texteingabe zu führen und abschließend eine strukturierte Übersicht über die ermittelten Daten zu erhalten (vgl. Abschnitt "Vollständig dynamische Dialogführung").

## Patientenumfrage

Relevante APs: AP1, AP4

Um einen umfassenden Einblick in die funktionalen Anforderungen aus Patient:innensicht zu erlangen, wurde in Kooperation mit mehreren Pilotpraxen eine umfangreiche Patient:innenbefragung durchgeführt.

Insgesamt haben fast 500 Personen an dieser Umfrage teilgenommen, wobei 277 Personen diese auch abgeschlossen haben (Stand: 22.08.24). Hierdurch ergibt sich eine Stichprobe von n=277 Patient:innen. Befragte wurden entweder digital über eine 'Landing Page' oder vor Ort über Flyer in den erwähnten ausgewählten Praxen auf die Umfrage aufmerksam gemacht und für eine Teilnahme gewonnen.

Es wurden demografische Daten sowie der Einfluss von Persönlichkeitseigenschaften und Basisemotionen und Meinungen zu bestimmten metakommunikativen Eigenschaften erfragt.

---

<sup>4</sup> <https://fastapi.tiangolo.com/>

## Ergebnisse

Generation	Verteilung
Z	5,05 %
Y	16,97 %
X	35,38 %
Babyboomer (W)	42,60 %

- Demografische Daten
  - Geschlechtszuweisungen, Altersgruppen (in Generationen), Berufsstatus und Berufsgruppen der Befragten wurden eruiert.
  - Geschlechterverteilung: Weiblich : 53,07% ; Männlich: 46,21%

Tabelle 1.1 Verteilung der Generationen

Z: 18-27 Jahre zum Erhebungszeitpunkt;

Y: 28-42 Jahre; X: 43-58 Jahre; W: >58 Jahre

Beschäftigung	Verteilung
Selbstständig	5,05 %
Angestelltenverhältnis	60,29 %
Student:in	1,44 %
In Ausbildung	1,81 %
Arbeitslos	4,33 %
In Rente	22,74 %
Arbeitsunfähig	4,33 %

Tabelle 1.2 Verteilung d. Berufsstatus befragter Patienten

Altersgruppen, Berufsgruppen bzw. -stati als auch (biologische) Geschlechterverteilung decken sich annähernd mit statistischen Daten zur Gesamtbevölkerung, sodass eine Repräsentativität der Stichprobe angenommen werden kann.

## Zu welcher Berufsgruppe gehören Sie ?

Beantwortet: 277 Übersprungen: 0

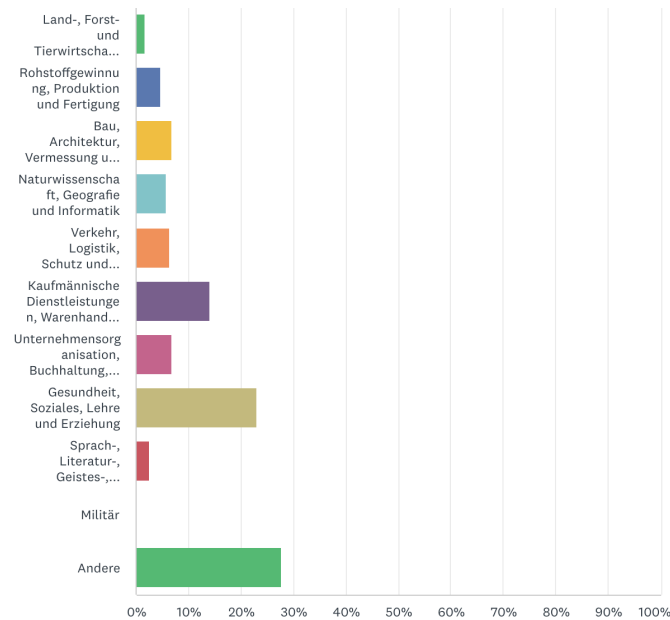


Abbildung 2.1 Berufsgruppenverteilung bei Patientenumfrage

- Einfluss von Persönlichkeitseigenschaften
  - Patient:innen sollten angeben, wie groß sie den Einfluss von Persönlichkeitseigenschaften von menschlichen Gesprächspartner:innen auf den Dialogerfolg mit dem Telefonassistenten einschätzen.
  - Um Teilnehmenden die Umfrage zu erleichtern, wurden etablierte umgangssprachliche Adjektive, die auf das Persönlichkeitsinventar nach McCrae und Costa (1997) abzielen, verwendet, um Persönlichkeitseigenschaften vereinfacht zu beschreiben.

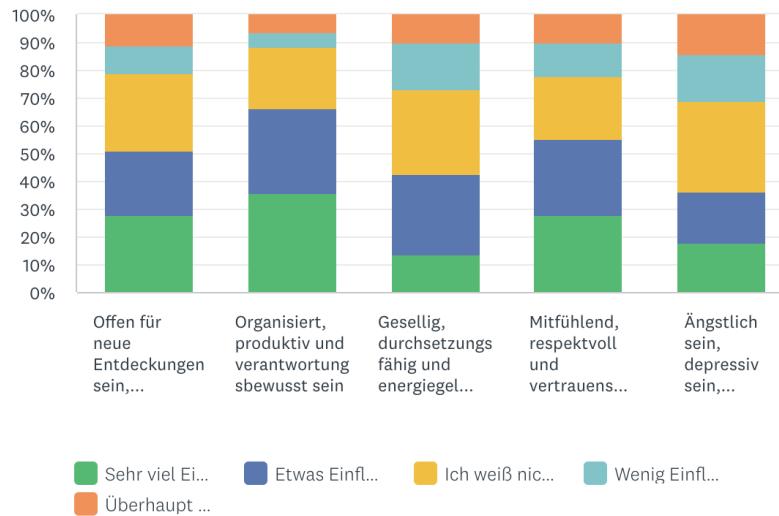


Abbildung 2.2 Einfluss von Persönlichkeitseigenschaften auf Dialogerfolg  
 1. Offenheit, 2. Gewissenhaftigkeit, 3. Extraversion, 4. Verträglichkeit, 5. Neurotizismus

- Einfluss von Emotionen
  - Befragte Patient:innen sollten angeben, wie groß sie den Einfluss von Basisemotionen von Patient:innen auf den Dialogerfolg mit dem Telefonassistenten einschätzen. Basisemotionen nach Ekman und Friesen (1971) wurden untersucht.

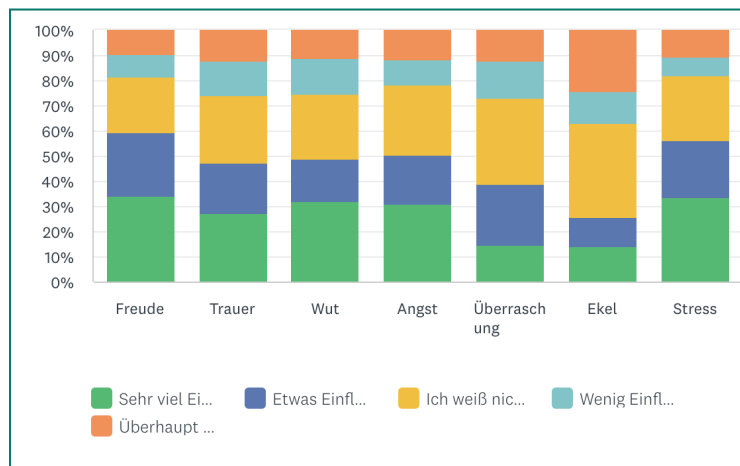


Abbildung 2.3 Einfluss von Emotionen auf Dialogerfolg

- Einfluss von metakommunikativen Signalen
  - Patient:innen sollten angeben, ob akustische Signale zu Beginn des Dialogs hilfreich sind. Zudem wurde der Einfluss der Sprechgeschwindigkeit und Stimme genauer untersucht.

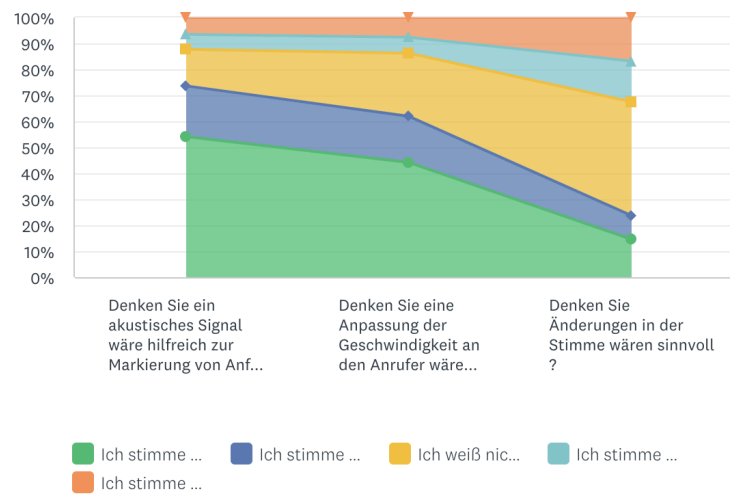


Abbildung 2.4 Einfluss metakommunikativer Signale

## Phase 2 - Kundenspezifische Dialogoptimierung

In dieser Phase wurden vorrangig quasi-statische Einflussfaktoren betrachtet, also kundenspezifische statische Konfigurationseigenschaften, die einen Einfluss auf den Dialogablauf haben, im Speziellen betrachtet. Motiviert durch die erfolgversprechenden Ergebnisse einer Durchführbarkeitsstudie, die die Aaron GmbH im Vorfeld des Forschungsprojektes erstellt hat, wurden vor allem die “Willkommensnachrichten” untersucht. Als “Willkommensnachricht” wird die erste Patient:innen-Ansprache bezeichnet, also der Text, mit dem ein Anrufer durch das System begrüßt wird, sobald eine Telefonverbindung aufgebaut wurde.

### Einfluss Willkommensnachricht auf Anruferfolg

Relevante APs: AP6

Ähnlich wie für die Dialogue Smoothness Metrik wurde für die Messung des Einflusses von Willkommensnachrichten auf den Dialogerfolg auf die in Phase 1 erstellten Trainingsdaten zurückgegriffen. Hierbei wurden Anrufe auf Praxis-Ebene aggregiert und die Anruferfolgsquote pro Praxis ermittelt. Der auf diese Weise erweiterte Trainingsdatensatz erlaubte nun die Modellierung von Text-Eingabe (die Willkommensnachricht) und Regression des Anruferfolgs als Zielvariable. Wie bereits zuvor wurde hierfür ein Modell auf Grundlage vortrainierter Embeddings verwendet und einem fine-tuning Prozess unterzogen.

Im Rahmen dieses AP wurde außerdem eine Masterarbeit betreut, die den Einfluss von Willkommensnachrichten auf den Anruferfolg untersuchte. Die dort ausgeführten Experimente bildeten die Grundlage für die Entscheidung zur Nutzung des oben beschriebenen Modells.

## Vorschlag einer verbesserten Willkommensnachricht

Relevante APs: AP7

Nach erfolgreicher Modellierung des Einfluss von Willkommensnachrichten auf den Anruferfolg wurden die spezifischen Bestandteile der Eingabe (also der Willkommensnachricht), die einen negativen Einfluss auf die Zielvariable aufweisen. Diese wurden mittels eines dafür entwickelten Prompts als Kontext einem generativen LLM zur Verfügung gestellt, zusammen mit der Aufgabe, auf Grundlage der vorhandenen Modellierung und dieses Kontextes eine verbesserte Willkommensnachricht zu formulieren. Diese Möglichkeit wurde als optionale Erweiterung in das bestehende Produkt integriert und soll zeitnah für interessierte Kunden zur Unterstützung bei der Konfiguration zur Verfügung gestellt werden.

## Phase 3 - Dynamische Dialogoptimierung

### Detektion meta-kommunikativer Signale

Relevante APs: AP8

Zusätzlich zum “Wann” (einer Optimierung) ist außerdem auch die Frage nach dem “Wie” zu beantworten. Ein erster Ansatz hierzu war es, Patient:innen in verschiedene generalistische Gruppen einzuordnen und anschließend auf dieser Grundlage optimierte Ansprachen zu generieren. Erste Experimente (vgl. (Soliman et al., 2024)) waren vor allem im Hinblick auf eine inklusivere Ansprache der Nutzenden vielversprechend (Nutzende wurden nach ihrem Alter und Geschlecht in Gruppen aufgeteilt und mit gezielt generierten Dialogbausteinen angesprochen). Bei der technischen Umsetzung zeigten sich allerdings vorher nicht antizipierte Schwierigkeiten:

- Eine Altersbestimmung der Anrufenden allein auf Grundlage der Audio-Daten konnte keine zufriedenstellenden Ergebnisse liefern, statistisch war zwischen Anwendung des Modells und einer randomisierten Zuordnung nicht zu unterscheiden. Da der Identifikationsschritt (bei dem unter anderem das Geburtsdatum des Anrufenden ermittelt wird) am Ende des Dialogs stattfindet, steht das Alter zur Optimierung des Dialogflusses auch durch direkte Abfrage nicht zur Verfügung.
- Eine Geschlechtsbestimmung (ebenfalls auf Grundlage des Audio-Signals) ist mit mehreren Schwierigkeiten behaftet. Zunächst besteht ein signifikanter Unterschied zwischen biologischem und sozialem Geschlecht, der durch simplifizierte (meist binäre) Klassifikatoren nicht abbildbar ist. Weiterhin waren die zur Verfügung stehenden Modelle nicht in der Lage, zufriedenstellende Klassifikationsergebnisse zu erzeugen (höchster F1 Wert 72%). Auch die auf öffentlich verfügbaren Datensätzen trainierten eigenen Ansätze kamen über einen F1 Wert von 78% nicht hinaus<sup>5</sup>.

---

<sup>5</sup> Während bei der Dialog Smoothness Metrik ein F1 Wert von 79% als akzeptabel angesehen wird, wurde in diesem Fall ein Klassifikator mit F1 Wert von 78% verworfen. Im ersten Falle wird die Smoothness Metrik als Indikator genutzt, um weitergehende Optimierungen auszulösen. Eine Fehlentscheidung dabei ist eher als unkritisch zu betrachten, da in diesem Fall Nutzer:innen auf die bestehende Dialoglogik zurückfielen. Eine Geschlechtererkennung hat hingegen direkte (inhaltliche) Auswirkungen auf den Dialogverlauf, die kontraproduktiv sein können (bei inkorrektter Ansprache).

## Anpassung der Nutzendenansprache

Relevante APs: AP8, AP9, AP10, AP11

Die Klassifikationsergebnisse des Audio-Signals dienen im Nachgang einer Anpassung der Nutzendenansprache durch das System, die adhoc, also während Dialogs, stattfindet. Anpassungen dieser Art wurden prototypisch in das bestehende Produktionssystem integriert und mit einer simpleren Adaptionmethode (Anpassung der Sprechgeschwindigkeit bei der Sprachsynthese) verglichen (siehe nächster Punkt "Stress Framework"). Durch die Vielzahl der durch die Aaron GmbH bearbeiteten Telefongesprächen und die Kostenstruktur bei der Nutzung kommerzieller Large Language Models ist ein gezielter Einsatz dieser Art von Dialoganpassung nicht nur aus Perspektive der User Experience<sup>6</sup> sondern auch aus der Perspektive eines effektiven Mitteleinsatzes und vertretbaren Einflusses auf die zu erwartende Gewinnmarge angezeigt.

## Stress Framework (Stresserfassung während Nutzung von Aaron)

Relevante APs: AP 10

Ganz generell kann festgehalten werden, dass auf physiologischer Ebene Stress mit der Erhöhung der Herzrate und einem erhöhten Hautleitwiderstand einhergeht (Ziegler, 2004). In Kooperation mit dem DFKI konnten wir dank freundlicher Leihgabe der Empatica Embrace Plus Uhr eine Erfassung der Herzrate und des Hautleitwiderstandes durchführen, um somit Rückschlüsse auf eine mögliche Belastung durch Stress ziehen zu können.

### Durchführung

Für die Durchführung des Experiments wurden anonymisierte Benutzerprofile mit einer Spezialsoftware erstellt, um später eine sachgemäße Auswertung der Sensordaten, die mit der Empatica Watch erhoben wurden, sicherstellen zu können.

Für das Experiment konnten 13 Personen verschiedenen Alters und Geschlechts gewonnen werden, wobei sichergestellt wurde, dass jede Person einen Computer und Telefon bedienen kann.

Mit Versuchsperson 1 wurde ein Testlauf durchgeführt, um zu prüfen, ob die Auswertungssoftware korrekt funktioniert und Daten aufzeichnet. Ein späterer Versuchsdurchgang musste als ungültig gewertet werden, da während der Durchführung technische Probleme in der Versuchspraxis auftraten, die zeitnah nicht gelöst werden konnten. Somit ergab sich eine Proband:innengruppe von 11 Personen.

---

<sup>6</sup> Der zusätzliche Anpassungsschritt erzeugt Verzögerung von ca. 500-700ms und ist damit klar bemerkbar. Nichtsdestotrotz erzielte das so konfigurierte System im Rahmen des Stress Framework klar die höchsten Zufriedenheits- und niedrigsten Stresswerte unter den Proband:innen.

Alle Versuchspersonen wurden über den groben Versuchsablauf informiert und darüber aufgeklärt, dass digitale Biomarker erhoben werden. Auf die Möglichkeit zur Einsicht in die eigenen digitalen Biomarker wurde hingewiesen (wovon von den Versuchspersonen auch Gebrauch gemacht wurde). Durch Befragung und explizite Einwilligung der Versuchspersonen wurde außerdem eine freiwillige Teilnahme sichergestellt.

In der Folge wurden den Versuchspersonen Empatica Watches ('The world's most advanced smartwatch for continuous health monitoring') angelegt wobei die korrekte und sachgemäße Erfassung der Biomarker durch die Sensorik überprüft und sichergestellt wurde. Hygienischen Grundregeln folgend, wurde die Uhr vor bzw. nach jedem Versuchsteilnehmenden desinfiziert. Jede Versuchsperson wurde vier Szenarien ausgesetzt. Die ersten Anrufe wurden mit dem vorhandenen statischen System und jeweils randomisierten Sprechgeschwindigkeiten (langsam, mittel oder schnell) durchgeführt. Der letzte Anruf war ein Gespräch mit dem System mit aktivierter dynamischer Adaption der Nutzendenansprache.

Versuchspersonen sollten in System 1 (statisch, randomisierte Sprechgeschwindigkeit) verschiedene Aufgaben erledigen. Anruf 1 beinhaltete eine Terminbuchung, Anruf 2 eine Rezeptbestellung und Anruf 3 die Bestellung einer Überweisung. Mit System 2 (Anruf 4) wurde ebenfalls ein Termin gebucht.

Nach Versuchsreihe 1 (statisches System) sowie nach Versuchsreihe 2 (dynamisches System) sollte ein Feedback eingeholt werden. (Siehe dazu die Auswertung in Abb. 4.4) Dieses Feedback diente als Self-Report der teilnehmenden Versuchspersonen.

Zudem wurde ein ATI Score (Affinity for Technology Interaction Scale) erhoben. Hierbei sind hohe Werte als eine hohe Affinität zu Technologie-Interaktionen zu deuten (0% als geringster Wert und 100% als Maximum).

Physiologisches Feedback konnte zudem durch die Sensorauswertung und Aufbereitung der Daten in Kooperation mit dem DFKI erfolgen. Beispiele dazu sind in Abb 4.3 zu sehen. Eine Zusammenfassung verwendbarer Daten ist in Tabelle 2.1 zu finden.

Die Personen wussten erst nach dem Versuch, dass spezifisch Stressparameter erfasst wurden und verschiedene konfigurierte Kommunikationssysteme (statisch vs. dynamisch) vorlagen, um eine diesbezügliche Voreingenommenheit (bias) zu vermeiden.

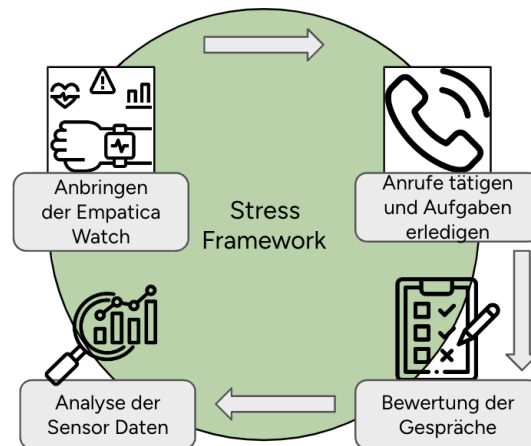


Abbildung 4.1 Versuchsdesign

## Ergebnisse

Abbildung 4.2 zeigt zwei Beispiele zur Auswertung der Sensordaten. Der Hautleitwiderstand bzw. die elektrodermale Aktivität (EDA) wurde über die Zeitspannen der verschiedenen Gespräche hinweg verglichen. Die Auswertungen zu Versuchsperson (VP) 9 und Versuchsperson (VP) 13 sollen hier beispielhaft genauer erläutert werden.

Wie alle VP sollten VP 9 und VP 13 im ersten Gespräch eine Terminbuchung absolvieren, im zweiten Gespräch ein Rezept beantragen und im dritten Gespräch eine Überweisung erfragen. Während des vierten Gesprächs sollte erneut ein Termin gebucht werden. Wie oben beschrieben, wurden die ersten drei Anrufe mit einem statischen System mit randomisierten Sprechgeschwindigkeiten durchgeführt, beim vierten Anruf wurde die Nutzendenansprache dynamisch adaptiert.

Für VP 9 wurden in System 1 (statisches System) für zwei Gespräche eine normaler Geschwindigkeit (violett) und für ein Gespräch eine schneller Geschwindigkeit (gelb) zufällig gewählt.

VP 13 führte Gespräche mit schneller (gelb), normaler (violett) und langsamer (rot) Geschwindigkeit der Sprachausgabe.

Tabelle 2.1 zeigt eine Zusammenfassung der auswertbaren Sensordaten.

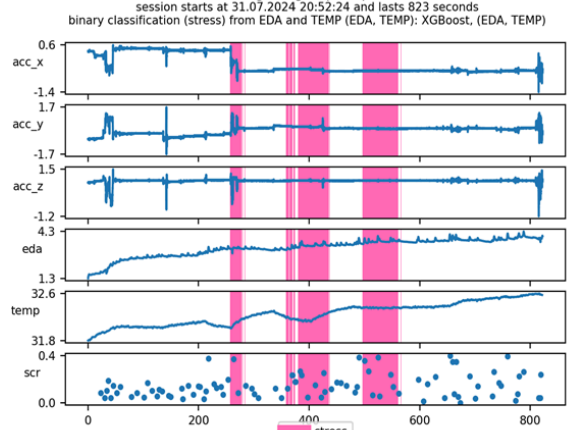
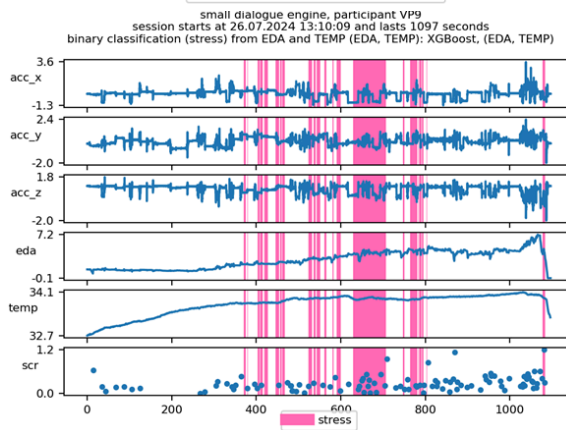
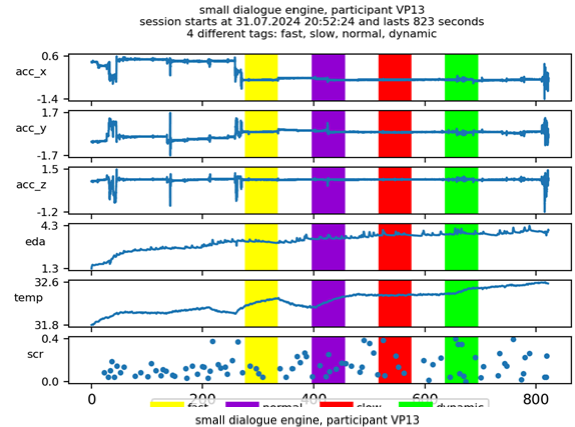
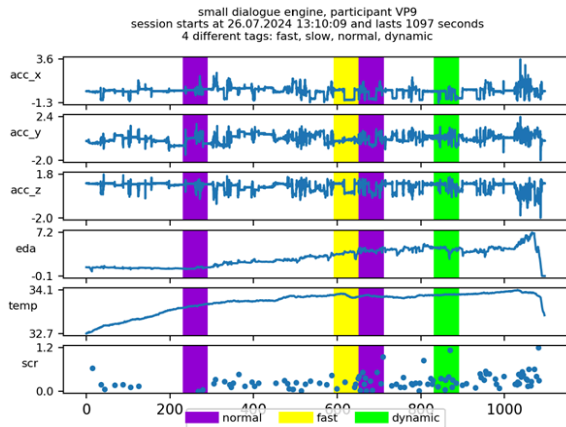


Abbildung 4.2 Beispiele zu Ergebnisse Stressframework

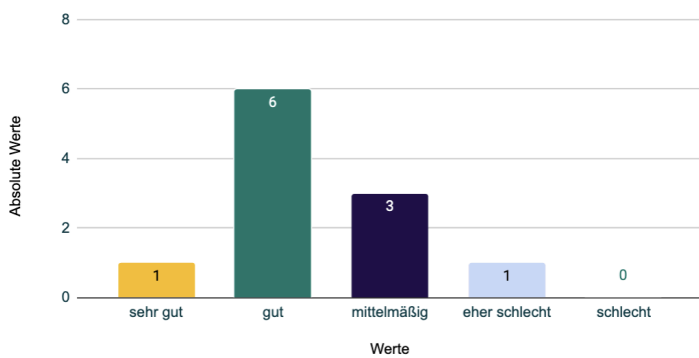
Links: Versuchsperson (VP) 9, Rechts VP 13

Oben: verschiedene Konversationstypen

Unten: Stressempfindung

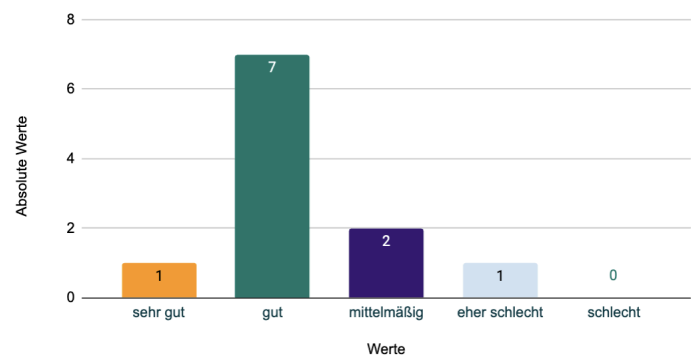
Wie würden Sie das Kommunikationserlebnis bewerten?

Statischer Assistent



Wie würden Sie das Kommunikationserlebnis bewerten?

Dynamischer Assistent



## Abbildung 4.3 Bewertung statisches vs dynamisches System

### Vollständig dynamische Dialogführung

Zusätzlich zu den bisher geschilderten Maßnahmen, Entwicklungen und Experimenten war es ein erklärtes Ziel der Aaron GmbH, weitere Innovations- und Forschungsansätze zu identifizieren, die nicht unmittelbar bzw. erst mit zeitlicher Verzögerung einem Verwertungsdruck unterliegen. Das Aufkommen einer kommerziellen Verfügbarkeit leistungsstarker generativer Textmodelle (LLM-Technologie, Large Language Models) stellt einen solchen Ansatz dar.

Im Lichte dieser Fortschritte wurde die Notwendigkeit identifiziert, den Ansatz zur Generierung von Dialogen bei der Interaktion mit menschlichen Nutzenden neu zu überdenken.

Es wurde festgestellt, dass die LLM-Technologie in ihrem aktuellen Zustand genügend Vorteile bietet, um ihre schrittweise Integration in den existierenden, produktiv verwendeten zustandsbasierten Ansatz (Finite State Machine, FSM) zur Dialogflussverarbeitung zu rechtfertigen. Es besteht eine nicht zu vernachlässigende Wahrscheinlichkeit, dass der FSM-basierte Ansatz zeitnah durch LLM-gesteuerte Dialogflüsse ersetzt wird, was jedoch umfangreiche Tests erfordert, die jedoch nicht im Rahmen des vorliegenden Projektes lagen.

Das übergeordnete technische Ziel ist es, eine modulare, anpassbare und erweiterbare Lösung zu entwerfen, die vorzugsweise auf einem Open-Source-basierten, nicht proprietären Anwendungsframework beruht.

Dies wird dazu beitragen, die formulierten Produkt- und Geschäftsziele zu erreichen, die wie folgt lauten:

- Reduzierung der Bearbeitungszeit für Anfragen durch Mitarbeiter:innen
- Erhöhung des Net Promoter Scores (NPS)
- Erzielung eines technologischen Vorteils gegenüber Wettbewerbern durch den Einsatz neuer Technologien, wobei gleichzeitig die Flexibilität erhalten bleibt, sich schnell anzupassen

Um die technischen, produktbezogenen und geschäftlichen Ziele umfassend zu unterstützen, werden folgende Anforderungen an die vorgeschlagene Lösung vorgeschlagen:

- *Abstraktes Framework:* Vermeidung API-spezifischer Implementierungen
- *Modulares Design und Architektur:*
  - Ermöglicht den einfachen Wechsel der zugrunde liegenden LLM-APIs
  - Ermöglicht die Abdeckung einzelner Schritte durch lokal ausgeführte oder gehostete LLMs
  - Ermöglicht die Abdeckung einzelner Schritte ohne LLM-Beteiligung (z.B. einfache NLP/IR-Aufgaben oder beliebige Funktionsaufrufe)

- Ermöglicht ein agentenbasiertes Design, d.h. zirkuläres Verhalten
- *Robustheit:*
  - Gegen API-Ausfälle
  - Gegen unerwünschtes/unerwartetes Verhalten oder Antworten von LLMs
  - Bereitstellung aufgabenspezifischer Fallback-Mechanismen
- *Überwachung/Testen:*
  - A/B-Testfähigkeit für aktualisierte Komponenten
    - Ermöglicht es, einen Bruchteil des Datenverkehrs durch die aktualisierte Komponente zu leiten
    - Markiert/protokolliert die Nutzung der aktualisierten Komponente im Zusammenhang mit der aktuellen Transaktion
    - Ermöglicht Berichte basierend auf der Nutzung aktualisierter vs. nicht aktualisierter Komponenten
    - Erleichtert das Auffinden von Verbesserungen/Verschlechterungen und potenziellen Kreuzkorrelationen
  - Kostenkontrolle und Erklärbarkeit von Entscheidungen/Ereignissen
    - Überwachung der Betriebskosten auf Kunden- bzw. Praxisbasis
    - Bereitstellung einer Quantifizierung des Nutzens im Verhältnis zu den Kosten (hinsichtlich der oben definierten Erfolgsmetriken)
    - Entscheidungsgrundlage für potenzielle Anforderungen an den Flusersatz

Diese Anforderungen bilden die Grundlage für die Entwicklung eines Systems zur vollständig dynamischen Generierung von Dialogen, das sowohl technologische als auch geschäftliche Vorteile bietet und gleichzeitig flexibel und anpassungsfähig bleibt.

Im Rahmen des Forschungsprojektes entstand hierbei ein prototypisch implementiertes System zur Dialoggenerierung.

# Voraussichtlicher Nutzen und Verwertbarkeit

## Nutzen für die Aaron GmbH

Die Aaron GmbH wird von den Projektergebnissen auf mehrere Weisen profitieren:

- Verbesserung der Transkriptionsqualität: Kurzfristig führt die massive Verbesserung der Transkriptionsqualität zu einer signifikant gesteigerten Kundenzufriedenheit, da die Transkription des gesprochenen Wortes die Grundlage für die Systemfunktionalität und die geplante Automatisierungs-Strategie, um Arbeitsalltag der Mitarbeitenden weiter zu entlasten bildet.
- Erhöhung der Nutzerakzeptanz und Anruferfolgsquote: Durch die im Projekt entwickelten Ansätze zur Erkennung und Anpassung an meta-kommunikative Signale wird eine höhere Nutzerakzeptanz und eine erhöhte Anruferfolgsquote erwartet. Dies wird durch die Ergebnisse der Stress-Experimente gestützt.
- Skalierbares und zukunftssicheres System: Mittel- bis langfristig stellen die Arbeiten zur vollständig dynamischen Dialoggenerierung die Grundlage für ein skalierbares und zukunftssicheres System dar. Dies ermöglicht unter anderem eine einfache Anpassung der Zielsprache und erfüllt damit unerlässliche Anforderungen für eine einfache Sprachlokalisierung.

## Nutzen für andere Anbieter von Voice-Lösungen

Andere Anbieter von Voice-Lösungen können ebenfalls von den Projektergebnissen profitieren:

- API-Endpunkt zur Messung der Gesprächsabbruchwahrscheinlichkeit: Über den exponierten API-Endpunkt könnten nach Bereitstellung und Bedarf auch andere Anbieter die Wahrscheinlichkeit eines Gesprächsabbruchs in der nächsten Sprach-Interaktion messen und ihre Gesprächsabläufe dahingehend optimieren. Dies ermöglicht es, während eines stattfindenden Dialogs Handlungsspielräume bei der Dialoggestaltung zu eröffnen, ohne hohen Ressourcenaufwand oder eigene Datenverfügbarkeit.

## Verwertbarkeit

Die Aaron GmbH plant die kommerzielle Verwertung der Projektergebnisse wie folgt:

- Integration der Erweiterungen am bestehenden Produkt: Die entwickelten Erweiterungen werden zeitnah in das bestehende Produkt integriert. Die antizipierten Verbesserungen in der Nutzerakzeptanz und im Anruferfolg sollen zuerst auf einer kleinen Gruppe von Pilotkunden validiert werden. Bei Erfolg werden die genannten Verbesserungen allen Kunden zur Verfügung gestellt.

Konkret erwartet sich die Aaron GmbH hier:

- SDE (mit adaptierter Ansprache) als Premium Add-On zu Aarons bestehendem Telefonassistenten (separat bepreist)

- Erhöhte Konversionsrate von Neu- und Trialkunden in 12-Monats Abonnements durch größere Kundenzufriedenheit
- Geringe Year-on-Year Churnrates für Kunden die das System nutzen
- Weiterentwicklung des dynamischen Dialogsystems: Im Falle des prototypischen dynamischen Dialogsystems sind bereits weitergehende Entwicklungen geplant, wie z.B. das sogenannte Guard-Railing, das verhindert, dass die zugrunde liegende LLM-Technologie vom eigentlichen Gesprächsinhalt abgebracht werden kann (z.B. durch zufällige Fragen).  
Abgesehen von einer, nach den Ergebnissen im Rahmen des Stress Framework zu erwartenden, geringeren Stressbelastung durch Nutzer, ergeben sich in diesem Szenario vor allem Synergieeffekte in der Skalierung und Internationalisierung:
  - Sprachanpassung: moderne LLMs sind meist multi-lingual trainiert, sodass eine Sprachanpassung in vielen Fällen durch entsprechende Instruktionen im Prompting erreicht werden kann.
  - Anpassbarkeit und Erweiterbarkeit: zukünftige funktionelle Anforderungen an den Dialog bzw. die Dialogführung können allein durch Konfigurationsanpassungen vorgenommen werden. Komplexe Entwicklungsarbeiten hierfür entfallen und entlasten Software-Entwickler:innen.
- Evaluation der Vermarktung und Monetarisierung der entwickelten API als SaaS-Lösung zur Messung und Überwachung der Dialoggüte für andere Voice-KI Anbieter

## Fortschritte bei anderen Stellen während des Projekts

Ursprünglich sollten ausschließlich auf Grundlage der vorhandenen Produktionsdaten ML-Modelle für die vorgesehenen Funktionen trainiert werden. Im Laufe des Projekts wurden jedoch neuartige Large Language Models (LLMs) (z.B. (OpenAI, 2024)) im kommerziellen Rahmen verfügbar, was dazu führte, dass bestimmte Teilaspekte neu bewertet wurden. Beispielsweise wurde davon abgesehen, Modelle zur Sentiment-Detection speziell zu trainieren und stattdessen LLMs für die Bestimmung von Sentimenten in Texten verwendet.

Zusätzlich eröffnete die freie Publikation des Transkriptionsmodells “Whisper” (Radford et al., 2023) die Möglichkeit einer signifikanten Verbesserung der Transkriptionsqualität und damit eine verbesserte Ausgangslage für alle weiteren, auf die Audiotranskription aufbauenden, Verarbeitungs-, Analyse- und Optimierungsprozesse.

## Veröffentlichungen

Im Rahmen des AP 8 (Meta-kommunikative Signale und Optimierung der Nutzendenansprache) wurde in Kooperation mit dem DFKI die Publikation (Soliman et al., 2024) erstellt und auf der

Konferenz “The 32nd ACM Conference on User Modeling, Adaptation and Personalization“ präsentiert:

Soliman, H., Kravcik, M., Basvoju, N., & Jaehnichen, P. (2024, June). Using Large Language Models for Adaptive Dialogue Management in Digital Telephone Assistants. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 399-405).

Verfügbar unter:

<https://dl.acm.org/doi/abs/10.1145/3631700.3664902>

## Quellen

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2), 124–129. <https://doi.org/10.1037/h0030377>

Ito, A. (2020). Human–machine metacommunication towards development of a human-like agent: A short review. *Acoustical Science and Technology*, 41(1), 166-169.  
10.1250/ast.41.166

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist*, 52(5), 509–516. <https://doi.org/10.1037/0003-066X.52.5.509>

OpenAI. (1, 1 2024). *OpenAI API*. OpenAI API overview. Retrieved September 26, 2024, from <https://openai.com/>

Parker, L. S., Sankar, P. L., Boyer, J., McEwen, J. J., & Kaufman, D. (2018). Normative and conceptual ELSI research: what it is, and why it's important. *Genetics in Medicine*, 21(2), 505–509. <https://doi.org/10.1038/s41436-018-0065-x>

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, 40(1182), 28492--28518.  
<https://dl.acm.org/doi/10.5555/3618408.3619590>

Soliman, H., Kravcik, M., Nagasandeepta, B., & Patrick, J. (2024). Using Large Language Models for Adaptive Dialogue Management in Digital Telephone Assistants. *Adjunct*

*Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and*

*Personalization*, 1(1), 399-405. <https://dl.acm.org/doi/abs/10.1145/3631700.3664902>

Suendermann, D., Liscombe, J., & Pieraccini, R. (2010). Optimize the obvious: Automatic call flow generation. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1(1). 10.1109/ICASSP.2010.5494936

Suhm, B., & Peterson, P. A. (2002). A Data-Driven Methodology for Evaluating and Optimizing Call Center IVRs. *International Journal of Speech Technology*, 5, 23-37.  
10.1023/A:1013674413897

Ziegler, M. G. (2004). Psychological stress and the autonomic nervous system. In *Elsevier eBooks* (pp. 189–190). <https://doi.org/10.1016/b978-012589762-4/50051-7>