Fachdidaktische Qualitätssicherung von Machine-Learning-Modellen in der Hochschuldidaktik – Individuelles KI-gestütztes Feedback im Videoanalyse-Tool ViviAn

TIM LUTZ, INNSBRUCK; MARC BASTIAN RIEGER, KARLSRUHE & JÜRGEN ROTH, LANDAU

***Zusammenfassung**:*
*Dieser Beitrag befasst sich im Rahmen einer Machbarkeitsstudie mit Forschungsfragen, die klären, (1) ob KI-Modelle, die auf Basis von Nutzenden-Selbst-Label-Ansätzen trainiert wurden, eine Klassifikationsqualität erreichen, die mit der von geschulten Mathematikdidaktiker*innen vergleichbar ist und (2) ob diese als Teil eines individuellen Feedbacksystems im webbasierten Videoanalyse-Tool ViviAn genutzt werden können.*
Der Mixed-Methods-Ansatz umfasst eine qualitative Inhaltsanalyse der Trainingsdaten durch Experten zur Erstellung eines validierten Trainingssets. Anschließend erfolgt eine quantitative Modellierung eines Klassifikationssystems durch supervised-learning, dessen Leistung mit statistischer Interrater-Reliabilität durch fachdidaktische Experten evaluiert wird. Die Analysen der vorliegenden Machbarkeitsstudie zeigen, dass dieses Vorgehen ausreichen kann, um KI-Modelle zu erstellen, die fachdidaktischen Experten-Ratings nahekommen.

***Abstract:** In the context of a feasibility study, this paper addresses research questions that clarify (1) whether AI models trained on the basis of user self-labeling approaches achieve a classification quality comparable to that of trained professionals in mathematics education and (2) whether they can be used as part of an individual feedback system in the web-based video analysis tool ViviAn.*
The mixed-methods approach includes a qualitative content analysis of the training data by experts to create a validated training set. This is followed by quantitative modeling of a classification system using supervised learning, the performance of which is evaluated by experts in mathematics education using statistical interrater reliability. The analyses of the present feasibility study show that this procedure is sufficient to create AI models that come close to ratings of experts in mathematics education.

1. Digitale Lern-/Trainings-Systeme und (individuelles) Feedback

Nutzerinnen und Nutzer arbeiten in digitalen Lern- bzw. Trainingssystemen in der Regel selbstständig an vom System dargebotenen Aufgaben und werden dabei nicht von menschlichen Trainern begleitet. Vor diesem Hintergrund stellt sich die Gabe von Feedback[1] als einer der größten Einflussfaktoren für erfolgreiches und nachhaltiges Lernen heraus (Hattie, 2015). Somit ist Feedbackgabe ein besonders wesentliches Instrument zur Steuerung des Lernens in derartigen Systemen. Feedback kann kognitive, metakognitive und motivationale Funktionen haben (Buttler & Winne, 1995; Narciss, 2006). Bei der Umsetzung digitaler Lern- bzw. Trainingsumgebungen, in denen mit individuellen Nutzerinnen und Nutzern gearbeitet werden soll, stellt sich für die Entwicklerteams diesbezüglich eine Reihe von Fragen: Sollen die Nutzerinnen und Nutzer Feedback zu ihren Aufgabenbearbeitungen erhalten? Wann sollen Nutzerinnen und Nutzer das Feedback ggf. erhalten? Auf welche Informationsarten soll das Feedback sich konzentrieren? (Hattie, 2015). Auch wenn die empirische Forschung hierzu noch uneindeutige und meist sehr inhaltsspezifische Aussagen liefert (Wisniewski et al., 2020) folgt der hier beschriebene Ansatz dem Wunsch nach unmittelbarem Feedback, wie es beispielsweise in fachmathematischen Kontexten die Software STACK ermöglicht. Dabei soll, wie bei STACK, die Unmittelbarkeit durch automatisierte Auswertung gewährleistet werden.

Feedback erfolgt in Reaktion auf einen Input (Narciss, 2006). Die Art des Inputs bestimmt maßgeblich, wie einfach es ist, eine automatisierte Auswertung umzusetzen. Fachdidaktische Kontexte erfordern häufig offene Beantwortungen, bei denen die Nutzerinnen und Nutzer ihre Antworten als frei formulierte Texte in ein digitales Tool eingeben. Der Einsatz von KI stellt hier eine Möglichkeit dar, automatisiert etwa Textkategorisierungen zu generieren. Der vorliegende Beitrag unternimmt eine fachdidaktische Qualitätssicherung von auf eigenen Daten trainierten und verwendeten KI-Modellen für die Hochschullehre, um den technischen Werten der Modellgüte auch inhaltlich fachdidaktische Abwägungen zur Seite zu stellen.

2. Das Videoanalyse-Tool ViviAn und KI

2.1 Was ist ViviAn und welche Zielsetzung verfolgt die vorliegende KI-Studie?

Das Videoanalyse-Tool ViviAn (das Akronym ViviAn steht für **Vi**deo**vi**gnetten zur **An**alyse von Unter-

richtsprozessen) wurde entwickelt, um diagnostische Fähigkeiten von (angehenden) Lehrkräften zu fördern (Bartel & Roth, 2017; https://vivian-training.de/). Die Wirksamkeit dieser Förderung durch ViviAn konnte in mehreren Studien nachgewiesen werden (Enenkiel et al., 2022; Hofmann & Roth, 2020; Bartel & Roth, 2020; Walz & Roth, 2019). In ViviAn werden kurze Videovignetten angeboten, die aus videografierten Schüler-Gruppenarbeiten im Mathematik-Labor „Mathe ist mehr" (https://mathe-labor.de) zugeschnitten und mit Kontextinformationen angereichert wurden, die Lehrkräften in einer entsprechenden Unterrichtssituation auch zur Verfügung stehen. (Angehende) Lehrkräfte erhalten in ViviAn Diagnoseaufträge, die diagnostische Fähigkeiten adressieren, operationalisiert durch die Teilkomponenten Beschreiben, Deuten, Ursachen finden und Konsequenzen ableiten (vgl. Bartel et al. 2018). Die (angehenden) Lehrkräfte beantworten die Diagnoseaufträge entweder in Freitextfeldern oder kreuzen zunächst eine Multiple-Choice-Auswahl an, die sie anschließend in einem Freitextfeld begründen. Nach der Beantwortung aller Diagnoseaufträge zu einer ViviAn-Vignette werden jedem Nutzer bzw. jeder Nutzerin die eigenen Antworten und direkt daneben validierte Expertenurteile dazu als Musterlösungen ausgegeben, sodass ein direkter Abgleich und damit ein Lerneffekt möglich ist (Bartel & Roth, 2020). Dieses Feedbacksystem unterstützt die Nutzerinnen und Nutzer dabei, ihre kognitiven und affektiv-motivationalen Fähigkeiten gezielt weiterzuentwickeln (Enenkiel et al., 2022). Die gegebenen Antworten können als anonymisiertes Trainingsmaterial für auf spezifische Fragen zugeschnittene KI-Modelle verwendet werden. Es entstehen so gelabelte Trainingsdaten für spezialisierte Analyse-Modelle, die eingesetzt werden können, um gegebene Antworten unmittelbar zu analysieren und Feedback generieren zu können.

ViviAn-Nutzende wünschen sich neben den validierten Expertenurteilen auch individuelle Rückmeldungen zu ihren getätigten Diagnosen (Scherb et al., 2023), die von Dozierenden in fachdidaktischen Großveranstaltungen mit über 100 Nutzerinnen und Nutzern nicht geleistet werden können. Aufgrund der aktuellen technischen Entwicklung und vor dem Hintergrund, dass Feedback ein wichtiger Erfolgsfaktor für Lernprozesse ist (Poulos & Mahony, 2008), stellt sich die Frage, inwiefern es möglich ist, auf Basis des Maschine-Learning (ML) den ViviAn-Nutzenden automatisiert ein passgenaues individuelles Feedback zu ihren Freitextantworten anzubieten. Dazu ist zunächst erforderlich in der hier vorgestellten Machbarkeitsstudie den Aspekt der Analyse von Freitextantworten, die vor der Ausgabe von konkretem Feedback in automatisierten Feedbacksystemen stattfindet, zu untersuchen und eine Möglichkeit zum Training von robusten, qualitätsgesicherten ML-Modellen für kleinere Datensätze (ca. 1.000) aufzuzeigen.

2.2 Umsetzungsaspekte von KI-Feedback

Die Entwicklung von ViviAn ist eng mit praxisorientierten Ausbildungsansätzen verknüpft. Die Plattform wurde mehrfach evaluiert, um sowohl ihre Usability als auch ihre Effektivität zu optimieren (Scherb et al., 2023). Ein besonderes Merkmal ist der Einsatz eines gestuften Feedbacksystems auf Basis des Response-to-Intervention (RTI)-Modells (Rieger & Roth, 2025). Dieses Modell besteht aus drei Stufen: Lernfortschritte überprüfen, inhaltliche Defizite erkennen und individuell fördern. Diese drei Stufen greifen im Hochschulkontext ständig ineinander, sind aber von einer einzigen Lehrperson nur schwer auf individueller Ebene leistbar. Eine KI-Integration in ViviAn bietet die Möglichkeit einer automatisierten Unterstützung auf allen drei Ebenen. Damit ist eine systematische Analyse der Datenmengen der Nutzerinnen und Nutzer möglich und erlaubt die Generierung von präzisen Rückmeldungen zu spezifischen Stärken und Schwächen (Rieger & Roth, 2025). KI-basiertes Feedback wird bereits als ebenso effektiv wie manuelles Feedback angesehen, allerdings stellt die Frage, wie dieses in Lehr-Lern-Kontexten eingesetzt werden soll bzw. kann nach wie vor eine große Hürde dar (Cavalcanti et al., 2021).

Das UNESCO AI competency framework for teachers (Miao & Cukurova, 2024) betont die Bedeutung der Nutzung von KI, um Lernprozesse gezielt zu unterstützen, ohne die Rolle der Lehrkraft zu ersetzen. Diese Prinzipien spiegeln sich in ViviAn wider, sodass nicht nur diagnostische Kompetenzen gefördert, sondern Nutzerinnen und Nutzer auch dazu befähigt werden, aktiv an ihrem eigenen Lernprozess mitzuwirken. So wird durch KI-Feedback in der Hochschullehre eine schnelle, personalisierte und kontinuierliche Rückmeldung ermöglicht, die Nutzerinnen und Nutzern dabei hilft, individuelle Wissenslücken frühzeitig zu identifizieren (Lindsay et al., 2023). Dies ist ein entscheidender Vorteil, insbesondere in großformatigen Lehrveranstaltungen, in denen personalisiertes Feedback ohne KI nur schwer umsetzbar wäre. Durch den Einsatz generativer KI-Modelle können zeitintensive manuelle Bewertungsprozesse au-

tomatisiert und Lehrkräfte dadurch signifikant entlastet werden (Lee & Moore, 2024). Aktuelle Untersuchungen belegen, dass KI-gestütztes Feedback positive Effekte auf die Lernmotivation und die individuelle Leistungsentwicklung hat, da es flexibel an die spezifischen Bedürfnisse der Nutzerinnen und Nutzer angepasst werden kann (Meyer et al., 2024). Insgesamt erweist sich die Integration von KI-gestütztem Feedback als wichtig, um den Herausforderungen der modernen Hochschullehre gerecht zu werden und die Qualität der Lehr- und Lernprozesse nachhaltig zu verbessern, aber auch als Gefahr, wenn Nutzerinnen und Nutzer im Umgang mit KI-generiertem Feedback nicht ausreichend geschult sind (Zhan et al., 2025). Ein großer Mangel besteht an Feedbacksystemen, die nicht nur Fehler identifizieren, sondern darauf ausgerichtet sind, Defizite zu beheben (Keuning et al., 2019). Die damit zusammenhängende Analyse von Freitextantworten und die Generierung von differenziertem Feedback stellt auch heute noch eine große Herausforderung dar, ebenso wie die automatisierte Evaluation komplexer Texte (Bernius et al., 2022). Es bedarf holistischer Systeme, die in der Lage sind, feinste Nuancen in Antwortvariationen herauszulesen und Deutungen richtig zu interpretieren. Dies ist konzeptionell komplex und technisch herausfordernd (Bai & Stede, 2022). Mit dem Aufkommen des Natural Language Processing (NLP) und großen Large Language Models (LLMs) wie ChatGPT existiert ein Trend zum Einbinden solcher Modelle, der allerdings die Risiken der sogenannten Halluzinationen (falsch Informationen) in sich birgt (Lindsay et al., 2023).

Um dies zu vermeiden und trotzdem NLP gewinnbringend zu nutzen, muss sichergestellt werden, dass die automatisierte Analyse der Antworten qualitätsgesichert stattfindet. Qualitätsgesichert meint in diesem Kontext, dass die Ausgaben der Modelle im Vorfeld fachdidaktisch geprüft werden. Im Hinblick auf den theoretischen Rahmen von ViviAn, in dem Expertenurteile bereits bei der Erstellung der verwendeten Vignetten besondere Bedeutung haben, scheint es sinnvoll, auch für die Maßnahmen der Qualitätssicherung Expertenurteile einzubinden. Der im Rahmen dieser Machbarkeitsstudie vorgestellte Ansatz sieht vor, dass die Ausgaben der automatisierten Auswertung mit KI zu einem sehr hohen Prozentsatz denen von Experten gleichen und Abweichungen untersucht werden. So kann replizierbares und nachvollziehbares Feedback den Nutzerinnen und Nutzern zur Verfügung gestellt werden. Bei der fachdidaktischen Prüfung der Modellausgaben werden Abweichungen von den Expertenratings berichtet. Dies geht bei kleinen spezialisierten Modellen, die natürliche Sprache verarbeiten, aber nicht produzieren, sehr effektiv. Deshalb werden für Textanalyseaufgaben, beispielsweise im Kontext der Retrieval-Augmented Generation, auch deutlich kleinere Modelle eingesetzt, die auf Analyse spezialisiert sind und entsprechend trainiert werden. Das Training von Experten zur Resourcenschonung wird neuerdings auch bei der Textgenerierung in der Idee Mixture of Experts (etwa Mixtral 8x7b, Qwen3-30B-A3B) wieder aufgegriffen, wobei für die dazu notwendigen Textanalysen ebenfalls auf kleine Modelle zurückgegriffen wird. In diesem Artikel wird zur Verbindung der Expertensysteme das Konzept der Cascading AI (CAI) genutzt werden, bei welchem individuelle Modelle als Kaskade genutzt werden, um einzelne spezialisierte Aspekte zu analysieren. Dies eignet sich besonders im Bereich des NLP, erhöht die Genauigkeit und verbessert die Berechnungseffizienz (Varshney & Baral, 2022). Die Nutzung von Modell-Kaskaden ist hilfreich, um exakte und transparentere Analysen mit ML-Modellen durchzuführen, wie bereits in mehreren Disziplinen gezeigt wurde (z. B. Mozannar et al., 2024, Mehralivand et al., 2022). Die in dieser Studie erforschten Modelle werden in weiteren Vorhaben für konkrete Feedbackgenerierung als Teil einer solchen Modell-Kaskade in ViviAn eingesetzt und das Evaluierungskonzept für weitere Use-Cases zur Modellgenerierung adaptiert. Am Ende einer solchen Kaskade können Analyseergebnisse durch ein LLM individuell verschriftlicht werden.

Beim Design einzelner Modelle oder ganzer Modell-Kaskaden für den Bildungsbereich sollte einem Human-Centered Design (HCD) gefolgt werden, denn es geht darum, den Bedürfnissen der Endnutzenden in der Praxis gerecht zu werden (Giacomin, 2014). Dieser Ansatz fordert, dass die Gestaltung und Erstellung iterative Prozesse sind, die ein Verständnis sowohl des Einsatzszenarios als auch der Bedürfnisse der Endnutzenden voraussetzen. Dies beinhaltet folglich auch die aktive Einbeziehung von Experten zur Entwicklung und Bewertung des Systems (z. B. Bødker et al., 2022). Die Experten für diese Studie bilden Mathematikdidaktiker, die aktiv in Lehre und Forschung tätig sind und somit sowohl das Interesse der Endnutzenden aus erster Hand kennen, als auch die qualitative Güte der Ausgaben bewerten können. Der HCD-Ansatz umfasst fünf Phasen: (1) Planung und Definition, (2) Erkundung und Design, (3) Konzept und Prototypen, (4) Auswertung und Verfeinerung sowie (5) Einführung und Monitoring (Hanington & Martin, 2019). Er startet also mit dem

Verstehen und Eingrenzen des Problems, um danach Lösungen und Prototypen zu konkretisieren und Annahmen testbar zu machen. Anschließend werden passende Lösungen evaluiert, iterativ verbessert und verfügbar gemacht. Durch ein begleitendes Monitoring kann weiter optimiert werden.

In der vorliegenden Machbarkeitsstudie werden die Phasen 1 bis 4 bearbeitet. Phase 5 ist aufgrund der Implementationsstrategie in eine Modell-Kaskade noch nicht möglich. Die häufig fehlende Einbeziehung von Fachexperten in die Überprüfung der Modellausgaben ist ein elementarer Mangel, der in schon bestehenden Systemen zu Problemen führt (Khosravi et al., 2022). Alfredo et al. (2024) zeigen einen klaren Mangel an HCD-Ansätzen in den Phasen 1 und 5 auf. Den ebenso großen Mangel an Experten versuchen manche Systeme mit Empfehlungssystemen zu überbrücken, die kritische Modellausgaben markieren, die einer Expertenbewertung bedürfen (Khosravi et al., 2022). Dies ist insofern kritisch zu betrachten, da die Rückmeldungen nur punktuell von Experten validiert werden, was eine große Menge an falschen Rückmeldungen ermöglichen kann, die das System nicht zuverlässig erkennt. Um solchen Problematiken vorzubeugen, ist die Evaluation der Klassifikationen der Modelle mit einer Interrater-Reliabilität zwischen ViviAn-KI und Experten sowie einer inhaltlichen Analyse der Abweichungen schon während des Designprozesses elementar. Khosravi et al. (2022) weisen deutlich darauf hin, dass die Erklärbarkeit der KI-Modelle im Bildungsbereich erweiterte Ansprüche stellt, die über eine Blackbox-Ausgabe von technischen Werten hinausgeht und dabei die authentische Bewertung eine zentrale Rolle spielt.

Vor diesem Hintergrund werden die Prinzipien der Explainable AI und des HCD-Ansatzes so kombiniert, dass die Erklärbarkeit nicht als nachgelagerte Modelldeutung (z. B. Feature-Attribution) verstanden wird, sondern als nutzungsnahe, aufgaben- und adressatengerechte Aufbereitung von Modellausgaben innerhalb realer Lern- und Feedbackprozesse. Die von uns geplante Kaskadenarchitektur übernimmt zwar das technische Ordnungsprinzip gestufter Modelle, als Entscheidungskriterien für Modellausgaben werden jedoch konsequent didaktische Kategorien genutzt (z. B. Validität im Nutzungskontext, Transparenz für Fachentscheidungen, anschlussfähige Begründungen). Im Unterschied zu technisch orientierten Kaskaden, etwa zur Detektionsleistung in der medizinischen Bildgebung (Mehralivand et al., 2022) oder der Verbesserung von Effizienz und Genauigkeit in NLP-Systemen (Varshney & Baral, 2022), steht hier nicht die Optimierung von Rechenaufwand oder Score-Metriken im Vordergrund, sondern die inhaltlich fundierte Prüfbarkeit und Nachvollziehbarkeit der Ausgaben durch Fachexpert:innen entlang des Designprozesses. In dieser Studie werden zu diesem Zweck Mathematikdidaktiker zur Bewertung der Modellausgaben hinzugezogen, um die Ausgaben inhaltlich zu prüfen und zu erklären.

2.3 KI-Integration in ViviAn

Ein zentrales Element von ViviAn ist die Generierung und Nutzung von Trainingsdaten aus authentischen Nutzerinteraktionen. Dieses Vorgehen stellt sicher, dass die zugrunde liegenden Machine-Learning-Modelle kontinuierlich an die realen Bedürfnisse der Nutzerinnen und Nutzer angepasst werden können. Die Nutzung echter Nutzerdaten ist entscheidend, da generische Modelle häufig nicht die spezifischen Anforderungen und Kontexte von Lernumgebungen widerspiegeln. Studien, wie die von Zhai et al. (2021) und Lee und Moore (2024), betonen die Notwendigkeit KI-Modelle mit domänenspezifischen Daten zu trainieren, um eine hohe Relevanz und Präzision sicherzustellen. In ViviAn ermöglicht diese Datenbasis nicht nur eine Verbesserung der Feedbackqualität, sondern auch eine skalierbare Anpassung der Plattform an neue didaktische Herausforderungen.

Besonders in der Lehrkräftebildung, wo die Entwicklung diagnostischer Fähigkeiten zentral ist, erweist sich die Nutzung eigener Daten als unverzichtbar. Durch die Analyse der Antworten und Interaktionen der Nutzerinnen und Nutzer kann in ViviAn ein datengestütztes Lernumfeld geschaffen werden, das über klassische Lehrmethoden hinausgeht. Diese Verknüpfung zwischen Technologie und Lernprozess wird durch den gezielten Einsatz von NLP verstärkt. In Echtzeit werden Freitextantworten analysiert und auf dieser Basis weitere Analysen und Entscheidungen, wie z. B. die Weitergabe von Modellergebnissen an weitere spezialisierte Modelle mit CAI, ermöglicht. So können personalisierte Lernpfade unterstützt werden.

Darüber hinaus ermöglicht der Einsatz eigener Daten nicht nur eine Optimierung der Modelle, sondern adressiert auch ethische und datenschutzrechtliche Anforderungen. Die Nutzung lokaler, kontrollierter Datenquellen minimiert das Risiko algorithmischer Verzerrungen und stellt sicher, dass die Plattform den Datenschutzanforderungen entspricht (Miao & Cukurova, 2024). Damit kann ViviAn zu einem Modell für einen verantwortungsvollen

und effektiven Einsatz von KI in Bildungsprozessen werden.

Durch die Möglichkeit der kontinuierlichen Verbesserung der KI-Modelle mit realen Nutzerdaten erfüllt ViviAn die Anforderungen an eine skalierbare, flexible und kontextangepasste Lernumgebung. Die Generierung eigener Modelle mit spezifischen Daten ist für den Bildungskontext nicht nur relevant, sondern notwendig. Denn im Gegensatz zu generischen KI-Ansätzen erlauben es solche eigenen Modelle, individuelle Lernprozesse in einer situationsangemessenen Tiefe und Qualität zu unterstützen.

3. Forschungsfragen

Im Rahmen der Machbarkeitsstudie, die untersucht, inwiefern es auf Basis des Maschine-Learning (ML) möglich ist, die Freitextantworten der Nutzerinnen und Nutzer didaktisch qualitätsgesichert zu klassifizieren, werden folgende Forschungsfragen (FF) beantwortet:

Forschungsfrage 1
Wie gut ermöglicht die bestehende Datenbasis von ca. 1.000 Bearbeitungen das Training robuster Machine-Learning-Modelle hinsichtlich der quantitativen Genauigkeit, um unbekannte Freitextantworten automatisch in diagnostisch relevante Kategorien einzuordnen?

Forschungsfrage 2
Wie hoch ist die Übereinstimmung zwischen den von der ViviAn-KI ermittelten Kategorien und den Zuordnungen diagnostisch geschulter Mathematikdidaktiker*innen?

Forschungsfrage 3
Welche systematischen Unterschiede bestehen zwischen den von der ViviAn-KI und den von diagnostisch geschulten Mathematikdidaktiker*innen erstellten Zuordnungen von Nutzendenantworten zu fachdidaktischen Kategorien?

4. Methodik

Um die Forschungsfragen im Rahmen der Machbarkeitsstudie sinnvoll bearbeiten zu können, wurde eine konkrete ViviAn-Vignette und zwei miteinander verbundene Diagnoseaufträge zu dieser Video-Vignette – im Folgenden Items genannt – genutzt, bei denen es sich um ein Multiple-Choice-Item und ein Freitext-Item handelt. Die Nutzerinnen und Nutzer dieser Studie, die die Vignette bearbeitet haben, sind Studierende der mathematikdidaktischen Lehrveranstaltungen.

4.1 Beschreibung der ViviAn-Vignette

Die vorliegende Studie bezieht sich auf eine ViviAn-Vignette, die im Vignetten-Bereich *Flächen- und Rauminhalte* verortet ist. In diesem Vignettenbereich geht es inhaltlich um Vergleichs- und Messstrategien sowie die Kernideen des Messens (vgl. Enenkiel et al., 2022). Das Video zur oben genannten Vignette zeigt innerhalb von insgesamt zwei Minuten und 35 Sekunden zwei Sequenzen, in denen vier männliche Schüler der Jahrgangsstufe 11 eines Gymnasiums in Gruppenarbeit versuchen, den Flächeninhalt der USA im Jahr 1913 zu bestimmen. Als Materialien stehen ihnen eine maßstabsgetreue zeitgenössischen Karte der USA sowie große blaue und kleine rote Quadrate zur Verfügung, deren maßstäbliche Kantenlängen anhand des auf der Karte abgedruckten Maßstabs als 500 km (blau) bzw. 200 km (rot) identifiziert werden können. Im Video wird die Gruppe bei der Bearbeitung der folgenden beiden Arbeitsaufträge gezeigt, auf die im Arbeitsheft, von dem jeder Schüler ein eigenes Exemplar vor sich liegen hat, jeweils ein leerer Kasten folgt, in die die Schüler individuell ihre Vorgehensweisen und Ergebnisse eintragen sollen:

1.2 Jetzt wollen wir es genauer wissen:
Können Sie die Fläche der USA von 1913 genauer bestimmen?
Neben den blauen Quadraten können Ihnen auch die roten Quadrate hilfreich sein.

1.4 Überlegen Sie gemeinsam:
Was müssen Sie tun, um die Fläche immer exakter eingrenzen zu können? Notieren Sie Ihre Lösungsvorschläge.

Die Vignette umfasst neben dem Video eine Reihe weiterer Materialien (unter anderem Schülerdokumente, Arbeitsaufträge und ein Foto der den Schülern zur Verfügung stehenden Materialien), die im Diagnoseprozess jederzeit abgerufen werden können.

4.2 Betrachtete Diagnoseaufträge zur Vignette

Zu dem Video der Gruppenarbeitssituation erhalten Nutzerinnen und Nutzer, die mit Hilfe der Lehr-Lern-Plattform ViviAn ihre Diagnosekompetenz trainieren, insgesamt sieben Diagnoseaufträge. Für die vorliegende Machbarkeitsstudie wurde der vierte Diagnoseauftrag (Aufgabe 4) ausgewählt. Er besteht aus zwei Teilfragen bezogen auf die „Kernideen des Messens“, die die Schüler in der videografierten Arbeitsphase nutzen. Bei der ersten Teilaufgabe handelt es sich um ein Multiple-Choice-Item, die zweite

ist ein Freitext-Item, in dem die getroffene Auswahl begründet werden soll.

4.1 Welche Aspekte der „Kernidee des Messens“ werden von den Schülern zum Lösen von Aufgabe 1.2 genutzt?
- ☐ Einheit nutzen
- ☐ Mit der Einheit auslegen
- ☐ Einheiten zählen
- ☐ Gegebenenfalls Einheit verfeinern
- ☐ Keine der oben genannten

4.2 Begründen Sie ihre Aussage unter 4.1.

In der bisherigen Umsetzung in ViviAn können die Nutzerinnen und Nutzer – nach Beantwortung aller Diagnoseaufträge zu einer Vignette – ihre eigenen Antworten mit Expertenantworten vergleichen, die in einem aufwändigen Prozess generiert wurden (vgl. Enenkiel et al., 2022). Die Expertenantworten zu den Teilaufgaben 4.1 und 4.2 sind im Folgenden abgedruckt und geben auch einen kleinen Einblick in die videografierte Situation:

Experten haben folgende Antworten gegeben:
- ☒ Einheit nutzen
- ☒ Mit der Einheit auslegen
- ☒ Einheiten zählen

Einheit nutzen/ Mit der Einheit auslegen: Die Schüler nutzen zunächst die blauen Quadrate als vorgegebene Einheiten, um die Fläche der USA auszulegen. Anschließend tauschen die Schüler teilweise die blauen Quadrate gegen kleinere rote Quadrate aus, um die Fläche (am Rand der USA) genauer auszulegen. Da die Schüler die blauen und roten Einheiten nur zum Auslegen verwenden und nicht direkt zueinander in Beziehung setzen (da dies auch aufgrund der Größe nur bedingt möglich ist), ist es in dieser Videosequenz fraglich, ob sie den Aspekt „Gegebenenfalls Einheiten verfeinern“ nutzen. Das Verfeinern der Einheit ist zwar im Ansatz vorhanden, wird aber nicht vollendet.

Einheit zählen: Nach dem Auslegen mit den Einheiten zählen sie jeweils die roten und blauen Quadrate und addieren (vermutlich nur in den Schülerdokumenten zu sehen) die Teilergebnisse zu einem Gesamtergebnis zusammen.

4.3 Qualitätskontrolle der ViviAn-KI-Modelle

In der vorliegenden Studie wird insbesondere im Rahmen der *Forschungsfragen 2 und 3* untersucht, ob es mit Hilfe von Maschine-Learning-Modellen möglich ist, in begrenztem Rahmen fachdidaktische Entscheidungen zu treffen. Für einen ethisch vertretbaren Einsatz solcher Modelle ist es zwingend erforderlich, eine fachdidaktische Qualitätssicherung durchzuführen (Lutz, 2022). Dies kann beispielsweise auf der Basis von qualitativen Inhaltsanalysen mithilfe stoffdidaktischer Expertise zum adressierten Inhaltsbereich gelingen. Dies sieht im hier vorgeschlagenen Vorgehen unter anderem eine Prüfung der Passung zwischen Selbst-Label und Antwort vor, wie auch eine Interrater-Reliabilitätsprüfung zwischen Experten- und Modellzuweisung. Für die qualitative Inhaltsanalyse wurden nach dem Rating der Antworten durch die Experten und durch das Modell alle Fälle individuell analysiert, bei denen Diskrepanzen bestehen zwischen der Bewertung der einzelnen Experten bzw. der vom Modell vorgenommenen Einschätzung. Es wurden sowohl orthografische, semantische als auch fachdidaktische Analysen durchgeführt, um diese Abweichungen eindeutig zu erklären. Im Sinne der internationalen Explainable AI (XAI)-Bestrebungen wurden für Analyse und Evaluation der Modelle verschiedene Werkzeuge entwickelt, um diese Analyse transparent, nachvollziehbar und reproduzierbar zu gestalten.

Eintrag 3: Die Schüler legen mit der Einheit "blaues Quadrat" und "rotes Quadrat" die Fläche der USA aus. Später zählen sie die Quadrate , um dann berechnen zu können wieviele die einzelnen gesamt ergeben.

User	Nutzen	Auslegen	Zählen	Verfeinern
Rater 1	Nicht vorhanden	Vorhanden	Vorhanden	Nicht vorhanden
Rater 2	Nicht vorhanden	Vorhanden	Vorhanden	Nicht vorhanden
Modelle	Vorhanden: 0.609 Nicht: 0.391 Confidence: 0.609 Class: vorhanden	Vorhanden: 1 Nicht: 0 Confidence: 1 Class: vorhanden	Vorhanden: 1 Nicht: 0 Confidence: 1 Class: vorhanden	Vorhanden: 0.001 Nicht: 0.999 Confidence: 0.999 Class: nichtvorhanden

Abb. 1: Custom-Tool für Transparenz und Nachvollziehbarkeit von Modellergebnissen der ViviAn-KI.

Abb. 1. zeigt ein Beispiel einer Abweichung zwischen Ratern und Modell. Die Abweichung ergibt sich dadurch, dass die Nutzung der Einheiten „blaues Quadrat“ und „rotes Quadrat“ im eingegebenen Text der Studentin bzw. des Studenten (fett gesetzt der Text in Abb. 1) nicht explizit genannt wird. Aus diesem Grund haben die beiden menschlichen Rater die Kernidee „Einheit nutzen“ als nicht angegeben und damit nicht vorhanden eingestuft. Das KI-Modell hat dagegen mit einem Vertrauen von 60,9 % das Vorhandensein dieser Kernidee im Text als gegeben bewertet. Diese Beispiele sind für den Entwicklungsprozess dokumentiert und können bei einem Einsatz der Modelle in der Lehre mit beigefügten Erklärungen veröffentlicht werden, sodass Nutzerinnen und Nutzer die Ergebnisse nachvollziehen können.

4.4 Güte der ViviAn-KI-Modelle

Um die Güte der entwickelten ViviAn-KI-Modelle zu testen, wurde der Testdatensatz genutzt, der während des Trainings weder der ViviAn-KI gezeigt

wurde, noch Einfluss auf die Modellauswahl hatte. Konkret wurden die 126 Antworten des Testdatensatzes zu der Freitextfrage 4.2 (vgl. Abschnitt 4.2) verwendet. Diese wurden ohne Kenntnis der Antworten der Nutzerinnen und Nutzer zu der Multiple-Choice-Frage 4.1 (vgl. Abschnitt 4.2) einerseits unabhängig von zwei Mathematikdidaktikern (nämlich dem ersten und dritten Autor dieses Beitrags) als auch durch die ViviAn-KI-Modelle geratet. Ziel des Ratings war es, dem jeweiligen Text der Nutzerinnen und Nutzer diejenigen der vier vorgegebenen Kategorien „Einheit nutzen (Nutzen)“, „mit der Einheit auslegen (Auslegen)“, „ggf. Einheiten verfeinern (Verfeinern)“ und „Einheiten zählen (Zählen)“ der Kernideen des Messens zuzuordnen, die im Text erkennbar sind.

Die Durchführung und Auswertung dieses Ratings erfolgten in folgenden sechs Schritten:

1) Die beiden Mathematikdidaktiker haben unabhängig voneinander in einem selbsterstellten Online-Ratingtool den 126 Texten die jeweils vorkommenden Kategorien zugeordnet.
2) Die für jede der vier Kategorien separat entwickelten ViviAn-KI-Modelle ordnen ebenfalls allen 126 Texten die jeweils vorkommenden Kategorien zu. Sie geben jeweils eine Zuordnung sowie die Sicherheit, mit der diese Zuordnung durchgeführt wurde, aus.
3) Die Übereinstimmung der Ratings der beiden Mathematikdidaktiker wurde untersucht und dabei festgehalten, bei welchen Nutzendentexten die Rater *nicht* übereingestimmt haben.
4) Die Übereinstimmung des Ratings der ViviAn-KI-Modelle mit den Ratings der beiden Mathematikdidaktiker wurde untersucht und dabei festgehalten, bei welchen Nutzendentexten die KI *nicht* mit einem oder beiden menschlichen Ratern übereingestimmt hat.
5) Die Texte, bei denen die Ratings der beiden Mathematikdidaktiker nicht übereingestimmt haben, wurden im Sinne einer qualitativen Inhaltsanalyse genauer betrachtet, um die Gründe für die Unterschiede zu identifizieren.
6) Die Texte, bei denen das Rating der ViviAn-KI-Modelle sich von einem oder beiden menschlichen Ratern unterschied, wurden im Sinne einer qualitativen Inhaltsanalyse genauer betrachtet, um die Gründe für die Unterschiede zu identifizieren.

Nach diesem kurzen Einblick in die Inhalte der Vignette und der Diagnoseaufgabe, zu der in der Machbarkeitsstudie KI-Modelle trainiert und analysiert werden, wird im Folgenden die Vorgehensweise beim Design und der Entwicklung dieser Modelle dargestellt.

5. Design und Entwicklung der eingesetzten KI-Modelle

In diesem Abschnitt wird berichtet, welche Designüberlegungen zu den KI-Modellen geführt haben, die im Rahmen der vorliegenden Machbarkeitsstudie eingesetzt wurden, wie diese entwickelt wurden und wie sie aufgebaut sind.

5.1 Vergleich des Designs mit früheren Arbeiten

In Lutz (2022) wurde ein word2Vec Embedding eingesetzt, welches die Zeichenkettenrepräsentation der Nutzertexte in einen Vektor wandelt, welcher dann zum Training eines einfachen Klassifikationsnetzes mit wenigen vollverbundenen (dense) Layern verwendet wird, um ein Modell zur Klassifikation zu erhalten. So basiert Lutz (2022) auf zwei getrennt betrachteten Modellen. Für ViviAn kommen im Gegensatz dazu verstärkt modernere Sprachmodellansätze zum Einsatz, die Sprachtraining und Klassifikationstraining kombinieren. Hierzu wird die Arbeitsweise aus Fahse und Lutz (2024), übertragen auf die neue Situation. Bei Fahse und Lutz sollte eine Klassifikation, also eine einzelne Klasse, Ergebnis der Codierung sein. Die Codierung gibt dort die Antwort auf die Frage: Welcher einen der folgenden Klassen entspricht die Antwort?

Zur Analyse der in ViviAn eingegebenen Diagnoseantworten müssen dagegen mehrere möglicherweise gleichzeitig vorhandene Kategorien in einem Text identifiziert werden. Es geht also primär um Fragen folgender Art: Welche der folgenden Aspekte sind in der Antwort abgebildet?

5.2 Grundlegende Struktur der Automatisierung

Um einen modularen Aufbau zu gewährleisten, der es ermöglicht, Teilmodelle auch einzeln einzusetzen und zu untersuchen, wird die Aufgabenstellung der Identifizierung von vier verschiedenen Kategorien (Einheiten nutzen, Auslegen, Zählen, Verfeinern) über das Zusammenspiel mehrerer getrennt voneinander trainierter Modelle realisiert. Für jede der vier zu codierenden Kategorien wird ein eigenes Modell entwickelt. Diese Modelle sollen dann für einen gegebenen Text aussagen können, ob die Kategorie, für die sie erstellt wurden, in diesem Text vorhanden

ist oder nicht. Neben der einfacheren didaktischen Untersuchbarkeit der Teilmodelle wird auf diese Weise auch sichergestellt, dass eine Kategorie-Entscheidung nicht systematisch von anderen Kategorie-Entscheidungen abhängen kann. Letztere Information soll der ViviAn-KI bewusst verwehrt werden, da sie konzeptionell nicht zur als unabhängig dargestellten händischen Codierungsentscheidung passt. Die händische Kodierung hat den Auftrag möglichst unabhängig von der Einschätzung der anderen Kategorien jede Kategorie einzeln zu kodieren. Daher soll die ViviAn-KI eben nicht beim Lernen immer Kenntnis von allen Nutzerausprägungen haben und daraus profitieren können, dass Kategorie A häufig mit Kategorie B genannt wird. Die ViviAn-KI soll per Design unabhängig von der Einschätzung zu den anderen Kategorien eine Einschätzung für die Zielkategorie abgeben.

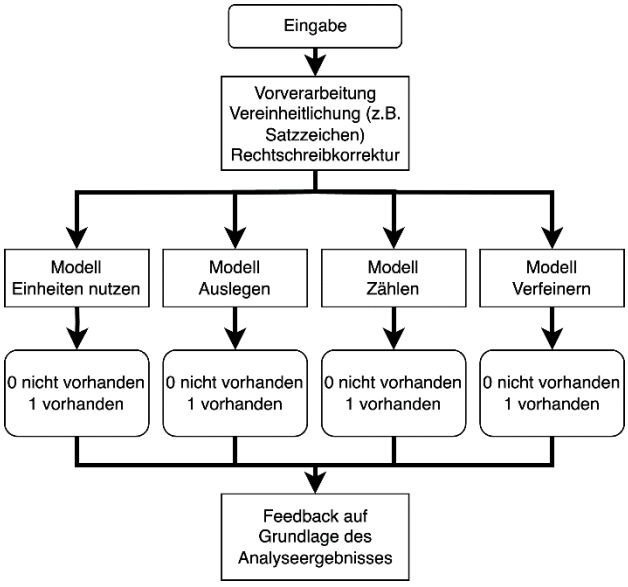

Abb 2.: Automatisiertes Auswertungsschema zur Generierung eines Feedbacks zur Diagnoseaufgabe 4.2

5.3 Aufbau der Einzelmodelle

Die Modelle der vier Kategorien sind alle strukturgleich aufgebaut. Alle verwendeten Modelle sind DistilBERT-Modelle (Sahn et al., 2020), die die Entscheidung „vorhanden“ bzw. „nicht vorhanden“ treffen. DistilBERT ist ein auf BERT (Devlin et al., 2018) basierender Modellansatz, welcher in Benchmark-Untersuchungen nur etwas schlechter als BERT abschneidet, aber sehr viel ressourcenschonender und schneller in der Anwendung ist. Dies macht DistilBERT attraktiv, wenn man bedenken muss, dass für die in diesem Artikel entstehenden Modelle, wie auch für viele vergangene, wie künftige Projekte, viele dieser Modelle gleichzeitig verfügbar sein müssen. Als initiale Gewichte werden auf deutschem Textkorpus trainierte Gewichte verwendet. Diese werden dann in einem Downstream Task, also unter Nutzung der vortrainierten Gewichte, als Ausgangspunkt eines Transfers zum Training der Klassifikation verwendet.

5.4 Vorverarbeitung der Texte vor dem Training

Die Vorverarbeitung der Texte erfolgte wie folgt (vgl. Fahse & Lutz, 2024): Alle Satzzeichen wurden entfernt, und eine automatische Rechtschreibkorrektur wurde vorgenommen. Die Rechtschreibkorrektur soll die Varianz wichtiger Ausdrücke im Text minimieren und zugleich dafür Sorge tragen, dass zumindest einfach erkennbare Rechtschreibfehler oder Zeichensetzung keinen Einfluss auf die Verarbeitung durch das KI-Modell haben. Ein Wort wird dabei technisch als Zeichenkette ohne Leerzeichen betrachtet. Zur Rechtschreibkorrektur wird zu einem Wort, das nicht in einem Wörterbuch (mitsamt Flexion) vorkommt, ein Abstand zu den Wörtern im Wörterbuch bestimmt und das unbekannte Wort durch ein bekanntes Wort mit möglichst wenig Abstand ersetzt.

5.5 Training der Modelle

Damit trotz der vierfachen Modellentwicklung eine gemeinsame Gütebeurteilung möglich ist, werden alle Modelle aufgrund der je Modell gleichen Datenbasis trainiert.

Für die Aufgabe 4 stehen 838 Texte zur Verfügung. Da die Nutzerinnen und Nutzer, wie in Abschnitt 4.2 dargestellt, in Aufgabe 4.1 bereits selbst ihre Antwort zu Aufgabe 4.2 den vier Kategorien zugeordnet haben, liegen auf diese Weise Codierungen aller Antworten vor. Mithilfe dieser Kodierung erfolgt das Training und die Modellprüfung. Dazu wird der vorhandene Gesamtdatensatz aus den Bearbeitungen in ViviAn gemäß Tabelle 1 in drei Teildatensätze zerlegt.

Datensätze	**Anteil in Prozent**
Trainingsdatensatz	70 %
Validierungsdatensatz	15 %
Testdatensatz	15 %
Gesamtdatensatz	**100 %**

Tab. 1: Aufteilung der Daten in drei Teildatensets

Der *Trainingsdatensatz* enthält 70 % der Texte. Mit diesen Texten trainieren alle Modelle. 15 % der Texte werden im Validierungsdatenset verwendet. Training führt zu einer immer besseren Anpassung des Modells an die Daten. Dabei kann es zu einem sogenannten *Overfitting* kommen, d. h. das Modell passt sich so gut an die Daten an, dass es zwar für die Trainingsdaten perfekt funktioniert, aber neue

Daten nicht gut zuordnen kann. Deshalb wird das beste Modell nicht aufgrund der Güte im Trainingsdatensatz ausgewählt, sondern aufgrund der Güte im Validierungsdatensatz, der während des Trainings nur zum Vergleich zur Verfügung steht. Die übrigen 15 % der Daten werden bis zur Wahl des finalen Modells überhaupt nicht betrachtet und als sogenannter Testdatensatz zunächst zur Seite gelegt. Nach der Wahl des finalen Modells wird die Güte des Modells in praktischen Anwendungen auf Basis des Trainingsdatensets beurteilt. Dieses dreiteilige Behandeln der Daten soll Probleme, die beim Training auftreten können, verhindern und so möglichst vielversprechende Modelle finden und gleichzeitig eine praxisnahe Einschätzung der Modellgüte sicherstellen. Die Entscheidung zur Datensatz-Unterteilung 70% Training und 30% Validation/Test entspricht einer von mehreren in der Praxis gängigen modellunabhängigen Möglichkeiten (Joseph, 2022).

6. Ergebnisse

Im Folgenden werden die Ergebnisse zu den einzelnen Forschungsfragen skizziert.

6.1 Ergebnisse zu Forschungsfrage 1

> **Forschungsfrage 1:** Wie gut ermöglicht die bestehende Datenbasis von ca. 1.000 Bearbeitungen das Training robuster Machine-Learning-Modelle hinsichtlich der quantitativen Genauigkeit, um unbekannte Freitextantworten automatisch in diagnostisch relevante Kategorien einzuordnen?

Zur Beantwortung der Forschungsfrage 1 werden in Tabelle 2 und 3 Kennwerte der vier KI-Modelle „Einheiten nutzen“, „Auslegen“, „Verfeinern“ und „Zählen“ für den weder beim Training noch bei der Modellwahl verwendeten Validierungsdatensatz präsentiert. Zunächst nachfolgend die Kennwerte Accuracy, Precision und Recall der einzelnen Modelle in Tabelle 2. Aus Darstellungsgründen wird auf die Auflistung des F1-Scores (Harmonisches Mittel zwischen Precision und Recall) verzichtet.

Modell	Accuracy	Precision	Recall
Einheiten nutzen	.7	.94	.7
Auslegen	.87	.89	.98
Verfeinern	.59	.87	.62
Zählen	.94	.99	.95

Tab. 2: Kennwerte (Accuracy, Precision, Recall) der Modelle. *Accuracy*: Anteil aller richtig klassifizierten Fälle; Precision: Anteil der tatsächlich positiven unter den als positiv vorhergesagten Fällen; *Recall*: Anteil der korrekt erkannten positiven unter allen tatsächlich positiven Fällen.

Modell	Anteil korrekter Zuordnungen
Einheiten nutzen	89 %
Auslegen	96 %
Verfeinern	94 %
Zählen	94 %

Tab. 3: Güte der KI-Modelle im Validierungsdatenset

Wie aus der Tabelle 3 hervorgeht, gelingt eine Modellerstellung in einem praxistauglichen Ausmaß. Damit ist die *Forschungsfrage 1* positiv beantwortet. Es kann also festgehalten werden: Die vorhandenen 838 Bearbeitungen der ViviAn-Diagnoseaufgaben aus 8 Jahren Vignetteneinsatz in der Hochschullehre reichen als Trainingsdaten für Maschine-Learning-Modelle aus. Es gelingt der ViviAn-KI mit Hilfe eines auf Basis der Trainingsdaten erstellten Maschine-Learning-Modells aus einer unbekannten Freitextantwort von Nutzerinnen und Nutzern, die Kategorien zu identifizieren, die dieselben Nutzerinnen und Nutzer vorher in einer Multiple-Choice-Auswahl als relevant angegeben haben.

Im Folgenden werden die Übereinstimmungen der verschiedenen Rater berichtet, die sich auf die Schritte 3 und 4 der in Abschnitt 4.3 angegebenen Schrittfolge beziehen.

6.2 Ergebnisse zu Forschungsfrage 2

In diesem Abschnitt werden die Ergebnisse zu Forschungsfrage 2 dargestellt.

> **Forschungsfrage 2:** Wie hoch ist die Übereinstimmung zwischen den von der ViviAn-KI ermittelten Kategorien und den Zuordnungen diagnostisch geschulter Mathematikdidaktiker*innen?

6.2.1 Übereinstimmung der menschlichen Rater

Die beiden Mathematikdidaktiker stimmten in hohem Maße in ihren Ratings überein. Konkret werden die Zuordnungen der Texte zu den vier Kategorien „Einheiten nutzen“, „Auslegen“, „Verfeinern“ und „Zählen“ bei 114 von 126 Texten von beiden menschlichen Ratern in identischer Weise vorgenommen. Dies entspricht einer Übereinstimmungsrate von 90,5 %.

Es ergeben sich folgende Interrater-Reliabilitäten:

Zwei menschliche Rater (Experte 1, Experte 2): „Einheiten nutzen“: κ=0,847; p_o=93,7 % (n=126), „Auslegen“: κ=0,883; p_o=97,6 % (n=126), „Verfeinern“ κ=0,966; p_o=98,4 % (n=126), „Zählen“: κ=1,000; p_o=100,0 % (n=126). Overall: κ=0,927; p_o=97,4 % (n=504).

6.2.2 Übereinstimmung zwischen den Ratings der ViviAn-KI und der Mathematikdidaktiker

Die ViviAn-KI stimmt mit beiden Ratern bei 96 von 126 Texten (76 %) in allen Details der vorgenommenen Zuordnung zu den vier Kategorien überein. In 11 Fällen stimmt die ViviAn-KI bei uneinigen menschlichen Ratern mit einem der Rater vollständig überein. Wenn man diese beiden Situationen zusammennimmt, um auszudrücken, wie gut es der ViviAn-KI gelingt, menschliche Rater zu imitieren, von denen die ViviAn-KI gar nicht gelernt hatte, so kommt man auf insgesamt 107 von 126 vollständigen Übereinstimmungen mit zumindest einem der beiden Rater (85 %).

Es ergeben sich folgende Interrater-Reliabilitäten:

Drei Rater (Experte 1, Experte 2, ViviAn-KI): „Einheiten nutzen“: $\kappa = 0{,}811$; $\bar{P} = 92{,}1$ % ($n = 126$), „Auslegen“: $\kappa = 0{,}559$; $\bar{P} = 92{,}6$ % ($n = 126$), „Verfeinern“: $\kappa = 0{,}921$; $\bar{P} = 96{,}3$ % ($n = 126$), „Zählen“: $\kappa = 0{,}911$; $\bar{P} = 97{,}9$ % ($n = 126$). Gesamt (Overall): $\kappa = 0{,}848$; $\bar{P} = 94{,}7$ % ($n = 504$).

Zwei gemischte Rater (Experte 1, ViviAn-KI): „Einheiten nutzen“: $\kappa=0{,}812$; $p_o=92{,}1$ % ($n=126$), „Auslegen“: $\kappa=0{,}316$; $p_o=88{,}9$ % ($n=126$), „Zählen“: $\kappa=0{,}864$; $p_o=96{,}8$ % ($n=126$), „Verfeinern“: $\kappa=0{,}915$; $p_o=96{,}0$ % ($n=126$). Overall: $\kappa=0{,}811$; $p_o=93{,}5$ % ($n=504$).

Zwei gemischte Rater (Experte 2, ViviAn-KI): „Einheiten nutzen“: $\kappa=0{,}774$; $p_o=90{,}5$ % ($n=126$), „Auslegen“: $\kappa=0{,}381$; $p_o=91{,}3$ % ($n=126$), „Zählen“: $\kappa=0{,}864$; $p_o=96{,}8$ % ($n=126$), „Verfeinern“: $\kappa=0{,}881$; $p_o=94{,}4$ % ($n=126$). Overall: $\kappa=0{,}804$; $p_o=93{,}3$ % ($n=504$).

Zusammenfassend lässt sich festhalten, dass sowohl die menschlichen Rater untereinander eine hohe Übereinstimmung aufweisen, als auch eine gute Übereinstimmung der ViviAn-KI mit menschlichen Ratern vorliegt.

Damit kann die *Forschungsfrage 2* positiv beantwortet und festgehalten werden, dass der ViviAn-KI eine ähnlich gute Zuordnung von Nutzerinnen und Nutzer Antworten zu vorgegebenen fachspezifischen Kategorien gelingt, wie diagnostisch geschulten Mathematikdidaktiker*innen.

6.3 Ergebnisse zu Forschungsfrage 3

Im Folgenden werden die Ergebnisse zu Forschungsfrage 3 berichtet.

> **Forschungsfrage 3**: Welche systematischen Unterschiede bestehen zwischen den von der ViviAn-KI und den von diagnostisch geschulten Mathematikdidaktiker*innen erstellten Zuordnungen von Nutzendenantworten zu fachdidaktischen Kategorien?

6.3.1 Texte mit unterschiedlicher Kategorie-Zuordnung: Fachdidaktische Analyse

Für eine mathematikdidaktische Qualitätskontrolle ist die Kenntnis der Übereinstimmung in den Ratings der ViviAn-KI mit menschlichen Experten nicht ausreichend. Es muss vielmehr zusätzlich untersucht werden, an welchen Stellen und warum Abweichungen in den Ratings auftreten. Erst auf der Grundlage dieser Kenntnis kann eingeschätzt werden, ob es ethisch und fachinhaltlich verantwortbar ist, den Nutzerinnen und Nutzern, die ViviAn nutzen, eine individuelle Rückmeldung zu ihren Antworttexten auf der Grundlage der Einschätzung der ViviAn-KI zu geben. Um die Abweichungen der Ratings der ViviAn-KI von denen der menschlichen Rater sinnvoll einschätzen zu können, müssen zunächst die Abweichungen der Ratings der beiden Mathematikdidaktiker voneinander analysiert und bewertet werden.

6.3.2 Unterschiede bei den menschlichen Ratern

Nur bei 11 von 126 Texten der Nutzerinnen und Nutzer gab es überhaupt einen Unterschied bzgl. der Zuordnung zu den Kategorien „Einheiten nutzen“, „Auslegen“, „Verfeinern“ und „Zählen“. Einer dieser elf Texte fällt aus dem Rahmen, weil er der einzige ist, bei dem sich die beiden Rater bei zwei Kategorien uneinig waren. Dies liegt an der unklaren Formulierung dieses Textes[2]:

> Sie messen mit der Einheit der Quadrate aus und verfeinern diese schlussendlich noch mal mit den kleineren Quadraten. Um den Flächeninhalt zu berechnen, zählen sie die Einheiten aus.

Es wird hier weder von „Auslegen“ noch von „Nutzen“ gesprochen und so ergibt es sich, dass ein Rater zwar „Einheiten nutzen“, aber nicht „Auslegen“ codiert, während der andere umgekehrt nicht „Einheiten nutzen“, aber „Auslegen“ codiert. Die beiden anderen Codierungen stimmen dagegen überein.

Bei den restlichen 10 Texten wurde nur eine der Zuordnungen von den Ratern unterschiedlich codiert, während die anderen drei übereinstimmen. Bei genauerer Analyse der Unterschiede ist festzustellen, dass es nur drei verschiedene Typen von unterschiedlichen Zuordnungen der beiden Rater gab, die im Folgenden einzeln beschrieben und anhand der Schülertexte belegt werden.

Auslegen: Umgang mit der Formulierung „gelegt“

Nur ein einziges Mal trat eine unterschiedliche Zuordnung mit Blick auf die Kategorie „Auslegen“ auf, nämlich bei folgendem Text:

> Es werden zuerst überall große blaue Quadrate gelegt, dann werden für die Einheit zu verfeinern an den Grenzen der Fläche anstatt große blaue Quadrate, rote kleine Quadrate genommen. Dann werden am Ende alle Quadrate zusammengezählt.

Ein Rater hat sich hier dafür entschieden, dass der Bezeichner „gelegt“ zusammen mit der konkreten Tätigkeit im Video, in dem zu sehen ist, wie die Schüler die Fläche der USA auf der Karte mit Quadraten auslegen, im Sinne des Auslegens mit einer Einheit zu verstehen ist und deshalb die Kategorie „Auslegen“ für den Text vergeben.

Der andere Rater hat die Kategorie „Auslegen“ für diesen Text nicht vergeben, weil für ihn „legen“ und alle ähnlichen Formulierungen nicht zwingend die Forderung nach Überlappungsfreiheit umfasst und damit nicht klar ist, ob das Prinzip des Auslegens von dem oder der Nutzerinnen und Nutzer erfasst wurde.

Verfeinern: Umgang mit dem Vorhandensein des Wortes „verfeinern“

Nur bei den folgenden beiden Texten waren sich die Rater bei der Zuordnung des Textes zur Kategorie „Verfeinern“ nicht einig:

> Sie nutzen die gegebene Einheit, sie legen die Fläche damit aus und zählen. Verfeinern tun sie nur ihr Handeln, sie zerschneiden nicht nocheinmal selber die vorhandenen Plättchen.

> Sie nutzten den würfel als einheit zum messen damit sie eine genaue anzahl bekommen. Dann legen die die komplette usa mit diesen aus und benutzten große und kleine dafür damit es genauer wird. Diese werden dann am ende gezählt und mal die vorgegebene Größe gerechnet. So kommen sie zu einem genaueren ergebnis

Beim ersten Text hat ein Rater die Kategorie „Verfeinern“ vergeben, weil das Wort „Verfeinern“ explizit auftritt. Der andere Rater hat sich bei diesem Text dafür entschieden, die Kategorie „Verfeinern“ *nicht* zuzuordnen, weil „Verfeinern“ zwar als Wort vorkommt, aber die Person, die sich äußert, einschränkend festhält, dass die Schüler die größeren Quadrate nicht „zerschneiden“. Dies wird vom zweiten Rater so gedeutet, dass die Schüler davon ausgehen, dass nicht wirklich im Sinne der entsprechenden Kernidee des Messens verfeinert wird.

Beim zweiten Text interpretiert ein Rater den Teilsatz „und benutzten große und kleine dafür damit es genauer wird“ nicht im Sinne der Kategorie „Verfeinern“, auch weil das Wort „Verfeinern“ nicht vorkommt. Dagegen vergibt der andere Rater die Kategorie „Verfeinern“, weil er denselben Teilsatz wegen der Phrase „damit es genauer wird“ als Idee des Verfeinerns interpretiert.

Einheiten nutzen: Umgang mit dem Auftreten des Wortes „nutzen“ ohne das Wort „Einheit“

Bei acht Texten unterscheiden sich die beiden Rater bzgl. der Zuordnung zur Kategorie „Einheiten nutzen“. Beim folgenden Text vergibt ein Rater die Kategorie „Einheiten nutzen“ nicht, weil im Text nicht konkret von Einheiten gesprochen wird. Der andere Rater geht in Kenntnis des Videos davon aus, dass „Quadrate nutzen“ im Sinne von „Einheiten nutzen“ gemeint ist und vergibt deswegen die entsprechende Kategorie:

> Die SuS wollen sinnvoll die Anzahl der gegebenen Quadrate benutzen um keine Ecke frei zu lassen.

Ein sprachlich besonderer Fall ist der folgende Text. Ein Rater sieht einen sprachlichen Unterschied zwischen „eine Einheit wird genutzt“ und wie hier formuliert „die Einheit nutzt“. Da letzteres im Sinne von „die Einheit ist nützlich“ zu verstehen ist, entscheidet er sich gegen die Kategorie „Einheiten nutzen“. Der andere Rater interpretiert das anders. Er stellt fest, dass die Einheit im Text als nützlich beschrieben und im Video verwendet wird. Zusammen betrachtet rechtfertigt das für ihn die Kategorie „Einheiten nutzen“.

> Als erstes legen die Schüler mit ihrer Einheit (Quadrate) die Fläche der USA von 1913 aus. Um die Einheit an sich zu ermitteln, hilft der beigefügte Maßstab. Die Einheit nutzt hier, da es somit einfacher ist, die flache zu bestimmen, da es so ersichtlicher wird. um den ganzen Flächeninhalt zu ermitteln müssen die Einheiten (blaue und rote Quadrate) gezählt werden, damit wie oben beschrieben der ganze Flächeninhalt der USA ermittelt werden kann.

Insgesamt lässt sich bei genauerer Betrachtung der wenigen nicht übereinstimmenden Ratings der beiden Mathematikdidaktiker feststellen, dass es für beide Betrachtungsweisen jeweils ein nachvollziehbares Argument gibt.

6.3.3 Unterschiede zwischen menschlichen Ratern und der ViviAn-KI

Im Folgenden soll analysiert werden, was geschieht, wenn die ViviAn-KI mit keinem der beiden Rater vollständig übereinstimmt. In 17 der 19 Fälle stimmen immerhin die Rater untereinander überein, die ViviAn-KI jedoch trifft mindestens in einer Kategorie eine andere Entscheidung als die Rater. Eine Analyse dieser 17 Fälle soll Aufschluss darüber geben, wie die ViviAn-KI stattdessen handelt.

In vier der 17 Fälle erkennt die ViviAn-KI anders als die Rater in Aussagen der Art: „Alle Kernideen des Messens kommen vor“, nicht alle Kategorien.

Bei vier Texten entscheidet sich die ViviAn-KI für die Kategorie „Auslegen“, während beide menschliche Rater diese Kategorie nicht wählen.

Beispiel:

> Schüler benutzen zuerst einmal nur blau Quadrate und verfeinern schließlich diese mit roten. Anschließend werden diese Quadrate gezählt.

Die ViviAn-KI kennt, genau wie die menschlichen Rater, die Antworten der Nutzerinnen und Nutzer auf das Multiple-Choice-Item im Testdatenset *nicht*.

Vergleicht man nun mit den von den Nutzerinnen und Nutzern jeweils angekreuzten Kategorien in Aufgabe 4.1, so stellt man fest, dass jeweils „mit der Einheit auslegen“ angekreuzt wurde. Dies kommt genau bei den hier angegebenen vier Fällen vor. Die KI, die auf Basis der Antworten im Trainingsdatenset trainiert wurde, imitiert also auch in diesen Fällen im Testdatenset erfolgreich Nutzendenantworten, stimmt dadurch aber leider nicht mit den Expertenantworten überein.

Ähnlich verhält sich dies je einmal bei der Kategorie „Verfeinern“ und der Kategorie „Einheiten nutzen“:

> Zuerst legen die Schüler die blauen Quadrate auf die Vorlage und entfernen dann die Quadrate, die überstehen. Danach werden dann diese Flächen durch die kleinen roten Quadrate ersetzt und zum Schluss zählen die Schüler wie viele Quadrate insgesamt auf der Vorlage liegen.

Im Gegensatz zu den menschlichen Ratern erkennt die ViviAn-KI bei obigem Text „Verfeinern“. Der Proband selbst hat hier Verfeinern angekreuzt.

> Die Schüler legen die Fläche zuerst mit großen Quadraten aus. Danach verfeinern sie die Einheit, indem sie einige der großen Einheiten wegnehmen und durch kleine Quadrate ersetzen. Als letztes zählen sie die Einheiten.

Die ViviAn-KI erkennt hier die Kategorie „Einheiten nutzen“, während diese Kategorie von den Experten-Ratern nicht vergeben wurde. Der Proband selbst wählt jedoch die Kategorie „Einheiten nutzen“.

Die ViviAn-KI entscheidet sich manchmal auch Kategorien nicht zu vergeben, die von den Experten-Ratern vergeben wurden:

> Die Einheit war meiner Meinung nach die Quadrate und diese wurde durch die verschiedenen Größen verfeinert.

Hier entscheidet sich die ViviAn-KI nicht für die Kategorie „Einheiten nutzen“ im Gegensatz zu den Ratern. Der Proband selbst wählt hier nicht die Kategorie „Einheiten nutzen“.

In seltenen Fällen sind sich sowohl Rater als auch Nutzer in einer Kategorie einig, die ViviAn-KI trifft jedoch eine andere Entscheidung und imitiert so leider weder die Rater noch die Nutzerinnen und Nutzer:

> Die Schüler haben die Quadrate als Einheit ausgewählt. Diese dann auf der Landkarte ausgelegt und anschließend gezählt.

Die ViviAn-KI sieht hier anders als die Rater nicht die Kategorie „Einheiten nutzen“. Der Proband stimmt jedoch mit den Ratern überein und wählt selbst auch die Kategorie „Einheiten nutzen“. Das „Auswählen“ einer Einheit ist eine in den übrigen Texten unübliche Formulierung, sie kommt im gesamten Datensatz nur zweimal als „Auswahl“ und zweimal als „ausgewählt“ vor. Im Testdatensatz tritt der Wortstamm von „Auswählen“ sogar nur genau einmal an dieser Stelle auf. Zwar wird beim Wortteil „Auswahl“ beides Mal die Kategorie „Einheiten nutzen“ von den Nutzerinnen und Nutzern angegeben, beim „ausgewählt“ jedoch tatsächlich nur bei der hier berichteten Stelle im Testdatensatz, sodass es nicht verwunderlich ist, dass die ViviAn-KI mit diesem Wort noch nicht umgehen kann.

Bei dem folgenden Beispiel ist es genau umgekehrt. Nur die ViviAn-KI äußert ein „Verfeinern“ erkannt zu haben und weder die Rater noch der Proband stimmen damit überein.

> Sie legen die große Landesinnenfläche mit großen blauen Quadraten aus und die übrig bleibende Fläche mit kleineren roten Quadraten aus und zählen diese anschließend um die Fläche so annähernd zu bestimmen.

Diese Abweichung könnte dadurch erklärt werden, dass die Nutzerinnen und Nutzer im Zusammenhang mit der Verwendung der Phrase „mit kleineren roten

Quadraten“ häufig mit „Verfeinern“ in Verbindung gebrachte Tätigkeiten schildern.

Da die Auswertung so konzipiert ist, dass die ViviAn-KI alle eingegebenen Antworten kategorisieren soll, hat die ViviAn-KI mit sehr kurzen Antworten Probleme, da sie wenig Beziehung zu den meist längeren Antworten anderer Nutzerinnen und Nutzer herstellen kann. So würde sie sich in vier Fällen für zumindest teilweise falsche Kategorien entscheiden.

Beispiel:

> habe es nicht mitbekommen

Beim letzten der oben angegebenen vier Beispiel-Antworten fällt auf, dass insbesondere Flexionen bei zu kurzen Eingaben noch nicht korrekt erkannt werden. Ein Teil dieser Antworten sollte gar nicht bewertet werden, etwa die Antwort „Schüler“.

Nun wird eine disjunkte Fallbeschreibung mit möglichen Konsequenzen als zusammenfassende Übersicht gegeben.

A: Bei 7 der 17 Antworten kommt die ViviAn-KI zwar nicht zum Ergebnis der Rater, aber durchaus in diesen Kategorien zur vom Nutzer bzw. der Nutzerin gewählten Kategorie. Da die ViviAn-KI an Nutzereinschätzungen trainiert wurde, kann man ihr letztlich nicht vorwerfen, diesen unter Umständen eher zu entsprechen als den Experten-Ratings. Wenn man Fälle dieser Art weiter ausschließen möchte, müsste man insbesondere für die Kategorie „Auslegen“ in den Trainingsdaten die Wertungen der Nutzerinnen und Nutzer schon im Training zumindest teilweise durch Experten-Ratings ersetzen. Damit die ViviAn-KI das Verhalten der Experten auch in dieser Kategorie noch besser imitieren lernt, müsste man folglich die ViviAn-KI schon an Expertenratings trainieren. Die Bewertungen der Nutzenden decken sich hier systematisch nicht so gut wie bei den anderen Kategorien mit den Expertenratings, was der Trainingsgrundlage schadet und so eine noch höhere Genauigkeit der ViviAn-KI im Abgleich mit den Expertenratings verhindert.

B: 4 der 17 Antworten können wohl aufgrund zu kurzer Antworttexte nur falsch zugeordnet werden. Durch ein Forcieren, mindestens eine gewisse Anzahl an Zeichen eingeben zu müssen, oder alternativ der Verweigerung einer Auswertung, wenn Texte zu kurz sind, könnte diesen Fällen sinnvoll begegnet werden.

C: Bei 4 der 17 Antworten scheitert die ViviAn-KI an kurz formulierten Aussagen, die die Kategorien nur indirekt nennen, indem sie von „allen Kernideen des Messens“ sprechen. Diese Art von Aussage kann über ein separates Modell zusätzlich und vorab erkannt werden. Alternativ müsste durch mehr Beispiele dieser Art von Antwort ein weiteres Training erfolgen.

D: 2 der 17 Antworten sind aus Sicht der Rater echte Fehlentscheidungen der ViviAn-KI, da sie zumindest teilweise eine Codierung vornimmt, die weder den Experten noch den Nutzenden-Codierungen entspricht. Für diese Fälle können jedoch anders als in den Fällen A bis C keine strukturell begründeten Strategien gefunden werden, um diese Fehler systematisch zu beheben. Um solche Ausnahmen weiter zu verringern, können die Autoren nur vornehmlich auf der Forderung nach der Erhebung weiterer Texte und damit der Ermöglichung besseren Trainings und möglicherweise der Verwendung komplexerer KI-Modelle bestehen. Es ist jedoch, wie bei menschlichen Ratern, nie davon auszugehen, dass bei unabhängiger Kodierung eine vollständige Übereinstimmung in allen Fällen erzielt werden kann.

Zusammenfassend kann man festhalten, dass eine bessere Passung der Modelle zu den Expertenratings durch kleine Anpassungen des Umgangs mit den Fällen B und C recht einfach möglich sein kann. Folglich bleiben noch 9 von 17 Antworten, deren Expertenkonformität nicht einfach hergestellt werden kann.

Eine solche Überarbeitung kann die vollkommene Übereinstimmung von mindestens einem Expertenrating mit der ViviAn-KI auf 117 von 126 Antworten anheben (93 %).

In den letzten 7 der 9 verbleibenden Antworten ist die nicht einfach auflösbare Abweichung aber immer noch kaum verwunderlich, weil die ViviAn-KI-Antwort in diesen Fällen die Antwort der Nutzenden korrekt voraussagt, und diese – wenn auch nicht die des Testdatensatzes – zum Training verwendet wurden. Die Modelle imitieren folglich besser die Antworten der Nutzenden als die Antworten der Experten.

Von den 2 verbleibenden Antwort-Einschätzungen der ViviAn-KI, die dadurch nicht erklärt werden können, ist zumindest eine Einschätzung wohl auf eine ungewöhnliche Wortwahl („Auswahl“) der Nutzerinnen und Nutzer zurückzuführen, die in den seltenen Fällen, in der sie im Datensatz vorkommt, auch noch unterschiedlich von den Nutzerinnen und Nutzern codiert wurde.

Im Rückblick auf Forschungsfrage 3 zeigen sich nur wenige systematische Unterschiede zwischen den Zuordnungen der ViviAn-KI und jenen diagnostisch

geschulter Mathematikdidaktiker. Diese Unterschiede lassen sich durch qualitative Inhaltsanalysen der entsprechenden Nutzendenantworten identifizieren und nachvollziehbar erklären bzw. deuten.

7. Fazit

Dieser Beitrag fokussiert auf die Beschreibung und exemplarische Durchführung einer neuen mathematikdidaktischen Methodik zur fachdidaktischen Qualitätssicherung von automatisierten KI-Auswertungssystemen, die argumentierende Textantworten ermöglichen soll (mit nur geringem Anteil algebraischer Formalismen). Diese Methodik lehnt sich in ihrer Vorgehensweise an in der Mathematikdidaktik übliche Methoden zur Auswertung von Textantworten an, indem etwa zunächst mehrere menschliche Rater:innen Kategoriezuweisungen vornehmen. Es wird dann jedoch grundsätzlich keine Konsenskodierung hergestellt. Widersprüchliche Labels der Rater:innen bleiben erhalten und dienen als Analysegrundlage für die Labels, die das automatisierte KI-Auswertungssystem vergibt. Durch das vorgeschlagene Vorgehen übernimmt der menschliche Kodiervorgang eine fachdidaktische Kontrollfunktion, suggeriert in der Bewertung jedoch zugleich, dass es immer ein einziges richtiges Label gibt, die ein Mensch immer vergeben würde und an der sich die KI daher messen muss. Die Unsicherheiten der Kodierung werden der KI dann nicht negativ angerechnet, wenn die KI so entscheidet, wie ein festzulegender relevanter Anteil der Rater:innen (in der exemplarischen Durchführung 50% der Rater). Damit ergänzt das beschriebene Verfahren übliche technische Qualitätskriterien, die, wie in 5 und 6.1 beispielhaft ausgeführt, nur „richtig“ und „falsch“ kennen. Besonders solche Fehler, bei denen die Expert:innenratings sich einig und anders als die KI äußern, werden in Analyse dieser Fälle genauer betrachtet. Ein auf die vorgeschlagene Weise untersuchtes KI-Modell wird so nicht nur technisch, sondern auch fachdidaktisch, mit an die Einigkeit der Rater:innen angepasster abgestufter Absolutheit beurteilt. So sollen angenommene Weiterentwicklungspotentiale der technischen Systeme identifiziert und das Produkt „KI Modell“ zugleich mit erläuternden Hinweisen aus der fachdidaktischen Analyse zur Benutzung versehen werden können.

Zusätzlich formuliert der Beitrag in der Folge beispielhaft einen Vorschlag für einen vollständigen Entwicklungsprozess einer Machine-Learning Modell-Kaskade, die ihrerseits dazu beiträgt, das beschriebene Vorgehen zu unterstützen, indem es zulässt, verwendete KI-Teilmodelle separat zu untersuchen.

Als Produkt der exemplarischen Anwendung des beschriebenen Verfahrens zur fachdidaktischen Qualitätssicherung von Machine-Learning Modellen entwickelte dieser Beitrag zugleich ein Set aus vier Machine-Learning-Modellen. Dieses Set hat ohne jegliche Modifikation in einem ihm vollkommen unbekannten Testdatensatz eine vollständige Übereinstimmung von 85 % mit je mindestens einem Rater erzielt, obwohl die Modelle nicht mit Daten der Rater trainiert wurden.

Die oben genannte Übereinstimmung kann mit kleineren Modifikationen auf Basis der Analyse der Nicht-Übereinstimmungen sogar auf eine Erkennungsrate von ca. 93 % aufgrund systematischer Ergänzungen der automatisierten Auswertung durch Umsetzung der Vorschläge zur Weiterentwicklung in Reaktion auf die Abweichungen B und C gebracht werden. Die Texte, die dann noch immer nicht entsprechend der Expertenratings zugewiesen würden, können im Wesentlichen nur durch eine Ersetzung der von den Nutzenden selbst erstellten Codes durch Experten-Codes in der Trainingsgrundlage erreicht werden, um der ViviAn-KI die Möglichkeit zu geben, direkt an Expertenratings zu erlernen, wie Experten codieren. Zu kleinen Teilen sind noch weitere Verbesserungen nur durch die Erhebung von neuen Daten erwartbar.

Texte, die Nutzerinnen und Nutzer schreiben und selbst labeln, sollten nicht unkritisch für das Training von KI-Modellen übernommen werden. Bei Label-Vorgängen, wie sie im Rahmen dieser Studie beschrieben wurden, kann es jedoch hilfreich sein, auf solche selbsterstellten Labels aufzusetzen, diese für KI-Trainings zu nutzen und anschließend mit für Mensch und ViviAn-KI neuen und unbekannten Testantworten zu prüfen. Praktische Voraussetzung dafür waren in unserem Fall die hohe Übereinstimmung der Experten untereinander und die bereits unmodifiziert hohe Übereinstimmung der ViviAn-KI mit den Experten. Unter dieser Voraussetzung können, wie im hier berichteten Beispiel, Selbst-Label-Vorgänge durch Nutzende durchaus zum Training von KI-Modellen genutzt werden.

Die KI-Modelle wurden darauf trainiert, die Nutzerinnen und Nutzer des Trainingsdatensatzes zu imitieren. Weder die KI-Modelle noch die menschlichen Rater hatten etwas anderes als den „Text“ des Trainingsdatensatzes vorliegen, also nicht die Kategorienzuordnungen der Nutzenden. Einerseits wurde

so grundsätzlich die Gefahr vermieden, dass die Rater sich von den Einschätzungen der Nutzerinnen und Nutzer beeinflussen lassen. Andererseits konnte die ViviAn-KI sich auch nicht direkt an die Rater anpassen, weil ihr als Trainingsdaten nur die Antworten der Nutzenden inklusive Kategorienzuordnungen im Trainingsdatenset vorlagen und zur Wahl der besten Modelle die Texte und Kategorien im Validierungsdatenset.

Eine Limitation der vorliegenden Machbarkeitsstudie besteht darin, dass nur mit zwei menschlichen Ratern gearbeitet wurde, die wiederum selbst die Daten zur Übereinstimmung der Rater analysiert haben. Dies erscheint vor dem Hintergrund, dass es sich hier um eine Machbarkeitsstudie handelt, in der es darum ging, zu identifizieren, inwiefern ein solches Vorgehen sinnvoll umsetzbar ist, als akzeptables Vorgehen. Bei einer Umsetzung als Grundlage für ein zu erzeugendes automatisiertes Feedback sollte man in der Zukunft noch weitere menschliche Rater hinzufügen. Dies erfordert dann zugleich eine Festlegung der beschriebenen nötigen Mindestübereinstimmung der Rater, um Stufen für den Grad der Konsensfähigkeit der Ratings festzulegen. Diese Festlegung konnte bei zwei Ratern trivial durch die Zustimmung von zwei bzw. einem Rater erfolgen. Diese Limitation wird angesichts der Forderung nach aktiver Einbindung von Stakeholdern (in diesem Fall die Didaktiker*innen, die die Modelle einsetzen) in den Designprozess sowie zur Validierung von Prototypen der Machine-Learning-Modelle (Alfredo et al., 2024) allerdings abgeschwächt, da es sich noch nicht um das finale Produkt handelt. Die Einbeziehung von Fachdidaktiker*innen zur inhaltlichen Evaluation der Modellausgaben schon während des Designprozesses wird in der Literatur gefordert, aber auch durch den Mangel an Personen relativiert (z. B. Khosravi et al., 2022). Die Konzeption der Machbarkeitsstudie unter Berücksichtigung von Transparenz und Praxisnutzen zeigt einen Weg auf, um schon während des Designprozesses KI Modelle zu evaluieren. Durch die Nutzung des HCD-Ansatzes wurden so bereits die Phasen 1 (Planung), Phase 2 (Design), Phase 3 (Konzept & Prototypeniteration) und Phase 4 (Bewertung & Verfeinerung) abgeschlossen. Anschließend an diese Machbarkeitsstudie und bei der Implementation in eine Modell-Kaskade muss Phase 5 (Einführung und Überwachung) angeschlossen werden, welche aufgrund des Gesamteinsatzkonzepts noch nicht durchführbar war. In Phase 5 werden auch die Endnutzenden als Stakeholder aktiv am Prozess beteiligt.

Die Machbarkeitsstudie zeigt, dass sich schon mit Modellen, die ausschließlich auf den von den Nutzenden erstellten Labels basieren, beeindruckende Übereinstimmungsraten mit Expertenbewertungen erzielen lassen. Da der Vergleich Mensch-ViviAn-KI bei den Texten aus dem Testdatensatz eine hohe Übereinstimmung aufweist, kann auch ohne die Betrachtung aller 838 Texte umgekehrt die Vermutung aufgestellt werden, dass die Nutzerinnen und Nutzer wohl mehrheitlich mit der Codierung der Experten-Rater übereinstimmen würden. Eine Vermutung, die ohne den Einsatz der KI-Modelle nur auf Basis der von den Experten codierten Texte getätigt werden könnte. KI-Modelle, die lernen, Nutzende zu imitieren, können folglich Ausgangspunkt weitreichender Analysen sein.

Als grundsätzliche Limitation muss beachtet werden, dass der hier formuliert Nachweis der Machbarkeit von einem expertenbasierten und empirischen Qualitätsverständnis ausgeht, für das, als Beleg der Machbarkeit, ein konkretes Verfahren an einem echten Beispiel beschrieben wird. Die Studie trifft keine Aussage darüber, ob andere Formen von Qualitätsverständnis, etwa rein theoriebasierte Ansätze, mit ähnlichen Verfahren untersucht werden können.

Die hier beispielhaft berichteten und für eine Machbarkeitsstudie bereits beachtlichen Ergebnisse sprechen dafür, dass ein Feedback auf der Grundlage eines Ratings der ViviAn-KI bei entsprechendem Training mit einem genügend großen Datensatz ethisch verantwortbar zu sein scheint. Damit eröffnen sich in Zukunft ganz neue Möglichkeiten für KI-generiertes individuelles Feedback auch zu Diagnoseaufgaben.

Anmerkungen

[1] „Feedback is information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, metacognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" (Butler & Winne 1995, S. 275).

[2] Alle zitierten Nutzendentexte werden im Original wiedergegeben und nicht verändert. Es wurde also keinerlei Korrektur bzgl. Satzbau, Zeichensetzung und Rechtschreibung vorgenommen.

Literatur

Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D. & Martinez-Maldonado, R. (2024). Human-centered learning analytics and AI in education: A systematic literature review. Computers and Education: Artificial Intelligence, 6, 100215. https://doi.org/10.1016/j.caeai.2024.100215

Bai, X. & Stede, M. (2022). A Survey of Current Machine Learning Approaches to Student Free-Text Evaluation for Intelligent Tutoring. *International journal of artificial intelligence in education*, 1–39. https://doi.org/10.1007/s40593-022-00323-0

Bartel, M. E. & Roth, J. (2017). Diagnostische Kompetenzen von Lehramtsstudierenden fördern. Das Videotool ViviAn. In J. Leuders, T. Leuders, S. Prediger & S. Ruwisch (Hrsg.), *Mit Heterogenität im Mathematikunterricht umgehen lernen – Konzepte und Perspektiven für eine zentrale Anforderung an die Lehrerbildung* (S. 43–52). Springer Spektrum. https://doi.org/10.1007/978-3-658-16903-9_4

Bartel, M. E. & Roth, J. (2020). Video- und Transkriptvignetten aus dem Lehr-Lern-Labor – Die Wahrnehmung von Studierenden. In B. Priemer & J. Roth (Hrsg.), *Lehr-Lern-Labore. Konzepte und deren Wirksamkeit in der MINT-Lehrpersonenbildung* (S. 299–315). Springer Spektrum. https://doi.org/10.1007/978-3-662-58913-7_19

Bartel, M.-E., Beretz, A.-K., Lengnink, K. & Roth, J. (2018). Prozessbegleitende Diagnose beim Mathematiklernen − Kompetenzentwicklung von Lehramtsstudierenden im Rahmen von Lehr-Lern-Laboren. *MNU 71*(6), 375-382.

Bernius, J. P., Krusche, S. & Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, *3*, 100081. https://doi.org/10.1016/j.caeai.2022.100081

Butler, D. L. & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research 65*(3), 245–281.

Bødker, S., Dindler, C., Iversen, O. S. & Smith, R. C. (2022). What Is Participatory Design? In S. Bødker, C. Dindler, O. S. Iversen & R. C. Smith (Hrsg.), Synthesis Lectures on Human-Centered Informatics. Participatory Design (S. 5–13). Springer International Publishing. https://doi.org/10.1007/978-3-031-02235-7_2

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D. & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Devlin, J., Chang M.-W., Lee, K. & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/arXiv.1810.04805

Enenkiel, P., Bartel, M.-E., Walz, M. & Roth, J. (2022). Diagnostische Fähigkeiten mit der videobasierten Lernumgebung ViviAn fördern. *Journal für Mathematik-Didaktik, 43*(1), 67-99. https://doi.org/10.1007/s13138-022-00204-y

Fahse, C. & Lutz, T. (2024 in press). Codieren mit KI - Herausforderungen und Chancen. *BzMU 2024*.

Giacomin, J. (2014). What Is Human Centred Design? The Design Journal, 17(4), 606–623. https://doi.org/10.2752/175630614X14056185480186

Hattie, J. (2015). *Lernen sichtbar machen. Überarbeitete deutschsprachige Ausgabe von "Visible Learning" von Wolfgang Beywl und Klaus Zierer*. Baltmannsweiler: Schneider

Hanington, B. & Martin, B. (2019). *Universal Methods of Design, Expanded and Revised: 125 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Rockport Publishers Inc.

Hofmann, R. & Roth, J. (2020). Arbeiten mit Funktionsgraphen – Zur Diagnose von Fehlern und Fehlvorstellungen beim Funktionalen Denken. *mathematica didactica, 43*(1), 1–17. https://doi.org/10.18716/ojs/md/2021.1165

Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(4), 531–538. https://doi.org/10.1002/sam.11583

Keuning, H., Jeuring, J. & Heeren, B. (2019). A Systematic Literature Review of Automated Feedback Generation for Programming Exercises. *ACM Transactions on Computing Education*, *19*(1), 1–43. https://doi.org/10.1145/3231711

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S. & Gašević, D. (2022). Explainable Artificial Intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Lee, M. & Moore, T. (2024). Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. *Journal of Educational Technology Research, 59*(2), 15–30. https://doi.org/10.24059/olj.v28i3.4593

Lindsay, E. D., Zhang, M., Johri, A. & Bjerva, J. (2023). *The Responsible Development of Automated Student Feedback with Generative AI.* https://doi.org/10.48550/arXiv.2308.15334

Lutz, T. (2022). Machine-Learning-Modelle zur automatisierten Textklassifikation von mathematischen Aufgabenbearbeitungen. In F. Reinhold & F. Schacht (Eds.), *Digitales Lernen in Distanz und Präsenz: Herbsttagung 2021 des Arbeitskreises Mathematikunterricht und digitale Werkzeuge in der Gesellschaft für Didaktik der Mathematik am 24.09.2021* (S. 89–98). https://doi.org/10.17185/duepublico/76037

Mehralivand, S., Yang, D., Harmon, S. A., Xu, D., Xu, Z., Roth, H., Masoudi, S., Sanford, T. H., Kesani, D., Lay, N. S., Merino, M. J., Wood, B. J., Pinto, P. A., Choyke, P. L. & Turkbey, B. (2022). A Cascaded Deep Learning-Based Artificial Intelligence Algorithm for Automated Lesion Detection and Classification on Biparametric Prostate Magnetic Resonance Imaging. *Academic radiology*, *29*(8), 1159–1168. https://doi.org/10.1016/j.acra.2021.08.019

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A. & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Miao, F. & Cukurova, M. (2024). *UNESCO AI competency framework for teachers*. United Nations Educational, Scientific and Cultural Organization. https://doi.org/10.54675/ZJTE2084

Mozannar, H., Bansal, G., Fourney, A. & Horvitz, E. (2024). When to Show a Suggestion? Integrating Human Feedback in AI-Assisted Programming. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(9), 10137–10144. https://doi.org/10.1609/aaai.v38i9.28878

Narciss, S. (2006). *Informatives tutorielles Feedback*. Münster: Waxmann.

Poulos, A. & Mahony, M. J. (2008). Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Education, 33*(2), 143-154. https://doi.org/10.1080/02602930601127869

Rieger, M. B. & Roth, J. (2025). Automatisiertes KI-Feedbacksystem zur Unterstützung individueller Lernprozesse. Konzeption und Anwendung im Videoanalysetool ViviAn. In L. Mrohs, J. Franz, D. Herrmann, K. Lindner, T. Staake (Hrsg.), *Digitales Lehren und Lernen in der Hochschule. Strategien-Bedingungen-Umsetzung*. Transcript Verlag.

Sahn, V., Debut, L., Chaumond, J. & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. https://doi.org/10.48550/arXiv.1910.01108

Scherb, C. A., Rieger, M. & Roth, J. (2023). Untersuchung von Usability und Design von Online-Lernplattformen am Beispiel des Video-Analysetools ViviAn. In J. Roth, M. Baum, K. Eilerts, G. Hornung, & T. Trefzger (Hrsg.), *Die Zukunft des MINT-Lernens – Band 1* (S. 105-121). Springer Spektrum. https://doi.org/10.1007/978-3-662-66131-4_6

Varshney, N. & Baral, C. (2022). Model Cascading: Towards Jointly Improving Efficiency and Accuracy of NLP Systems. In Y. Goldberg, Z. Kozareva & Y. Zhang (Hrsg.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (S. 11007–11021). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.756

Walz, M. & Roth, J. (2019). Interventionen in Schülergruppenarbeitsprozesse und Reflexion von Studierenden − Einfluss diagnostischer Fähigkeiten. In A. Frank, S. Krauss & K. Binder (Hrsg.), *Beiträge zum Mathematikunterricht 2019* (S. 1099–1102). WTM-Verlag. https://dx.doi.org/10.17877/DE290R-20735

Wisniewski, B., Zierer, K. & Hattie J. (2020) The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology,* 10:3087. https://doi.org/10.3389/fpsyg.2019.03087

Zhai, X., Chu, X., Wang, M. & He, W. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020: Progress and Trends. *Complexity*. https://doi.org/10.1155/2021/8812542

Zhan, Y., Boud, D., Dawson, P. & Yan, Z. (2025). Generative artificial intelligence as an enabler of student feedback engagement: a framework. *Higher Education Research & Development*, 1–16. https://doi.org/10.1080/07294360.2025.2476513

Anschrift der Verfasser

Tim Lutz
Pädagogische Hochschule Tirol
Institut für Sekundarpädagogik
Pastorstraße 7
A-6010 Innsbruck, Austria
tim.lutz@ph-tirol.ac.at

Marc Bastian Rieger
Pädagogische Hochschule Karlsruhe
Karlsruhe School of Education
Bismarckstraße 10
76133 Karlsruhe
marc.rieger@ph-karlsruhe.de

Jürgen Roth
RPTU Kaiserslautern-Landau
Institut für Mathematik
Didaktik der Mathematik (Sekundarstufen)
Fortstraße 7
76829 Landau
j.roth@rptu.de