Mathis – Lernförderliches und individuelles Feedback mit KI

Studentischer DINI-Wettbewerb 2023/24

Niclas Dern[1], Niels Glodny[2], Marius Jacobs[2], Samuel Leßmann[2]

1 Einleitung

Individuelles Feedback ist ein wesentlicher Bestandteil des Lernprozesses von Schülerinnen und Schülern (Hattie & Timperley, 2007) (Shute, 2008). In vielen Lernumgebungen im Schulbereich sind die Möglichkeiten für detailliertes individuelles Feedback jedoch stark eingeschränkt. Ein erheblicher Teil des Feedbacks, das wir während unserer Schullaufbahn erhalten, erfolgt in Form von Noten und kurzen Kommentaren in Arbeiten oder Prüfungen, die unseren Lernstand bewerten sollen (Bransford, 2000). Gleichzeitig ist individuelles Feedback aus eigener Erfahrung am nützlichsten, wenn es aktiv in den täglichen Lernprozess integriert wird.

In Deutschland hängt der erreichte Bildungsabschluss nach wie vor von der sozialen Herkunft und insbesondere von einem eventuellen Migrationshintergrund ab (Autor:innengruppe Bildungsberichterstattung, 2022, S. 340) (Helbig & Nikolai, 2015) (ifo Institut, 2023).
Die soziale Herkunft bestimmt insbesondere, inwieweit individuelles Feedback außerhalb der Schule verfügbar ist: Sprechen die Eltern beispielsweise nur wenig Deutsch oder haben sie keine allgemeine Hochschulreife, können sie häufig nur wenig oder gar nicht bei Schulaufgaben bzw. diesbezüglichen Fragen helfen.

Vor dem Hintergrund des beschriebenen Chancengefälles und weiterer möglicher positiver Effekte wie besseres Verständnis von Sachverhalten, schnellere Korrektur falscher Herangehensweisen und höherer Lernmotivation Shute (2008) erscheint es sehr sinnvoll, mehr und qualitativ hochwertiges individuelles Feedback innerhalb und außerhalb des Unterrichts bereitzustellen. Allerdings wird dies in den nächsten Jahren nicht allein durch mehr Lehrkräfte zu bewerkstelligen sein — schon jetzt fehlen allein an den allgemeinbildenden Schulen ca. 17.000 Lehrkräfte (Autor:innengruppe Bildungsberichterstattung, 2022). Gleichzeitig gab es in den letzten Jahren bahnbrechende Entwicklungen im Bereich der Künstlichen Intelligenz (KI) (OpenAI et al., 2024) (Gemini Team et al., 2024) (Anthropic, 2024). Mit diesen Technologien scheint es zunehmend möglich, individuelles Feedback breitflächig verfügbar zu machen (Köller et al., 2024) (Middendorf, 2024).

Im folgenden Wettbewerbsbeitrag stellen wir den KI-gestützten Mathe-Tutor Mathis vor, den wir auf Basis der Erkenntnisse aus einer Studie zu individuellem und lernförderlichem KI-Feedback zu geschriebenen Texten entwickelt haben.

[1]Technische Universität München
[2]Ludwig-Maximilians-Universität München

Unser Beitrag ist folgendermaßen aufgebaut:

1. In Kapitel 2 stellen wir das Studiendesign und die Ergebnisse der gemeinsam mit Prof. A. Kinder durchgeführten Studie vor und diskutieren diese.
2. In Kapitel 3 gehen wir auf die Wahl des Fachbereichs Mathematik und der Zielgruppe für Mathis ein. Weiterhin gehen wir auf die Bereitstellung von Feedback unter Nutzung von Mathis und die Ergebnisse unserer Praxistests ein, mit denen wir insgesamt über 150 Schülerinnen und Schüler erreichen konnten.

2 Studie an der FU-Berlin zu personalisiertem KI-Feedback

Im ersten Teil unseres Projektes haben wir in Zusammenarbeit mit Prof. A. Kinder von der Freien Universität Berlin (Fachbereich Erziehungswissenschaft und Psychologie) eine Studie zu personalisiertem KI-Feedback durchgeführt. Dabei haben wir untersucht, inwiefern unter Nutzung des generativen Textmodells „GPT-4“ erstelltes Feedback sich auf den Lernfortschritt von Lehramtsstudierenden bei der Analyse eines Fallbeispiels auswirkt. Zum Vergleich (d. h. als alternatives Feedback) wurde eine Musterlösung verwendet.[1]

Unser Team hat folgende Beiträge zu dieser Studie geleistet:

- Wir haben die technische Umsetzung der Studie übernommen, indem wir eine Web-Plattform entwickelt haben, über die die Studierenden an der Studie teilnehmen konnten (zu finden unter feedback-studie-fu.de).
- Wir haben die automatische Generierung des Feedbacks durch GPT-4 entwickelt, so dass personalisiertes Feedback zu den Analysen der Studierenden zu den Fallbeispielen gegeben werden konnte.
- Zuletzt haben wir bei der statistischen Auswertung der Studie unterstützt.

Studiendesign

Die Studie wurde online durchgeführt und umfasste 270 Lehramtsstudierende, die in zwei Phasen (1. Phase: „Felix“; 2. Phase: „Sarah“) jeweils eine Fallbeschreibung zu einem Schüler erhielten. In diesen Texten wurden die Lernschwierigkeiten von Felix und Sarah in verschiedenen Fächern qualitativ sowie die Ergebnisse in standardisierten Tests beschrieben. Die Aufgabe der Studierenden bestand dann darin, jeweils ein Fördergutachten zu analysieren, das eine fiktive Lehrkraft für Felix und Sarah verfasst hatte. Konkret sollten sie argumentieren, ob sie dem Fördervorschlag zustimmen oder ihn ablehnen würden. Eine Übersicht über das Studiendesign gibt Abbildung 2.1.

[1] Ein wissenschaftliches Poster zu dieser Studie wurde bereits für die SIG 20 & 26 Conference 2024 an der Humboldt-Universität zu Berlin akzeptiert. Ein wissenschaftlicher Artikel, bei dem wir Mitverfasser sind, wird voraussichtlich Ende des Jahres veröffentlicht.

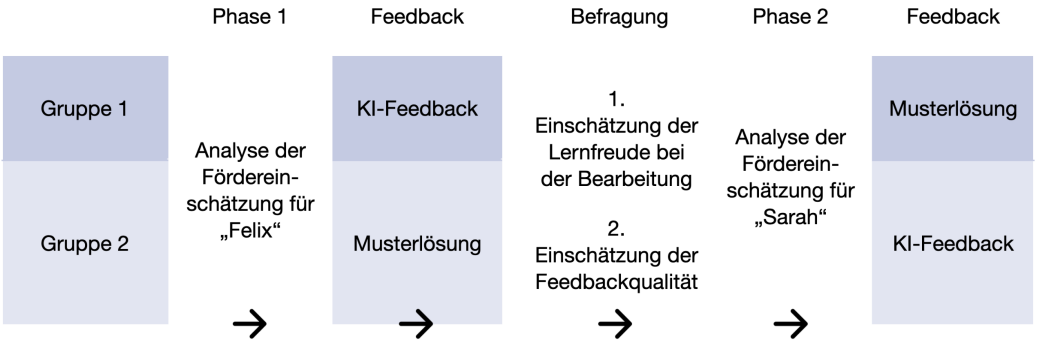

Abbildung 2.1: Design der KI-Feedback-Studie von Prof. A. Kinder

Um zu testen, inwiefern sich Feedback von GPT-4 auf die Analyse eines zweiten Fallbeispiels gleicher Art auswirkt, wurden die Studierenden zufällig in zwei Gruppen eingeteilt.

- Die erste Gruppe erhielt in der ersten Phase (nach einer Wartezeit von 5–10 s) von GPT-4 generiertes personalisiertes Feedback zu ihrer Analyse des Fallbeispiels.
- Im Gegensatz dazu erhielt die zweite Gruppe nach einer künstlichen Wartezeit von 5–10 s lediglich die Musterlösung.

Anschließend füllten beide Gruppen einen Fragebogen aus, in dem auf zwei Likert-Skalen (Likert, 1932) mit jeweils vier Items sowohl die Qualität des Feedbacks als auch die Lernfreude in Bezug auf das Feedback abgefragt wurden.

In der zweiten Phase erhielten dann beide Gruppen den entsprechenden Text und Fördervorschlag für den Fall Sarah, den sie ebenfalls analysierten, um den Einfluss der Musterlösung und des KI-Feedbacks nach der ersten Phase auf die schriftlichen Analysen in der zweiten Phase untersuchen zu können.

Ergebnisse

In der statistischen Auswertung wurden sowohl die Befragung der Studierenden zur Qualität des Feedbacks als auch die Analysen der Fallbeispiele durch die Studierenden genauer untersucht.

Für die Befragung wurde sowohl für die Feedbackqualität als auch für die Lernfreude der Durchschnitt der Likert-Skala (1–4) gebildet (siehe Abb. 2.2) und zur Prüfung der Signifikanz der Abweichung ein statistischer Test durchgeführt.[2] Hinsichtlich der Lernfreude konnte kein signifikanter Unterschied zwischen den beiden Gruppen festgestellt werden. Dahingegen war die Bewertung der Qualität des Feedbacks in der Gruppe, die das KI-basierte Feedback erhalten hatte, signifikant besser als die Studierenden der Gruppe, die nur die Musterlösung erhalten hatte.

[2] Da die Ergebnisse der Befragung nicht normalverteilt waren, wurde hier der nichtparametrische Wilcoxon-Mann-Whitney-Test verwendet.

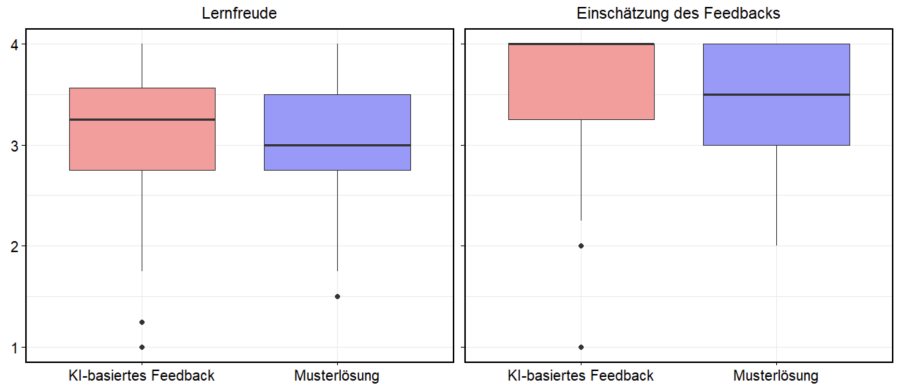

Abbildung 2.2: Box-Plots für Lernfreude und Einschätzung des Feedbacks für beide Gruppen (1–4). Während die Einschätzung zur Lernfreude keinen signifikanten Unterschied zwischen den Gruppen aufwies ($p = 0{,}138, n = 270$), war die Einschätzung der Feedbackqualität in der Gruppe mit KI-Feedback in der ersten Phase signifikant besser ($p = 0{,}004, n = 270$).

Darüber hinaus wurden die von den Studierenden verfassten Analysen von zwei geschulten Hilfswissenschaftlerinnen von Prof. A. Kinder unabhängig begutachtet. Diese untersuchten mithilfe von vorgegebenen Kriterien, wie gut die Studierenden ihre Entscheidung begründet hatten. Das Ergebnis dieser Auswertung war, dass die Studierenden, die in der ersten Phase das KI-Feedback erhalten hatten, in der zweiten Phase signifikant besser argumentierten als die Studierenden der Gruppe, die nur die Musterlösung erhalten hatten (siehe Tabelle 2.1).

Die bessere Argumentationsqualität der Gruppe mit KI-Feedback und die signifikant bessere Bewertung des KI-Feedbacks in der Befragung zeigen das Potenzial von personalisiertem KI-Feedback in der Lehre. Gleichzeitig haben wir in der Studie nur einen Aufgabentyp abgedeckt, weshalb die Effekte solcher und anderer Tools weiter untersucht werden sollten.

Für uns als Team war diese Studie insbesondere auch ein erster Test, der gezeigt hat, dass es technisch möglich ist, lernförderliches Feedback durch KI-Modelle zu generieren.

Gruppe	Argumentation Text „Felix“	Argumentation Text „Sarah“
KI-Feedback	1.86	**2.18**
Musterlösung	1.80	1.82

Tabelle 2.1: Durchschnittliche Bewertung der Argumentationsqualität der Texte (0–5 Punkte) durch die beiden Hilfswissenschaftlerinnen. Die Argumentationsqualität im zweiten Text „Sarah“ war signifikant besser in der Gruppe, die KI-Feedback zum ersten Text erhalten hatte ($p = 0{,}0234, n = 270$).

3 KI-basiertes Feedback für Mathematikaufgaben auf Abiturniveau

Ideenfindung

Nachdem wir in der Studie erste Erfahrungen mit der Generierung und Rezeption von KI-basiertem, individuellem Feedback gemacht hatten, waren wir motiviert, eine skalierbare Lösung zu entwickeln, die entsprechendes Feedback in den Schulalltag integriert.

Auswahl des Fachbereichs. Nach einiger Diskussion entschieden wir uns dafür, Mathematik als Fachbereich für die Entwicklung unseres Feedbacksystems zu wählen. Unsere Gründe für die Wahl des Fachbereichs Mathematik waren insbesondere:

- Sowohl wissenschaftliche Analysen (Dohmen, Erbes, Fuchs & Günzel, 2008) als auch unsere eigenen Erfahrungen[1] zeigen, dass Mathematik ein Fach ist, in dem Schülerinnen und Schüler besonders häufig Nachhilfe benötigen.
- Mathematik ist ein Fach, in dem Lösungen bzw. Lösungsschritte i. d. R. richtig oder falsch sind, was – so hofften wir – die automatische Korrektur erleichtert.
- Der Sprung von Mathematik zu anderen naturwissenschaftlichen Fächern ist eher klein, sodass wir bei Erfolg unser System auch auf andere Fächer ausweiten können.
- Wir studieren Mathematik bzw. Informatik und fühlen uns daher in diesen Bereichen am besten in der Lage, die Feedbackqualität zu beurteilen.

Zielgruppe. Als Zielgruppe für unser Feedbacksystem wählten wir Schülerinnen und Schüler der gymnasialen Oberstufe, die sich auf ihr Mathe-Abitur vorbereiten. Die Hauptgründe dafür waren, dass das Mathe-Abitur deutschlandweit relativ einheitlich ist, so dass wir unser System für eine breite Zielgruppe entwickeln konnten, und dass Schülerinnen und Schüler in der Oberstufe bereits ein hohes Maß an Selbstständigkeit und Eigenverantwortung besitzen, was die Nutzung unseres Systems vereinfacht. Wir haben uns dafür entschieden, unsere Lösung direkt den Schülerinnen und Schüler zur Verfügung zu stellen, um schnell herauszufinden, wie hilfreich unser Feedback ist.

Bereitstellung von Feedback mit Mathis

Der Prozess zur Bereitstellung von Feedback mit Mathis sieht wie folgt aus:

1. Wir oder eine Lehrkraft wählen eine Klausur, die von Schülerinnen und Schülern bearbeitet werden soll.
2. Die Schülerinnen und Schüler lösen die Klausur entweder auf Papier oder digital und laden ihre Lösungen auf einer von uns bereitgestellten Plattform hoch (unter Verwendung eines Codes, der die Pseudonymität der Schülerinnen und Schüler gewährleistet[2]).

[1] Samuel Leßmann gibt Nachhilfe bei der Studenteninitiative für Kinder e. V., die mit Abstand die meisten Anfragen für Mathematiknachhilfe erhält.

[2] Entsprechende Codes werden bei Einsatz in der Schule durch die Lehrkraft verteilt.

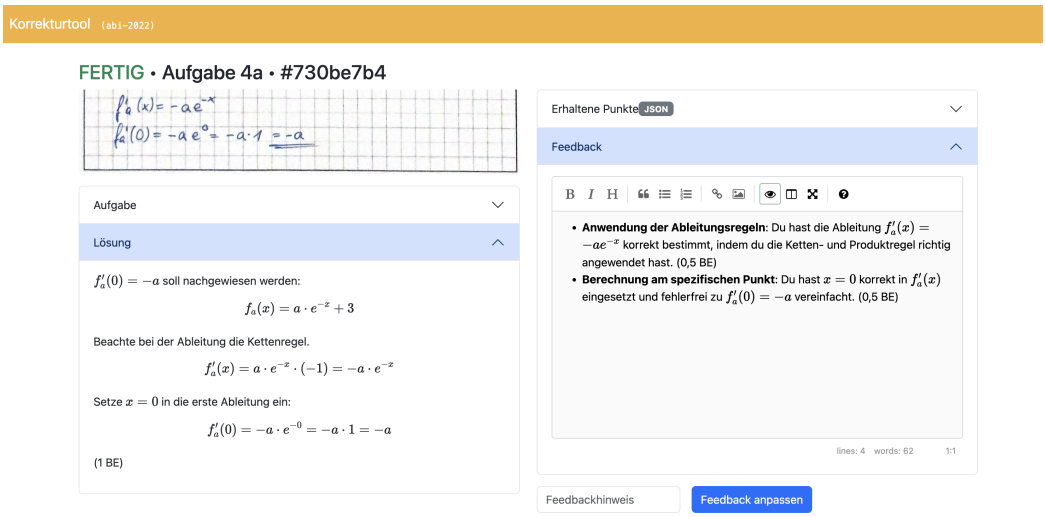

Abbildung 3.1: Lehrkräfte können Einblick in den Korrekturprozess nehmen und das Feedback einfach anpassen.

3. Die Lösungen werden automatisch korrigiert und ein erstes Feedback wird generiert.
4. Lehrkräfte können das Feedback einsehen und bei Bedarf anpassen, wie in Abbildung 3.1 dargestellt.
5. Schülerinnen und Schüler können das Feedback herunterladen und einsehen (unter Verwendung ihres Codes).

Die jeweiligen Schritte in diesem Prozess wurden von uns so konzipiert, dass der (zusätzliche) Aufwand für die beteiligten Personen möglichst gering ist: So können die Lernenden die Klausur – falls gewünscht – beispielsweise weiterhin klassisch auf Papier lösen und einfach mit dem Smartphone einscannen. Für die Lehrkräfte gibt es dabei keinerlei Einschränkungen bei der Erstellung der Klausur, so dass auch Textaufgaben und Skizzen problemlos korrigiert werden können. Es ist jedoch in der praktischen Anwendung von Vorteil, wenn jede Teilaufgabe einen klar erkennbaren Bereich zum Lösen enthält, wofür wir eine LaTeX-Vorlage entwickelt haben.

Der Korrekturprozess. Nachdem die Lösungen auf unsere Plattform hochgeladen wurden (entweder durch die Schülerinnen und Schüler selbst oder durch die Lehrkraft), werden sie automatisch korrigiert. Die Korrektur erfolgt in mehreren Schritten (siehe auch Abbildung 3.2):

1. Ausschnitte der Abgabe (d. h. Screenshots davon) werden Teilaufgaben zugeordnet: Die Erstellung der Screenshots muss in der aktuellen Version von einer Person durchgeführt werden. Wenn diese in der Reihenfolge der Teilaufgaben erstellt werden, erfolgt die weitere Zuordnung automatisch.
2. Die Abgabe wird eingelesen und der Text erkannt: Das für die Texterkennung verwendete KI-Modell kennt eine Musterlösung und den mathematischen Hintergrund der Aufgabe, so dass es auch z. B. auch unleserlich geschriebene Zahlen in der Regel korrekt

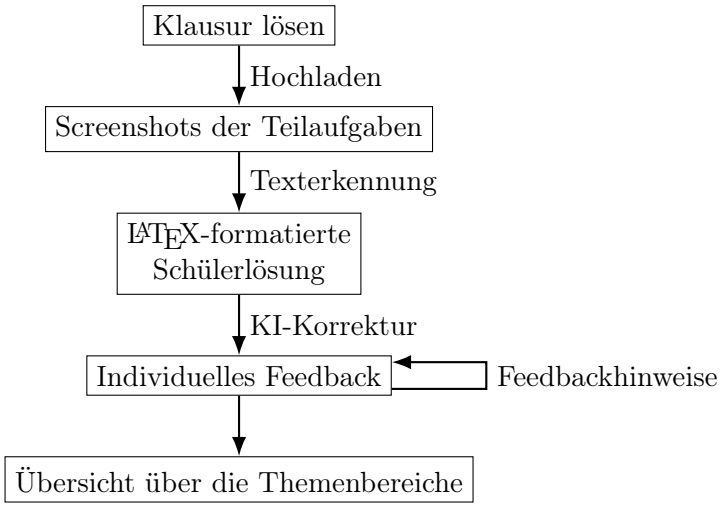

Abbildung 3.2: Überblick über den Korrekturprozess

interpretieren kann. Auch Skizzen, Zeichnungen und Randnotizen können weitgehend korrekt interpretiert werden.

3. Erstellung einer ersten Korrektur: Ein weiteres KI-Modell erstellt eine erste Korrektur. Dazu stehen dem Modell eine Musterlösung und ein Erwartungshorizont zur Verfügung. So wird sichergestellt, dass die Korrektur fachlich korrekt und einheitlich ist.
4. Nach der ersten Korrektur wird das Ergebnis von der Lehrkraft überprüft: Diese kann Mathis (auch mehrfach) Hinweise zur Anpassung der Korrektur geben, die Mathis dann entsprechend umsetzt oder die Korrektur direkt manuell anpassen.
5. Abschließend wird ein Feedback-Dokument (PDF) erstellt, das die Schülerinnen und Schüler herunterladen können.

Für Details zur technischen Umsetzung siehe Anhang.

Aufbau des Feedbacks. Das Feedback für die Lernenden ist in zwei Teile gegliedert: einen Überblick und eine detaillierte Korrektur.

Die Übersicht enthält Grafiken darüber, in welchen Themenbereichen die Person besonders gut war oder noch Potenzial hat. Für letztere werden in der Übersicht themenspezifische Übungsaufgaben von der kostenlosen Lernplattform Serlo empfohlen.

Die detaillierte Korrektur enthält für jede Teilaufgabe die erreichte Punktzahl und ein detailliertes Feedback zur Abgabe, falls nicht die volle Punktzahl erreicht wurde. Bei Fehlern in der Bearbeitung wird insbesondere erläutert, wo diese genau lagen und wie diese Schritte – unter Berücksichtigung des gewählten Lösungsweges – korrekt hätten durchgeführt werden müssen. Eine Beispielkorrektur inklusive Übersicht und Feedback ist im Anhang „Beispielkorrektur“ zu finden.

Praxistests

Erste Tests in Mathe-Oberstufenkursen. Wir haben die erste Version von Mathis zunächst in Zusammenarbeit mit zwei Lehrkräften von Mathematikkursen der Jahrgangsstufe 12 an zwei bayerischen Gymnasien getestet. Die für zwei Stunden konzipierten Probeklausuren wurden in Absprache mit den jeweiligen Lehrkräften erstellt.

So haben wir von den beiden Kursen insgesamt 20 Klausuren (freiwillige Abgaben) erhalten, die wir mithilfe von Mathis korrigiert haben. Die beiden Lehrkräfte haben das Feedback als hilfreich empfunden und waren von der hohen Qualität überrascht. Uns sind während der Tests jeweils einige Schwachstellen von Mathis aufgefallen. Insbesondere haben wir noch relativ lange gebraucht, um Klausuren unter Nutzung von Mathis zu korrigieren (ca. 1,5 Stunden pro Klausur). Wir haben daraufhin den Korrekturprozess beschleunigt, unter anderem durch die Möglichkeit, nachträgliche Feedbackhinweise als Freitext zu geben.

Einsatz in einem Mathe-Crashkurs. Einige Wochen nach unserem ersten Test mit einem Mathekurs haben wir unser System dann (mit einigen Modifikationen) in einem einwöchigen Abivorbereitungskurs in den Osterferien eingesetzt.[3] Mit Hilfe von Mathis konnte dort im Vorfeld ein Einstufungstest für die knapp 100 Kursteilnehmenden durchgeführt und effizient korrigiert werden. Dadurch war es möglich, die Kursteilnehmer nach Themengebiet und Lernfortschritt in Kleingruppen einzuteilen und während des Kurses gezielter auf die einzelnen Schüler einzugehen.

Insbesondere sind wir im Rahmen dieses Tests erstmals auf die Idee gekommen, den Schülerinnen und Schülern Übungsaufgaben aus alten Abiturklausuren und von Serlo zu empfehlen, um so gezielt an ihren Schwächen zu arbeiten. Durch die große Anzahl an korrigierten Klausuren konnten wir auch eine ganze Reihe an Schwachstellen des Korrektursystems finden. Zu diesem Zeitpunkt waren wir mit ca. 75 % des generierten Feedbacks so zufrieden, dass wir es nicht mehr verändert haben. Bei diesem Test kamen wir auch auf die Idee einen *Prompt-Chaining*-Ansatz (siehe Kapitel 4) zu verwenden, der später die Qualität der Korrekturen deutlich verbesserte.

Veröffentlichung für Abiturienten. Im letzten Schritt war unser Ziel, Mathis noch breiter zu testen und für Schülerinnen und Schüler in der Abiturphase zugänglich zu machen. Dies haben wir am 16. April 2024 mit der Veröffentlichung der Website abimitmathis.de getan. Dort konnten Schülerinnen und Schüler entweder eine kostenlose Testkorrektur (ein Analysis-A-Teil aus dem Abitur 2022), oder die Korrektur eines halben bzw. eines vollständigen Abiturs (jeweils 2022 oder 2023) erwerben. Der Nutzungsablauf für die Schülerinnen und Schüler ist in Abbildung 3.3 dargestellt.

Der Preis für die Korrektur einer vollständigen Abiturklausur betrug 4,95 € und für eine halbe Abiturklausur 2,95 €. Wir haben uns nach einiger Diskussion bewusst für diese Preise entschieden, um zu testen, wie sich die Nachfrage im Bezug zu den Kosten verhält. Bei technischen Korrekturkosten von damals etwa 10 € (heute etwa 5 €) pro Abitur, haben wir also ungefähr die Hälfte unserer Kosten an die Lernenden weitergegeben.

[3] `https://www.dermatheabicrashkurs.de/`

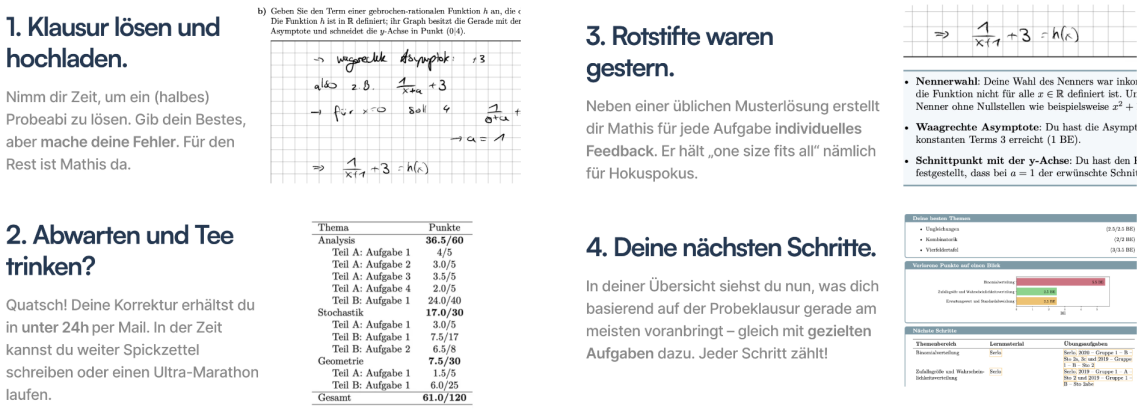

Abbildung 3.3: Der Nutzungsablauf von Mathis wie auf unserer Website abimitmathis.de dargestellt.

Um auf Mathis aufmerksam zu machen, wurden wir in den rund drei Wochen bis zum Mathe-Abitur auf verschiedene Weise aktiv: Wir sprachen Schülerinnen und Schüler in Bibliotheken an und verteilten dort Flyer. Außerdem versuchten wir mit authentischen und witzigen Inhalten auf Instagram und TikTok auf Mathis aufmerksam zu machen und teilten unsere Website in verschiedenen Foren und Gruppen. Insgesamt haben sich in dieser Zeit ca. 30 Schülerinnen und Schüler für eines der drei Angebote entschieden (nur in drei Fällen handelte es sich um eine kostenpflichtige Version), von denen knapp zehn ihre Klausur auch abgegeben haben.

Unserer Einschätzung nach, basierend auf Gesprächen mit Schülerinnen und Schülern, war ein Grund, warum wir nicht noch mehr Abgaben hatten, dass die Schülerinnen und Schüler mehrere Abiturprüfungen in kurzer Abfolge hatten, so dass sich viele erst wenige Tage vor der Prüfung auf Mathematik konzentrierten. Zudem war die Nutzung unseres noch unbekannten Angebots kurz vor der Prüfung für einige mit zu viel Unsicherheit verbunden.

Gleichzeitig waren die Rückmeldungen der Nutzerinnen und Nutzer, die wir erhalten haben, durchweg positiv. Auch waren wir am Ende mit 80–90 % des automatisch generierten Feedbacks so zufrieden, dass wir daran keine weiteren Änderungen vorgenommen haben. Dadurch konnten wir detailliertes Feedback für ein komplettes Abitur in unter einer Stunde erstellen.

4 Fazit

Wir haben mit diesem Wettbewerbsbeitrag gezeigt, dass es mit den neuesten Entwicklungen im Bereich der Künstlichen Intelligenz möglich ist, qualitativ hochwertiges und individuelles Feedback sowohl zu Texten als auch zu mathematischen Lösungen zu erstellen.

Die gemeinsam mit Prof. A. Kinder durchgeführte Studie hat dabei gezeigt, dass personalisiertes KI-Feedback das Potenzial hat, den Lernfortschritt von Studierenden signifikant zu verbessern. Basierend auf diesen Erkenntnissen und der Erfahrung, dass der Bedarf an

individuellem Feedback gerade im Fach Mathematik sehr hoch ist, haben wir mit Mathis ein Korrektursystem entwickelt, das speziell an die Anforderungen von Mathe-Abituraufgaben angepasst ist.

Unsere Praxistests mit Mathis haben gezeigt, dass entsprechendes individuelles Feedback auch bei den Schülerinnen und Schülern auf Interesse stößt. Darüber hinaus haben wir festgestellt, dass Mathis in einer realen Umgebung bereits 80–90 % der Lösungen richtig einschätzen kann und somit solche Ansätze bereits kurzfristig das Potenzial haben, den Zeitaufwand für die Erstellung von lernförderlichem, individuellem Feedback für Lehrkräfte deutlich zu reduzieren.

Zukünftig streben wir an, Mathis weiter zu verbessern und zugänglicher zu machen, z. B. durch die Integration in bestehende Lernsysteme. Wir wollen damit einen Beitrag zu einem individuelleren Bildungssystem leisten, Lehrkräfte entlasten bzw. ihnen mehr Zeit für den direkten Austausch mit Schülerinnen und Schülern geben und gleichzeitig die Chancengleichheit im Bildungsbereich fördern.

Relevante Aspekte, die es dafür in Zukunft zu klären gilt, umfassen insbesondere:

- Die weitere Verbesserung der Qualität des lernförderlichen Feedbacks und die effektive Integration von Mathis in bereits bestehende Lernplattformen.
- Die Entwicklung unkomplizierter Nutzungsmöglichkeiten für Schulen und Lehrkräfte, damit KI-basierte Korrektursysteme wie Mathis eingesetzt werden können, ohne dass langwierige bürokratische Prozesse an jeder Schule durchlaufen werden müssen.
- Ausarbeitung einer Finanzierungsstruktur für Mathis bspw. durch Kultusministerien, die einen gleichberechtigten Zugang für alle ermöglicht und damit Bildungsgerechtigkeit fördert.

Danksagung

Unser herzlicher Dank gilt Prof. A. Kinder, mit der wir gemeinsam die vorgestellte Studie durchführen konnten, Frau Quarg und Herrn Pattusch, die die Durchführung der Praxistests in ihren Mathe-Kursen ermöglicht haben, Lennart Redl und Hafez Pirzadeh, die den Einsatz von Mathis bei ihrem Mathe-Abi-Crashkurs ermöglicht haben, Magnus Saurbier, der uns bei der technischen Umsetzung des Promptings unterstützt hat, Magnus Saurbier, Martin Ehmler und Paskal Thomas, die uns bei der Bewertung von Klausuren während der Praxistests geholfen haben, sowie Cederik Höfs, Lucy Strüfing, Nils Reichardt, Lucas Scherberger, Britt Besch, Pia Leßmann, Kilian Pschierer, Maria Matveev, Marvin von Hagen, Adrian Cipriani, Peter Spitznagel, Stephan Kaspar, Georg Petermann und vielen weiteren die uns im gesamten Prozess hilfreiches Feedback gegeben und so das gesamte Projekt erst ermöglicht haben.

Literaturverzeichnis

Anthropic. (2024). *Introducing the next generation of Claude.* Zugriff auf `https://www.anthropic.com/news/claude-3-family`

Autor:innengruppe Bildungsberichterstattung. (2022). *Bildung in Deutschland 2022 – Ein indikatorengestützter Bericht mit einer Analyse zum Bildungspersonal.*

Bransford, J. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition (2000).*

Dohmen, D., Erbes, A., Fuchs, K. & Günzel, J. (2008). *Was wissen wir über Nachhilfe? Sachstand und Auswertung der Forschungsliteratur zu Angebot, Nachfrage und Wirkungen.* (A. Knauf, Hrsg.). Köln: FIBS. Zugriff auf `https://tu-dresden.de/die_tu_dresden/fakultaeten/erzw/erzwibf/sp/forschung/ganztagsschule/FIBS.pdf`

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., ... Vinyals, O. (2024). *Gemini: A family of highly capable multimodal models.*

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77* (1), 81-112. Zugriff auf `https://doi.org/10.3102/003465430298487` doi: 10.3102/003465430298487

Helbig, M. & Nikolai, R. (2015). *Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949.* Bad Heilbrunn: Klinkhardt.

ifo Institut. (2023, April 18). *Chancenmonitor 2023: Bildungschancen hängen stark vom Elternhaus ab.* Pressemitteilung.

Köller, O., Thiel, F., van Ackeren-Mindl, I., Anders, Y., Becker-Mrotzek, M., Cress, U., ... Stanat, P. (2024). *Large Language Models und ihre Potenziale im Bildungssystem. Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz.* Bonn: SWK. doi: https://doi.org/10.25656/01:28303

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55–55.

Middendorf, W. (2024). *Zur Möglichkeit digital gestützter Leistungsbewertung mithilfe Künstlicher Intelligenz in der Schulpraxis; On the possibility of digitally supported performance assessment using artificial intelligence in school practice.* doi: https://doi.org/10.25656/01:28739

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., ... Zoph, B. (2024). *Gpt-4 technical report.*

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78* (1), 153-189. Zugriff auf https://doi.org/10.3102/0034654307313795 doi: 10.3102/0034654307313795

Anhang

Technische Details zu Mathis

Entwickelt haben wir eine Abgabeplattform für Schülerinnen und Schüler, das eigentliche KI-basierte Korrektursystem und eine Plattform für Lehrpersonal, um den Korrekturprozess zu überwachen und beeinflussen zu können.

Die Abgabeplattform. Die Abgabeplattform ist eine in Python geschriebene Webanwendung, die es Schülerinnen und Schülern ermöglicht, ihre Lösung hochzuladen und später ihr Feedback herunterzuladen. Sowohl zum Hoch- als auch Herunterladen verwenden die Schülerinnen und Schüler einen zuvor ausgehändigten Korrekturcode, sodass die gesamte Verarbeitung der Daten pseudonymisiert erfolgen kann. Zum zusätzlichen Schutz der Schülerdaten werden die Abgaben auf einem von uns verwalteten und in Deutschland gehosteten V-Server gespeichert und verarbeitet.

Das Korrektursystem. Der erste Schritt der Korrektur ist das Einlesen der Abgaben und die damit verbundene Texterkennung (OCR). Dies geschieht für jede Teilaufgabe einzeln. Um die Fehlerhäufigkeit gering zu halten, kombinieren wir die Texterkennungssysteme von Mathpix und GPT-4-Vision. Dabei ist insbesondere von Bedeutung, dass GPT-4-Vision den Kontext der Aufgabe erhält und so auch bei unleserlich geschriebenen Zahlen und Formeln gute Ergebnisse liefert. Als Ergebnis dieses OCR-Prozesses erhalten wir die Abgabe im LaTeX-Format für die weitere Korrektur.

Die eigentliche Korrektur erfolgt ebenfalls auf Basis eines KI-Modells. Während unserer Praxistests und für die Beispiele in der Beispielkorrektur wurde GPT 4-Turbo von OpenAI verwendet. Das konkret verwendete Modell lässt sich jedoch einfach austauschen.

Ein Kernpunkt unserer Lösung ist die Entwicklung einer Abfolge von Anweisungen (*Prompts*), mit denen wir das KI-Modell dazu bringen, möglichst gute Korrekturen zu erstellen (sog. *Prompt-Chaining*). Schließlich verwenden wir die folgenden Schritte:

1. Anpassung der Musterlösung an den individuellen Lösungsweg: Hat die Schülerin oder der Schüler einen anderen Lösungsweg als die Musterlösung gewählt, so wird für diesen Lösungsweg eine neue Musterlösung erstellt.
2. Aufteilung der Musterlösung in Teilschritte: Die angepasste Musterlösung wird in einzelne Rechenschritte unterteilt und diesen wird jeweils eine zu erreichende Punktzahl zugewiesen.[1]
3. Die Bearbeitung des Schülers wird in Schritte aufgeteilt, mit der Musterlösung abgeglichen und es werden Punkte vergeben: Hier wird die Korrektheit der Schritte überprüft und Fehler sowie Folgefehler erkannt.

[1] Ein Aspekt, der dem Modell besonders geholfen hat, die Punkte richtig zu verteilen, war, dass es bei jedem Schritt auch die bisher erreichte Gesamtpunktzahl ausgeben sollte.

4. Um dem Schüler bzw. der Schülerin eine Übersicht über seine bzw. ihre Stärken und Schwächen zu geben, werden alle Schritte einem Themengebiet zugeordnet.[2]

5. Aus den bepunkteten Schritten wird ein Feedbacktext generiert.

Die einzelnen Anweisungen profitieren insbesondere von der Technik des *Chain-of-Thought-Prompting*. Hierbei werden die einzelnen Schritte intern nochmals in Teilschritte unterteilt. Dies kann beispielsweise zunächst der Vergleich des Lösungswegs des Schülers mit der Musterlösung sein, dann die Erstellung eines Plans für die neue Musterlösung und schließlich das Schreiben der neuen Musterlösung.

Abschließend erstellt das Korrektursystem die Übersichtsgrafiken und stellt die Feedbacktexte in einem Feedbackdokument (als PDF) zusammen, das an die Klausur angelehnt ist. Dieses Dokument wird mit Python und LaTeX erstellt.

Die Lehrpersonen-Plattform. Die Lehrpersonen-Plattform (siehe Abbildung 3.1) ermöglicht es, den Korrekturprozess zu starten und die einzelnen Schritte zu überwachen und gegebenenfalls zu korrigieren. Die Anforderungen an das System sind:

1. Die KI-Ausgaben sollen im Detail einsehbar und veränderbar sein, damit eventuelle Fehler korrigiert werden können. Insbesondere sollen die OCR-Ausgabe und der Feedbacktext auch ohne technische Kenntnisse leicht korrigierbar sein.

2. Die Feedbackdaten der Schülerinnen und Schüler, sowie deren Abgaben müssen sicher verwaltet werden.

3. Es sollte möglich sein, gleichzeitig an verschiedenen Aufgaben zu arbeiten. Dies ist besonders wichtig, da die KI-Schnittstelle für die Korrektur einer einzelnen Teilaufgabe ca. eine Minute benötigt und man daher manchmal den Korrekturprozess für mehrere Aufgaben gleichzeitig starten möchte.

4. Am Ende soll ein Feedback-Dokument für die Schülerinnen und Schüler erstellt werden können.

Umgesetzt haben wir dies mit einer weiteren Python-Webanwendung, die nur intern zugänglich ist. Diese ist ebenfalls über einen von uns verwalteten und in Deutschland gehosteten V-Server bereitgestellt.

Die Benutzerfreundlichkeit (1) haben wir im Entwicklungsprozess deutlich verbessert, indem sowohl für Abgaben als auch für den Feedbacktext direkt auf der Lehrpersonal-Plattform eine LaTeX-Vorschau zur Verfügung steht. Durch die Möglichkeit, Feedbackhinweise an das Korrektursystem zu geben, wurde das Anpassen von Feedback ebenfalls noch einmal deutlich einfacher. Beim Test mit dem Mathe-Crashkurs haben uns einige Tutorinnen und Tutoren geholfen, die schnell in die Plattform eingearbeitet werden konnten.

Für die Datenverwaltung (2) haben wir eine SQLite-Datenbank implementiert. So können wir die Daten gut strukturiert speichern, sichern und die Ergebnisse für die Feedback-Dokumente verwenden. Die Datenbank enthält auch die Aufgabenstellungen und Musterlösungen, so

[2] Diese müssen nicht unbedingt dem Themengebiet der Aufgabe entsprechen. Es kann beispielsweise bei einer Aufgabe aus dem Analysisteil ein lineares Gleichungssystem falsch gelöst werden.

dass die Lehrpersonen und das KI-System diese bei der Erstellung des Feedbacks verwenden können.

Den gleichzeitigen Zugriff von mehreren Korrigierenden auf mehrere Aufgaben (3) gewährleisten wir über ein Task-Queue-System (Celery und Redis). So kann das System das Feedback für viele Aufgaben gleichzeitig generieren, was den Korrekturprozess erheblich beschleunigt.

Schlussendlich können die Feedback-Dokumente generiert und heruntergeladen werden (4). Dabei werden zuerst Statistiken über die in der Datenbank enthaltenen Korrekturergebnisse berechnet (beispielsweise verlorene Punkte pro Themengebiet) und mit Hilfe der Softwarebibliothek *Matplotlib* Grafiken erzeugt. Die Dokumente werden zunächst als LaTeX-Code generiert, dann kompiliert und zum Download angeboten.