

# **KI-gestützte Spracherkennung für automatisierte Notfall- protokolle**

**Eine Literaturstudie zu Notfall Information and Communication Technology-Work-  
flows in der präklinischen Versorgung**

Fakultät Angewandte Gesundheitswissenschaften

Prüfungs- und Studienarbeit (PStA)

Studiengang: Bachelor Management im Rettungswesen

Betreuer

Marc Lechner

Vorgelegt von

Julian, Linden

Stephingerberg 1,

86152 Augsburg

Matr.-Nr.: 22410991

Augsburg, den 30.01.2026

## Inhalt

|                                                                      |     |
|----------------------------------------------------------------------|-----|
| Abkürzungsverzeichnis .....                                          | iii |
| Abbildungsverzeichnis.....                                           | iv  |
| Tabellenverzeichnis.....                                             | iv  |
| 1. Einleitung.....                                                   | 1   |
| 1.1. Problemstellung und Relevanz .....                              | 1   |
| 1.2. Forschungsfrage und Zielsetzung .....                           | 1   |
| 2. Grundlagen und State-of-the-Art .....                             | 2   |
| 2.1. NICT-Herausforderungen in der präklinischen Dokumentation ..... | 2   |
| 2.2. KI-Technologien für medizinische Spracherkennung.....           | 2   |
| 2.2.1. Automatic-Speech-Recognition (ASR) .....                      | 2   |
| 2.2.2. Large-Language-Models (LLMs).....                             | 3   |
| 2.2.3. Multi-Agenten-Systeme .....                                   | 4   |
| 3. Methodik.....                                                     | 4   |
| 4. Konzeptioneller Workflow und Prototyp.....                        | 5   |
| 4.1. Systemarchitektur .....                                         | 5   |
| 4.2. Workflow-Beschreibung.....                                      | 5   |
| 4.3. Vergleich: Cloud vs. Lokale Verarbeitung .....                  | 6   |
| 5. Regulatorik und Ethik.....                                        | 7   |
| 5.1. Rechtlicher Rahmen .....                                        | 7   |
| 5.2. Ethische Erwägungen und Risikomanagement .....                  | 8   |
| 6. Diskussion und Ausblick.....                                      | 9   |
| 6.1. Potenziale der KI-gestützten Protokollierung .....              | 9   |
| 6.2. Herausforderungen und Limitationen .....                        | 9   |
| 7. Exkurs: KI als klinische Entscheidungsunterstützung .....         | 10  |
| 8. Fazit .....                                                       | 10  |
| Literaturverzeichnis.....                                            | 12  |
| Anlage .....                                                         | 15  |
| Eigenständigkeitserklärung.....                                      | i   |

## Abkürzungsverzeichnis

|         |                                                                                                                           |
|---------|---------------------------------------------------------------------------------------------------------------------------|
| ACS     | Acute Coronary Syndrome / Akutes Koronarsyndrom                                                                           |
| AI      | Artificial Intelligence / Künstliche Intelligenz                                                                          |
| ASR     | Automatic Speech Recognition / Automatische Spracherkennung                                                               |
| ASS     | Acetylsalicylic Acid / Acetylsalicylsäure                                                                                 |
| CDSS    | Clinical Decision Support System / Klinisches Entscheidungsunterstützungssystem                                           |
| CREOLA  | Clinical Risk Evaluation Of LLM-based Automation / Framework zur Bewertung von Halluzinationen                            |
| DIVI    | Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin                                                   |
| DSGVO   | Datenschutz-Grundverordnung                                                                                               |
| EMS     | Emergency Medical Services / Rettungsdienst                                                                               |
| GMLP    | Good Machine Learning Practice / Gute Praxis für maschinelles Lernen                                                      |
| IMDRF   | International Medical Device Regulators Forum                                                                             |
| KHK     | Koronare Herzkrankheit                                                                                                    |
| KI-VO   | Verordnung über künstliche Intelligenz (EU-KI-Verordnung)                                                                 |
| LLM     | Large Language Model / Großes Sprachmodell                                                                                |
| (m)WER  | (Medical) Word Error Rate / Medizinische Wortfehlerrate                                                                   |
| MDR     | Medical Device Regulation / Verordnung (EU) 2017/745                                                                      |
| MIND    | Minimaler Notfalldatensatz                                                                                                |
| NICT    | Notfall Information and Communication Technology                                                                          |
| RD      | Rettungsdienst                                                                                                            |
| RTW     | Rettungswagen                                                                                                             |
| SAMPLER | Symptome, Allergien, Medikation, Patientenvorgeschichte (Past medical history), Letzte Mahlzeit, Ereignis, Risikofaktoren |
| SNR     | Signal-to-Noise Ratio / Signal-Rausch-Verhältnis                                                                          |
| UI      | User Interface / Benutzeroberfläche                                                                                       |

**Abbildungsverzeichnis**

Figure 1 Systemarchitektur (Schema) ..... 5

**Tabellenverzeichnis**

Table 1 NICT-Herausforderungen und Implikationen für KI-Systeme..... 2

Table 2 SST-Engine-Vergleich unter RD-Bedingungen ..... 3

Table 3 Ergebnisse zu Halluzinationsrisiken ..... 3

Table 4 Verarbeitungs-Optionen für den deutschen Rettungsdienst ..... 6

Table 5 Regulatorische Anforderungen im Überblick..... 7

Table 6 GMLP-Prinzipien und praktische Umsetzung ..... 7

## **1. Einleitung**

### **1.1. Problemstellung und Relevanz**

Die Dokumentation medizinischer Maßnahmen im Rettungsdienst stellt eine der komplexesten Aufgaben in der präklinischen Notfallversorgung dar (Hedderson et al., 2025). Rettungsfachpersonal ist mit multiplen simultanen Anforderungen konfrontiert: unmittelbare Patientenversorgung, Kommunikation mit der Leitstelle, Koordination mit weiteren Einsatzkräften sowie parallele Dokumentation – unter denkbar ungünstigen Bedingungen, wie engen Räumlichkeiten, Vibrationen, variierenden Lichtverhältnissen und inhärentem Zeitdruck.

Die Konsequenzen unvollständiger Protokolle betreffen mehrere Dimensionen: Medizinisch gefährdet lückenhafte Übergabedokumentation die Behandlungskontinuität, da kritische Informationen verloren gehen können. Rechtlich bildet das Notfallprotokoll ein zentrales Beweismittel mit forensischer Relevanz. Qualitätssicherungssysteme basieren maßgeblich auf der Analyse vollständiger Protokolldaten gemäß dem Minimalen Notfalldatensatz (MIND) der Deutschen Interdisziplinären Vereinigung für Intensivmedizin und Notfallmedizin (DIVI) (2024).

Notfall Information and Communication Technology (NICT) bezeichnet die Gesamtheit digitaler Systeme zur Dokumentation und Verarbeitung von Informationen während Rettungseinsätzen. Die Integration von Künstlicher Intelligenz (KI), insbesondere durch Automatic Speech Recognition (ASR) und einem Large Language Model (LLM), verspricht eine Transformation der Dokumentationsprozesse (Menzies et al., 2024). Diese Technologien könnten es ermöglichen, die Dokumentation von einer konkurrierenden Tätigkeit zur Patientenversorgung in eine begleitende, weitgehend automatisierte Funktion zu transformieren.

### **1.2. Forschungsfrage und Zielsetzung**

Die zentrale Forschungsfrage dieser Arbeit lautet: Inwiefern können KI-gestützte Spracherkennungssysteme die Erstellung von Notfallprotokollen im präklinischen Setting unterstützen und welche technologischen, regulatorischen sowie ethischen Anforderungen sind dabei zu berücksichtigen?

Zur Beantwortung dieser Frage werden folgende Teilaspekte untersucht:

1. Welchen technologischen Reifegrad haben aktuelle ASR- und LLM-Systeme für medizinische Domänen erreicht?
2. Welche spezifischen Herausforderungen ergeben sich aus dem präklinischen Einsatzkontext?
3. Welche regulatorischen Rahmenbedingungen gelten für KI-basierte Dokumentationssysteme, insbesondere die Good Machine Learning Practice (GMLP)-Prinzipien?
4. Wie kann ein datenschutzkonformer Workflow nach DIVI-MIND-Standard konzipiert werden?

## 2. Grundlagen und State-of-the-Art

### 2.1. NICT-Herausforderungen in der präklinischen Dokumentation

Die Herausforderungen lassen sich in drei Kategorien, akustische, normative und kognitive Faktoren, einteilen. Akustische Umgebungsbedingungen stellen eine der gravierendsten Herausforderungen dar (Moser et al., 2025). Die Benchmark-Studie von Moser, Stanic und Sariyar (2025) quantifizierte erstmals systematisch die Auswirkungen verschiedener Lärm-typen: 99 synthetische Notfalldialoge wurden mit Geräuschtypen (Menschenmengen, Ver-kehr, öffentliche Räume, Rettungswageninnenraum) bei fünf SNR-Stufen überlagert, was 1.980 verrauschte Audio-Samples ergab. "Inside crowded"-Lärm hatte den stärksten nega-tiven Einfluss auf die Erkennung medizinischer Fachbegriffe, wobei vor allem im Rettungs-dienst-Kontext Wert auf die Medical Word Error Ratio (mWER) gelegt wird.

Der DIVI-MIND-Standard bildet den normativen Rahmen für die präklinische Dokumenta-tion. MIND 4.0 ermöglicht modulare, situationsgerechte Dokumentation (DIVI, 2024). Dieser bundesweit akzeptierte Standard bildet die Grundlagen für medizinisches Qualitätsma-nagement. Darüber hinaus limitiert kognitive Belastung die Dokumentationskapazität er-heblich (Asgari et al., 2023). Sprachbasierte Systeme versprechen Entlastung durch Elim-inierung motorischer Schreibanforderungen.

Die nachfolgende Tabelle fasst die größten Herausforderungen in der präklinischen Doku-mentation mit jeweiliger Quantifizierung und daraus resultierender Anforderung an ASR/LLM-Systeme zusammen.

Table 1 NICT-Herausforderungen und Implikationen für KI-Systeme

| Herausforderung       | Quantifizierung           | Implikation für ASR/LLM           |
|-----------------------|---------------------------|-----------------------------------|
| Umgebungs-lärm        | SNR -10 bis +20 dB        | Robuste Geräuschunterdrückung     |
| "Inside crowded" Lärm | Stärkster Effekt auf mWER | Spezifisches Training             |
| Kognitive Last        | Multitasking unter Stress | Minimale Interaktionskomplexität  |
| Standardheterogenität | DIVI-MIND 4.0             | Strukturierte Output-Formatierung |
| Medizinvokabular      | mWER >> WER               | Domänenadaptation essentiell      |

### 2.2. KI-Technologien für medizinische Spracherkennung

#### 2.2.1. Automatic-Speech-Recognition (ASR)

Moderne ASR-Systeme nutzen End-to-End-Deep-Learning-Architekturen. OpenAI Whisper ist eines dieser Systeme und wurde auf 680.000 Stunden mehrsprachiger Audiodaten trai-niert und unterstützt 99 Sprachen; allerdings ist deutschsprachiges, medizinisches Fachvo-kabular nicht proportional repräsentiert (Radford et al., 2022). Die Medical Word Error Rate (mWER) misst spezifisch die Fehlerquote bei der Erkennung von medizinischen Fachbe-griffen, wie Diagnosen, Prozeduren und Medikamentennamen (Moser et al., 2025). Diese

Differenzierung ist klinisch relevant, da ein falsch erkannter Medikamentenname schwerwiegendere Konsequenzen hat als ein falsch erkanntes Füllwort.

Nachfolgende Tabelle zeigt einen Vergleich der Speech-to-Text-Engine-Performance unter Rettungsdienst-Bedingungen basierend auf aktuellen kontextualisierten Benchmark-Studien.

Table 2 SST-Engine-Vergleich unter RD-Bedingungen

| System               | WER (sauer) | WER (realistisch noisy)  | mWER    | Stärken                            | Schwächen                                |
|----------------------|-------------|--------------------------|---------|------------------------------------|------------------------------------------|
| Whisper large-v3     | ~5%         | ~15-20% (Inside Crowded) | hoch    | Multilingual, Open-Source          | Rechenintensiv, keine RD-Adaptation      |
| Whisper v3 Turbo     | low         | ~12-15%                  | niedrig | Beste Open-Source für EMS, schnell | Braucht Fine-Tuning für Deutsch          |
| Google ASR           | ~4%         | ~12%                     | niedrig | Beste Scores in EMS/RD-Kategorien  | Cloud-only, keine mediz. Spezialisierung |
| recapp (kommerziell) | ~4%         | ~12%                     | niedrig | Beste RD-Performance, robust       | Proprietär, Cloud-abhängig               |

Basierend auf Moser, Hofecker und Luo (Hofecker, 2024; Luo et al., 2024; Moser et al., 2025).

2.2.2. Large-Language-Models (LLMs)

LLMs bilden die zweite Säule KI-gestützter Dokumentationssysteme, indem sie die unstrukturierten ASR-Outputs in formatierte, kohärente Protokolltexte transformieren (Menzies et al., 2024). Das im Kontext vielleicht größte aktuelle Problem von LLMs sind Halluzinationsrisiken, also dass inhaltliche Fehler entstehen durch beispielsweise erfundene, aus dem Kontext gerissene oder fehlinterpretierte Wörter. Asgari et al. (2025) entwickelten das sog. CREOLA-Framework zur Quantifizierung und Klassifikation von Halluzinationen in LLM-generierten medizinischen Texten. Zentrale Ergebnisse aus 12.999 Sätzen (450 klinische Notizen) sind nachfolgend zusammengefasst:

Table 3 Ergebnisse zu Halluzinationsrisiken

| Metrik                         | Wert  | Bedeutung                   |
|--------------------------------|-------|-----------------------------|
| Halluzinationsrate             | 1,47% | Vergleichsweise niedrig     |
| Major Errors (Halluzinationen) | 44%   | Erhebliches Risikopotenzial |

Die verschiedenen Halluzinationstypen nach Häufigkeit sind: 43% Fabricated (erfunden), 30% Negations (Verneinungen), 17% Contextual (Kontextfehler), 10% Causality (Kausalitätsfehler). Major Halluzinationen traten am häufigsten in den Abschnitten Plan (21%), Assessment (10,5%) und Symptoms (5,2%) auf (Asgari et al., 2025). Diese Verteilung ist für die Notfallprotokollierung besonders relevant, da gerade der Behandlungsplan und die Befundbeurteilung kritische Dokumentationselemente darstellen. Diese Daten zeigen, dass trotz einer vergleichsweise niedrigen absoluten Halluzinationsrate fast die Hälfte dieser Fehler klinisch schwerwiegend ist (Asgari et al., 2025). Für die Implikation in der Praxis unterstreicht das die obligatorische Notwendigkeit menschlicher Verifikation vor der Finalisierung jedes protokollierten Dokuments.

Lokale LLMs adressieren Datenschutzbedenken, indem sie Inferenz auf lokaler Hardware ermöglichen. Modelle wie LLaMA (Touvron et al., 2023) und Mistral 7B (Jiang et al., 2023) bieten Open-Source-Alternativen für den deutschen Rettungsdienstkontext, wo DSGVO-Konformität essentiell ist.

### **2.2.3. Multi-Agenten-Systeme**

Multi-Agenten-Systeme sind KI-Architekturen, bei denen mehrere spezialisierte Agenten (z.B. ASR-Agent, Strukturierungs-Agent, Validierungs-Agent) parallel oder sequenziell zusammenarbeiten, um komplexe Aufgaben wie Echtzeit-Dokumentation zu bewältigen. Diese Systeme bieten gegenüber einzelnen LLM-Ansätzen den Vorteil der Modularität und Fehlertoleranz: Während ein LLM bei Halluzinationen oder ASR-Fehlern das gesamte Protokoll kompromittiert, können spezialisierte Agenten mit integrierten Validierungsschleifen kritische Fehler in Echtzeit erkennen und korrigieren – ein Ansatz, der für sicherheitskritische RD-Dokumentation ebenfalls vielversprechend ist. Das Fraunhofer IAIS entwickelt im Projekt „TraumaInterfaces“ ein Multi-Agenten-System für die Echtzeitdokumentation im Schockraum. Ein Prototyp soll im Sommer 2026 einsatzfähig sein (Fraunhofer IAIS, o. J.)

## **3. Methodik**

Die vorliegende Arbeit basiert auf einer im Januar 2026 durchgeführten narrativen, KI-gestützten Literaturrecherche in den Datenbanken PubMed, IEEE Xplore, arXiv sowie Google Scholar. Der Suchzeitraum umfasst hauptsächlich Publikationen von 2020 bis Januar 2026, um den rapiden technologischen Fortschritt im Bereich generativer KI adäquat abzubilden.

Suchstring: (speech recognition OR ASR OR Whisper OR LLM OR "large language model") AND (emergency OR EMS OR "prehospital care" OR Rettungsdienst OR präklinisch) AND (documentation OR protocol OR transcription OR "clinical note"), Zeitraum 2020-2026, Sprachen Deutsch/Englisch.

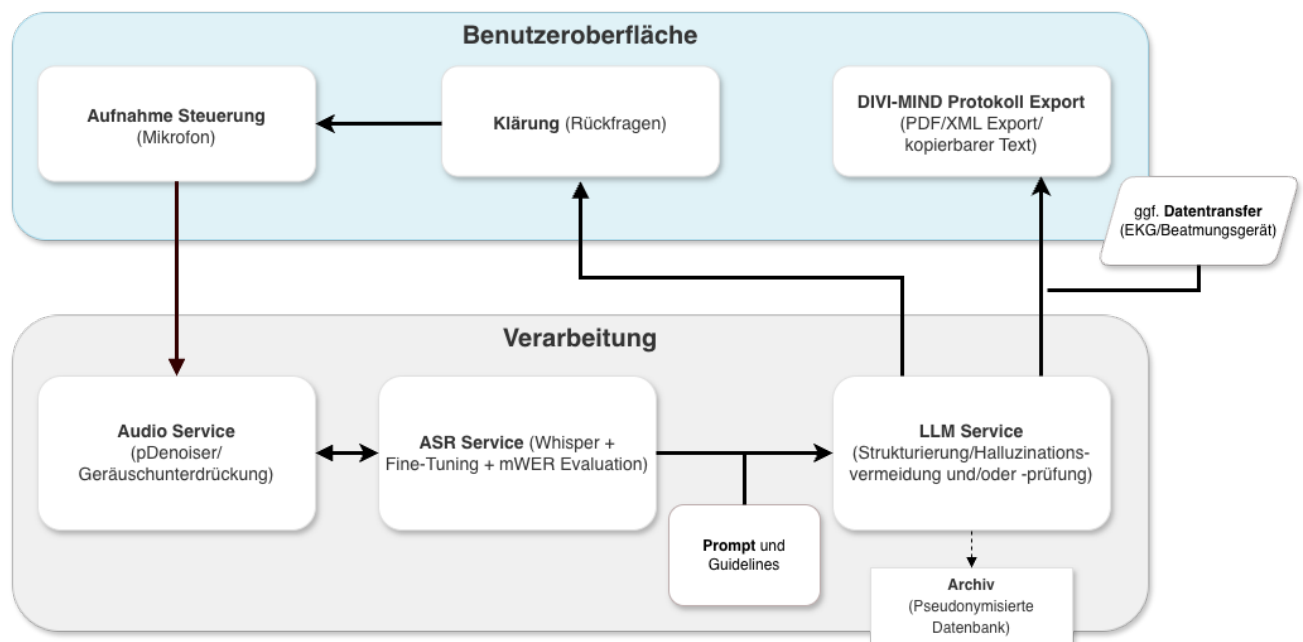
Mit Unterstützung von KI-Tools (Claude für Datenbankauswahl, Consensus für Abstract-Filterung) wurden 10 Kernquellen und 8 weitere Quellen identifiziert. Aufgrund der rasanten KI-Entwicklung – insbesondere seit ChatGPT Ende 2022 – erfordert eine aktuelle Literaturstudie eine Erweiterung des klassischen Quellenspektrums. Peer-reviewed Publikationen weisen typischerweise eine Latenz von 12-24 Monaten auf, weshalb auch Preprints, regulatorische Primärquellen und Grey Literature einbezogen wurden (Paez, 2017)

## 4. Konzeptioneller Workflow und Prototyp

### 4.1. Systemarchitektur

Zur Umsetzung der theoretischen Überlegungen mit den bereits beschriebenen Anforderungen wurde eine iOS-App entwickelt, die aus der Erzählung des Einsatzgeschehens durch das Rettungsdienstpersonal ein strukturiertes schriftliches Protokoll mithilfe eines LLM anfertigt. Screenshots der App-Oberfläche und der Nutzung finden sich in der Anlage. Die Systemarchitektur folgt einem modularen Aufbau von der Sprachaufnahme bis zum fertigen Protokoll. Nachfolgende Abbildung zeigt schematisch den Ablauf hiervon sowie den Zusammenhang zwischen Benutzeroberfläche und interner System-Verarbeitung.

Figure 1 Systemarchitektur (Schema)



### 4.2. Workflow-Beschreibung

Der Workflow umfasst fünf aufeinanderfolgende Phasen mit klaren Übergaben, sodass Fehler früh erkannt und korrigiert werden können.

Phase 1: Sprachaufnahme mit Geräuschunterdrückung. Das Diktat wird per Headset oder Gerätemikrofon aufgenommen. Eine personalisierte Sprachverbesserung (z.B. pDenoiser) trennt die Stimme des Dokumentierenden von typischen Störquellen im RTW (Motor, Sirene, Nebengespräche) (Zhang et al., 2025). Die Verarbeitung sollte echtzeitfähig sein und die Aufnahmequalität über einen visuellen Indikator (Signal-Rausch-Verhältnis) rückmelden, um ggf. Position oder Aufnahme zu korrigieren.

Phase 2: Transkription mit domänenadaptiertem ASR. Die bereinigte Audiospur wird mit einem lokal ausführbaren, medizinisch feinjustierten deutschen Whisper-Modell transkribiert. Das schützt Patientendaten und erlaubt Dokumentation auch ohne Netzabdeckung. Zur Qualitätssicherung werden WER/mWER und Konfidenzen überwacht; unsichere Passagen werden markiert. Kritische Begriffe (Medikamente, Dosierungen, Diagnosen) können eine explizite Bestätigung erfordern.

Phase 3: LLM-Strukturierung nach DIVI-MIND. Die Transkription wird durch ein LLM in das DIVI-MIND-Format (4.0) überführt. Relevante Informationen werden extrahiert und den passenden Protokollfeldern zugeordnet; deterministische Zuordnungslogik (Prompting/Regeln) reduziert Variabilität. Inhalte lassen sich u.a. an das SAMPLER-Fragenschema anlehnen (z.B. Symptome/Ereignis, Vorerkrankungen/Medikation/Allergien).

Phase 4: CREOLA-basierte Qualitätsprüfung. Vor der Finalisierung prüft ein CREOLA-Check das Protokoll gegen die Transkription und markiert Abweichungen risikobasiert (Asgari et al., 2025). So wird die menschliche Prüfung auf potenziell kritische Stellen fokussiert.

Phase 5: Obligatorische menschliche Verifikation. Die Freigabe durch Rettungsfachpersonal ist systemkritisch. Das Protokoll wird vollständig geprüft, ggf. korrigiert und erst dann final exportiert. Das User Interface (UI) unterstützt durch Hervorhebung der Markierungen und einfache Bearbeitung. Idealerweise erfolgt die Verifikation noch während oder unmittelbar nach dem Einsatz; unbestätigte Protokolle bleiben als „Entwurf“ gekennzeichnet und werden erinnert.

**4.3. Vergleich: Cloud vs. Lokale Verarbeitung**

Es folgt ein systematischer Vergleich zwischen Cloud und lokaler Verarbeitung anhand verschiedener relevanter Kriterien, um die Verarbeitungsoptionen hinreichend zu beleuchten und eine Empfehlung für den eventuell überlegenen Verarbeitungsdienst auszusprechen.

Table 4 Verarbeitungs-Optionen für den deutschen Rettungsdienst

| Kriterium         | Cloud-Dienste                                                                                  | Lokale Verarbeitung               |
|-------------------|------------------------------------------------------------------------------------------------|-----------------------------------|
| mWER-Performance  | Überlegen aufgrund der serverseitigen Rechenleistung                                           | Akzeptabel (mit context Training) |
| DSGVO-Konformität | Aufwendig (Auftragsverarbeitungsvertrag, technische und organisatorische Maßnahmen, Drittland) | Vereinfacht                       |
| Offline-Fähigkeit | Nein                                                                                           | Ja (kritisch für Funklücken)      |
| DIVI-MIND-Export  | Möglich                                                                                        | Möglich                           |
| GMLP-Compliance   | Drittanbieter-Abhängigkeit                                                                     | Volle Kontrolle                   |
| Latenz            | Netzwerkabhängig                                                                               | Konsistent                        |

Für den produktiven Einsatz im deutschen Rettungsdienst wird eine lokale Verarbeitung mit Kontext-trainiertem Whisper-Modell empfohlen, um DSGVO-Anforderungen zu erfüllen und Offline-Fähigkeiten in Funklücken zu gewährleisten. Die technische Konzeption allein reicht jedoch nicht aus, um ein KI-gestütztes Dokumentationssystem in den Rettungsdienst zu bringen. Der folgende Abschnitt widmet sich den regulatorischen Rahmenbedingungen

und ethischen Erwägungen, die bei der Entwicklung und Zulassung solcher Systeme zu berücksichtigen sind.

5. Regulatorik und Ethik

5.1. Rechtlicher Rahmen

Der rechtliche Rahmen für KI-basierte medizinische Dokumentationssysteme wird durch mehrere europäischen Verordnungen definiert. Tabelle 5 bietet einen Überblick über die relevanten Anforderungen.

Table 5 Regulatorische Anforderungen im Überblick

| Rechtsakt                               | Anforderung                                | Status für Notfall-ASR            |
|-----------------------------------------|--------------------------------------------|-----------------------------------|
| Verordnung (EU) 2017/745 (MDR)          | Klassifikation als Software-Medizinprodukt | Prüfung je nach Funktionsumfang   |
| Verordnung (EU) 2024/1689 (KI-VO)       | Hochrisiko-Kategorisierung                 | Wahrscheinlich, wenn diagnostisch |
| IMDRF N88 GMLP                          | 10 Prinzipien                              | Implementierbar                   |
| Art. 9 Verordnung (EU) 2016/679 (DSGVO) | Gesundheitsdaten                           | Lokale Verarbeitung empfohlen     |
| DIVI-MIND 4.0                           | Dokumentationsstandard                     | Output-Format                     |

Die Good Machine Learning Practice (GMLP)-Prinzipien definieren gemäß IMDRF N88:2025 international anerkannte Qualitätsanforderungen für den gesamten Lebenszyklus medizinischer KI-Systeme – von der Datenerhebung über das Modelltraining bis zum klinischen Einsatz. Ziel dieser Prinzipien ist es, die Sicherheit, Wirksamkeit, Nachvollziehbarkeit und kontinuierliche Überwachung KI-basierter Medizinprodukte sicherzustellen und sie klar in bestehende Software- und Medizinprodukte-Normen einzuordnen. Im Kontext KI-gestützter Rettungsdienst-Dokumentation geben die GMLP-Prinzipien damit den Rahmen vor, wie ASR- und LLM-Systeme entwickelt, validiert und im täglichen Einsatz so betrieben werden müssen, dass Patientensicherheit und regulatorische Anforderungen gewahrt bleiben. Tabelle 6 zeigt die praktische Umsetzung dieser Prinzipien für Notfall-ASR-Systeme:

Table 6 GMLP-Prinzipien und praktische Umsetzung

| Nr. | GMLP-Prinzip                 | Praktische Umsetzung für Notfall-ASR                    |
|-----|------------------------------|---------------------------------------------------------|
| 1   | Multi-disziplinäre Expertise | Notfallmediziner + ML-Ingenieure + Regulatorik-Experten |

|    |                             |                                                           |
|----|-----------------------------|-----------------------------------------------------------|
| 2  | Good Software Engineering   | ISO 62304 für medizinische Software                       |
| 3  | Repräsentative Daten        | Deutschsprachige Rettungsdienst-Daten (vgl. MedASR-DE-AT) |
| 4  | Datentrennung               | Strikte Training/Validation/Test-Splits                   |
| 5  | Referenzdaten-Qualität      | Verifizierte medizinische Transkriptionen                 |
| 6  | Modell-Design               | Dokumentierte Architektur (Whisper-Variante)              |
| 7  | Fokus auf Human-AI-Teaming  | Obligatorische Verifikation (vgl. Abschnitt 2.2.2)        |
| 8  | Klinische Validierung       | Prospektive Evaluation im EMS-Setting                     |
| 9  | Transparenz                 | Erklärbare Outputs, Konfidenz-Scores                      |
| 10 | Kontinuierliches Monitoring | Post-Market mWER-Tracking                                 |

## 5.2. Ethische Erwägungen und Risikomanagement

Das CREOLA-Framework (Asgari et al., 2025) bietet einen strukturierten Ansatz zur ethischen Bewertung. Basierend auf den in Abschnitt 2.2.2 dokumentierten Halluzinationstypen ergeben sich folgende Handlungsempfehlungen:

Table 7 Risikomatrix (Halluzinationen)

| Halluzinationstyp | Häufigkeit | Major Error Rate | Handlungsempfehlung      |
|-------------------|------------|------------------|--------------------------|
| Fabricated        | 43%        | Hoch             | Immer manuell prüfen     |
| Negation          | 30%        | Hoch             | Vitalzeichen-Cross-Check |
| Contextual        | 17%        | Mittel           | Plausibilitätsprüfung    |
| Causality         | 10%        | Mittel           | Klinisches Review        |

Die dokumentierten Halluzinationsrisiken von LLM-Systemen – mit 44% der Fehler als klinisch schwerwiegend – erfordern klare ethische Leitlinien und Verantwortlichkeitsstrukturen. Daraus ergeben sich vier zentrale ethische Anforderungen:

1. Transparenz: Patienten über KI-Unterstützung informieren
2. Verantwortlichkeit: Dokumentierende Person bleibt verantwortlich
3. Human-in-the-Loop: Obligatorische humane Verifikation vor Finalisierung
4. Bias-Monitoring: Regelmäßige Evaluation auf systematische Verzerrungen

## **6. Diskussion und Ausblick**

### **6.1. Potenziale der KI-gestützten Protokollierung**

Die Literatur zeigt, dass KI-gestützte Spracherkennung das Potenzial hat, die Rettungsdienst-Dokumentation deutlich zu verbessern. Dabei lassen sich messbare technische Fortschritte von qualitativen Entlastungseffekten unterscheiden.

Domänenadaptierte ASR-Systeme können die Fehlerraten medizinischer Transkription spürbar senken; die Wiener Studie berichtet eine wesentliche WER-Verbesserung (vgl. Abschnitt 2.2.1). Das ist klinisch relevant, weil eine geringere Fehlerrate bei Fachbegriffen das Risiko von Missverständnissen bei kritischen Angaben (z. B. Dosierungen, Diagnosen) reduziert. Zudem erlauben systematische Benchmarks zur Geräuschrobustheit (Moser et al., 2025) eine evidenzbasierte Auswahl und Optimierung von ASR-Systemen für typische Lärmprofile im Rettungswagen. Die automatisierbare, DIVI-MIND-konforme Strukturierung verspricht darüber hinaus eine Standardisierung der Protokollqualität, da ein KI-System konsistent erforderliche Felder befüllen und auf Informationslücken hinweisen kann.

Neben den technischen Effekten kann KI-gestützte Protokollierung die Arbeitssituation qualitativ verbessern, indem kognitive Last vom parallelen Dokumentieren auf die Patientenversorgung verlagert wird (Asgari et al., 2023). Diktatbasierte Dokumentation lässt sich oft besser in den Einsatzablauf integrieren als Formulararbeit. Gleichzeitig kann ein verlässliches System, das Informationen erfasst und strukturiert, Dokumentationsstress und die Sorge, relevante Angaben zu vergessen, reduzieren.

### **6.2. Herausforderungen und Limitationen**

Den Potenzialen stehen erhebliche Herausforderungen gegenüber, die für eine erfolgreiche Implementierung gezielt adressiert werden müssen.

Das Halluzinationsrisiko bleibt zentral: Die in Abschnitt 2.2.2 beschriebenen Befunde deuten darauf hin, dass ein relevanter Anteil der Halluzinationen klinisch schwerwiegend ist. Für den Rettungsdienst bedeutet das, dass bereits seltene Fehlgenerierungen potenziell sicherheitskritisch sein können – weshalb eine obligatorische menschliche Verifikation erforderlich bleibt. Frameworks wie CREOLA können das Risiko durch Abgleich von Protokoll und Transkription reduzieren, benötigen jedoch zusätzliche Validierung im präklinischen Setting.

Zudem ist die Geräuschrobustheit von ASR-Systemen weiterhin limitierend: Benchmarks zeigen unter realistischen Lärmbedingungen deutlich erhöhte mWER-Werte (Moser et al., 2025). Ansätze wie personalisierte Sprachverbesserung (z. B. pDenoiser) und Training auf domänenspezifisch verrauschten Daten sind vielversprechend, aber für den Routinebetrieb noch nicht ausreichend abgesichert.

Eine weitere Hürde sind Dialektale und soziolinguistische Variationen: Regionale Sprachmuster können die Erkennungsleistung messbar beeinflussen. Für den deutschen Rettungsdienst kann das robuste Basismodelle oder regionale Adaptationen erforderlich machen (Hofecker, 2024).

Eine mögliche Zertifizierung als Medizinprodukt kann ebenfalls erheblichen Aufwand bedeuten; je nach Einstufung sind Konformitätsbewertungen im Bereich von etwa 12-18 Monaten realistisch. Eine frühe Umsetzung von GMLP-Prinzipien kann spätere Nacharbeiten reduzieren, setzt aber regulatorische Expertise im Team voraus. Parallel ist das Umfeld dynamisch: Die EU-KI-Verordnung (EU) 2024/1689 und fortlaufend aktualisierte Leitlinien erfordern kontinuierliches Compliance-Monitoring und vorausschauende Anpassungsfähigkeit.

Für den Rettungsdienst ist Offline-Fähigkeit kritisch: In ländlichen Funklöchern, Tunneln oder bei überlasteten Netzen in Großschadenslagen muss das System zuverlässig funktionieren. Das spricht für lokale Verarbeitung – mit entsprechenden Anforderungen an Hardware, Energieverbrauch und Wartbarkeit auf mobilen Endgeräten.

## **7. Exkurs: KI als klinische Entscheidungsunterstützung**

Die in dieser Arbeit betrachtete Sprach-zu-Text- und Protokollpipeline ist nicht nur für die Dokumentation relevant, sondern kann auch als Grundlage für weiterführende KI-Anwendungen dienen. Sobald ein Einsatzdiktat zuverlässig transkribiert und in ein strukturiertes Format (z. B. DIVI-MIND) überführt wird, liegen die Informationen maschinenlesbar vor und könnten prinzipiell für klinische Entscheidungsunterstützung genutzt werden. Ob und in welchem Umfang das sinnvoll ist, erfordert jedoch eine separate wissenschaftliche und regulatorische Bewertung. Ein begleitender iOS-Prototyp (Anlage) demonstriert die technische Machbarkeit (Whisper, strukturierte Protokolle, regelbasierte Checks, Warnhinweise), ist aber nicht klinisch validiert und nicht für den Routineeinsatz gedacht.

Mögliche Anwendungsfelder sind beispielsweise die Markierung typischer Konstellationen (z. B. Brustschmerz + Dyspnoe + KHK → Verdacht ACS), der Abgleich von Allergien und Dauermedikation mit geplanten Notfallmedikamenten (z. B. ASS-Allergie trotz diktiertem ASS-Gabe) oder die Triage/Priorisierung (Unterstützung bei vielen Patienten oder knappen Ressourcen). In dieser Arbeit liegt der Fokus auf automatisierter Dokumentation, nicht auf Therapie- oder Diagnostikempfehlungen. Für das Clinical Decision Support (CDSS) würden die bekannten Herausforderungen noch gewichtiger: Halluzinationen und ASR-Fehler hätten unmittelbare klinische Folgen. Gleichzeitig bleiben die Datenschutzanforderungen gleich hoch und die regulatorischen Hürden wären deutlich strenger (u. a. im Kontext EU-KI-VO und MDR). Zukünftige Forschung sollte prospektive Validierungsstudien, ein belastungsarmes UI-Design, den regulatorischen Pfad sowie Haftungs- und Verantwortlichkeitsfragen systematisch untersuchen.

## **8. Fazit**

Die vorliegende Literaturstudie untersuchte das Potenzial KI-gestützter Spracherkennung für die automatisierte Erstellung von Notfallprotokollen. Die systematische Analyse von 20 Kernquellen führt zu folgenden Kernerkenntnissen:

Moderne ASR-Systeme bezüglich ihrer technologischen Reife erreichen unter kontrollierten Bedingungen akzeptable Erkennungsraten. Die mWER als neue Metrik zeigt jedoch, dass medizinische Fachbegriffe weiterhin problematisch sind. Domänenspezifisches Fine-

Tuning (Feinabstimmung) auf deutschsprachigen medizinischen Daten kann die Erkennungsleistung signifikant verbessern. Außerdem bestehen klinisch relevante Halluzinationsrisiken für LLM-generierte medizinische Texte, wie die CREOLA-Studie dokumentiert. Obligatorische menschliche Verifikation ist daher nicht optional. Des Weiteren sind einheitliche Regulatorien wichtig. Die GMLP-Prinzipien (IMDRF N88, 2025) und vordefinierte Änderungskontrollpläne bieten hier einen praktikablen Rahmen für die Entwicklung und Zertifizierung KI-basierter Medizinprodukte. Im Allgemeinen wird für die Anwendung im deutschen Rettungsdienst eine Lösung empfohlen, die lokale Verarbeitung mit domänenadaptiertem Whisper-Modell, DIVI-MIND-konformes Output-Format und CREOLA-basiertes Qualitätsmonitoring kombiniert. Die KI-gestützte Protokollierung im Rettungsdienst steht an der Schwelle zur klinischen Anwendung. Die vorliegenden Benchmark-Daten und Frameworks bilden einen Baustein zur evidenzbasierten Implementierung.

## Literaturverzeichnis

- Asgari, E., Kaur, J., Nuredini, G., Balloch, J., Taylor, A. M., Sebire, N., Robinson, R., Peters, C., Sridharan, S., & Pimenta, D. (2023). *Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review (Preprint)*. JMIR Medical Informatics. <https://doi.org/10.2196/preprints.55499>
- Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *Npj Digital Medicine*, 8(1), 274. <https://doi.org/10.1038/s41746-025-01670-7>
- Bieß, A. (2026, Januar 30). *MIND*. DIVI – Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin. <https://www.divi.de/service/https%3A%2F%2Fwww.divi.de%2Fservice%2Fmind>
- Frauenhofer IAIS. (o. J.). *AI agents to assist in saving lives in emergency rooms—Frauenhofer IAIS*. Abgerufen 30. Januar 2026, von [https://www.iais.fraunhofer.de/en/press-events/press-releases/AI\\_agents\\_in\\_emergency\\_rooms.html](https://www.iais.fraunhofer.de/en/press-events/press-releases/AI_agents_in_emergency_rooms.html)
- Hedderson, D., Courtney, K. L., Monkman, H., & Blanchard, I. E. (2025). Speech recognition technology in prehospital documentation: A scoping review. *International Journal of Medical Informatics*, 193, 105662. <https://doi.org/10.1016/j.ijmedinf.2024.105662>
- Hofecker, V. (2024). *Domain adaptation of whisper models to German medical automatic speech recognition*. <https://doi.org/10.25365/THESIS.76227>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (arXiv:2310.06825). arXiv. <https://doi.org/10.48550/arXiv.2310.06825>

- Luo, X., Zhou, L., Adelgais, K., & Zhang, Z. (2024). Assessing the Effectiveness of Automatic Speech Recognition Technology in Emergency Medicine Settings: A Comparative Study of Four AI-powered Engines. *Research Square*, rs.3.rs-4727659. <https://doi.org/10.21203/rs.3.rs-4727659/v1>
- Menzies, D., Kirwan, S., & Albarqawi, A. (2024). *AI Managed Emergency Documentation with a Pretrained Model* (arXiv:2408.09193). arXiv. <https://doi.org/10.48550/arXiv.2408.09193>
- Moser, D., Stanic, N., & Sariyar, M. (2025). Benchmarking speech-to-text robustness in noisy emergency medical dialogues: An evaluation of models under realistic acoustic conditions. *JAMIA Open*, 8(6), ooaf147. <https://doi.org/10.1093/jamiaopen/ooaf147>
- Paez, A. (2017). Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, 10(3), 233–240. <https://doi.org/10.1111/jebm.12266>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (arXiv:2212.04356). arXiv. <https://doi.org/10.48550/arXiv.2212.04356>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. <https://doi.org/10.48550/arXiv.2307.09288>
- Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (Text von Bedeutung für den EWR), 119 OJ L (2016). <http://data.europa.eu/eli/reg/2016/679/oj>

Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte, zur Änderung der Richtlinie 2001/83/EG, der Verordnung (EG) Nr. 178/2002 und der Verordnung (EG) Nr. 1223/2009 und zur Aufhebung der Richtlinien 90/385/EWG und 93/42/EWG des Rates (Text von Bedeutung für den EWR. ), 117 OJ L (2017). <http://data.europa.eu/eli/reg/2017/745/oj>

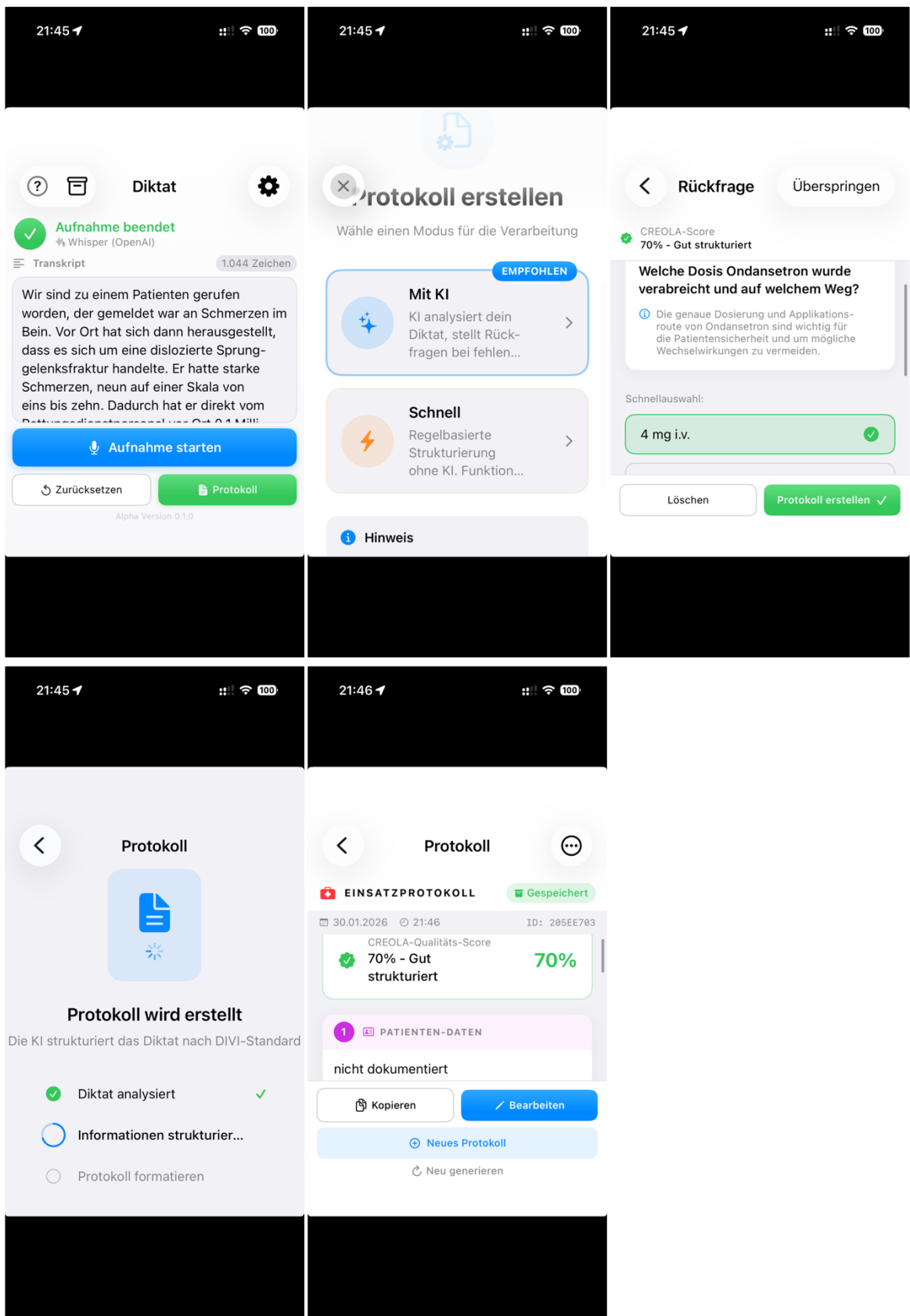
Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz) (Text von Bedeutung für den EWR) (2024). <http://data.europa.eu/eli/reg/2024/1689/oj>

Wg, A. (2025). *Good machine learning practice for medical device development: Guiding principles*.

Zhang, Z., Tian, Y., Zhou, T., Zhao, Y., Li, Jinpeng, Zhang, J., & Li, Jingsong. (2025). pDenoiser: A Personalized Speech Enhancement Neural Network for Pre-Hospital Emergency Medical Services. *IEEE Journal of Biomedical and Health Informatics*, 29(12), 9198–9209. <https://doi.org/10.1109/JBHI.2024.3395832>

Anlage

Beispielsbilder aus dem IOS-App Prototypen



### **Eigenständigkeitserklärung**

Ich erkläre hiermit, dass ich die vorliegende Prüfungs- und Studienarbeit (PStA) mit dem Titel „KI-gestützte Spracherkennung für automatisierte Notfallprotokolle – Eine Literaturstudie zu Notfall Information and Communication Technology-Workflows in der präklinischen Versorgung“ selbstständig und ohne fremde Hilfe verfasst habe. Alle verwendeten Quellen und Hilfsmittel sind vollständig angegeben und gekennzeichnet. Ich bestätige, dass ich alle in dieser Arbeit verwendeten Literaturquellen, Daten, Tabellen, Abbildungen und sonstigen Inhalte kenntlich gemacht habe und dass keine Teile dieser Arbeit wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, ohne dies durch entsprechende Zitierung zu kennzeichnen. Ich bestätige weiterhin, dass ich KI-Assistenzsysteme (Claude AI) zur Unterstützung bei folgenden Aufgaben genutzt habe: Datenbankauswahl, Abstract-Filterung und -analyse, Strukturierung von Inhalten, Verbesserung und Korrektur der deutschen Sprache sowie Formatierung von Tabellen und Quellenangaben. Alle mit KI-Unterstützung erstellten oder überarbeiteten Inhalte wurden von mir inhaltlich überprüft, validiert und beantwortet. Diese Arbeit wurde in der vorliegenden Form oder einer ähnlichen Form noch nicht als Prüfungsleistung eingereicht. Mir ist bewusst, dass eine falsche Erklärung rechtliche Konsequenzen haben kann.

Augsburg, den 30. Januar 2026



Julian Linden