Michael B. Strecker, Susanne Hähnchen, Martin Heidebach, Marie Herberger, Simon Alexander Nonn, Sarah Großkopf, Gökhan Erol, Menaf Erol, Ilja Garber, Constantin Höhmann, Enci Huang, Clemens Hufeld, Louisa Zachmann[1]

KI-Unterstützung und Rohpunkteschemata: Die Zukunft der juristischen Klausurkorrektur?

Übersicht

1 Die AutorInnen danken Tobias Erhan, Christoph Schramm und Patricia Sommer für die Analyse des formativen Feedbacks in Phase 1 (Bielefeld), Andreas Bartholomä, Melanie Lang und den zahlreichen beteiligten Mitgliedern von MLTech für die Unterstützung der Durchführung an der LMU München sowie den beteiligten Studierenden für die Bereitschaft, ihre Klausur digital zu schreiben. Zusätzlicher Dank geht an Sven Dümpelmann und Felix Kaiser für den Input bei der Teilnahme an einer digitalen Besprechung des DigitalProjekts, sowie Hedwig Vogel und Hannah Morlin Wigand für das Korrekturlesen der Veröffentlichung.

A. Einleitung

„Über 80 % der Befragten (81,63 %) halten die Bewertung juristischer Klausuren nicht für objektiv. […] Die Bewertung juristischer Klausuren wird von den meisten Befragten als hochgradig subjektiv, intransparent und teilweise willkürlich empfunden."[2]

Rund fünf Jahre lang legen Jurastudierende fast ausschließlich schriftliche Prüfungen ab.[3] Während des Studiums und für die Zulassung zum Ersten Juristischen Staatsexamen muss je nach Studienordnung der jeweiligen Fakultät eine bestimmte Zahl von Klausuren und ggf. Hausarbeiten[4] bestanden werden. Die Endnote setzt sich aber nicht – wie in vielen anderen Studiengängen – aus dem Mittel aller erbrachten Leistungen zusammen. In die Gesamtwertung des Staatsexamens fließen nur zu 30 % die Schwerpunktbereichsprüfung am Ende des universitären Studiums und zu 70 % die Prüfung bei den Justizprüfungsämtern ein. Zu letzterer gehören neben einer mehrstündigen mündlichen Prüfung[5] – je nach Bundesland – sechs bis sieben Klausuren, die wie in kaum einem anderen Berufszweig den gesamten weiteren Lebensweg determinieren. Noch wichtiger ist das Zweite Staatsexamen zum Abschluss des Referendariats, wobei hier ausschließlich die Klausuren und die mündliche Prüfung unter staatlicher Aufsicht für die Gesamtnote zählen.[6]

Der Bedeutung der Noten im Studium und in den Examina zum Trotz ist ihr Zustandekommen alles andere als klar und objektiv nachvollziehbar. Standards zur Bewertung von juristischen Prüfungsleistungen beschränken sich auf vage Beschreibungen von Notenstufen. Diese sind weder klar, noch wird die vorhandene Skala ausgeschöpft: „eine Leistung, die trotz ihrer Mängel durchschnittlichen Anforderungen noch entspricht" ist etwa mit 4 bis 6 Punkten auf einer Notenskala bis 18 Punkte (sog. summatives Feedback) zu bewerten und damit gerade bestanden.[7] Ein „sehr gut" wird im Ersten Examen von 0,4 % der PrüfungskandidatInnen erreicht, im Zweiten Examen sogar nur von 0,1 %.[8] Einheitliche Kriterien, wie viel Prozent der Aufgabe „richtig" bearbeitet sein müssen, um zu bestehen, gibt es nicht. Auch die Frage, ob die Verteilung der oberen Hälfte der Notenskala linear oder anderweitig erfolgt, ist offen. Qualitative Maßstäbe, wie z. B. Punktetabellen mit „Rohpunkten"[9] für Teilaspekte der Prüfungsleistung, die später in eine juristische Note umgerechnet werden, kommen

2 Bundesverband rechtswissenschaftlicher Fachschaften e.V. (BRF), Sechste bundesweite Absolvent:innenbefragung 2025, https://bundesfachschaft.de/wp-content/uploads/2025/10/Sechste-Absolventinnenbefragung-2025.pdf, S. 100 f. Alle Online-Quellen wurden zuletzt abgerufen am 17.10.2025.

3 Bundesamt für Justiz (BfJ), Ausbildungsstatistik Juristenausbildung 2023, Stand 11.06.2025, S. 20: durchschnittlich 10,3 Semester, https://www.bundesjustizamt.de/SharedDocs/Downloads/DE/Justizstatistiken/Juristenausbildung_2023.pdf.

4 Die Verwendung von Hausarbeiten in Zeiten von KI-Nutzung bei deren Erstellung ist jedoch weniger sinnhaft als bisher, vgl. *Sebastian Dötterl*, Der Elefant im Raum – Generative Künstliche Intelligenz und die Zukunft der juristischen Ausbildung, Ordnung der Wissenschaft 3/2025, S. 155 ff. (164) mwN, https://ordnungderwissenschaft.de/wp-content/uploads/2025/06/Doetterl.pdf. Tatsächlich ist auch zu beobachten, dass immer mehr Fakultäten Hausarbeiten abschaffen – jedenfalls im Grund- und Hauptstudium, weniger in den Schwerpunktbereichen.

5 Die Länge der Prüfung ergibt sich aus der Anzahl der an der Prüfung teilnehmenden Prüflinge; die inhaltliche Prüfung fällt für den Einzelnen somit deutlich kürzer aus.

6 § 5 DRiG spricht von der ersten Prüfung (diese wiederum ist in die universitäre Schwerpunktbereichsprüfung und eine staatliche Pflichtfachprüfung unterteilt) und von der zweiten Staatsprüfung. Während im Studium und im Ersten Examen der juristisch zu beurteilende Sachverhalt als feststehend und beweisbar zu unterstellen ist, muss er im Zweiten Examen oftmals erst erarbeitet werden. Dies birgt zusätzliche Probleme – noch stärker für die autonome Entscheidungsfindung durch KI in der Praxis. Der vorliegende Beitrag konzentriert sich auf Klausuren vor und im Ersten Staatsexamen.

7 Die gesamte Notenskala mit ihren Beschreibungen findet sich in § 1 der JurPrNotSkV: www.gesetze-im-internet.de/jurprnotskv/JurPrNotSkV.pdf.

8 Vgl. BfJ (Fn. 3), S. 2 und 10. Im obersten Drittel wird die Notenskala generell kaum ausgeschöpft.

9 Ausführlich dazu unter C. III. 2 b).

nach der Erfahrung der AutorInnen selten zum Einsatz. Empirische Forschung zur juristischen Notengebung ist kaum vorhanden. Eine der seltenen Studien aus dem Jahr 2024 zeigt das Ergebnis einer solchen „Bauchentscheidung": Bei 23 JuristInnen, welche dieselben 15 Klausuren korrigierten, lag die Abweichung zwischen niedrigster und höchster Note nicht in Extremfällen, sondern im Durchschnitt bei 6,47 Punkten.[10]

Neben diesem summativen Feedback erhalten Studierende im Laufe ihrer Ausbildung zudem regelmäßig Hinweise (sog. formatives Feedback) zu ihren Klausurleistungen. Vorgaben hierfür existieren ebenfalls selten. Am Rand der Klausuren finden sich in der Praxis teilweise nur Haken oder andere Korrekturzeichen, teilweise jedoch auch detailliertere Randbemerkungen zu einzelnen Abschnitten in Textform. Am Ende stehen zum Teil Gesamtvoten, manchmal aber auch nur die Noten. Es profitieren diejenigen Studierenden, die eine ausführlichere Rückmeldung erhalten. Sie haben die Möglichkeit, ihre Fehler nachzuvollziehen, aus ihnen zu lernen und sich gezielt zu verbessern, während andere mit knappen Anmerkungen oder bloßen Korrekturzeichen weniger Unterstützung für ihren Lernprozess erhalten.

Die Personen, die Klausuren im Studienverlauf bis zum Ersten Examen korrigieren, sind wegen der Fülle der zu korrigierenden Arbeiten in der Regel Externe und erhalten eine fixe Vergütung pro Arbeit. Meist müssen sie lediglich eine Note melden, was Anreize schafft, möglichst schnell „durchzukorrigieren". Strengere Vorgaben würden zu mehr Arbeit in der Vorbereitung und Durchführung der Korrekturen und damit zu einer relativen Absenkung der ohnehin geringen Vergütung führen. Es besteht die Befürchtung, dass sich dann gar keine Personen mehr für Korrekturtätigkeiten finden lassen. Ein weiterer Grund, wenig formatives Feedback zu geben, könnte darin liegen, sich – mangels Nachvollziehbarkeit und damit Angreifbarkeit der einer Notengebung zugrundeliegenden Erwägungsgründe – einem geringeren Remonstrationsrisiko auszusetzen. Hinzu kommen die Einflüsse, die nicht bewertungsrelevant sein sollten, es bei der menschlichen Korrektur aber nachweislich sind. Solche Faktoren sind z. B. psychische oder physiologische Zustände wie Stress und Hunger oder die Reihenfolge, in welcher Klausuren korrigiert werden.[11] Angesichts dieser Umstände überrascht es nicht, dass die Mehrheit der angehenden JuristInnen (2025: 81,63 %[12]; 2022: 80,71 %[13]) die Korrektur juristischer Klausuren nicht für objektiv hält.

Kann Künstliche Intelligenz (KI) das besser? Was als rein empirische Untersuchung dieser Frage startete, wurde zu einer umfassenden Untersuchung der grundlegenden rechtsdidaktischen Frage, was „besser" überhaupt bedeuten könnte.

In diesem Beitrag geht es darum, was wir aus dem Aufbau von KI-Korrekturassistenzen und ausführlichen Diskussionen dazu gelernt haben und wie dadurch die juristische Notengebung durch KI-Systeme sowie durch Menschen objektiver gestaltet werden kann. Er soll ein Gedankenanstoß für weitere Diskussionen und zukünftige (empirische) Forschung sein.

Die nachfolgende Untersuchung beginnt mit einer Projektbeschreibung (B.) in der die Projektbeteiligten und die Finanzierung offengelegt werden (I.), es folgen eine Darstellung der unterschiedlichen Projektphasen (II.) und ein – durch einen umfassenderen Annex hierzu ergänzter – Einblick in die technische Vorgehensweise (III.). Die Auswertung (C.) beginnt mit einem allgemeinen Vergleich zwischen menschlicher und KI-Korrektur anhand von Kriterien wie Kosten, Zeit und Detailtiefe (I.). Anschließend werden nacheinander die jeweiligen Korrekturen der Klausuren in Bielefeld (II.) und München (III.) ausgewertet. Den Schwerpunkt dieser Untersuchung bildet die Analyse der zahlreichen parallel erfolgten menschlichen Klausurkorrekturen sowie der unterschiedlichen KI-Korrekturen des Durchgangs in München. Dabei wird zunächst eruiert, wie man den Vergleich verobjektivieren kann, insbesondere wie eine menschliche Basislinie als Vergleichsmaßstab aussehen könnte. Hierfür werden zunächst menschliche Korrekturen mit und ohne Rohpunkteschema miteinander verglichen. Anschließend erfolgt ein Vergleich der menschlichen Korrektur mit der KI-Korrektur. Der Beitrag schließt – nach der Darlegung künftiger Verbesserungsmöglichkeiten in Bezug auf die KI-Systeme und die For-

10 *Clemens Hufeld*, Jede Korrektur eine andere Note: Quantitative Untersuchung der Objektivität juristischer Klausurbewertungen, ZDRW 2024, S. 59 ff., https://doi.org/10.5771/2196-7261-2024-1-59.

11 Grundlegend zu jenen Einflüssen auf juristische Urteile *Danziger/Levav/Avnaim-Pesso*, Extraneous factors in judicial decisions, *Proc. Natl. Acad. Sci. U.S.A.* 17/2011, S. 6889 ff., https://www.pnas.org/doi/abs/10.1073/pnas.1018033108.

12 BRF 2025 (Fn. 2).

13 Bundesverband rechtswissenschaftlicher Fachschaften e.V. (BRF), Absolvent:innen Befragung - Abschlussbericht, 2022, https://bundesfachschaft.de/wp-content/uploads/2025/05/Fuenfte-bundesweite-AbsolventInnenbefragung-2022.pdf, S. 90.

schung zu ihnen (D.) – mit einem Ausblick auf die (rechts)didaktischen und gesellschaftlichen Implikationen und einer Zusammenfassung der Ergebnisse (E.).

B. Projektbeschreibung

I. Projektbeteiligte und Finanzierung

Das vorliegende Projekt (DigitalProjekt) ist ein loser Zusammenschluss rechtsdidaktisch und digital affiner JuristInnen – von Studierenden über Doktoranden und Habilitanden bis zu Professorinnen. Es existieren weder eine feste Organisationsstruktur noch stehen Fördermittel zur Verfügung. Die erforderlichen Finanzmittel für die zahlreich beteiligten menschlichen KorrekturassistentInnen wurden dankenswerterweise von der studentischen Initiative Munich Legal Tech Student Association e. V. aus München (MLTech) bereitgestellt.

Um möglichst vielfältige Erkenntnisse zu den in der Einleitung beschriebenen Problemen und Aspekten zu gewinnen, diese abgleichen und Ergebnisse weiter optimieren zu können, werden vom DigitalProjekt verschiedene Ansätze verfolgt, sowohl inhaltlich zur Bewertung juristischer Arbeiten als auch technisch:

Es werden zwei hinsichtlich ihrer Arbeitsweise näher beschriebene[14] KI-Korrektursysteme eingesetzt. Das eine wurde vom aus der Universität Köln ausgegründeten Start-up KlausurenKIste, das andere vom BMFTR-geförderten Forschungsprojekt DeepWrite der Universität Passau entwickelt. Die verwendeten Large Language Models (LLMs) fallen unter den Begriff der KI, ohne dass hier der bisher noch nicht gelungene Versuch unternommen werden soll, eine Definition für KI zu geben.

Es erfolgt im DigitalProjekt eine ständige Diskussion der Ergebnisse mit anschließender Optimierung der Systeme. Über die technischen Fragen hinaus besteht ein intensiver Austausch zwischen Studierenden und Forschenden/Lehrenden über die Kriterien, nach denen juristische Arbeiten beurteilt werden sollten. Denn wie in der Einleitung (A.) beschrieben, existieren diesbezüglich viele Ungereimtheiten. Nur bei klaren Vorgaben können Fortschritte sowohl beim Einsatz von KI als auch in der menschlichen Korrektur erzielt werden.

II. Projektphasen

Stand Oktober 2025 wurden bereits zwei Praxisphasen durchgeführt (Phase I und II) und mindestens eine weitere (Phase III) ist geplant:

1. Phase I: Bielefeld (WiSe 2024/2025)

An der Universität Bielefeld wurde im Wintersemester 2024/2025 die Vorlesung „Zwangsvollstreckungsrecht" von Marie Herberger unter Verwendung des Online-Gesetzbuches LexMea[15] gehalten. Am 16.12.2024 durften die Studierenden eine Probeklausur im Rahmen dieser Vorlesung unter Verwendung der digitalen Gesetze an Computern der Universität schreiben. Dabei konnten sie neben der weiterhin stattfindenden menschlichen Klausurkorrektur optional auch einer parallelen KI-Korrektur zustimmen. Jene erfolgte getrennt durch die KI-Systeme von KlausurenKIste und DeepWrite. Es handelte sich soweit ersichtlich um die erste volldigitalisierte juristische Klausur Deutschlands.[16]

Bei der KI-Korrektur sollten die von den Studierenden erstellten Gutachten mit einer vorgegebenen Musterlösung automatisiert abgeglichen und bewertet werden. Neben einer formativen Rückmeldung – für die in dieser Phase keine von der Aufgabenstellerin oder dem DigitalProjekt definierten Vorgaben existierten – sollte durch die KI eine summative Bewertung – also eine Note auf der juristischen Punkteskala zwischen 0 und 18 Punkten – generiert werden. Auch für die technische Erarbeitung der Notengebung gab es zunächst keine zentralen Vorgaben durch das DigitalProjekt – insbesondere kein zentral zur Verfügung gestelltes Rohpunkteschema.

2. Phase II: München (SoSe 2025)

Am 18.06.2025 wurde an der Ludwig-Maximilians-Universität München (LMU) im Kurs von Martin Heidebach mit Unterstützung von MLTech eine zweite voll-

14 Im Folgenden unter B. III. sowie im Annex.

15 Das Online-Gesetzbuch LexMea ist unter https://lexmea.de kostenfrei nutzbar.

16 Näher dazu etwa Universität Bielefeld, Digitale Zukunft der juristischen Ausbildung, Pressemitteilung, 16.12.2024, https://aktuell.uni-bielefeld.de/2024/12/16/digitale-zukunft-der-juristischen-ausbildung/; *Mathilde Harenberg*, Die erste komplett digitalisierte Juraklausur Deutschlands, LTO, 18.12.2024, https://www.lto.de/karriere/jura-studium/stories/detail/klausur-digitalisierung-jurastudium-pilot-online-gesetzbuch; *Jannina Schäffer*, Erste volldigitale Präsenzklausur an der Uni Bielefeld, JURios, 27.12.2024, https://jurios.de/2024/12/27/erste-volldigitale-praesenzklausur-an-der-uni-bielefeld/.

digitalisierte Probeklausur im Tutorium Allgemeines Verwaltungsrecht und Verwaltungsprozessrecht organisiert.

Auch in Phase II wurden erneut alle Klausuren am PC digital verfasst und anschließend in einem ersten Durchgang von dem Kursverantwortlichen Martin Heidebach selbst korrigiert. Anschließend wurden anhand dieser Noten 16 möglichst breit über das gesamte Notenspektrum gestreute Klausuren ausgewählt. Die Begrenzung auf 16 Klausuren beruhte auf dem verfügbaren Budget für die Klausurkorrektur. Diese vorausgewählten Klausuren wurden anschließend von 14 KorrekturassistentInnen korrigiert. Acht davon wurde ein Rohpunkteschema mit Einzelpunkten für die jeweiligen Lösungsabschnitte und einer Umrechnungstabelle ausgehändigt.[17] Sechs der KorrekturassistentInnen erhielten lediglich die Lösungsskizze – sie korrigierten (traditionell) ohne Rohpunkteeschema. Die Ausgestaltung dieses Rohpunkteschemas profitierte im Vorfeld u. a. von der langjährigen Erfahrung im Einsatz eines solchen durch Susanne Hähnchen und von der Diskussion innerhalb des DigitalProjekts.

Alle KorrekturassistentInnen erhielten jeweils 8 Euro pro korrigierte Klausur. Sie selbst hatten mindestens 9 Punkte in der Ersten Juristischen Staatsprüfung und gingen unterschiedlichen Tätigkeiten nach (Wissenschaftliche Mitarbeitende, RechtsreferendarInnen, RechtsanwältInnen u.a.).

Zudem wurde jede Klausur durch die beiden Teams DeepWrite und KlausurenKIste einer KI-Korrektur unterzogen. Auch die beiden Teams ließen die KI dabei einmal nach ihrem – einzig durch einen Prompt mit allgemeinen Korrekturanweisungen gesteuerten – „Bauchgefühl" (ohne Rohpunkteschema; oRPS) bewerten und einmal anhand des auch der KI zur Verfügung gestellten Rohpunkteschemas (mit Rohpunkteschema; mRPS). Dabei wurden je zwei LLMs verwendet, nämlich GPT-4o von OpenAI und Gemini 2.5 Pro von Google. Insgesamt entstanden somit acht KI-Korrekturdatensätze (zwei Teams x zwei LLMs x zwei Durchgänge je mRPS und oRPS).

3. Phase III: Potsdam (WiSe 2025/2026)

Im Wintersemester 2025/2026 wird die nächste digitale Probeklausur an der Universität Potsdam im Rahmen der Vorlesung BGB AT organisiert, um weiter an der Optimierung der KI-Korrektursysteme zu arbeiten. Außerdem soll eine im Vergleich zu Phase II andere Vorgehensweise bei der Nutzung des Rohpunkteschemas getestet werden. Geplant ist auch eine Analyse des formativen Feedbacks. Die Einzelheiten sind noch offen.

III. Technische Vorgehensweise

Während die beiden KI-Korrektur-Projektpartner KlausurenKIste und DeepWrite stets mit denselben Ausgangsdaten arbeiten (1.), werden zwei unterschiedliche technische Ansätze verfolgt (2.).

1. Ausgangsdaten

Die dem DigitalProjekt zugrundeliegenden Klausuren, Lösungsskizzen, Rohpunkteschemata sowie Ergebnisse der KI-Korrektur können – ebenso wie die Tabelle mit allen Korrekturen und statistischen Auswertungen – auf Anfrage zu wissenschaftlichen Zwecken zur Verfügung gestellt werden.

a) Klausuren

Die Klausuren wurden ausgedruckt ausgeteilt und bestanden jeweils aus einem umfassenden zusammenhängenden Fall, zu dem ein juristisches Gutachten verfasst werden musste.

Die Klausuren wurden in Phase I und II am PC geschrieben. In Phase I standen stationäre Rechner im Computerraum der Universität Bielefeld zur Verfügung. In Phase II hatten die Studierenden die Wahl zwischen dem Schreiben an den stationären Rechnern im Computerraum der LMU München und dem Schreiben auf dem eigenen Laptop (sog. „bring your own device" [BYOD]).[18] Die Unterteilung in unterschiedliche Klausurabschnitte und Gliederungsebenen musste von den Studierenden – wie auch bei analogen Klausuren – selbst vorgenommen werden, war aber nicht verpflichtend. Es standen keine vorformatierten Überschriftenebenen zur Verfügung. Genutzt werden konnten Absätze, frei eingegebene Gliederungsebenen (z. B. I., II., III.) und Fettungen.

Die abgegebenen Klausuren wurden als PDF-Dateien exportiert. Vor der weiteren Verarbeitung wurden die Klausuren bereinigt: Alle personenbezogenen Angaben (z. B. Name, Matrikelnummer, Universität) sowie nicht bewertungsrelevante Bestandteile wie kopierte Sachverhaltstexte oder Kopfzeilen wurden aus dem Text ent-

17 Näher dazu unter C. III. 2. b) .

18 Siehe zu den Erfahrungen aus München und künftigen Herausforderungen *Michael B. Strecker*, LMU digitalisiert und Deutschland schaut zu: Warum die juristische E-Klausur keine föderale Vision hat, Legal Tech Verzeichnis, 10.07.2025, https://legal-tech-verzeichnis.de/fachartikel/lmu-digitalisiert-und-deutschland-schaut-zu-warum-die-juristische-e-klausur-keine-foederale-vision-hat/.

fernt, sodass nur der reine Gutachtentext übriggeblieben ist.

b) Lösungsskizze

Für die Korrektur wurde eine umfassende (in Phase I stichpunktartig formulierte und in Phase II ausformulierte) Lösungsskizze als Grundlage für beide KI-Korrektur-Projektpartner zur Verfügung gestellt.

Um die Lösungsskizze für die KI-Korrektur verarbeiten zu können, wurden zunächst alle essenziellen Informationen aus der Lösungsskizze herausdestilliert sowie Passagen manuell sprachlich für die KI angepasst. Dabei wurden insbesondere lange Sätze von Füllwörtern befreit und sprachlich vereinfacht sowie teilweise ausformuliert, um die Kontexteinbettung zu gewährleisten – inhaltlich blieb die Lösung dabei allerdings unverändert.

Zentraler Unterschied zwischen den Phasen I und II war, dass die Lösungsskizze in Phase II für manche KorrekturassistentInnen und KI-Systeme von einem Rohpunkteschema flankiert wurde.

2. Technische Ansätze

Eine ausführliche Darstellung der jeweiligen technischen Ansätze findet sich im Annex. Für die weitere Interpretation der Ergebnisse sind insbesondere die unterschiedlichen technischen Ansätze zwischen der „Absatzkorrektur" (KlausurenKIste) und der „Gesamtkorrektur" (DeepWrite) bedeutsam:

a) „Absatzkorrektur" (KlausurenKIste)

Beim Ansatz von KlausurenKIste werden die Klausuren entlang der von den Studierenden selbst gesetzten Absätze (Gliederungselemente + Überschrift + zugehöriger Text) in kurze Abschnitte unterteilt und automatisiert mit den entsprechenden Lösungsabschnitten abgeglichen – „Absatzkorrektur".[19] Dieser Ansatz wurde gewählt, um präziser an spezifischen Absätzen auf das studentische Gutachten eingehen zu können.

b) „Gesamtkorrektur" (DeepWrite)

Der technische Ansatz von DeepWrite bestand zunächst darin, der KI die gesamte Klausur zur Bewertung zu geben und diese auch in Gänze bewerten zu lassen – „Gesamtkorrektur". In späteren Iterationsschleifen wurde zumindest grob in größere Sinnabschnitte aufgeteilt (etwa: Zulässigkeit und Begründetheit).[20] Dieser Ansatz wurde gewählt, um das „Kontextverständnis" in der jeweiligen Bearbeitung besser zu erfassen.

C. Auswertung

Eine große Herausforderung der Bewertung der KI-generierten Ergebnisse liegt in der Bestimmung objektiver Vergleichsmaßstäbe (Benchmarks). Dabei ist zwischen Verfahrens- und Ergebnisvergleich zu differenzieren.

Zumindest für das Verfahren finden sich allgemeine objektivierbare und quantifizierbare Kriterien wie die Kosten der Korrektur (in Geld), die für sie benötigte Zeit (in Tagen bis zum Feedback), ihr Umfang (in Prozent der Klausurteile, mit denen sich vertieft auseinandergesetzt wird) sowie der Detailgrad der Korrekturanmerkungen (in Zeichen). Die Variabilität dieser Faktoren darf als allgemein bekannt vorausgesetzt werden und ist daher hier nicht detailliert zu erörtern. Diese Faktoren werden nachfolgend für einen allgemeinen Vergleich zwischen menschlicher und KI-Korrektur fruchtbar gemacht (I.).

Nach einer kompakten Auswertung der KI-Korrekturen der Phase I in Bielefeld (II.) beschäftigt sich die Auswertung der Korrekturen der Phase II in München (III.) zunächst intensiv mit dem Finden eines objektiven Bewertungsmaßstabes. Dabei wird insbesondere der Einsatz eines Rohpunkteschemas für mehr Objektivität eruiert, bevor die diversen Korrekturergebnisse miteinander verglichen werden.

I. Allgemeiner Vergleich zwischen menschlicher und KI-Korrektur

1. Kosten der Korrektur

In der Regel werden derzeit 8 Euro pro Klausur innerhalb des Studiums an externe Korrekturkräfte gezahlt. Die Korrektur von Hausarbeiten wird besser vergütet. Dies mag zwar zunächst viel erscheinen, jedoch sollte bedacht werden, dass der Zeitaufwand für eine gründliche Korrektur erheblich ist und die erzielten Einnahmen unter Umständen auch noch zu versteuern sind.[21] Soweit eine Vergütung in den Examina erfolgt, was in manchen

19 Ausführlich dazu im Annex unter II.

20 Ausführlich dazu im Annex unter I.

21 Bis 3.000 € pro Jahr sind solche Einnahmen nach der sog. „Übungsleiterpauschale" des § 3 Nr. 26 EStG steuerfrei. So können ca. 230 fünfstündige Klausuren steuerfrei korrigiert werden.

Bundesländern nicht allgemein gilt, fällt diese mit 13 Euro in der Regel höher aus.

Jedes Jahr sind bei ca. 115.000 Jura-Studierenden[22] hunderttausende juristische Prüfungsarbeiten zu korrigieren. Dies kostet die Universitäten und staatliche Prüfungsämter viel Geld, welches für Bildung ohnehin (zu) knapp vorhanden ist. Daher ist in Zukunft eine bessere, dem Aufwand für eine gründliche Korrektur angemessene Vergütung nicht zu erwarten. Zu den Kosten für externe KorrekturassistentInnen kommen an den Universitäten noch solche Korrekturen hinzu, die mittelbar über die Gehälter des wissenschaftlichen Personals vergütet werden – beispielsweise in den Schwerpunktbereichen, für welche der Einsatz von KI ebenfalls interessant sein dürfte.

Bei KI-Systemen spielen die Kosten weniger für deren Anwendung[23] als vielmehr für deren Entwicklung und Erwerb eine Rolle. Frei zugängliche KI-Korrektursysteme, deren Aufbau und Betrieb finanziell unterstützt wird und die auch auf europäischen Open-Source-Modellen beruhen, können hier Vorteile bieten. Beispielsweise wurde DeepWrite vom Bundesministerium für Forschung, Technologie und Raumfahrt über die Bund-Länder-Initiative „Digitale Hochschulbildung" mit knapp 2 Mio. € (entspricht etwa den Kosten für 250.000 Korrekturen) für eine Projektlaufzeit von vier Jahren gefördert, wovon jedoch nicht alle Mittel in die juristische KI-Korrektur flossen.[24] KlausurenKIste wurde durch das NRW-Gründerstipendium mit ca. 42.000 € (entspricht etwa den Kosten für 5.250 Korrekturen) bezuschusst. Es stellt sich somit die Frage, ob Lizenzmodelle – gerade, wenn sie die deutsche Start-Up-Wirtschaft stärken – aus systemischer Perspektive nicht ähnlich kosteneffektiv sind.

Je nach eigenen Entwicklungs- oder fremden Lizenzkosten bietet die KI-Korrektur angesichts der enormen Anzahl in Deutschland korrigierter juristischer Klausuren erhebliche Kostenvorteile.

2. Korrekturzeit

Studierende und ExamenskandidatInnen wünschen sich oft, dass die Korrekturen schneller durchgeführt werden. Hier punkten kommerzielle Repetitorien, die Probeklausuren teilweise mit großem personellem und finanziellem Aufwand korrigieren lassen und schon am selben Tag zurückgeben. Bei schnellerem Feedback ist die Klausuraufgabe zum Zeitpunkt einer stattfindenden Besprechung noch präsent, Lernprozesse werden nicht unterbrochen und (angesichts hoher Durchfallquoten nicht unrealistische) Versagensängste in der Wartezeit verringert.

An den Universitäten und in den staatlichen Examina dauert die Notenbekanntgabe in der Regel mehrere Wochen bis Monate. Ursache dafür sind die internen Verwaltungsabläufe und – u. a. wegen der Bezahlung – das Fehlen einer hohen Zahl von KorrektorInnen. Die aktuell zunehmende Digitalisierung des Klausurenschreibens („eKlausur") führt immerhin zur Beschleunigung der Abläufe, insbesondere müssen nicht mehr große Mengen von Klausuren wiederholt transportiert werden. Von einer weiteren (digitalisierten) Optimierung der Prozesse in der Verwaltung würden alle Formen der Korrektur profitieren.

KI-Systeme benötigen demgegenüber nur Sekunden bis Minuten für die Bewertung einer einzelnen Prüfungsarbeit, also deutlich weniger als eine menschliche Korrekturkraft (geschätzt 15-120 Minuten je nach Umfang der Klausur oder Hausarbeit und Gründlichkeit) und sind damit klar im Vorteil.

3. Umfang der Korrektur

Wichtig für die inhaltliche Bewertung einer Prüfungsleistung ist, dass sie zunächst voll erfasst wird.

Die menschliche Korrektur erfordert dafür keine technischen Vorbereitungen (über die allgemeinen Verwaltungsabläufe hinaus), wohl aber eine besondere fachliche Qualifikation. Eine optimale Korrekturkraft würde alle Abschnitte einer Prüfungsleistung sehen und in die Bewertung einbeziehen. Tatsächlich beklagen sich Studierende aber oft, dass sie den Eindruck haben, dass etwas übersehen wurde. Aus den Randbemerkungen und abschließenden Voten kann dies nicht immer klar entnommen werden. In den meisten Fällen ist davon auszugehen, dass jedenfalls die wesentlichen Teile der Prüfungsleistung in die Bewertung einfließen.

KI-Systeme bergen unter Umständen technische Herausforderungen. Die exakte Zuordnung von Klausur und Lösungsskizze auf Absatzebene (KlausurenKIste) ist technisch besonders komplex. Gering ist hingegen der technische Aufwand bei einer Gesamtkorrektur (Deep-

22 Angabe für das Jahr 2024 in: Statistisches Bundesamt, Studierende Studienfach Rechtswissenschaft Deutschland, Stand: 13.08.2025, https://www.destatis.de/DE/Themen/Wirtschaft/Konjunkturindikatoren/Lange-Reihen/Bildung/lrbil03a.html.

23 Zu berücksichtigen sind natürlich auch Kosten für die zu beschaffenden Server, den anfallenden Strom für Server und die Kühlung etc.

24 Für die juristisch-inhaltliche Arbeit im Teilbereich Jura wurde eine Vollzeit TV-L E13-Stelle für die Dauer von 3 Jahren veranschlagt. Für die Entwicklung des KI-Korrektursystems waren jedoch auch Stellenanteile aus den Bereichen Data Science, classEx-Entwicklung, Didaktik, User-Experience und IT notwendig.

Write), wo diese Zuordnung nicht erforderlich ist. Hinzu kommen auch bei der Verwendung eines KI-Systems die generelle Unsicherheit und fehlende Nachprüfbarkeit, ob alle Informationen erfasst und berücksichtigt wurden. Durch die spezifischeren, absatzbezogenen Antworten bei KlausurenKIste kann man allerdings die tatsächliche Berücksichtigung der jeweiligen Abschnitte besser erkennen. Beide KI-Systeme haben also in diesem Punkt Vor- und Nachteile. Bei richtiger technischer Gestaltung dürfte die KI-Korrektur jedoch vollständiger sein als eine durchschnittliche menschliche Korrektur.

4. Detailgrad der Korrekturanmerkungen

Um aus einer Prüfungsleistung Rückschlüsse auf den eigenen Lernstand ziehen zu können, benötigen Studierende ein detailliertes Feedback. Auch die Akzeptanz[25] der Noten ist umso höher, je nachvollziehbarer sie begründet werden. Der BRF schlussfolgert aus seiner AbsolventInnenbefragung 2022 den „Wunsch nach ausführlicheren Voten […], um zum einen die Nachvollziehbarkeit der Noten zu erhöhen und zum anderen ein Lernen aus den eigenen Fehlern zu ermöglichen".[26]

Auch bei einer etwaigen Nachkontrolle oder offenen Zweitkorrektur müssen die Bewertungen durch die Korrekturkräfte nachvollzogen werden können. Andererseits kann ein Feedback auch zu lang und detailreich sein, um im Regelfall zur menschlichen Kenntnis genommen zu werden.

Menschliche Korrekturen haben unter Studierenden teilweise den Ruf, dass die Ergebnisse ‚gewürfelt' werden. Das ist nicht nur ironisch hinsichtlich der fehlenden Nachvollziehbarkeit gemeint, sondern auch Ausdruck des Gefühls, einer willkürlichen Bewertung ausgeliefert zu sein.[27] Die grundsätzlichen Probleme fehlender Standards wurden bereits beschrieben und es wird darauf später zurückzukommen sein.[28]

Die Anzahl an Randbemerkungen und die Ausführlichkeit der abschließenden menschlichen Voten ist nach der Erfahrung der Projektbeteiligten tatsächlich sehr unterschiedlich. Wie genau die Korrekturkraft zu ihrer eigentlichen Bewertung kommt, kann in der Regel nicht nachvollzogen werden. Besser sieht es bei der Verwendung von Rohpunkteschemata[29] aus, zumal wenn diese – entweder auch für die KandidatInnen oder nur für die Aufgabenstellenden sichtbar – vermerkt werden.

Auch wie genau eine KI zu ihrem Ergebnis kommt, ist ein bereits angesprochenes allgemeines Problem. Die Absatzkorrektur (KlausurenKIste) ist insofern transparenter als die Gesamtkorrektur (DeepWrite), weil Feedback zu jedem einzelnen Abschnitt gegeben wird und der Lösungsskizze nicht zuordenbare Abschnitte explizit ausgesondert werden. Zudem erfolgen zu jedem Gliederungspunkt umfassende Anmerkungen. Bei der Gesamtkorrektur werden hingegen längerer Sinnabschnitte zusammenfassend verglichen, wodurch zwar ein geringerer Detailgrad des Feedbacks, aber auch ein höheres Kontextverständnis bedingt ist. Umfassende Anmerkungen gibt es jedoch auch hier, allerdings zu jedem Sinnabschnitt.

Beide KI-Systeme geben also umfangreicheres und detaillierteres Feedback als eine durchschnittliche menschliche Korrekturkraft. Dadurch erhöht sich auch die Nachvollziehbarkeit der Bewertung. Von der Verwendung eines Rohpunkteschemas zwecks Erhöhung der Transparenz der Notenfindung profitieren jedoch beide Arten der Korrektur.

5. Verständlichkeit und Motivationsfunktion der Korrekturanmerkungen

Inhaltlich finden sich nach den Erfahrungen der Projektbeteiligten – neben der Verwendung von Zeichen am Rand und Linien im Text der Bearbeitung – zumeist lediglich knappe, direkte Anmerkungen wie „falsch", „korrekt", „…fehlt", „vertretbar". Die abschließenden Voten sind oft eine Aufzählung der Mängel. Lücken und Fehler müssen zwar benannt werden – die meisten Korrekturen gehen darüber aber kaum hinaus.[30] Lernförderliche und motivierende Ausführungen sind nur bei über-

25 Ein „mulmiges Gefühl" hatte *Sebastian Dötterl* (Fn. 4), S. 165, der allerdings nicht die Ausführlichkeit und Qualität des Feedbacks kannte, das von den für das DigitalProjekt verwendeten KI-Systemen ausgegeben wurde.

26 BRF 2022 (Fn. 13), S. 91.

27 Siehe dazu BRF 2025 (Fn. 2), S. 100 f. sowie BRF 2022 (Fn. 13), S. 90. Dieser Vorwurf ist für ein Verwaltungsverfahren, um das es sich bei dem Bewertungsprozess handelt, besonders problematisch; dazu *Martin Heidebach*, Prüfen im rechtswissenschaftlichen Studium: Die Korrektur juristischer Hausarbeiten anhand eines verbindlichen Bewertungseinheiten-Systems, ZDRW 2015, S. 205 (213), https://doi.org/10.5771/2196-7261-2015-3-205.

28 S. dazu bereits in der Einleitung oben sowie unten unter C. III. 2. a).

29 Ausführlich dazu unten unter C. III. 2. b).

30 Anders aber das Projekt zur Videokorrektur von *Martin Heidebach*, vgl. https://www.lmu.de/de/newsroom/newsuebersicht/news/lehrfoerderung-an-der-lmu-fuer-eine-kultur-der-guten-lehre.html. Das gesprochene Format bietet viele Möglichkeiten, etwa im Ton, die im geschriebenen nicht zur Verfügung stehen.

durchschnittlich guten Klausuren oder von besonders engagierten Korrigierenden zu erwarten. Kommerzielle Korrekturdienstleister sind hingegen darauf angewiesen, dass ihre Nutzenden (u. a. durch freundliche Sprache) motiviert bleiben und das Angebot weiter in Anspruch nehmen.

Der AbsolventInnenbefragung des BRF 2025 zufolge besonders kritisch bewertet wurden im offenen Anmerkungsfeld zur Abfrage der Objektivität juristischer Klausuren „pauschale oder beleidigende Korrekturanmerkungen, die nicht nur wenig hilfreich, sondern auch entmutigend sind".[31] Die ErstellerInnen der Umfrage kommen zu dem Ergebnis:

> „Die emotionale Intensität der Anmerkungen macht deutlich, dass es sich hier nicht nur um eine fachliche, sondern auch um eine psychologisch tiefgreifende Erfahrung handelt. Die (subjektiv empfundene) Ungerechtigkeit bei der Notenvergabe beeinflusst nicht nur die Motivation, sondern kann auch Zweifel an den eigenen Fähigkeiten auslösen."[32]

KI-Systeme, wie die im DigitalProjekt verwendeten, geben bei Unterteilung z. B. in „Treffer oder Lücken" (mit dezisiven, direkten Hinweisen) und in „Verbesserungspotenzial" (mit motivierend formulierten, lernförderlichen Vorschlägen) ein tendenziell verständlicheres und motivierenderes Feedback.

II. Auswertung der KI-Korrekturen in Phase I (Bielefeld)

1. Formatives Feedback

Das formative Feedback der Korrekturen aus Phase I wurde aus Kapazitätsgründen nicht systematisch ausgewertet. Stattdessen wurden drei studentische Hilfskräfte um Bewertung des formativen KI-Feedbacks gebeten. In Summe lässt sich anhand dieser Auswertung festhalten, dass das Feedback der KI-Systeme ausführlicher ausgefallen ist als das des menschlichen Korrektors. Die Studierenden waren auf der einen Seite der Auffassung, dass Fehler so gut nachvollzogen werden können und die KI-Korrektur somit transparenter erscheint als die menschliche Korrektur. Auf der anderen Seite wurde aber auch angemerkt, dass die Korrektur bei manchen Inhalten zu umfangreich gewesen sei.

Zudem entstand der subjektive Eindruck, dass die KI-Systeme teilweise höhere Anforderungen an die Bearbeitung stellten als die Lösungsskizze. So hieß es in der Lösungsskizze beispielsweise, dass die KandidatInnen „in gebotener Kürze" feststellen sollten, dass es um eine Erinnerung i.S.d. § 766 ZPO geht – sodass eine umfangreiche Abgrenzung zu den anderen Rechtsbehelfen des 8. Buches der ZPO nicht notwendig war, aber auch nicht für schädlich erklärt wurde. Dennoch kritisierte etwa das damalige KI-System[33] von KlausurenKIste mehrfach, dass die verschiedenen Rechtsbehelfe der Zwangsvollstreckung (insb. §§ 766, 767, 771 ZPO) nicht voneinander abgegrenzt worden seien.

Kleinere Schwächen zeigten sich auch bei der Differenzierung zwischen Definition und Subsumtion bei den im juristischen Gutachtenstil zu bearbeitenden Teilen der Klausur.

Beide KI-Systeme verteilten in dieser Phase zudem Punkte für einen Prüfungspunkt, der von der bearbeitenden Person überhaupt nicht geprüft wurde.

Insgesamt lässt sich eine sehr starre Orientierung der KI-Systeme an der Musterlösung erkennen. Diese Feststellungen decken sich jedoch mit Erfahrungen der Projektbeteiligten bezüglich menschlicher Korrekturkräfte. Dabei zeigt sich, welche Bedeutung einer eindeutigen Formulierung der Lösungsskizze für die KI-Korrektur zukommt.

Das nach der Phase I erhaltene Feedback wurde zur Verbesserung der Prompts sowie der Architektur der KI-Systeme genutzt.

2. Summatives Feedback

Sowohl für die Gesamtkorrektur (DeepWrite) als auch für die Absatzkorrektur (KlausurenKIste) stand hier zwar die Lösungsskizze als Grundlage für die Notenvergabe, jedoch kein einheitlich vorgegebenes Gewichtungsschema (Rohpunkteschema) zur Verfügung. Obwohl KlausurenKIste einen umfassenden Prompt als Anweisung für die Korrektur entwarf, entsprach die Notenvergabe in diesem Durchlauf noch dem „KI-Bauchgefühl". Durch DeepWrite wurde hingegen ein erstes rudimentäres Rohpunkteschema erarbeitet, um der KI konkretere Korrekturanweisungen geben zu können. Die Klausuren wurden zudem von lediglich einer menschlichen Korrekturkraft mit Erstem Juristischem Staatsexamen korrigiert.

Ein Vergleich der Ergebnisse beider KI-Systeme sowie einer menschlichen Korrekturkraft ist der folgenden Tabelle (Abb. 1) zu entnehmen.

Die Noten von KlausurenKIste liegen im Mittelwert / „Durchschnitt" (Ø 7,79) nur etwa einen Notenpunkt von der menschlichen Korrektur (Ø 8,71) entfernt, während

31 BRF 2025 (Fn. 2), S. 101.

32 Ebd.

33 Zum Einsatz kam damals GPT-4o (Stand: Dezember 2024), das kein Reasoning-Modell war – ein solches Modell hätte dieses Problem vermutlich vermeiden können.

Klausur #	Seiten	Mensch Note	KlausurenKIste Note	KlausurenKIste Diff. Mensch	DeepWrite Note	DeepWrite Diff. Mensch
10	4,75	3	8	5	3	0
14	4,00	5	8	3	2	-3
1	5,50	6	7	1	3	-3
13	9,80	7	8	1	7	0
5	8,00	7	8	1	4	-3
12	5,00	9	8	-1	10	1
3	3,00	9	11	2	3	-6
7	5,25	9	8	-1	4	-5
6	4,40	10	11	1	7	-3
2	3,60	10	7	-3	7	-3
11	6,50	11	7	-4	17	6
9	4,50	11	6	-5	4	-7
8	7,10	12	7	-5	7	-5
4	3,60	13	5	-8	7	-6
		MW Note: 8,71	MW Note: 7,79	MW Diff Mensch: 2,93	MW Note: 6,07	MW Diff Mensch: 3,64
				Summe Diff Mensch: 41,00		Summe Diff Mensch: 51,00

Abb. 1: Bewertung der in Phase I korrigierten Klausuren durch eine menschliche Korrekturkraft (Spalte 3) sowie durch die KI-Teams KlausurenKIste und DeepWrite inkl. der jeweiligen Abweichung zur menschlichen Korrektur – sortiert nach Note der menschlichen Korrektur (aufsteigend). In der zweiten Spalte ist die Länge der jeweiligen Klausur in Seiten angegeben.

die Differenz der DeepWrite-Noten (Ø 6,07) zur menschlichen Korrektur etwa 2,64 Notenpunkte beträgt.

Nimmt man den Mittelwert der Differenz zwischen allen menschlichen und KI-Korrekturen liegt KlausurenKIste mit nur 2,93 mittlerer Differenz näher an der menschlichen Korrektur als DeepWrite mit 3,64 Differenz im Mittel.

Die Noten von KlausurenKIste weisen dabei allerdings auch die geringste Varianz auf. Nur drei von 14 Klausuren liegen mehr als einen Notenpunkt von sieben Punkten entfernt. Die von *Dötterl* in eigenen Experimenten bereits beobachtete „Konvergenz zur Mitte" hin[34] zeigt sich auch hier.

DeepWrite korrigiert strenger als der Mensch. Lediglich zwei von 14 Klausuren liegen über der menschlichen Note. Eine (in Phase II des DigitalProjekts nach zahlreichen Experimenten mit unterschiedlichen Prompts vermutete) Erklärung könnte darin liegen, dass diese KI nicht nur mehr Fehler findet, sondern auch die Formu-

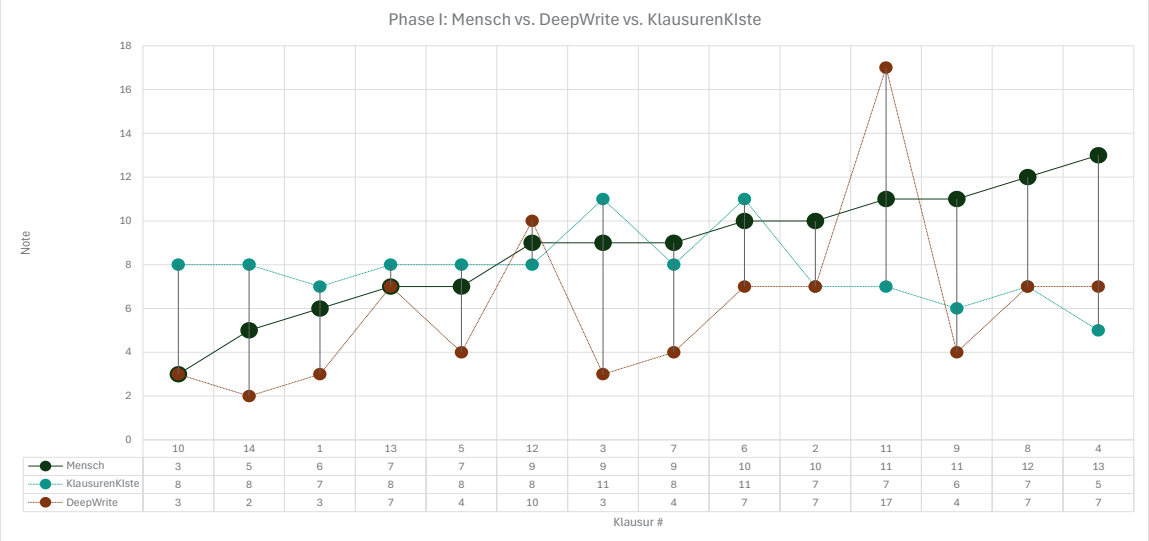


Abb. 2: Bewertungen in Notenpunkten von 0 bis 18 durch die menschliche Korrekturkraft (dunkelgrün) sowie durch die KI-Systeme von KlausurenKIste (dunkeltürkis) und DeepWrite (dunkelorange). Die Punkte der Klausurnoten durch eine Korrektur wurden zur besseren Verdeutlichung miteinander verbunden, ohne dass zwischen den einzelnen Klausurbewertungen ein Zusammenhang besteht.

34 *Sebastian Dötterl* (Fn. 4), S. 165.

lierung im Prompt ursächlich für eine generell kritischere Bewertung ist. Zur Verdeutlichung der Abweichungen in der Punktevergabe dient die nachfolgende Grafik (Abb. 2).

Insgesamt ist der Mittelwert der Abweichungen beider KI-Korrekturen zur menschlichen Korrektur noch recht hoch.

III. Auswertung der Korrekturen in Phase II (LMU)

Wegen des erheblichen hierfür benötigten individuellen, menschlichen Ressourcenaufwandes erfolgte nur eine knappe Auswertung des formativen KI-Feedbacks (1.).

Aufgrund der in Phase I noch recht hohen Abweichungen beider KI-Korrekturen zur menschlichen Korrektur stand im weiteren Projektverlauf im Mittelpunkt, wie die Abweichung verringert werden kann und zu welchem Maßstab genau die Abweichung verringert werden soll – mithin die Frage, was eine „objektive" Bewertung in Jura überhaupt ist. Dem wurde in der Vorbereitung der Phase II in zahlreichen mehrstündigen Videokonferenzen nachgegangen. Die detaillierte Auswertung des summativen Feedbacks (2.) beginnt daher zunächst mit den Diskussionsergebnissen sowie den ermittelten offenen Forschungsfragen bei der Suche nach einem objektiven Bewertungsmaßstab.

1. Formatives Feedback

Eine fundierte Einschätzung der Qualität des formativen Feedbacks würde eine inhaltliche Auseinandersetzung mit den KI-Bewertungen erfordern, idealerweise durch die menschlichen KorrektorInnen derselben Klausuren. Dies konnte im Rahmen dieser Studie nicht geleistet werden. Es war lediglich möglich, eine (kleine) Stichprobe genauer zu analysieren.[35] Dabei hat sich ein sehr positiver erster Eindruck ergeben: Das Feedback ist in einem überzeugend professionellen Ton verfasst, die Rückmeldungen sind verständlich und inhaltlich nachvollziehbar erläutert. Generell sind die juristischen Einordnungen treffend und zum Teil in ihrer Detailschärfe sogar beeindruckend. Über einzelne Bewertungen lässt sich streiten – aber auch nicht mehr als bei den Bewertungen verschiedener menschlicher KorrektorInnen. Nur vereinzelt waren die Einschätzungen tatsächlich fehlerhaft.

Die Absatzkorrektur (KlausurenKIste) ermöglicht zwar präziseres, sehr umfangreiches formatives Feedback (ähnlich angebrachter Randbemerkungen), macht jedoch die Aufteilung und Zuordnung von Lösungsskizze und Klausuren deutlich komplexer und aufwändiger. Die Gesamtkorrektur (DeepWrite) führte zu einem umfangreichen, in die Sinnabschnitte aufgeteilten Schlussvotum. Das formative Feedback war entsprechend generischer, das Kontextverständnis hingegen erhöht.

Hinsichtlich des Umfangs variierte das Feedback der Gesamtkorrektur (DeepWrite) zwischen den verschiedenen Lösungsaspekten, wobei nicht immer erkenntlich war, nach welchem Prinzip der Schwerpunkt der Rückmeldungen gesetzt wurde. Das Feedback der Absatzkorrektur (KlausurenKIste) war äußerst umfangreich, zum Teil aber auch redundant und die Zuordnung der Aspekte zu einem bestimmten Absatz nicht durchgehend nachvollziehbar. Jedenfalls ist durch beide KI-Systeme ein weit umfangreicheres und detaillierteres Feedback zu erlangen als durch eine (durchschnittliche) menschliche Korrektur, was den Studierenden besonders dabei helfen kann, aus ihren Fehlern zu lernen.

Das KI-System von KlausurenKIste hat aktuell eine generell geringe Kontextsensitivität, da einzelne Abschnitte isoliert betrachtet werden. Dies führt dazu, dass zwar inhaltlich korrekte Ausführungen, die aber an einer logisch falschen Stelle erfolgen, als richtig bewertet werden. Die Gesamtkorrektur von DeepWrite ist hier stärker, da sie einzelne Sinnabschnitte betrachtet und zusätzlich eine zusammenfassende Endbewertung vornimmt. Beide Systeme sind aber stark von den Vorgaben der Lösungsskizze abhängig; so wirken sich beispielsweise Ungenauigkeiten in der Lösungsskizze (z.B. weiche Formulierungen wie „kann" oder „soll") auch auf die KI-Korrektur aus.

2. Summatives Feedback

a) Herausforderungen beim Finden eines objektiven Bewertungsmaßstabs (Benchmarking)

Aus dem nach Phase I vorgenommenen schlichten Vergleich zwischen einer bestimmten menschlichen Korrektur und der entsprechenden KI-Korrekturen[36] lassen sich nur begrenzt Folgerungen darüber treffen, wie „gut" die KI tatsächlich korrigiert. Denn es ist nicht gesagt, dass die eine menschliche Korrektur den „objektiven" Bewertungsmaßstab trifft. Das führt zu der Frage, ob sich ein solcher Maßstab finden lässt.

In der Rechtswissenschaft existiert selten ein einziges richtiges Ergebnis. Von der herrschenden Meinung abweichende Ansichten sind bei entsprechender Begründung „vertretbar", nicht falsch. Hintergrund dafür

35 Diese Analyse hat *Martin Heidebach* (Klausursteller) vorgenommen.

36 Dazu oben unter C. II. 2.

ist, dass die Normen, die in der Klausur zu prüfen sind, oft vom Gesetzgeber nicht klar „programmiert" sind, sondern der Interpretation unter Verwendung der anerkannten rechtswissenschaftlichen Methoden bedürfen. Gerade der korrekte Einsatz dieser Methoden ist eine der wichtigsten juristischen Kompetenzen. Deshalb spielt die argumentative Stärke und Stringenz (auch mit Blick auf die Auswertung der im Sachverhalt gegebenen Informationen) neben der Anwendung einer logisch richtigen Prüfungsreihenfolge für eine Norm (Schema) und dem Auffinden des in der Lösungsskizze aufgezeigten (vermeintlich) richtigen Ergebnisses eine zentrale Rolle für eine gute Prüfungsleistung. Zu dieser Ausgangslage kommt hinzu, dass bislang keine eindeutigen, objektivierbaren Maßstäbe für die Gewichtung dieser Aspekte bei der Benotung herausgearbeitet wurden (siehe Einleitung). Außerdem ist der allgemeine Umstand zu bedenken, dass menschliche Urteile bei gleichen Fällen zumeist eine unerwünschte, oft affektbedingte Variabilität aufweisen (sog. Noise).[37] Der bisherigen menschlichen Notengebung bei juristischen Korrekturen ist deshalb eine erhebliche Subjektivität inhärent.

Für die Phase II wurde zunächst angedacht, die Korrekturen des langjährigen Kursverantwortlichen Martin Heidebach zum „objektivsten subjektiven Standard" zu erklären und als Vergleichsmaßstab heranzuziehen. Wie bereits angesprochen, stellte sich jedoch die Frage, ob eine einzige Korrektur überhaupt den Anspruch der Objektivität erheben kann. Es wurde entschieden, dass zumindest ein Vergleich mit weiteren menschlichen Korrekturen erforderlich sei. Die dabei entstandene Debatte über die Gewährleistung möglichst objektiver Vorgaben für etwaige weitere KorrekturassistentInnen mündete in einer umfassenderen Diskussion zum Einsatz von Rohpunkteschemata. Ob die Korrektur anhand von Rohpunkten zu einheitlicheren Ergebnissen führt, ist auch unmittelbar für die Korrektur durch KI-Systeme relevant, denn wenn das der Fall ist, dann wäre es sinnvoll, KI-Systeme mit Rohpunkten Korrekturen durchführen zu lassen.

Wie in anderen Bereichen zwingt der Einsatz und die Gestaltung automatischer Systeme somit dazu, über die Festlegung von Verfahrenskriterien für zuvor stark vom „Bauchgefühl" geprägte Entscheidungen nachzudenken.[38] Und so kommt das viel zu lange ungeklärte Problem der fehlenden Einheitlichkeit juristischer Korrekturen mit voller Wucht auf die Tagesordnung der am Einsatz von KI-Korrektursystemen arbeitenden Personen.

b) Rohpunkteschema für eine objektivere Bewertung?

Können weniger zufällige, homogenere Bewertungen für das summative Feedback durch die Verwendung von Rohpunkteschemata erreicht werden? Gilt dies gleichermaßen für die menschliche wie für die KI-Korrektur? Diese rechtsdidaktisch bedeutenden Fragen bildeten einen wichtigen Teil des Projekts in Phase II.

Bei der Beschreibung der Phase II des DigitalProjekts wurde bereits erwähnt, dass ein Rohpunkteschema bei der Korrektur zum Einsatz kam.[39] Die Hypothese der Beteiligten war, dass der Einsatz eines Rohpunkteschemas zu einer einheitlicheren Benotungspraxis und darüber zu mehr Objektivität führt. Diese Methode zur Erzielung eines summativen Feedbacks ist nicht neu, aber auch nicht sehr verbreitet. Sie wird zunächst kurz erklärt und anschließend werden Einwände und offene Fragen andiskutiert. Dabei sei vorab angemerkt, dass – gerade angesichts der beobachteten Konvergenz zur Mitte – eine einheitliche Benotungspraxis keineswegs möglichst einheitliche Noten für mehrere verschiedene Klausuren, sondern möglichst einheitliche Noten über mehrere Korrekturen derselben Klausur hinweg bedeutet.

aa) Wie funktioniert ein Rohpunkteschema?

Bei Verwendung eines solchen Schemas verteilen die AufgabenstellerInnen zunächst eine bestimmte Anzahl an Teilpunkten (z. B. 100 oder 180) auf die einzelnen Gliederungspunkte der Lösung ("Korrekturbogen").[40] Das kann sehr kleinteilig geschehen (jeder Rohpunkt[41] wird einzeln zugeordnet) oder summarisch (z. B. 10 Rohpunkte für einen Sinnzusammenhang). In einem zweiten Schritt muss festgelegt werden, wie viele Rohpunkte welche juristische Note – also „Jura-Punkte" – ergeben ("Umrechnungstabelle"). Dabei können die Noten schlicht linear anhand der Anzahl der Rohpunkte vergeben werden; genauso ist aber auch eine nicht lineare Umrechnung denkbar (siehe dazu Abb. 3 unten).

Ähnlich ist die Methode der Verwendung von Prozentsätzen der Gesamtnote für die einzelnen Gliede-

37 *Kahneman/Sibony/Sunstein*, Noise: A Flaw in Human Judgment, 2021.

38 Siehe beispielhaft für methodische Lösungsansätze der Problematik von durch „Bauchgefühl-Entscheidungen" in Einstellungsverfahren entstandenen „Biases" und „Verzerrungen" in Trainingsdaten und die daraus resultierenden Diskriminierungen im Recruiting-Sektor https://algorithmwatch.ch/de/findhr/#toolkits.

39 S. dazu bereits oben unter B. II. 2.

40 Ansätze dieser Richtung verfolgt die Ausbildungszeitschrift JuS seit einiger Zeit. Allerdings ist die Punkteverteilung sehr grob (sie arbeitet nur mit ganzen juristischen Punkten) und ohne Berücksichtigung der methodischen Kompetenzen. Vgl. die Bewertungsbögen unter https://rsw.beck.de/zeitschriften/jus/jus-klausurbewertungsb%C3%B6gen.

41 Möglich ist auch die Bezeichnung als „Bewertungseinheiten" um den Unterschied zu den Noten, die im juristischen Kontext als „Punkte" bezeichnet werden, zu verdeutlichen; s. z. B. *Martin Heidebach* (Fn. 27), S. 205 ff.

rungspunkte der Lösungsskizze, die hinterher zusammengerechnet werden. Bei dieser Methode fällt die anschließende Umrechnung in juristische Notenpunkte weg. Die Noten werden hier also im Ergebnis stets linear errechnet.

bb) Bedenken und offene Fragen der Verwendung

(1) Flexibilität der Bewertung

Zunächst besteht „bei den Aufgabenstellenden die [Befürchtung], die Flexibilität der Korrektur könnte bei (zu vielen oder zu genauen) Gewichtungsvorgaben verloren gehen".[42] Ein Rohpunkteschema dürfte sich ohnehin nicht für jede Klausur gleichermaßen eignen – je mehr wertende Fragen in der Klausur den Schwerpunkt bilden, umso schwieriger stellt es sich dar, diese Aspekte in ein entsprechendes Korsett zu zwingen. Allerdings kann die Folge nicht sein, dass in solchen Fällen völlig ohne Maßstab korrigiert wird. Auch bei derartigen Klausuren ist es zumindest möglich, eine Gewichtung der verschiedenen Teile zueinander vorzugeben, so dass der grobe Maßstab vereinheitlicht ist, wenn es auch schwerfallen mag, die Gewichtung im Detail allgemein festzulegen. Bei anderen Klausuren, die stärker an einem bestimmten Lösungsschema orientiert sind (wie beispielsweise die Klausuren im öffentlichen Recht im Regelfall), können auch detaillierte Rohpunkte-Vorgaben möglich sein. Dabei muss sich vergegenwärtigt werden, dass in der Praxis vermutlich jede korrigierende Person – zumindest unterbewusst – irgendeinen Maßstab im Kopf oder auch tatsächlich „auf dem Zettel" hat, denn wie soll sonst eine Korrektur überhaupt möglich sein.

Insgesamt ist angesichts der beschriebenen Besonderheiten des Prüfungsgegenstands der bestehende Beurteilungsspielraum der Prüfenden kein zu beseitigendes Übel, sondern liegt in der Natur der Sache. Völlig identische Korrekturen verschiedener Prüfender sind deshalb zwar anzustreben, jedoch wohl aufgrund des stets verbleibenden menschlichen Faktors nicht realistischerweise erreichbar. Vielmehr geht es darum, Vereinheitlichung zu schaffen, wo sie möglich ist, um – vermeidbare – Schwankungen in der Notengebung zu reduzieren. Das Rohpunkteschema könnte dafür ein geeignetes Instrument sein.

(2) Fachliches Wissen („Sachpunkte") und methodische Kompetenzen („Strukturpunkte")

Eng verbunden mit dem Problem der Flexibilität der Korrektur ist die wichtige Frage, ob und wie neben dem fachlichen Wissen die weiteren zu prüfenden Kompetenzen bewertet werden sollen.

(a) Grundsätzliche Probleme

Gem. § 5a II 3 DRiG („Studium") zählen zu den Pflichtfächern des juristischen Studiums „die Kernbereiche des Bürgerlichen Rechts, des Strafrechts, des Öffentlichen Rechts und des Verfahrensrechts einschließlich der […] rechtswissenschaftlichen Methoden". Der sehr spärliche § 5d DRiG („Prüfungen") hingegen macht die Methodenkompetenz nicht zwingend zum Prüfungsgegenstand, sondern überlässt die nähere Regelung in seinem Absatz 6 dem (sehr unterschiedlich ausgestalteten) Landesrecht.

Schlechte Methodenkompetenz der zu Prüfenden wird seit langem immer wieder bemängelt[43] und jedenfalls ist eine exzellente juristische Leistung ohne diese Kompetenz schwer denkbar.

Die AutorInnen sprechen sich daher dafür aus, neben einer bestimmten Anzahl an Rohpunkten für die rechtsdogmatische Fachkompetenz (zu der auch „Faktenwissen" gehört, das oftmals auswendig gelernt wird) auch eine Anzahl an Rohpunkten für methodische Kompetenzen zu vergeben.

(b) Kriterien für die Vergabe von Strukturpunkten

Bevor über eine (verallgemeinerbare) Verteilung der Strukturpunkte nachgedacht werden kann, müsste zunächst ein einheitliches Verständnis über diese Kompetenzen bestehen. Schon daran fehlt es aber, so dass auf keine existierende Definition für den Begriff der „methodischen Kompetenz" zurückgegriffen werden kann. Für die Beteiligten des DigitalProjekts zählen dazu (für die Verteilung der Strukturpunkte) insbesondere: (Sinnvol-

42 *Urs Kramer/Bernadette Hauser*, Die Korrektur juristischer Arbeiten – ist sie heute schon auf Examensniveau? Ergebnisse eines „Feldversuches", in: Brockmann/Pilniok (Hrsg.), Prüfen in der Rechtswissenschaft, 2013, S. 144 (154), https://www.nomos-elibrary.de/de/10.5771/9783845245171-144/die-korrektur-juristischer-arbeiten-ist-sie-heute-schon-auf-examensniveau-ergebnisse-eines-feldversuches?page=1 (paywall). Zu den verschiedenen Bedenken und möglichen Gegenargumenten auch *Martin Heidebach* (Fn.27), S. 207 ff.

43 Vgl. etwa den Bericht des Ausschusses der Konferenz der Justizministerinnen und Justizminister zur Koordinierung der Juristenausbildung, „Juristin und Jurist der Zukunft", Frühjahr 2024, S. 90, https://www.mj.niedersachsen.de/download/208142/zu_TOP_I.4_Anlage_Bericht_zur_Koordinierung_der_Juristenausbildung_nicht_barrierefrei_.pdf.

le) Anwendung der juristischen Gliederung, präzise und korrekte Normzitate, Anwendung des juristischen Gutachtenstils (im Ersten Juristischen Staatsexamen), Arbeit mit den juristischen Auslegungsmethoden und Normkollisionsregeln, gute Argumentationstechnik (insbesondere die Auswahl spezifisch juristischer Argumente, das Sortieren und Gewichten von Argumenten; die Logik/Kohärenz der Argumentation), Schwerpunktsetzung (etwa dass Unwichtiges nicht zu breit behandelt wird), Problembewusstsein und Arbeit mit dem Sachverhalt, Struktur/Reihenfolge der Prüfung sowie sprachliche Aspekte (korrekte Verwendung juristischer Fachtermini und allgemein korrekte Sprache[44]).

(c) Integrierte oder separate Vergabe der Strukturpunkte

Außerdem muss entschieden werden, wie die Vergabe der Strukturpunkte im Rohpunkteschema umgesetzt wird. Dafür bestehen zwei Möglichkeiten: In Modell 1 ist neben der Verteilung der Rohpunkte im Korrekturbogen ein separater „Block" an Rohpunkten für die Methodenkompetenz zu reservieren. Bei Modell 2 wird die Bewertung der Methodenkompetenz jeweils in die Einzelbewertung eines jeden Abschnittes integriert, das heißt bei jeder konkreten Verteilung der einem bestimmten Abschnitt im Korrekturbogen zugewiesenen Punkte berücksichtigt.

Bei der separaten Ausweisung (Modell 1) lässt sich einfacher festlegen, in welchem Umfang fachliche und methodische Aspekte zueinander gewichtet werden sollen, weil klar ist, wie viele Rohpunkte insgesamt als „Strukturpunkte" für die Bewertung der Methodenkompetenz zur Verfügung stehen. Für einen derartigen separaten Bewertungsblock sprechen die einfachere Erstellung einer Lösungsskizze und die höhere Flexibilität bei der Vergabe der Strukturpunkte. Zudem kann es Aspekte geben, die nur im Gesamtkontext und nicht auf Absatzebene bewertet werden können – etwa ob im Strafrecht sinnvolle Tatkomplexe gebildet wurden. Diese könnten so besser bei der Ermittlung der Gesamtnote berücksichtigt werden.[45]

Demgegenüber liegt ein Vorteil der Integration in die Einzelbewertung eines jeden Abschnittes (Modell 2) darin, dass sich Aspekte der Fach- und Methodenkompetenz teilweise nur schwer voneinander trennen lassen. Außerdem wird so besser nachvollziehbar, an welcher Stelle die Methoden gut oder schlecht angewendet wurden. Denkbar wäre es sogar, die Rohpunkte bei jedem einzelnen Abschnitt in Sach- und Strukturpunkte aufzuteilen (denn nicht überall ist etwa der Gutachtenstil einzuhalten und nicht überall muss mit den Auslegungsmethoden gearbeitet werden). Dies würde den Aufwand der Erstellung des Korrekturbogens jedoch weiter erhöhen und die Flexibilität der Korrektur stärker beschränken, so dass sich die Frage nach der Praktikabilität stellt. Die Schwerpunktsetzung wird bei beiden Modellen bereits darüber mitbewertet, dass für unterschiedlich wichtige Aspekte der Lösung unterschiedlich viele Rohpunkte vergeben werden.

Der Verfasser der Klausur in Phase II (München), Martin Heidebach, folgte bei dem von ihm erstellten und in langjähriger Praxis erprobten Rohpunkteschema dem Modell 2 der integrierten Vergabe von Strukturpunkten. Dabei sah er innerhalb der Abschnitte, bei denen es besonders auf die Anwendung der Methodenkompetenzen ankam, einen hohen Bewertungsspielraum für die Korrektur vor und machte gleichzeitig für die konkrete Vergabe der Strukturpunkte keine Vorgaben. Dadurch nähern sich beide Modelle im konkreten Fall stark an. Die Verfasserin der Klausur in Phase III (Potsdam), Susanne Hähnchen, wird das Modell 1 verwenden.

(d) Anteil der Strukturpunkte

Wie viel Prozent der gesamten Rohpunkte als Strukturpunkte vergeben werden sollten, lässt sich nicht pauschal festlegen. In Anfängerklausuren wird möglicherweise ein größerer Fokus auf dem Gutachtenstil und damit dem Methodenaspekt liegen, während in den Examensklausuren die Fachkompetenz weiter in den Vordergrund rückt. Eine „Standardklausur" verlangt mehr Wissen, eine „exotischere" Klausur mehr methodische Kompetenzen. Teilweise wird vertreten, dass man ohne korrekte Verwendung des Gutachtenstils gar nicht bestehen kann.[46]

Den AutorInnen erscheint eine für die jeweiligen Klausurtypen zu konkretisierende Spannbreite von 15 bis 25 Prozent Anteil von Strukturpunkten an den Rohpunkten sinnvoll.

44 Ob, ab wann und wie sprachliche Mängel in die Bewertung einfließen, ist ein Thema für sich, zu dem verschiedene Positionen vertreten werden und das hier nicht näher diskutiert wird. Jedenfalls ist Recht heute ohne (präzise) Sprache unmöglich. Einer ggf. vertieften, separaten Analyse bedarf die damit verbundene Gerechtigkeitsfrage, welche Auswirkungen dies auf Personen mit anderen sprachlichen Wurzeln hat.

45 S. dazu unten unter C. III.2. b) bb) und unten C. III.2. b) bb) (3) (b).

46 Zu dem damit zusammenhängenden Problem der "Kardinalfehler" s. unten C. III.2. b) bb) (3) (b).

(3) Wie kommt die Gesamtnote zustande?

Unabhängig davon, ob man Rohpunkte weiter in Sach- und Strukturpunkte unterteilt, muss aus den erreichten Rohpunkten eine juristische Gesamtnote (summatives Feedback) gebildet werden. Für die Festlegung der „Umrechnungstabelle" stellen sich vor allem zwei Fragen.

(a) Erforderlicher Anteil an Rohpunkten für das Bestehen

Aus der Anerkennung von im Ausland erbrachten juristischen Leistungen ist das Problem der Umrechnung bereits bekannt. Hier werden typischerweise – zumindest innerhalb einer Fakultät – einheitliche Umrechnungstabellen verwendet.

Auch beim Einsatz eines Rohpunkteschemas geht es um die allgemein-rechtsdidaktische Frage:

Welche Leistung ist erforderlich, um zu bestehen, also wie viele Rohpunkte werden für 4 „Jura-Punkte" benötigt? Sind für jeden juristischen Punkt gleich viele Rohpunkte erforderlich (lineare Skalierung der Rohpunkte auf die juristischen Punkte) oder benötigt man für das Bestehen mehr Rohpunkte als z.B. für die Differenz zwischen 15 und 16 juristischen Punkten? Abb. 3 verdeutlicht den Unterschied der beiden Ansätze. Die nichtlineare Umrechnungstabelle geht dabei von der – ebenfalls zur Diskussion zu stellenden – Annahme aus, für das Bestehen seien 40 Prozent der Rohpunkte erforderlich. Hier erfolgt die Verteilung ab der Bestehensgrenze linear. Denkbar wäre auch, die Umrechnung im weiteren Verlauf auszudifferenzieren, etwa durch stärke Abstände bei den mittleren Noten, so dass für das Erreichen der guten Noten eine besonders hohe Zahl an Rohpunkten notwendig wäre.

Ein Anliegen dieses Forschungsprojekts war und ist es, zu untersuchen, inwiefern der Einsatz von KI-Unterstützung im Korrekturverfahren dazu beitragen kann, gruppenbasierte Verzerrungen in der Bewertung juristischer Klausuren zu verringern. Die KI-Korrektur könnte die Möglichkeit eröffnen, systematische Tendenzen zur Ergebnisnormierung – etwa bedingt durch implizite Vergleichsmaßstäbe oder subjektive Erwartungshaltungen menschlicher KorrektorInnen – zu identifizieren und durch ein vergleichbares, konsistentes Anforderungsprofil zu ersetzen. Dabei kann die KI eine doppelte Funktion übernehmen: Einerseits könnte sie als Analyseinstrument zur Sichtbarmachung bestehender Bewertungsungleichgewichte dienen und implizite Bewertungskriterien offenlegen, andererseits könnte sie in Zukunft als potenzielles Werkzeug zur Objektivierung und Standardisierung der Klausurbewertung genutzt werden. Offen bleibt dabei einstweilen die (allgemein ungeklärte) Frage, was eine wünschenswerte Notenverteilung wäre – die aktuell übliche oder eine gleichmäßigere Ausschöpfung aller juristischer Notenpunkte.

(b) Reine Addition der Rohpunkte?

Die zweite Frage ist, ob sich die Gesamtnote am Ende tatsächlich einfach aus der Addition der Rohpunkte ergeben kann. Dabei geht es zunächst darum, wie das Vorhandensein eines „Kardinalfehlers" zu bewerten ist, danach um das spiegelbildliche Problem des Fehlens „unverzichtbarer Teile" einer Klausur. Es ist offensichtlich, dass es massive Folgen für die Einheitlichkeit einer Korrektur haben muss, wenn die verschiedenen KorrektorInnen mit unterschiedlichen Vorstellungen über diese Fragen arbeiten.

Als „Kardinalfehler" lassen sich Fehler bezeichnen, deren Vorliegen für sich genommen dazu führt, dass eine Klausur nicht bestanden wird. Was dazu zählt, wird aber in der Praxis sehr unterschiedlich gehandhabt. Für manche Korrekturkraft ist es unverzeihlich, wenn gesetzliche Vorschriften nicht genaustens mit Absatz, Satz

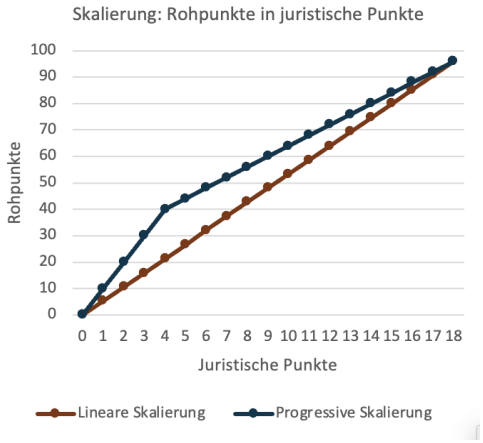


Abb. 3: Vergleich zwischen linearer und progressiver Verteilung der Rohpunkte bei der Umrechnung in juristische Notenpunkte.

etc. oder gar überhaupt nicht zitiert werden – andere sehen das großzügiger. Unter ZivilrechtlerInnen besteht etwa grundsätzlicher Konsens darüber, dass der Verstoß gegen das Trennungs- und Abstraktionsprinzip ein grober Fehler ist. Aber nicht jede oder jeder würde deshalb einen Prüfling durchfallen lassen. Teilweise wird nach dem Stand der Ausbildung differenziert, teilweise wird großzügig eine nur sprachliche Ungenauigkeit vermutet. Problematisch ist es z.B. auch, zwar eine inhaltlich sinnvolle Prüfung, aber entlang der falschen Anspruchsgrundlage vorzunehmen („Folgefehler"). Es zeigt sich also, dass es „den" Kardinalfehler so nicht gibt. Von daher wäre es wünschenswert, dass im Korrekturbogen definiert wird, was als Kardinalfehler für die vorliegende Prüfung gilt. Zu überlegen ist, ob und wie sich diese Besonderheiten im Rohpunkteschema angemessen abbilden lassen. Denn in der Regel dürfte es nicht genügen, lediglich die für den Abschnitt vorgesehenen Rohpunkte nicht zu vergeben, um derartige Fehler ausreichend (negativ) zu berücksichtigen. Entweder muss für diesen Fall die Option eröffnet werden, eine von der errechneten Anzahl der Rohpunkte abweichende, schlechtere Note zu vergeben oder für diese Fehler müssen Minus-Rohpunkte vorgesehen sein.

Was als „unverzichtbarer Teil" einer Klausur gilt, lässt sich ebenfalls nur klausurspezifisch beantworten – und auch dort nur schwer: Reicht es beispielsweise für das Bestehen einer Klausur, in der eine Verfassungsbeschwerde zu prüfen ist, aus, wenn eine Person alle Aspekte der Begründetheit vollständig anspricht, aber nichts zur Zulässigkeit schreibt? Oder muss zumindest alles in Grundzügen erkannt und bearbeitet worden sein? Eine Möglichkeit, dieses Problem im Rohpunkteschema zu berücksichtigen, wäre, für den unverzichtbaren Teil eine bestimmte Mindestpunktzahl festzulegen, die erreicht werden muss, damit man a) bestehen kann und b) alle Punkte gewertet werden.

(4) Kommunikation der Bewertungen (Transparenz)

„Transparenz ist problematisch" kann das letzte hier zu thematisierende Argument (nicht nur) bei der Vergabe von Rohpunkten sein. Oft wird argumentiert, man mache sich mit der (aufwendigen und umfangreichen) Offenlegung aller für die Bewertung relevanten Aspekte (für Remonstration oder Widerspruch) angreifbarer. Es könnten die „Studierenden versuchen, um jede BE [Bewertungseinheit] zu feilschen".[47] Bei Verwendung von Rohpunkten ist diese Sorge nicht abwegig, gerade wenn nur ein einziger Rohpunkt fehlt, um den nächsten juristischen Punkt zu erreichen. Daher muss man sich fragen, ob die vergebenen Rohpunkte transparent kommuniziert werden sollten.

Aktuell ist die Intransparenz – wie bereits dargestellt[48] – ein besonderes Problem der juristischen Notengebung. Sie darf – gerade im Rechtsstaat, den die PrüferInnen an der Universität repräsentieren – kein Mittel sein, um Studierende durch das Fehlen von Angriffsfläche von in der Sache berechtigter Kritik an einer konkreten Note abzuhalten. Angst vor Kritik ist also schon im Ausgangspunkt kein berechtigter Einwand.

Andererseits ist das reine „Berechnen" von Noten[49] in einem Fach, welches keine Naturwissenschaft ist und weniger Fakten abprüft als vielmehr sehr spezifische Anforderungen hat, nicht unproblematisch. Eine hundertprozentig exakte Bewertung wäre wohl nur bei (möglichst einfach gestalteten) Multiple-Choice-Tests möglich. Diese können jedoch sinnvoll primär zur Wissensabfrage eingesetzt werden, weniger um die – auch in der Praxis relevante – Anwendung methodischer Kompetenzen zu prüfen.[50]

Ein Kompromiss könnte es sein, zwischen dem (im Vorstehenden beschriebenen und empfohlenen) Prozess der Notenfindung und der Kommunikation wie folgt zu differenzieren: Das Vertrauen in den Prozess sollte durch Erläuterung des Zustandekommens von juristischen Noten verbessert werden. An die Prüflinge sollte jedoch nur die Gesamtnote und ein ausführliches formatives Feedback anstelle der Rohpunkte ausgegeben werden. Damit wäre im Vergleich zum aktuellen Zustand schon viel gewonnen.

(5) Fazit zum Einsatz des Rohpunkteschemas

Die geschilderten Einwände und etwaige weitere könnte man noch detaillierter diskutieren, was hier jedoch den Rahmen sprengen würde. Sie führen jedenfalls nicht dazu, das Hauptargument für den Einsatz des Rohpunk-

47 *Martin Heidebach* (Fn. 27), S. 211. Diese Befürchtung wird dort allerdings widerlegt und für volle Transparenz plädiert.

48 S. dazu bereits oben unter A. sowie unter C. I. 4.

49 S. dazu bereits oben unter C. III. 2. b) bb) und (3) (b).

50 Multiple-Choice-Tests spiegeln nach *Sebastian Dötterl* (Fn. 4), S. 165: „offensichtlich nicht die Praxis", bieten jedoch Vorteile bei Aufwand/Kosten und Chancengerechtigkeit. Differenzierend zum Einsatz hingegen *Stefan Griller/Monika Stöggl*, Multiple Choice als Prüfungsform bei rechtswissenschaftlichen Fachprüfungen, ZDRW 2018, 148 ff., https://doi.org/10.5771/2196-7261-2018-2-148.

teschemas in Frage zu stellen: Erreichen von mehr Objektivität und damit gerechterer Benotung. Die Beteiligten des DigitalProjekts sind von dessen grundsätzlicher Nützlichkeit überzeugt.

Ob die Verwendung eines Rohpunkteschemas auch tatsächlich zu einer „besseren" Notenverteilung führt, sich die bisherige Hypothese also empirisch bestätigt, ist Gegenstand der nachfolgenden quantitativen Auswertung.

c) Statistische Auswertung des summativen Feedbacks

Von besonderem Interesse für die Beteiligten – gerade in den abschließenden Prüfungen – ist das Gesamtergebnis in Form des summativen Feedbacks, also die Note.

aa) Vergleich Mensch ohne Rohpunkteschema vs. Mensch mit Rohpunkteschema

Um zu untersuchen, ob und wie sich die menschliche Korrektur mit Rohpunkteschema auf die Notengebung auswirkt, wurden die KorrektorInnen – ohne dies zu wissen – in zwei Gruppen eingeteilt. Die Korrekturergebnisse finden sich in Abb. 4. Alle KorrektorInnen korrigierten jeweils die gleichen sechzehn pseudonymisierten[51] Klausuren (Pseudonym-Nummern in Spalte 1 der Abb. 4).

Zunächst wurden die Klausuren durch den jahrelangen Kursverantwortlichen Martin Heidebach (MH) selbst mit Rohpunkteschema korrigiert (Spalte 2 der Abb. 4).

Sechs KorrektorInnen korrigierten ohne Rohpunkteschema (im Folgenden: oRPS; in Abb. 4 die KorrektorInnen 3, 5, 6, 9, 12 und 13 in den hellblauen Spalten) und acht KorrektorInnen mit Rohpunkteschema (im Folgenden: mRPS; in Abb. 4 die KorrektorInnen 1, 2, 4, 7, 8, 10, 11 und 14 in hellgrün).

Abb. 4 zeigt auch die Durchschnittsnote („MW Note") der jeweiligen KorrektorInnen über alle Klausuren hinweg. Die Differenz zwischen dem Durchschnitt der „strengsten" und „wohlwollendsten" oRPS korrigierenden Person liegt bei 3 Notenpunkten, bei den mRPS korrigierenden Personen lediglich bei 1,94 (je „MW Diff Max-Min" in Abb. 4).

Vergleicht man in (Abb. 5) die von Martin Heidebach vergebenen Noten (MH, dunkelrot) mit dem Mittelwert aller Korrekturen ohne Rohpunkteschema (Ø oRPS, hellblau) sowie aller Korrekturen mit Rohpunkteschema (Ø mRPS, hellgrün), so zeichnet sich insgesamt ein recht homogenes Bild. Bei elf Klausuren (101; 108; 94; 99; 97; 98; 95; 93; 105; 105; 107) ist die Differenz zwischen MH und oRPS größer als die Differenz zwischen MH und

	MH	Externe Korrektoren													
		3	5	6	9	12	13	1	2	4	7	8	10	11	14
#	mRPS	oRPS	oRPS	oRPS	oRPS	oRPS	oRPS	mRPS	mRPS	mRPS	mRPS	mRPS	mRPS	mRPS	mRPS
101	1,00	2	2	3	4	1	1	1	2	1	1	1	1	2	1
88	2,00	2	1	2	3	3	1	2	2	2	2	2	2	2	2
102	2,00	2	3	3	6	1	0	2	3	2	2	2	2	2	3
94	3,00	5	3	5	8	3	1	2	3	4	3	3	2	3	3
99	3,00	4	6	5	5	3	2	3	5	3	3	3	3	4	3
100	5,00	5	4	6	6	2	5	3	6	5	6	3	2	6	6
91	5,00	4	6	8	9	5	4	5	4	5	5	4	3	3	3
97	5,00	8	4	6	6	4	4	4	4	5	7	4	4	8	6
98	7,00	6	5	7	6	4	5	5	4	8	8	6	4	10	5
92	7,00	5	8	6	14	3	4	6	6	7	8	4	4	6	6
95	7,00	7	7	7	6	4	3	7	7	6	8	6	8	8	7
87	10,00	7	7	7	6	8	6	6	6	7	7	3	4	6	6
93	10,00	7	12	7	7	6	6	8	8	8	9	8	5	8	9
105	10,00	10	11	9	7	7	7	9	10	10	11	10	8	10	11
104	11,00	12	15	12	14	14	6	10	13	11	12	10	12	12	10
107	15,00	16	17	12	7	17	11	14	16	14	15	14	12	16	17
	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:	MW Note:
	6,44	6,38	6,94	6,56	7,13	5,31	4,13	5,44	6,19	6,13	6,69	5,19	4,75	6,63	6,13
		MW Diff Max-Min:						MW Diff Max-Min:							
		3,00						1,94							

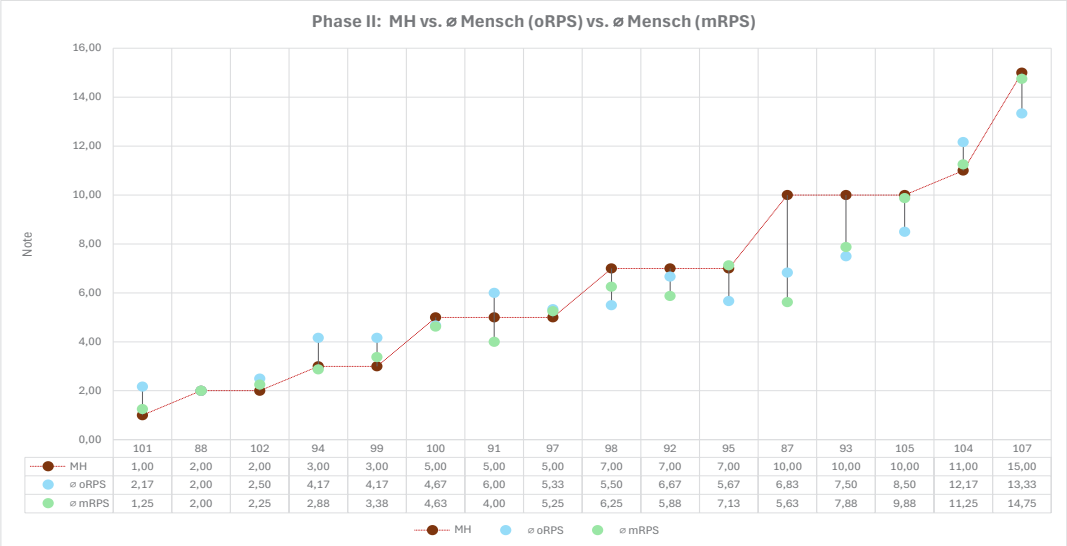
Abb. 4: Übersicht über alle vergebenen Einzelnoten für die 16 Klausuren (Spalte 1) durch Martin Heidebach (Spalte 2), die KorrektorInnen ohne Rohpunkteschema (Spalten 3 – 8) und die KorrektorInnen mit Rohpunkteschema (Spalten 9 – 16).

51 Im späteren Verlauf des Projektes fiel auf, dass die bearbeitenden Personen zwar nicht über die Namen oder den Text der Klausuren einsehbar waren, jedoch mit weiteren Klicks über die Metadaten der Dateien.

mRPS. Im unteren und oberen Notenspektrum sind die Differenzen zwischen MH und oRPS besonders hoch. Insbesondere im mittleren Notenspektrum liegen die Noten von MH vermehrt über den beiden Durchschnittswerten der Vergleichsgruppen.

Besonders aussagekräftige Einblicke ergeben sich aus den statistischen Auswertungen der Korrekturergebnisse der beiden Vergleichsgruppen mRPS und oRPS (Abb. 6). Im Mittelwert aller Noten (MW Note) liegen beide Gruppen mit einem Gesamtdurchschnitt von 6,07 Punkten (oRPS) und 5,89 Punkten (mRPS) zunächst erneut nah beieinander. Die Abweichung der Durchschnittsnote zur Korrektur von MH liegt für die oRPS-Gruppe im gesamten Mittelwert über alle Klausuren hinweg (MW Diff MH) bei 1,18 Punkten und für die mRPS-Gruppe bei 0,73 Punkten.

Der Notendurchschnitt aller Korrektoren entfaltet auf Einzelfallebene jedoch wenig Bedeutung. Klausuren werden in der juristischen Ausbildung nie von vier oder mehr Personen parallel korrigiert. Sich in der Vielzahl mittelnde Werte können durch einzelne Ausreißer nach oben und unten zu erheblichen Notenungerechtigkeiten



	101	88	102	94	99	100	91	97	98	92	95	87	93	105	104	107
MH	1,00	2,00	2,00	3,00	3,00	5,00	5,00	5,00	7,00	7,00	7,00	10,00	10,00	10,00	11,00	15,00
⌀ oRPS	2,17	2,00	2,50	4,17	4,17	4,67	6,00	5,33	5,50	6,67	5,67	6,83	7,50	8,50	12,17	13,33
⌀ mRPS	1,25	2,00	2,25	2,88	3,38	4,63	4,00	5,25	6,25	5,88	7,13	5,63	7,88	9,88	11,25	14,75

Abb. 5: Vergleich der von Martin Heidebach (MH) vergebenen Noten (dunkelrot) mit dem Durchschnitt aller Korrekturen oRPS (hellblau) und dem Durchschnitt aller Korrekturen mRPS (hellgrün); sortiert nach Klausur in aufsteigender Reihenfolge der Noten von MH. (STABW.S) sowie die Varianz (VAR.S) der Korrekturen untereinander.

	Auswertung										
	oRPS					mRPS					Alle
#	⌀	Diff MH	Diff Max-Min	STABW.S	VAR.S	⌀	Diff MH	Diff Max-Min	STABW.S	VAR.S	⌀
101	2,17	1,17	3,00	1,17	1,37	1,25	0,25	1,00	0,46	0,21	1,56
88	2,00	0,00	2,00	0,89	0,80	2,00	0,00	0,00	0,00	0,00	2,00
102	2,50	0,50	6,00	2,07	4,30	2,25	0,25	1,00	0,46	0,21	2,31
94	4,17	1,17	7,00	2,40	5,77	2,88	0,13	2,00	0,64	0,41	3,38
99	4,17	1,17	4,00	1,47	2,17	3,38	0,38	2,00	0,74	0,55	3,63
100	4,67	0,33	4,00	1,51	2,27	4,63	0,38	4,00	1,69	2,84	4,69
91	6,00	1,00	5,00	2,10	4,40	4,00	1,00	2,00	0,93	0,86	4,88
97	5,33	0,33	4,00	1,63	2,67	5,25	0,25	4,00	1,58	2,50	5,25
98	5,50	1,50	3,00	1,05	1,10	6,25	0,75	6,00	2,19	4,79	6,06
92	6,67	0,33	11,00	3,98	15,87	5,88	1,13	4,00	1,36	1,84	6,31
95	5,67	1,33	4,00	1,75	3,07	7,13	0,13	2,00	0,83	0,70	6,56
87	6,83	3,17	2,00	0,75	0,57	5,63	4,38	4,00	1,41	1,98	6,63
93	7,50	2,50	6,00	2,26	5,10	7,88	2,13	4,00	1,25	1,55	8,00
105	8,50	1,50	4,00	1,76	3,10	9,88	0,13	3,00	0,99	0,98	9,38
104	12,17	1,17	9,00	3,25	10,57	11,25	0,25	3,00	1,16	1,36	11,56
107	13,33	1,67	10,00	4,03	16,27	14,75	0,25	5,00	1,58	2,50	14,25
	MW Note:	MW Diff MH:	MW Diff Max-Min:	MW STABW.S	MW VAR.S	MW Note:	MW Diff MH:	MW Diff Max-Min:	MW STABW.S	MW VAR.S	MW Note:
	6,07	1,18	5,25	2,01	4,96	5,89	0,73	2,94	1,08	1,46	6,03
		Summe Diff ⌀MH:	Range:				Summe Diff ⌀MH:	Range:			
		18,83	2-11				11,75	0-6			

Abb. 6: Statistische Auswertung der Korrekturergebnisse unterteilt nach Korrekturen oRPS (hellblau) und mRPS (hellgrün). Je Vergleichsgruppe zu sehen ist u.a. die Durchschnittsnote (Ø), die Differenz der Durchschnittsnote zur von Martin Heidebach vergebenen Note (Diff MH), die Abweichung zwischen der besten und der schlechtesten Note („Diff Max-Min"), die Standardabweichung (STABW.S) sowie die Varianz (VAR.S) der Korrekturen untereinander.

führen. Aufgrund des Umstandes, dass sich die Examensnote zu großen Teilen aus wenigen Examensklausuren und gerade nicht auch aus allen universitären Klausuren zusammensetzt, findet im Einzelfall auch nur eine geringe Mittelung über den Ausbildungsweg hinweg statt. Aufgrund der Bedeutung einzelner Korrekturen für den individuellen Lebensweg kommen daher Metriken wie der Range (Spannweite / Max-Min-Differenz) und der Standardabweichung erhebliche Bedeutung zu.

Blickt man in Abb. 6 auf die Range, also die Abweichung zwischen der besten und der schlechtesten vergebenen Note („Diff Max-Min"), so lassen sich enorme Differenzen erkennen. Bei den Korrekturen oRPS lag die maximale Range bei Klausur 92 bei 11 Punkten und bei keiner Klausur unter 2 Punkten. Bei den Korrekturen mRPS lag die maximale Range bei Klausur 98 bei 6 Punkten und bei Klausur 88 gar bei 0 (was bedeutet, dass alle acht KorrektorInnen die exakt gleiche Note vergeben haben). Der Mittelwert der Range aller Klausuren liegt für die Gruppe oRPS bei 5,25 Punkten. Dies deckt sich in etwa mit den Ergebnissen der Studie von Clemens Hufeld, in der dieser Wert sogar bei 6,47 Punkten lag. Bei der Gruppe mRPS liegt dieser Wert lediglich bei 2,94 Punkten. Im Schnitt hat man bei KorrektorInnen, die mit Rohpunkteschema korrigieren also mit 2,31 Punkten (Differenz zwischen 5,25 Punkten oRPS und 2,94 Punkten mRPS) weniger Notendifferenz zu rechnen. Die KorrektorInnen mRPS hatten somit etwa 44 % weniger Range als die KorrektorInnen oRPS.

Abb. 7 zeigt die Range der Gruppen oRPS und mRPS je Klausur (sortiert nach aufsteigenden Durchschnittsnoten) im Vergleich. Mit Ausnahme des Ausreißers der Klausur 92 lässt sich erkennen, dass die Range der Korrekturen oRPS (hellblaue Balken) im unteren sowie im oberen Notenspektrum (insb. bei den Klausuren 102, 94, 104 und 107) besonders hoch ist. Die Range der Korrekturen mRPS (hellgrüne Balken) ist im unteren Notenspektrum hingegen besonders gering. Bei allen Klausuren, die im Gesamtdurchschnitt unter vier Punkten lagen (Klausuren 101, 88, 102, 94 und 99) liegt die Range der Korrekturen mRPS im Schnitt bei lediglich 1,2 Punkten. Für die Vergleichsgruppe oRPS liegt dieser Wert hingegen bei 4,4 Punkten. Gerade im unteren und oberen Notenspektrum fällt die Bewertung mRPS also deutlich homogener aus.

Anhand der Klausur 102 dargestellt bedeutet dies etwa Folgendes: Die acht KorrektorInnen mRPS haben durchgängig entweder 2 oder 3 Punkte (Range: 1) vergeben. Die sechs KorrektorInnen oRPS haben hingegen Noten zwischen 0 und 6 Punkten vergeben (Range: 6). Allein der Blick auf den Durchschnitt hätte diese Differenzen nicht offenbart. Der Durchschnitt der Korrektoren oRPS liegt bei 2,5 Punkten und der Durchschnitt der Korrektoren mRPS bei 2,25 Punkten.

Als weiterer Kennwert kann die in Abb. 8 detailliert dargestellte Standardabweichung herangezogen werden. Die Standardabweichung misst, wie stark die Werte im Durchschnitt vom Mittelwert abweichen. Sie ist im Unterschied zur Range weniger anfällig für Ausreißer (also einzelne Extremwerte). Eine kleine Standardabweichung bedeutet, dass die Werte eng am Mittelwert liegen und somit weniger Schwankungen unterliegen. Der Mittelwert der Standardabweichungen der Korrekturen oRPS

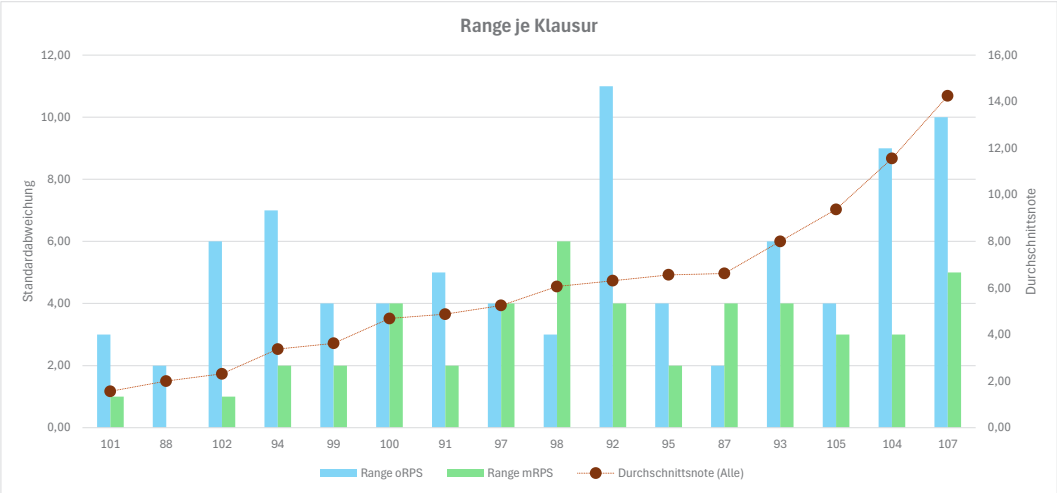


Abb. 7: Range (Differenz zwischen schlechtester und bester Note) der Korrekturen oRPS (hellblaue Balken) und mRPS (hell-grüne Balken) je Klausur; sortiert nach aufsteigender durchschnittlicher Klausurnote (rote Punkte).

liegt bei 2,01 und jener der Korrekturen mRPS bei lediglich 1,08. In Abb. 8 ist diese – analog zur Range – nochmals für die Gruppen oRPS und mRPS je Klausur dargestellt. Es ergibt sich ein nahezu identisches Bild zur Range, was die dort getroffenen Erkenntnisse – insbesondere dazu, dass die Abweichung der Korrekturen oRPS im unteren und oberen Notenbereich besonders hoch ist – stützt.

Auch die durchschnittliche Stichprobenvarianz der Korrekturen (siehe Abb. 6) oRPS liegt bei 4,96 und die der Klausuren mRPS bei 1,46. Das bedeutet: oRPS-Korrektoren streuen stärker in ihren Bewertungen als mRPS-Korrektoren.

Im Ergebnis weisen die Noten der Korrektoren, die ein Rohpunkteschema ausgehändigt bekommen haben (mRPS) im Vergleich zur Testgruppe an Korrektoren, die ohne Rohpunkteschema korrigierten (oRPS) eine deutlich geringere mittlere Standardabweichung (1,08 zu 2,01), eine um etwa 44 % geringere durchschnittliche Range (2,94 zu 5,25) und eine erheblich geringere mittlere Varianz (1,46 zu 4,96) auf. Im Notenbereich unter vier Punkten ist die durchschnittliche Range der Korrekturen mRPS gar um ca. 72% geringer als die der Korrekturen oRPS (1,2 zu 4,4). KorrektorInnen mit Rohpunkteschema bewerten erheblich[52] homogener als KorrektorInnen ohne Rohpunkteschema.

Die empirische Auswertung zeigt daher: Die Verwendung eines Rohpunkteschemas führt zu einer objektiveren Notengebung.

Denkbar sind nach Ansicht der Autoren daher künftig unterschiedliche Stadien der „Verbindlichkeit" von Rohpunkteschemata. Diese können lediglich als „interne Arbeitshilfe" an die Korrigierenden ausgegeben werden, die jedoch weiterhin nur eine juristische Note abgeben. Die Rohpunkte können aber auch den Kursverantwortlichen/Prüfungsämtern übermittelt werden, um die Notenvergabe intern besser nachkontrollierbar zu gestalten. Oder die Rohpunkteschemata werden zusätzlich zur juristischen Note ausgefüllt an die Studierenden zurückgegeben. Der Abgleich von Lösungsskizze und Klausur und die Vergabe der Rohpunkte sorgen somit für maximale Transparenz und einen höchstmöglichen Lerneffekt.

bb) Vergleich Mensch vs. KI

KlausurenKIste und DeepWrite haben jede Klausur zweimal mit Gemini 2.5 Pro und zweimal mit GPT-4o korrigieren lassen. In beiden Durchgängen wurde dem System die Lösungsskizze mitgegeben, es erfolgte jedoch jeweils ein Durchgang mit zusätzlicher Eingabe eines Rohpunkteschemas (mRPS) als Inputdaten und einer ohne (oRPS).

Die Abweichung (Diff) zwischen den Noten von MH und der jeweiligen Durchschnittsnote der menschlichen Korrekturgruppe oRPS liegt im gesamten Mittelwert über alle Klausuren hinweg bei 1,18 Punkten und für die mRPS-Gruppe bei 0,73 Punkten.[53] Der Mittelwert der Abweichung der KI-Korrekturen zur Korrektur von

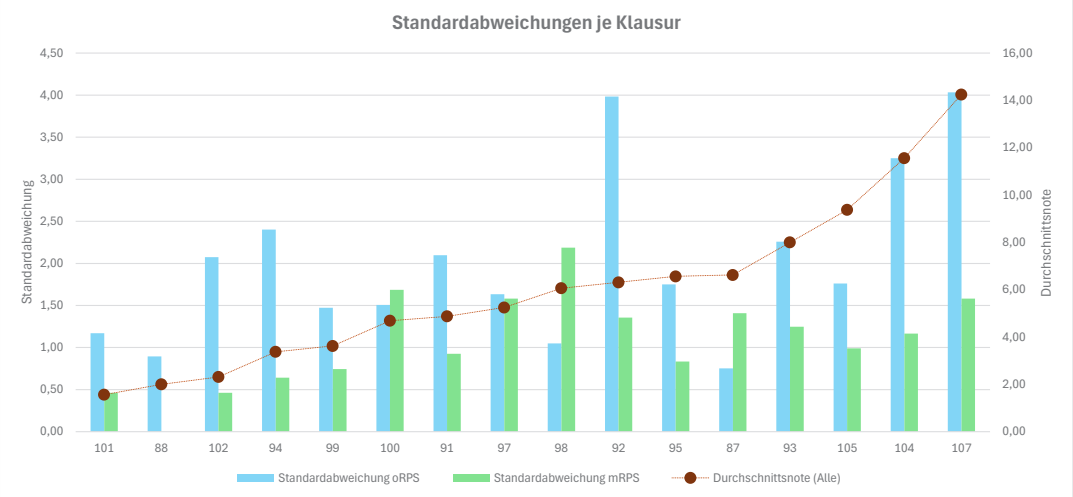


Abb. 8: Standardabweichung der Korrekturen oRPS (hellblaue Balken) und mRPS (hellgrüne Balken) je Klausur; sortiert nach aufsteigender durchschnittlicher Klausurnote (rote Punkte).

52 Ob dieser Unterschied auch statistisch signifikant ist, ist aus juristischer Perspektive nur bedingt relevant. Ein gepaarter t-Test mit einer hypothetischen Differenz der Mittelwerte von 0 und einem Signifikanzniveau (Alpha) von 0,05 ergab beidseitig p-Werte (einseitig = 0,0067; zweiseitig = 0,0134) von deutlich unter 0,05. Ob die vorliegenden Daten jedoch hinreichend normalverteilt sind und keine zu extremen Ausreißer für einen belastbaren t-Test enthalten wurde nicht untersucht, weshalb nicht von „signifikanten" Unterschieden gesprochen wird.

53 Siehe oben Abb. 6.

MH liegt für beide Teams, beide Modelle und beide Verfahren (KI-Korrektur mRPS und oRPS) jeweils über diesem Wert (Abb. 9). Die geringste durchschnittliche Differenz zur Korrektur von MH (MW Diff MH) weist bei DeepWrite die Korrektur durch GPT-4o mit Rohpunkteschema auf (2,06 Punkte) und bei KlausurenKIste die Korrektur durch Gemini 2.5 pro mit Rohpunkteschema (1,69 Punkte). Beide Werte liegen – nach zahlreichen zwischenzeitlich erfolgten Architekturverbesserungen – jedoch deutlich unter der Differenz der Ergebnisse zur menschlichen Vergleichskorrektur in Phase I (Bielefeld). Dort betrug der Mittelwert der Abweichung bei KlausurenKIste noch 2,93 und bei DeepWrite noch bei 3,64 (DeepWrite).[54]

Aus bereits dargestellten Gründen[55] wird nachfolgend nicht lediglich die Korrektur durch MH als Vergleichsmaßstab herangezogen. Hierfür wurde stattdessen ein durch die Projektbeteiligten gewichteter Durchschnittswert gebildet. Der Wert „Alle Ø" in Abb. 9 setzt sich daher zusammen aus allen einfach gewichteten Einzelkorrekturen (mRPS und oRPS) sowie der zweifach mit einberechneten Korrektur MH (im Folgenden: gewichteter menschlicher Notendurchschnitt).

Vergleicht man die KI mit dem Durchschnitt einer Vielzahl an menschlichen KorrektorInnen mit und ohne Rohpunkteschema, so liegen die Durchschnittswerte der menschlichen Korrekturen noch jeweils näher an den Noten des Kursbetreuers MH. In der Praxis wird eine Klausur jedoch nahezu nie von mehr als drei Personen mit Bildung eines Durchschnitts korrigiert, weshalb dieser Vergleich wenig Aussagekraft hat.

Betrachtet man hingegen die Streuung der einzelnen Korrekturwerte um den gewichteten menschlichen Notendurchschnitt (Abb. 10), so fällt diese bei den jeweils besten Ergebnissen der beiden KI-Systeme (oben) geringer aus als bei den menschlichen Korrekturen nach dem derzeit etablierten System oRPS (unten). Für diesen Vergleich wurden die jeweils „besten" Korrekturen von DeepWrite oRPS sowie von KlausurenKIste mRPS herangezogen. Zu beachten gilt jedoch, dass die „besten" KI-Modelle hier anhand der geringsten Abweichung zum gewichteten menschlichen Gesamtdurchschnitt ausgewählt werden konnten. Dies wird nicht bei jeder universitären Korrektur möglich sein. In Phase III dieses Projektes wird daher untersucht werden, ob sich erneut die gleichen Modelle als „beste" herauskristallisieren. In Zukunft wird sich die Frage stellen, in welchen Abständen solche Vergleiche mit menschlichen Korrekturen angesichts sich ständig verändernder Modelle stattfinden müssten.

Sowohl bei der menschlichen Korrektur als auch bei der KI-Korrektur gibt es Ausreißer nach oben und nach unten, die Ausreißer bei der KI-Korrektur fallen jedoch geringer aus als die bei der menschlichen Korrektur. Im

	Feedback DeepWrite												Feedback KlausurenKIste											
	Gemini 2.5 Pro						GPT-4o						Gemini 2.5 pro						GPT-4o					
	Feedback mRPS			Feedback oRPS			Feedback mRPS			Feedback oRPS			Feedback mRPS			Feedback oRPS			Feedback mRPS			Feedback oRPS		
#	Note	Diff MH	Diff ⌀ Alle	Note	Diff MH	Diff ⌀ Alle	Note	Diff MH	Diff ⌀ Alle	Note	Diff MH	Diff ⌀ Alle	Note	Diff MH	Diff ⌀ Alle	Note	Diff MH	Diff ⌀ Alle	Note	Diff MH	Diff ⌀ Alle	Note	Diff MH	Diff ⌀ Alle
101	1	0	0,56	2	1	0,44	2	1	0,44	3	2	1,44	3	2	1,44	6	5	4,44	3	2	1,44	7	6	5,44
88	0	2	2,00	3	1	1,00	5	3	3,00	6	4	4,00	3	1	1,00	4	2	2,00	7	5	5,00	4	2	2,00
102	1	1	1,31	2	0	0,31	5	3	2,69	4	2	1,69	3	1	0,69	9	7	6,69	4	2	1,69	3	1	0,69
94	1	2	2,38	2	1	1,38	8	5	4,63	5	2	1,63	6	3	2,63	9	6	5,63	6	3	2,63	8	5	4,63
99	2	1	1,63	2	1	1,63	10	7	6,38	6	3	2,38	4	1	0,38	7	4	3,38	4	1	0,38	7	4	3,38
100	4	1	0,69	2	3	2,69	5	0	0,31	4	1	0,69	5	0	0,31	3	2	1,69	4	1	0,69	4	1	0,69
91	2	3	2,88	2	3	2,88	5	0	0,13	5	0	0,13	6	1	1,13	6	1	1,13	3	2	1,88	4	1	0,88
97	6	1	0,75	2	3	3,25	5	0	0,25	4	1	1,25	6	1	0,75	10	5	4,75	2	3	3,25	5	0	0,25
98	3	4	3,06	3	4	3,06	10	3	3,94	6	1	0,06	7	0	0,94	7	0	0,94	5	2	1,06	6	1	0,06
92	4	3	2,31	3	4	3,31	10	3	3,69	7	0	0,69	9	2	2,69	15	8	8,69	7	0	0,69	8	1	1,69
95	5	2	1,56	3	4	3,56	12	5	5,44	7	0	0,44	8	1	1,44	7	0	0,44	9	2	2,44	5	2	1,56
87	4	6	2,63	3	7	3,63	10	0	3,38	5	5	1,63	8	2	1,38	3	7	3,63	8	2	1,38	10	0	3,38
93	6	4	2,00	2	8	6,00	10	0	2,00	5	5	3,00	5	5	3,00	9	1	1,00	4	6	4,00	9	1	1,00
105	4	6	5,38	3	7	6,38	11	1	1,63	8	2	1,38	10	0	0,63	7	3	2,38	9	1	0,38	4	6	5,38
104	10	1	1,56	2	9	9,56	13	2	1,44	8	3	3,56	8	3	3,56	9	2	2,56	4	7	7,56	3	8	8,56
107	11	4	3,25	5	10	9,25	15	0	0,75	7	8	7,25	11	4	3,25	3	12	11,25	10	5	4,25	10	5	4,25
	MW Note: 4,00	MW Diff MH: 2,56	MW Diff ⌀Alle: 2,12	MW Note: 2,56	MW Diff MH: 4,13	MW Diff ⌀Alle: 3,64	MW Note: 8,50	MW Diff MH: 2,06	MW Diff ⌀Alle: 2,50	MW Note: 5,63	MW Diff MH: 2,44	MW Diff ⌀Alle: 1,95	MW Note: 6,38	MW Diff MH: 1,69	MW Diff ⌀Alle: 1,57	MW Note: 7,13	MW Diff MH: 4,06	MW Diff ⌀Alle: 3,79	MW Note: 5,56	MW Diff MH: 2,75	MW Diff ⌀Alle: 2,42	MW Note: 6,06	MW Diff MH: 2,75	MW Diff ⌀Alle: 2,74
		Summe Diff MH: 41	Summe Diff ⌀Alle: 33,94		Summe Diff MH: 66	Summe Diff ⌀Alle: 58,31		Summe Diff MH: 33	Summe Diff ⌀Alle: 40,06		Summe Diff MH: 39	Summe Diff ⌀Alle: 31,19		Summe Diff MH: 27	Summe Diff ⌀Alle: 25,19		Summe Diff MH: 65	Summe Diff ⌀Alle: 60,56		Summe Diff MH: 44	Summe Diff ⌀Alle: 38,69		Summe Diff MH: 44	Summe Diff ⌀Alle: 43,81

Abb. 9: Ergebnisse der KI-Korrektur unterteilt in DeepWrite (linke Hälfte) und KlausurenKIste (rechte Hälfte). Dargestellt ist die Differenz der jeweiligen Note zur Note von Martin Heidebach (Diff MH) sowie zum gewichteten menschlichen Notendurchschnitt (Diff Ø Alle).

54 Siehe oben Abb. 1.

55 Siehe etwa oben C. III. 2. c) aa).

unteren und oberen Notenbereich ist die Streuung der Korrekturwerte größer als im mittleren Notenbereich, wobei sie im oberen am größten ist.

Die durchschnittliche Differenz zur Notengebung von Martin Heidebach (siehe die jeweiligen Mittelwerte Diff MH in Abb. 9) ist bei beiden Teams bei beiden Modellen mit Rohpunkteschema (mRPS) jeweils geringer oder zumindest gleich, als wenn die Systeme ohne Rohpunkte korrigiert haben (oRPS). Dies legt zunächst den Schluss nahe, dass auch KI-Systeme mit Rohpunkteschemata zu konstanteren Ergebnissen kommen.

Betrachtet man jedoch die Differenz zum gewichteten menschlichen Notendurchschnitt (siehe die jeweiligen Mittelwerte Diff Ø Alle), so zeigt sich ein uneinheitlicheres Bild. Bei KlausurenKIste ist die Differenz zum gewichteten menschlichen Notendurchschnitt erneut geringer, wenn die KI-Systeme das Rohpunkteschema zur Verfügung gestellt bekommen haben, als wenn dies nicht der Fall war: nur 1,57 (mRPS) statt 3,79 (oRPS) Punkte Differenz im Schnitt bei Gemini 2.5 Pro und nur 2,42 (mRPS) statt 2,74 (oRPS) Punkte Differenz im Schnitt bei GPT-4o. Bei DeepWrite liegt Gemini 2.5 Pro mRPS mit 2,12 Punkten Differenz im Schnitt ebenfalls näher am gewichteten menschlichen Notendurchschnitt als oRPS mit 3,64 Punkten Differenz im Schnitt. Die geringste Differenz weist bei DeepWrite mit lediglich 1,95 Punkten Differenz im Schnitt jedoch das GPT-4o-Modell auf, das das Rohpunkteschema nicht zur Verfügung gestellt bekommen hat (oRPS). Mit Rohpunkteschema (mRPS) lag die Differenz dieses Modells im Schnitt hingegen bei 2,5 Punkten.

Ein Erklärungsansatz hierfür könnte darin liegen, dass das System von KlausurenKIste aufgrund der gewählten Architektur der Absatzkorrektur besser zum System der auf der entsprechenden Absatzebene vergebenen Rohpunkte passt, als das System der Gesamtkor-

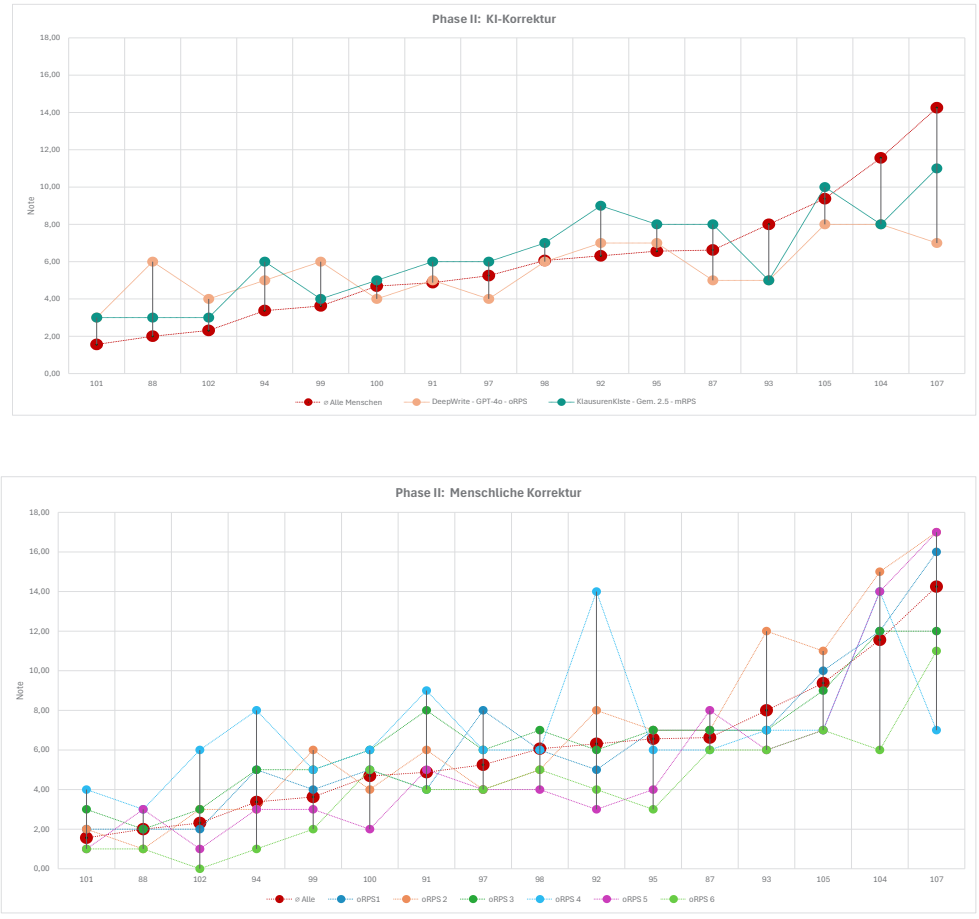


Abb. 10: Abweichung der KI-Korrektur (oben) sowie der menschlichen Korrekturen oRPS (unten) vom gewichteten menschlichen Notendurchschnitt (jeweils dunkelrot) im Vergleich.

rektur von DeepWrite. Diese Annahme bedarf jedoch vertiefter Untersuchungen.

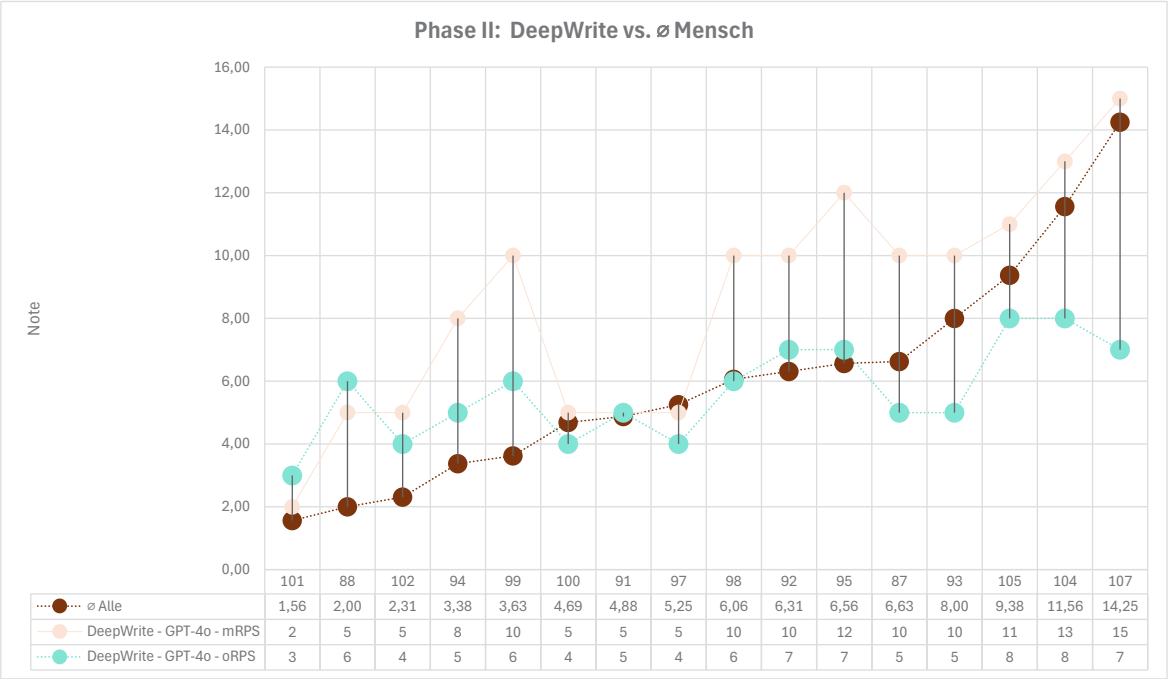
Die insgesamt geringste Abweichung zum gewichteten menschlichen Notendurchschnitt weist jedenfalls bei DeepWrite das Model GPT-4o oRPS auf (Abb. 12) und bei KlausurenKIste das Model Gemini 2.5 pro mRPS (Abb. 13).

Betrachtet man die Punktedifferenzen zwischen DeepWrite GPT-4o oRPS mit dem gewichteten menschlichen Notendurchschnitt genauer (Abb. 12), fällt auf, dass DeepWrite im unteren Notenspektrum (für alle fünf Klausuren, die im gewichteten menschlichen Notendurchschnitt unter 4 Punkten liegen) durchweg bessere Noten vergibt und im oberen Notenspektrum (für alle fünf Klausuren, die im gewichteten menschlichen Notendurchschnitt über 6,5 Punkten liegen) durchweg schlechtere Noten vergibt. Es zeigt sich auch hier eine starke Konvergenz zur Mitte hin.

Das von KlausurenKIste mit Rohpunkteschema zum Einsatz gebrachte Gemini 2.5 pro bewertet bei allen Klausuren unter 7 Punkten ebenfalls durchgehend besser als der gewichtete menschliche Durchschnitt (Abb. 13). Auch dieses System könnte Schwierigkeiten damit haben, im oberen Notenspektrum sehr gute Klausuren als solche zu erkennen und dort in der Tendenz schlechter als der gewichtete menschliche Notendurchschnitt benoten.

Es gibt jedoch eine andere Erklärungsmöglichkeit für diese Ergebnisse als die Konvergenz der KI-Systeme zur Mitte. Angesichts der hohen Durchfallquoten bei Klausuren[56] könnte man vermuten, dass diese bisher zu streng bewertet werden. Dann wäre eine KI-Korrektur sogar besser, zumindest aus der Sicht der Prüflinge. Auf der anderen Seite beobachten die Projektbeteiligten die (menschliche) Neigung, eine über dem Durchschnitt liegende Klausur besser zu bewerten, als es proportional angemessen wäre, um die Notenskala wenigstens einigermaßen auszuschöpfen. Dass ein KI-System, welches sich an die Vorgaben der Lösungsskizze hält, hier zu einer kritischeren Bewertung kommt, ist dann nicht verwunderlich. Aber auch diesbezüglich besteht noch Forschungsbedarf.

Zusammenfassend lässt sich festhalten, dass die KI-Korrektur bereits jetzt objektiver sein kann als eine einzelne (durchschnittliche) menschliche Korrekturkraft oRPS.



	101	88	102	94	99	100	91	97	98	92	95	87	93	105	104	107
⌀ Alle	1,56	2,00	2,31	3,38	3,63	4,69	4,88	5,25	6,06	6,31	6,56	6,63	8,00	9,38	11,56	14,25
DeepWrite - GPT-4o - mRPS	2	5	5	8	10	5	5	5	10	10	12	10	10	11	13	15
DeepWrite - GPT-4o - oRPS	3	6	4	5	6	4	5	4	6	7	7	5	5	8	8	7

Abb. 12: Vergleich der Korrekturergebnisse des Models GPT-4o mit Rohpunkteschema bei DeepWrite (mRPS; hellrot) und ohne Rohpunkteschema (oRPS, dunkelrot) mit der gewichteten menschlichen Durchschnittsnote (Ø Alle, dunkelgrün).

56 BfJ (Fn. 3), S. 3 f.: Die Durchfallquote im Ersten Juristischen Staatsexamen liegt im Bundesdurchschnitt bei 27,5%; in Brandenburg gar bei 44%.

3. Notenunterschiede nach Endgerät und Länge der Klausur

Von den sechzehn für dieses Forschungsprojekt ausgewählten Klausuren wurden sechs am eigenen Laptop verfasst (BYOD-Klausur) und zehn im PC-Raum der LMU München. Werden Klausuren am PC geschrieben, fallen die Verzerrungen hinsichtlich der Anzahl der abgegebenen Seiten (z. B. durch ein großes Schriftbild) im Vergleich zu handschriftlichen Klausuren weg. Allerdings ließ sich feststellen, dass der Mittelwert der abgegebenen Wortanzahl bei BYOD-Klausuren (2219) um ca. 21,5 % über dem Mittelwert der im PC-Pool geschriebenen Klausuren (1826) lag. Dies deckt sich mit dem zahlreich geäußerten Feedback der Teilnehmenden, dass Klausuren lieber an der eigenen, gewohnten Tastatur verfasst werden. Es konnte jedoch keine Korrelation zwischen dem Umfang der Klausuren in Wörtern und ihrer Benotung festgestellt werden (Abb. 14).

4. Zusammenfassung

Im Folgenden werden die Erkenntnisse und Diskussionen innerhalb des DigitalProjekts zusammengefasst sowie Thesen zu den Vor- und Nachteilen der Korrektur durch Menschen und KI-Systeme aufgestellt. Dabei wird zunächst von einer durchschnittlichen[57], rein menschlichen Korrektur oder einer solchen allein durch ein KI-System ausgegangen. Dass und warum die Beteiligten des DigitalProjekts eine Synergie für optimal halten, wird im Ausblick[58] dargestellt.

Sowohl bei der durchschnittlichen menschlichen Korrektur als auch bei den KI-Systemen sind Stärken

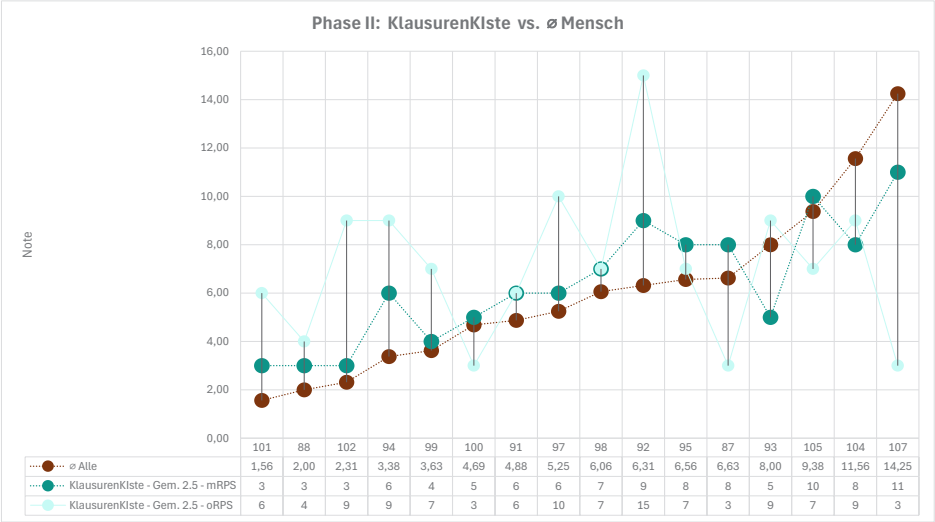


Abb. 13: Vergleich der Korrekturergebnisse des Models Gemini 2.5 pro mit Rohpunkteschema bei KlausurenKIste (mRPS; dunkeltürkis) und ohne Rohpunkteschema (oRPS, helltürkis) mit der gewichteten menschlichen Durchschnittsnote (Ø Alle, dunkelgrün).

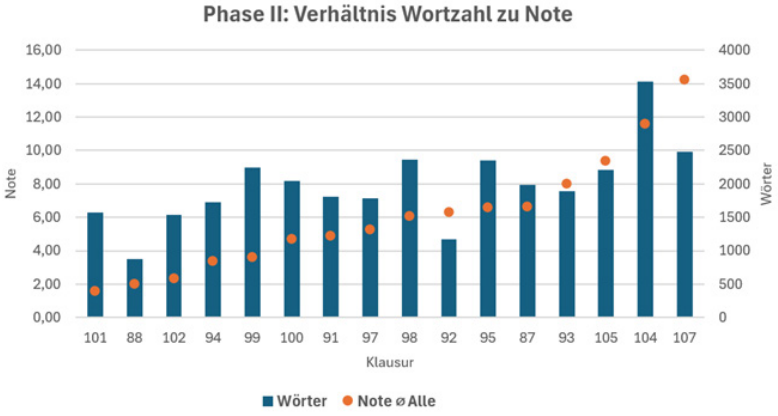


Abb. 14: Durchschnittsnote der jeweiligen Klausur (orangene Punkte; Skala links) im Verhältnis zur Länge der Klausur in Wörtern (dunkelblaue Balken; Skala rechts).

57 Es gibt in der Praxis unstreitig sehr gute und weniger gründliche KorrektorInnen.

58 S. dazu unten unter E.

und Schwächen erkennbar. Bei beiden besteht noch Optimierungspotential.

Die Partner des DigitalProjekts kommen zur folgenden Bewertung ihres Vergleichs (Abb. 15):

KI-Systeme können vor allem bei der Kosten- und Zeitersparnis punkten.[59] Außerdem führen sie zu einer im Vergleich zur Streuung der einzelnen menschlichen Korrekturen homogeneren und damit - in formeller Hinsicht - gerechteren Benotung (summatives Feedback).

Das formative Feedback der KI-Systeme ist ausführlicher und damit die Bewertung besser nachvollziehbar als bei einer durchschnittlichen menschlichen Korrektur. Dies erhöht nicht nur die Akzeptanz der Noten, sondern ermöglicht es den Studierenden auch, aus ihren Fehlern zu lernen.

Um die inhaltliche Qualität der Korrekturen zu bewerten, bedarf es weiterer vertiefter und damit sehr aufwändiger Analyse, die im Rahmen des DigitalProjekts nur exemplarisch geleistet werden konnte.

Im Übrigen gleichen sich die Vor- und Nachteile in etwa aus. Dabei ist davon auszugehen, dass die Vor- und Nachteile der menschlichen Korrektur recht stabil sind, während bei der KI-Korrektur bereits während des Forschungszeitraums Verbesserungen erkennbar waren. Daher stellt diese Bewertung eine Momentaufnahme dar. Es wird erwartet, dass die Fähigkeiten der KI zur Berücksichtigung des Inhaltes die Nachvollziehbarkeit und die Gesamtqualität weiter steigern werden.

Angesichts der zu erwartenden weiteren Entwicklung von KI-Systemen sowie besserer Prompts sollte die flächendeckende KI-Korrektur ernsthaft in Betracht gezogen werden. Für Probeklausuren und Selbsttests der Studierenden könnte die Bewertung endgültig sein und darüber hinaus sind Synergien mit einer menschlichen (Nach-)Korrektur für echte Prüfungen denkbar. Auch eine Vorsortierung durch die KI-Systeme kommt in Betracht; dadurch könnten den einzelnen Korrekturkräften potenziell breiter gestreute Klausuren vorgelegt werden, was wiederum zu einer objektiveren Notenvergabe beitragen könnte.[60]

D. Verbesserungsmöglichkeiten

I. Forschung und Bewertung

KorrektorInnen haben mindestens das Erste Juristische Staatsexamen absolviert. Damit haben sie eine gewisse Basis an juristischen Kenntnissen, sind jedoch nicht unbedingt ExpertInnen für das Korrigieren im Allgemeinen[61] und das konkrete Thema der zu bewertenden Prüfungsleistung im Besonderen. Ob die extremen Abweichungen hinsichtlich der vergebenen Noten bei den menschlichen Korrekturen[62] u.a. auf juristische Fehler zurückzuführen sind, wurde – so weit ersichtlich –

	Menschliche Korrektur	KI-System
Kosten der Korrektur	★☆☆	★★★
Zeit bis zum Ergebnis	★☆☆	★★★
Berücksichtigung des Inhalts	★★☆	★★☆
Nachvollziehbarkeit	★☆☆	★★☆
Qualität der Korrektur	★★☆	★★☆

Abb. 15: Ergebnisbewertung der Korrekturen durch Menschen und KI-Systeme in Phase I und II. Die Bewertung der menschlichen Korrektur bezieht sich auf eine durchschnittliche Korrektur. Die Bewertung erfolgt mit Sternen von 1 (schwach/schlecht) bis 3 (stark/gut).

59 Dies allerdings nur, nachdem zuvor viel Geld und Zeit in die Entwicklung gesteckt wurde, siehe dazu oben C. I. 1.

60 Einen anderen Weg zur Erreichung dieses Ziels geht – allerdings mit einem Fokus auf textlich deutlich umfangreichere Hausarbeiten – das Bayreuther Projekt, welches anhand von textlichen Unterschieden mit Hilfe von Cluster-Algorithmen möglichst unterschiedliche Arbeiten identifiziert. Siehe dazu etwa den Bericht von Jana Borochowitsch, Mehr Gerechtigkeit bei der Notenvergabe: KI-Software an der Universität Bayreuth, 07.08.2025, https://jurios.de/2025/08/07/mehr-gerechtigkeit-bei-der-notenvergabe-ki-software-an-der-universitaet-bayreuth/. Die Projektbeteiligten danken Felix Kaiser für den spannenden Austausch und die Clusterbildung für die in Phase II untersuchten Klausuren. Vermutlich aufgrund der Kürze der Klausurtexte sowie aufgrund der lediglich geringen Anzahl an untersuchten Klausuren ließen sich jedoch keine Korrelationen zwischen den gefundenen Clustern und den vergebenen Noten finden. In zu kurzen juristischen Texten wie Klausuren scheint sich nicht von der lexikalischen auf die semantische Ähnlichkeit und somit eine vergleichbare Prüfungsleistung schließen zu lassen.

61 An der Universität Passau existiert daher über „Lehre+ Hochschuldidaktik“ ein Seminar von *Prof. Kramer* und *Prof. Kuhn* zum Thema „Korrektur juristischer Arbeiten“, siehe https://www.uni-passau.de/lehreplus-weiterbildungsangebot.

62 Dazu oben C.III.2.c) aa).

bisher noch nicht untersucht. Die anekdotische Evidenz zeigt aber, dass nicht alle Korrekturkräfte alles selbst wissen und dadurch z.B. auch Fehler in Prüfungsarbeiten übersehen. Lösungsskizzen werden zudem gelegentlich zu stark als Orientierung genutzt und alternative, vertretbare Ansätze[63] in Prüfungsarbeiten nicht nachvollzogen. Während – wie einleitend dargestellt – schon kaum einheitliche Maßstäbe für die Bewertung juristischer Klausuren existieren, gibt es erst recht keine Maßstäbe für die Bewertung juristischer Korrekturen. Quantitative Metriken werden sich nach Auffassung der AutorInnen kaum finden lassen – hier existiert jedoch erhebliches Forschungspotenzial. Die Bewertung und Verbesserung der Qualität von sowohl menschlichen als auch von KI-Korrekturen wird auf absehbare Zeit mühsahme qualitative Einzelarbeit bleiben, die zudem nur von Personen übernommen werden kann, welche die Rechtsmaterie noch tiefer beherrschen als die ohnehin schon hoch qualifizierten Korrekturkräfte. In einem weiteren Forschungsdesign könnten also die menschlichen sowie die KI-Korrekturen „korrigiert" und bewertet werden.

II. Möglichkeiten der technischen Optimierung der KI-Modelle

Bei der Verwendung von KI-Systemen hängt die juristische Präzision des „Wissens" vor allem vom eingesetzten Modell und dem erstellten Prompt ab. Je präziser das Feedback sein soll, desto mehr Kontextwissen ist erforderlich.

1. Datenqualität: Lösungsskizze und Trainingsdaten

Lösungsskizzen allein haben für das Erstellen sehr detaillierter Korrekturhinweise oft nicht die notwendige Detailtiefe – u. a. was den Umgang mit (noch) vertretbaren anderen Ansichten angeht –, sodass die KI auf generisches Wissen aus dem Gesamtumfang ihrer bisherigen Trainingsdaten zurückgreift. Diese Daten sind jedoch fast nie ausschließlich auf das deutsche Recht zugeschnitten, stammen auch aus öffentlichen Blogs, Foren etc. und sind mithin nicht immer zuverlässig.[64] Wie bei jeder KI gilt, dass mit der Optimierung der Trainingsdaten die Qualität der Ergebnisse steigt.

Eine besonders hervorzuhebende Problematik ist die Kontextsensitivität. In juristischen Gutachten ist es entscheidend, dass eine logisch strukturierte Prüfung erfolgt und nicht nur „Textbausteine" an irgendeiner Stelle verwendet werden. Die optimale menschliche Korrektur hat dies im Rahmen einer Gesamtbetrachtung im Blick. Typischerweise wird in der Lösungsskizze eine entsprechende Struktur vorgegeben. Allerdings gibt es – wie bei den inhaltlich vertretbaren anderen Ansichten – auch hier nicht immer nur einen Weg, also ein einziges richtiges Schema. Eine Möglichkeit bestünde etwa darin, die Lösungshinweise auch mit alternativ zulässigen Aufbauvarianten zu versehen.

Außerdem wäre es denkbar, die Modelle anhand kuratierter juristischer Daten (etwa Lehrbücher und Aufsätze zu Methodenlehre, korrekten Normzitaten, juristischer Gliederung etc.) nachzutrainieren (fine tuning). Hier stellen sich jedoch urheberrechtliche Fragen.

Zudem könnten spezifische Jura-Korrektur-KI-Systeme mit dem vorhandenen Datenschatz der öffentlichen Hand gefüttert werden. Beispielsweise werden allein im Probeexamen an der Universität Passau pro Jahr ca. 2.000 Klausuren korrigiert, im Münchner Examenstraining ca. 10.000 Klausuren. Würde man alle Klausuren deutschlandweit zusammenfassen, ließe sich auf diese Weise binnen kurzer Zeit ein entsprechend großer Trainingsdatensatz generieren. Dabei stellt sich die Frage, wie die Klausuren und Korrekturen strukturiert sein müssten, wie viele Daten man benötigt, um zu sinnvollen Ergebnissen zu kommen und ein Under- oder Overfit-

63 Ein alternativer, vertretbarer Ansatz (sog. Mindermeinung) ist eine Position, die zwar von der allgemeinen Ansicht oder der überwiegenden, sog. „herrschenden Meinung" (h.M.) abweicht, jedoch mit den anerkannten Methoden begründet wird und daher vertretbar, d.h. nicht „falsch" ist. Dass eine solche Meinung in einer Musterlösung nicht vorkommt, kann verschiedene Ursachen haben. Z.B. kann der/die VerfasserIn der Lösung diese Position für irrelevant gehalten haben, da sie kaum/nicht mehr in Literatur und/oder Rechtsprechung vertreten wird und es für unwahrscheinlich gehalten wurde, dass Studierende diese kennen oder argumentativ herleiten. Denkbar ist auch, dass BearbeiterInnen in der Fallbearbeitung eine solche Ansicht selbst (neu) entwickeln.

64 Dies gilt bereits für menschlich geschriebene Blogs. In neuerer Zeit werden zunehmend KI generierte – teils falsche – Inhalte wieder als Quelle herangezogen, etwa wenn eine Onlinequelle (z.B. ein Kanzleiblog) ein von ChatGPT halluziniertes Urteil übernimmt und diesem damit den Schein von Legitimität verschafft. Diese Online-Quelle wird dann wiederum von ChatGPT für eine Aussage als Quelle für das Urteil herangezogen, ohne dass das Urteil existiert. Siehe zum Problem selbstreferenzieller KI-Modelle unter dem Begriff des „model collapse" etwa *Shumailov/Shumaylov/Zhao/Papernot/Anderson/Gal*, AI models collapse when trained on recursively generated data, 24.07.2024, 631 Nature 2024, S. 755 ff., https://doi.org/10.1038/s41586-024-07566-y.

ting[65] zu verhindern. Zudem dürfte sich die datenschutz- und urheberrechtliche Ausgestaltung als komplex erweisen.

2. Prompt-Engineering

Eine wichtige erste Weichenstellung für die Qualität stellt die Formulierung des Prompts für die Korrektur dar. Dieser wurde im DigitalProjekt durch beide Teams in mehreren Iterationsschritten angepasst und nach subjektiven Einzelbewertungen verbessert. Ein so optimierter Prompt kann erheblich zu besseren Ergebnissen beitragen. Verschiedene Methoden, wie z. B. die Vorgabe eines schrittweisen Vorgehens (chain of thought-prompting) oder das Einbauen von Beispielen (few shot-prompting) können dabei helfen, die „Gedanken" des Modells in geordnete Bahnen zu lenken, um mehr Einfluss auf das Ergebnis zu haben. So kann es auch hilfreich sein, Kontextwissen z. B. zum Gutachtenstil im Prompt zu erklären, sodass die Schritte bei der Bewertung durch die KI erkannt werden und Berücksichtigung finden. Ob und wie sich Veränderungen im Prompt auf die Qualität der erzeugten Lösung auswirken, kann jedoch derzeit erneut lediglich subjektiv bewertet werden. Hier besteht erheblicher weiterer Forschungsbedarf.

3. Vergleich unterschiedlicher KI-Modelle

Beide Ansätze nutzten in Phase I Open AIs GPT-4o sowie in Phase II sowohl GPT-4o als auch Googles Gemini 2.5 Pro. Bei KlausurenKIste generierte Gemini und bei DeepWrite GPT-4o die näher am menschlichen Durchschnitt liegenden Noten.[66] Hier zeigte sich das komplexe Zusammenspiel zwischen Modell und Prompt.

In späteren Projektphasen bestünde die Möglichkeit, die verwendeten Modelle tiefergehend zu vergleichen, weitere Modelle zu erproben und diese Ergebnisse bei gleichen Eingaben zu analysieren und miteinander zu vergleichen.

Zudem könnte man Vergleiche mit einer KI-Korrektur ohne konkrete juristische Vorgaben im Prompt anstellen.[67] Allgemeine generative KI-Modelle (wie z. B. GPT, Gemini o.a.) könnten spezifischer „Jura-KI" (etwa Harvey oder Noxtua) gegenübergestellt werden.

4. Architekturverbesserungen

Die technische Architektur der Gesamtkorrektur (DeepWrite) und der Absatzkorrektur (KlausurenKIste) unterscheidet sich grundlegend. Dies hat etwa Auswirkungen auf das Kontextwissen der jeweiligen KI-Abfrage. Auch der Detailgrad der Antworten unterscheidet sich dadurch. Was aus Qualitätsgesichtspunkten und was aus Perspektive der Nutzenden vorzugswürdig ist, erfordert weitere Untersuchungen.

Es ließen sich zudem gänzlich andere Architekturansätze untersuchen. Denkbar wäre etwa ein Modell, das mit der Normalverteilung der juristischen Noten aus der amtlichen Ausbildungsstatistik trainiert wird und alle Klausuren eines Durchgangs gleichzeitig korrigiert und entsprechend dieser Normalverteilung einordnet. Das Kontextwissen würde somit über die einzelne Klausur hinaus auf einen gesamten Durchgang erweitert.

5. Umgang mit nicht von der Lösungsskizze umfassten Klausurelementen

In der Forschung gilt es näher zu eruieren, wie mit Elementen einer Prüfungsleistung umgegangen werden soll, die keinem Teil einer Lösungsskizze zugeordnet werden können. Erforderlich ist dafür u. a. die – auch für Menschen schwierige – Entscheidung, ob es sich um eine vertretbare andere Ansicht oder sogar um einen in der Lösungsskizze übersehenen Aspekt handelt. Für die Bewertung von Elementen, die an der logisch falschen Stelle oder jedenfalls nicht entsprechend der Lösungsskizze platziert wurden, besteht ebenfalls weiterer Forschungsbedarf.

Ein erster Lösungsansatz zur Erreichung höherer Zuordnungsraten läge in der verstärkten Aufnahme alternativer Lösungswege in die vom Aufgabenstellenden vorgegebenen Lösungsskizzen – etwa auch zu Mindermeinungen etc. – was jedoch wahrscheinlich allein auf Grund der Anzahl der vertretbaren Lösungen nur mit

65 Overfitting (Überanpassung) liegt vor, wenn ein Modell die Trainingsdaten (inklusive Zufallsrauschen) zu genau „auswendig lernt" und auf neue Daten schlecht generalisiert, während Underfitting (Unteranpassung) bedeutet, dass das Modell zu simpel ist und schon die grundlegenden Muster der Trainingsdaten nicht richtig erfasst. Instruktiv dazu etwa https://developers.google.com/machine-learning/crash-course/overfitting/overfitting?hl=de.

66 S.o. unter C. III. 2. c) bb).

67 *Sebastian Dötter*l (Fn. 4), S. 165 zieht aus seinen Experimenten mit Standard-KI-Systemen den Schluss, dass sie für die Korrektur nicht zu gebrauchen seien. Er vermutet bessere Ergebnisse, wenn eine Musterlösung hinterlegt werde. Die im DigitalProjekt gemachten Erfahrungen sprechen dafür, aber ein konkreter Vergleich wurde bisher nicht angestellt.

exorbitant hohem Aufwand oder sogar nur theoretisch betrieben werden kann. Doch selbst dann werden einzelne (vertretbare oder nicht vertretbare) Elemente nicht oder nur selten von der Lösungsskizze umfasst sein, nämlich wenn sie von den Prüflingen neu kreiert wurden.

Es kommen folgende Optionen in Betracht: Man könnte diese Elemente ignorieren, wie es auch eine unsichere menschliche Korrekturkraft vermutlich macht. Man könnte eine menschliche (Nach-)Korrektur durchführen, im Idealfall durch eine besonders qualifizierte Person – was auch aktuell wünschenswert wäre. Aus Kapazitätsgründen ist dies jedoch nur begrenzt möglich und wird erleichtert, wenn das KI-System auf das Problem hinweist. Würde eine menschliche Korrekturkraft hingegen sämtliche Zweifelsfälle vorlegen, würde sie unter Umständen nicht wieder eingesetzt werden.

Aus den Erfahrungen der Projektbeteiligten sind Totalabweichungen vom Lösungsweg eher selten. Deshalb wäre es eine vernünftige Lösung, derartige Klausuren durch das KI-System herauszufiltern und menschlich zu korrigieren – die ersparten Kosten wären immer noch hoch.

E. Zusammenfassung und Ausblick

Nicht nur die Arbeitswelt, sondern auch die Ausbildung verändert sich durch die Existenz von KI.[68] Als erster Schritt vor dem Einsatz von KI-Systemen wäre es dringend notwendig, das juristische Prüfungswesen in seinen Abläufen beim Verfassen der Arbeiten und in der Verwaltung der Korrekturen vollständig zu digitalisieren.[69] Mit EDUTIEK und LexMea stehen den Universitäten jedoch hierfür bereits kostenfreie Lösungen zur Verfügung, die nur genutzt werden müssten.

I. Korrekturen mit Rohpunkteschemata sind objektiver

Viele Probleme der Korrektur und Bewertung im juristischen Kontext wurden bisher noch nicht einmal bei menschlichen Korrekturen ausreichend erörtert. Die Digitalisierungswelle aufgrund von KI sollte daher zum Anlass genommen werden, generelle Maßstäbe für jede Form der Korrektur zu diskutieren und zu erarbeiten.
Der vorliegenden Forschung liegt die Annahme zugrunde, dass ein Orientierungsmaßstab für Mensch und KI existieren könnte. Die empirischen Ergebnisse haben gezeigt, dass Rohpunkteschemata zumindest für mehr Objektivität sorgen.

Im direkten Vergleich der menschlichen Korrektur mit und ohne Rohpunkteschema zeigten die Korrekturen mit Rohpunkteschema (mRPS) erheblich weniger Abweichungen sowohl voneinander als auch zur Korrektur des Klausurerstellers.

Der Mittelwert der Range aller Klausuren liegt für die Gruppe ohne ein solches Schema (oRPS) bei 5,25 Punkten. Dies deckt sich in etwa mit den Ergebnissen der Studie von Clemens Hufeld, in der dieser Wert bei 6,47 Punkten lag. Bei der Gruppe mRPS liegt dieser Wert lediglich bei 2,94 Punkten. Im Schnitt hat man bei KorrektorInnen, die mit Rohpunkteschema korrigieren, also mit 2,31 Punkten weniger Notendifferenz zu rechnen. Die KorrektorInnen mRPS haben demnach etwa 44 % weniger Range als die KorrektorInnen oRPS.

Damit ist die Korrektur mRPS vorzugswürdig, sofern man Wert auf eine verbesserte Objektivität legt. Zur genauen Ausgestaltung der Schemata besteht weiterer Forschungsbedarf.

II. KI-Korrekturen bieten zahlreiche Vorteile

KI-Korrekturen bieten erhebliche Vorteile für die Studierenden. Besonders deutlich wird das im Hinblick auf die zeitliche Komponente und die Möglichkeit der „Massenkorrektur".[70] Individuelle Korrekturen können so binnen weniger Minuten generiert werden. Hinzu kommt, dass die Studierenden bei der KI-Korrektur ein ausführlicheres formatives Feedback erhalten, was bei menschlichen KorrektorInnen eher selten der Fall ist. Dies ermöglicht eine Beschäftigung mit den eigenen Fehlern daher zu einem Zeitpunkt, an dem das Wissen um die relevanten Klausurprobleme noch frisch ist.

68 Siehe dazu nur *Fries/Gössl/Hähnchen/Heidebach/Krönke/Strecker/Wischmeyer/Zwickel*, Juristisches Prüfen 2030, Verfassungsblog, 23.09.2025, https://verfassungsblog.de/juristisches-prufen-2030.

69 So schon *Sebastian Dötterl* (Fn. 4), S. 163 f.

70 So auch die Ergebnisse in *Hackl/Braun/Großkopf/Nonn/Müller/Zwickel,* KI-Feedback in der Rechtslehre: Eine explorative Studie zur Wahrnehmung und Bewertung durch Studierende, ZDRW 2024, S. 320 ff., https://doi.org/10.5771/2196-7261-2024-4-320. In Bezug auf Faktoren wie Verständlichkeit, Eindeutigkeit, Formulierung, Hilfestellung und Motivationswirkung deuten die quantitativen Auswertungen hingegen auf eine „Präferenz für das Tutor-Feedback" hin. Im bedeutenden Unterschied zur vorliegenden Studie wurde dort jedoch durch die KI kein summatives Feedback gegeben und das formative KI-Feedback beschränkte sich auf die Frage, ob die vier Schritte des Gutachtenstils innerhalb der jeweiligen Gliederungsebenen eingehalten wurden. Die Ergebnisse lassen sich daher weder verallgemeinern noch auf die vorliegende Untersuchung übertragen.

Außerdem hat man im Studium öfter die Chance, Feedback zu erhalten und dieses auch im Selbststudium für die eigene Verbesserung zu nutzen.

Für die korrigierenden und aufgabenstellenden Personen ergeben sich erhebliche Vorteile. KI-Korrekturen könnten etwa dafür eingesetzt werden, jeder Korrekturkraft eine das Leistungsspektrum möglichst breit abdeckende Auswahl an Klausuren vorzulegen.[71] Zudem könnten sie unterstützend eingesetzt werden, um bei Klausuren, die nur von einer Person korrigiert werden (wie die Mehrzahl der universitären Klausuren) verdeckte, nur für diesen Menschen einsehbare Benchmarks auszugeben. Dies ermöglicht eine Einschätzung der „Strenge" und Homogenität der eigenen Bewertungen. Ferner kann im Falle signifikant abweichender Ergebnisse einer menschlichen Korrekturkraft von den anderen eine verdeckte Zweitkorrektur durchgeführt werden – mit oder ohne Benachrichtigung des Erstkorrektors. Auch im Rahmen der AbsolventInnenbefragung 2025 des BRF waren „einheitliche Bewertungsmaßstäbe und Erwartungshorizonte […] sowie unabhängige Drittkorrekturen bei größeren Abweichungen" häufig genannte Lösungsvorschläge für das empfundene Problem der mangelnden Objektivität juristischer Bewertungen.[72]

Die Streuung der Noten beim summativen Feedback der derzeit besten KI-Systeme liegt deutlich unter jener der menschlichen Korrektoren oRPS. Im Regelfall werden Übungsfälle und Sachverhalte im Studium – wohl vor allem aufgrund des Arbeitsaufwands – ohne ein Rohpunkteschema gestellt, was gerade für einen Einsatz von KI-Systemen zum Selbststudium oder im Rahmen von Probeklausuren spricht. Die KI-Systeme zeigen jedoch je nach verwendeter System-Architektur (Gesamtkorrektur oder Absatzkorrektur), nach verwendetem Modell (GPT 4.0 oder Gemini 2.5 Pro o.a.) und nach Aufbereitung der Lösungsskizze (mit oder ohne Rohpunkteschema) eigenes Verhalten. Hier besteht weiterer Forschungs- und Verbesserungsbedarf – insbesondere zur Übertragbarkeit der vorliegenden Ergebnisse auf andere Klausuren sowie nach der Veröffentlichung neuer Modelle.

KI-Korrekturen sind bei weitem noch nicht perfekt, aber sie bieten bereits jetzt erhebliche Vorteile beim formativen Feedback. Viele Klausuren an der Universität dienen der Übung, sind also Probeklausuren, keine Prüfungen. Es geht zumindest bei diesen also nicht so sehr um eine exakt „richtige" Note (summatives Feedback), sondern vor allem um ein hilfreiches inhaltliches Feedback.[73] Um die eingangs gestellte Frage zu beantworten: Ja, das können KI-Systeme schon jetzt besser – vor allem als unter finanziellem und zeitlichem Druck arbeitende Menschen ohne ausführliche, zentrale Bewertungsvorgaben. Dies sollte zu einer hohen Akzeptanz der KI-Systeme führen.

Letztlich geht es auch nicht darum, dass KI-Korrekturen perfekt sind, sondern insbesondere darum, dass sie objektiver[74] sind als menschliche: Machine Bias vs. Human Bias.

III. Grenzen der KI-Korrektur

Wie in allen Bereichen muss jedoch – gerade bei so lebensweisenden Bewertungen – eine menschliche Letztentscheidung[75] für echte Prüfungsleistungen („scharfe" Klausuren) gesichert sein. Aufgrund des Automation Bias – also dem psychologischen Effekt, dass Menschen sich zu stark auf maschinelle Entscheidungen verlassen und diese unreflektiert übernehmen[76] – empfiehlt es sich zudem jedenfalls derzeit, Menschen zuerst unvoreingenommen bewerten zu lassen und ihre Ergebnisse in einem zweiten Schritt mit den maschinellen abzugleichen.

In rechtlicher Hinsicht stellen sich, gerade bei der Vollautomatisierung, verfassungsrechtliche Fragen, auf die an dieser Stelle nicht vertieft eingegangen werden kann.[77]

Neben die Ergebnisgerechtigkeit tritt stets die Komponente der Verfahrensgerechtigkeit. KI-Systeme gelten verbreitet als „Black Box". Sie basieren nicht auf nach-

71 Beachte zu diesem Ansatz auch das Projekt von Felix Kaiser an der Universität Bayreuth, s.o. Fn. 60.

72 BRF 2025 (Fn. 2), S. 101.

73 Tatsächlich im Einsatz ist KI in diesem Zusammenhang bereits durch *Thomas Hoeren* „ChatGPT – RechtsMentor ITM GPT", https://chatgpt.com/g/g-67c0248dd54c819191618b99577cc7c7-rechtsmentor-itm-gpt.

74 Dafür, dass dem so ist, auch *Sebastian Dötterl* (Fn. 4), S. 165.

75 Zu den rechtlichen Grundlagen, insb. Art. 22 DSGVO, vgl. *Sebastian Dötterl* (Fn. 4), S. 165 f., wobei dessen Einordnung als Hochrisiko-KI i.S.d. Art. 6 II KI-VO iVm Anhang III Nr. 3b diskussionswürdig ist, was hier jedoch nicht vertieft werden soll.

76 Umfassend hierzu etwa *Hannah Ruschemeier/Lukas J. Hondrich*, Automation bias in public administration – an interdisciplinary perspective from law and psychology, 41 Government Information Quarterly 3/2024, Article 101953, https://doi.org/10.1016/j.giq.2024.101953. Kritisch zur tatsächlichen Relevanz dieses Effektes hingegen etwa die Meta-Studie Saar *Alon-Barkat/Madalina Busuioc*, Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice, 33 J. Public Adm. Res. Theory, 1/2023, S. 153 ff., https://doi.org/10.1093/jopart/muac007.

77 Exemplarisch sei etwa verwiesen auf das Grundrecht auf informationelle Selbstbestimmung (Art. 2 I i.V.m. Art. 1 I GG), den allgemeinen Gleichheitssatz (Art. 3 I GG) sowie den Funktionsvorbehalt (Art. 33 IV GG). Eingehende Untersuchungen hierzu finden sich etwa bei *Robert Kreyßing*, Verwaltungsentscheidungen durch Künstliche Intelligenz, 2025, S. 149 ff., https://link.springer.com/book/10.1007/978-3-658-48413-2.

vollziehbaren menschlich gecodeten „wenn-dann"-Funktionen, sondern bestehen aus Milliarden von Parametern, die auf statistische Muster im Trainingsdatensatz reagieren. Ihre Entscheidung treffen sie auf der Grundlage hochdimensionaler Wahrscheinlichkeitsverteilungen. Auch Korrektur-Systeme nutzen (aktuell) kein explizites Regelwerk oder juristisches Wissenssystem. Es ist daher nicht immer nachvollziehbar, auf welche Rechtsnorm, Argumentation oder Fallstruktur sich ein Modell intern bei seiner Bewertung beruft.

Hinzu kommt, dass regelmäßig neue Modellversionen veröffentlicht werden, die auf unterschiedlichen Daten trainiert wurden und andere „Entscheidungsverfahren" einsetzen, die für den Anwender nicht einsehbar oder zumindest nicht nachvollziehbar sind.

Ähnliche Einwände lassen sich allerdings auch gegen eine menschliche Bewertung erheben. Auch jene unterliegt stetig neuen externen Einflüssen, die ein „Weiterlernen" bedingen (oder auch nicht) und Entscheidungen lassen sich nicht immer als kausal logische Gedankenketten nachzeichnen. Vergleichsmaßstab sollte die durchschnittliche menschliche Klausur sein. Wenn die KI-Korrektur jener gegenüber nachvollziehbarer ist, spricht dies für ihren Einsatz.[78]

Jedenfalls sollten schon heute die Vorteile menschlicher Korrektur und solcher durch KI-Systeme kombiniert werden, um daraus Synergien zu ziehen. Den Anfang dafür könnten Probeklausuren und Selbstlern-Klausuren bilden, da sie keine („echten") Prüfungen sind.

Annex: Vertiefte Erläuterungen zu den technischen Ansätzen

I. „Gesamtkorrektur" (DeepWrite)

1. Hinzufügen von Klarstellungen für die KI

Für die KI-Korrektur wurden einige Klarstellungen in die Lösungsskizzen eingearbeitet. Dies erschien primär in Passagen nötig, die besonders viel juristisches Kontext- und Vorwissen erforderten – etwa, wenn unter einer Überschrift lediglich der Korrekturhinweis „Keine Besonderheiten" stand. Dies geschah durch wissenschaftliche Mitarbeitende.

2. Aufteilung und Zuordnung von Lösungsskizze- und Klausurabschnitten

Als Front-End für die KI-Korrektur wurde das Audience-Response-Tool classEx[79] genutzt (siehe in Abb. 16 ein Beispiel-Screenshot).

Zur Verarbeitung in der Umgebung erfolgte eine Unterteilung in Sinnabschnitte (hier für Phase I: Zulässigkeit, Begründetheit Teil I und II; für Phase II waren es lediglich Zulässigkeit und Begründetheit). Dabei wurden die Abschnitte so gewählt, dass die Gesamtkohärenz des Falles nicht oder nur minimal beeinträchtigt wird. Jeder dieser Sinnabschnitte enthält in classEx zwei sog. Stages: die Writing-Stage zur Eingabe der studentischen Lösung und die Feedback-Stage zur Generierung des KI-Feedbacks. Hinzu kommt am Ende noch eine Endbewertung (EB-Stage), um die KI-Feedbacks der Sinnabschnitte miteinander zu vergleichen und den Gesamteindruck der Klausur abbilden zu können. Der Aufbau sieht sodann wie folgt aus:
Sinnabschnitt 1:

Die Feedback-Stage basiert jeweils auf demselben „Grund-Prompt", in den der jeweilige Teil der Musterlösung für den entsprechenden Sinnabschnitt eingefügt wurde. Die Klausuren wurden für das vorliegende Projekt händisch (Copy & Paste) in den jeweiligen Writing-Stages in classEx eingegeben. Im Echtzeitbetrieb können die Studierenden ihre Klausurlösung auch selbst unterteilen und in die jeweiligen Eingabefelder kopieren.

Am Ende erfolgt, wie bereits erwähnt, noch eine zusammenfassende Endbewertung (EB-Stage), um trotz der vorherigen Einteilung in Sinnabschnitte der Gesamtstruktur des Falles gerecht zu werden und „den Gesamteindruck einzufangen". Diese Vorgehensweise ist an die

Abb. 16: Screenshot von classEx. Stage 2 bis 8 enthalten die inhaltlichen Stages zum Schreiben, zum Feedback und zur Endbewertung

78 Detailliert zur „'Human box' und ‚black box' im Leistungsvergleich" *Jürgen Lorse*, Entscheidungsfindung durch künstliche Intelligenz - Zukunft der öffentlichen Verwaltung?, NVwZ 2021, 1657 (1661 f.).

79 Bei classEx handelt es sich um ein etabliertes Hörsaaltool der Universität Passau, das weltweit erfolgreich in der Lehre eingesetzt wird. Es ermöglicht eine direkte Interaktion im Hörsaal sowie in der synchronen oder asynchronen (Online)-Lehre.

menschliche Korrektur angelehnt. Außerdem ermöglicht diese Stage den Studierenden, ihr gesamtes Gutachten sowie die Einzelfeedbacks zu den Sinnabschnitten und die Endbewertung mit Note auf einen Blick zu sehen. Eine weitere Eingabe ist dafür nicht erforderlich, die vorherigen Angaben und KI-Feedbacks werden über classEx im System gespeichert.

Als LLM wurde OpenAIs GPT-4o (in Phase II zusätzlich Gemini 2.5 Pro) verwendet, das über ein Application Programming Interface (API) an classEx angebunden ist.

3. KI-gestütztes Feedback

Im Laufe des seit 2021 bestehenden Forschungsprojektes DeepWrite wurde bereits ein „Grund-Prompt" entwickelt, der die nachfolgende Struktur enthielt und für die Feedback-Stages aller Sinnabschnitte verwendet wurde. Zunächst ist ein Kontextgeber enthalten. Dieser „Einleitungssatz" dient dazu, der KI eine klare Rolle und Perspektive vorzugeben. Dadurch sollen die Antworten kontextgerecht, fachlich präzise und auf die Zielgruppe abgestimmt werden. Sodann sind generelle didaktische Anweisungen für das Erstellen des Feedbacks beschrieben (z. B. sollte Feedback zwar ehrlich und motivierend sein, aber auch direkt und klare Anweisungen geben). Anschließend ist der entsprechende Teil der Musterlösung integriert.

Für das Feedback sind verschiedene (Gedanken-) Schritte definiert[80], die der KI als Leitfaden dienen sollen und sowohl formative als auch summative Aspekte des Feedbacks enthalten. Schritt 1 enthält die Anweisungen für die Vorgehensweise beim Erstellen eines inhaltlichen Feedbacks, während Schritt 2 solche zum stilistischen Feedback (inkl. Erklärungen zum Gutachtenstil) enthält. Schritt 3 gibt vor, welche Rohpunkte für welchen Gliederungspunkt der Lösung maximal zu vergeben sind und umfasst auch den Hinweis zu einer Begründung der jeweils vergebenen Rohpunkte. Abschließend sind noch allgemeine Hinweise zu „Fehlantworten", wie z. B. zufälligen Buchstaben (wie „wjbfkwjbfkwbfwe") oder zur ausschließlichen Wiedergabe der Frage, vorhanden.

Die Zuordnung von Rohpunkten zu den jeweiligen Gliederungsebenen erfolgte in Phase I autonom durch die an diesem Ansatz arbeitenden wissenschaftlichen Mitarbeitenden unter Anwendung ihres eigenen juristischen Wissens sowie ihrer bisherigen Korrekturerfahrung. Als Orientierungshilfe für die Vergabe der konkreten Rohpunkte für eine bestimmte Gliederungsebene in Phase I dienten lediglich einleitende Sätze zur erwarteten Schwerpunktsetzung zu Beginn der Lösungsskizze.

Der Prompt in der letzten Stage der Endbewertung (EB-Stage) enthielt die Anweisung für eine Zusammenfassung der positiven und negativen Aspekte sowie für eine Addition der Rohpunkte aus den vorherigen KI-Feedbacks. Er enthielt in Phase I ferner eine durch die wissenschaftlichen Mitarbeitenden von DeepWrite selbst nach eigener langjähriger Korrekturerfahrung erstellte Umrechnungstabelle (siehe Abb. 18) von Rohpunkten in juristische Punkte, wobei 37 % zum Bestehen (4 Punkte) ausreichten.

Die Tabelle für die Umrechnung wurde in Phase I grundsätzlich anhand des Schwierigkeitsgrades der Klausur ausgewählt und nach wiederholtem Test der KI-Korrektur im Nachhinein angepasst, um eine einer menschlichen Korrektur im Durchschnitt vergleichbare Notenverteilung zu erzielen.

In Phase II wurden Rohpunkte für die Teilabschnitte sowie eine Umrechnungstabelle durch den Klausurersteller zur Verfügung gestellt.

Sinnabschnitt 1: Zulässigkeit		**Sinnabschnitt 2: Begründetheit I**		**Sinnabschnitt 3: Begründetheit II**		**End-bewertung**
(Menschliche) Writing-Stage	(KI) Feedback-Stage: Prompt + Teil der Lösungs-skizze	(Menschliche) Writing-Stage	(KI) Feedback-Stage: Prompt + Teil der Lösungs-skizze	(Menschliche) Writing-Stage	(KI) Feedback-Stage: Prompt + Teil der Lösungs-skizze	Zusammenfassende (KI-)Feedback-Stage inkl. Note: Besonderer Prompt

Abb. 17: Aufbau der verschiedenen Stages exemplarisch für Phase I

80 Bekannt als chain of thought-prompting.

Anhand des beschriebenen Aufbaus wurde der Prompt durch Überprüfung und subjektive Bewertung des KI-Outputs durch die wissenschaftlichen Mitarbeitenden schleifenweise getestet und verbessert. Hier konnte bereits auf die Vorerfahrungen zurückgegriffen werden, die in der Vergangenheit zu einer stetigen Verbesserung des Grund-Prompts geführt hatten und durch verschiedene andere Projekte vorhanden waren. Der Fokus lag vorliegend daher insbesondere auf der Optimierung der Leistung des Tools, gemessen an der Fähigkeit, auf untypische Studierendenantworten und unterschiedliche Eingabeformulierungen (inkl. etwaiger Buchstabenwiederholungen und ähnlicher „Fehlantworten") mit höherer Wahrscheinlichkeit reagieren zu können. Dies erfolgte ebenfalls durch wissenschaftliche Mitarbeitende.

Aufgrund des beschriebenen und überarbeiteten Prompts erhalten die Studierenden KI-Feedback zu jeder Gliederungsebene, wobei – wie bereits im Grund-Prompt angelegt – formative und summative Elemente im Feedback enthalten sind.

a) Formatives Feedback

Der entsprechende Musterlösungsteil, der im Prompt enthalten ist, dient als Referenzrahmen für die Bewertung und beugt Halluzinationen vor. Die Studierenden erhalten dabei Feedback zu ihrem Gutachten im Sinne einer Einteilung in abgehandelte und nicht vorhandene Elemente der Lösung sowie eine Bewertung der Qualität des Inhalts. Zudem werden Verbesserungsvorschläge (teilweise mit Formulierungsbeispielen) ausgegeben. Das Feedback bezieht sich auch auf weitere Aspekte (Gutachtenstil, sprachliche Qualität etc.) der studentischen Lösung.

Im letzten Schritt finden die Studierenden eine Darstellung des Feedbacks anhand der in der Lösungsskizze vorgesehenen Gliederungsebenen mit einer kurzen Begründung.

b) Summatives Feedback

Mit dem formativen Feedback erhalten die Studierenden auch summatives Feedback, indem sie die erreichten Rohpunkte zum jeweiligen Prüfungspunkt für jeden Unterabschnitt angezeigt bekommen.

Das für die Unterabschnitte erstellte formative Feedback wird im Rahmen der Endbewertung genutzt, um – mit einem Endvotum vergleichbar – die positiven und negativen Aspekte des Gutachtens darzulegen und eine Gesamtnote anhand der errechneten Rohpunkte auszugeben. Somit wird die Gesamtleistung des Studierenden zusammengefasst und abschließend bewertet.

II. „Absatzkorrektur" (KlausurenKIste)

Im Ausgangspunkt dieses Ansatzes standen ähnliche Überlegungen wie bei der Gesamtkorrektur (DeepWrite). Jedoch wurde eine andere technische Herangehensweise genutzt.

1. Segmentierung der Texte und Umwandlung der Absätze in numerische Werte

Nachdem die Texte in die Software hochgeladen wurden, unterteilte die Software die Musterlösung und das studentische Gutachten nach ihren Gliederungsabsätzen. Die Gliederungsebenen/-absätze wurden vorher hän-

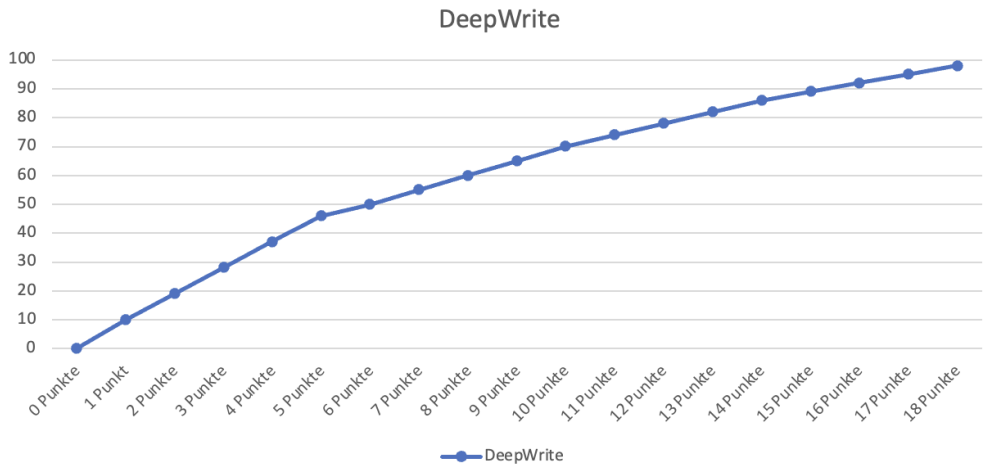


Abb. 18: Umrechnungstabelle in Phase I, Rohpunkte (max. 100) in juristische Noten (0-18 Punkte) für das summative Feedback

disch angepasst, um technisch verarbeitet werden zu können.

Jeder Gliederungsabsatz wird in eine mathematische Repräsentation umgewandelt, durch Speicherung als Vektorzahl. Dies geschieht mittels eines separaten text embedding models.

2. Zuordnung von Musterlösung- und Gutachtenabschnitten

Nach der Segmentierung und Vektorisierung der Musterlösung und des studentischen Gutachtens erfolgt die automatisierte Zuordnung der Gliederungsabsätze. Die Software analysiert dazu die numerischen Repräsentationen der Absätze und bestimmt für jeden Abschnitt der Musterlösung denjenigen im Gutachten, der ihm semantisch am nächsten steht. Dabei orientiert sich der Abgleich ausschließlich an der inhaltlichen Ähnlichkeit. Ziel ist es, die passenden Abschnitte aus der Musterlösung mit den entsprechenden der studentischen Arbeit zu verknüpfen, sodass sinnvolle Vergleichspaare entstehen.

a) Semantischer Abgleich und Paarbildung

Die Gliederungsabsätze aus der Musterlösung und jene Absätze im Gutachten mit der höchsten semantischen Übereinstimmung werden als zusammengehörige Paare identifiziert und gespeichert.

Um dieser Variabilität gerecht zu werden, erfolgt der Abgleich nicht starr von Absatz zu Absatz, sondern flexibel. Es können mehrere Teile eines Gutachtens mit einem oder mehreren der Musterlösung verknüpft werden und umgekehrt.

b) Grenzen der Zuordnung von Absätzen

Nicht immer lassen sich Vergleichspärchen bilden. In der Folge bleiben bestimmte Passagen des Gutachtens ohne Randbemerkungen. Regelmäßig ist das der Fall, wenn inhaltliche Differenzen bestehen. Dies könnte etwa geschehen, wenn Studierende in der Lösungsskizze nicht berücksichtigte „Mindermeinungen“[81] vertreten, an einer entscheidenden Stelle „falsch abbiegen“, nur entfernt relevante Aspekte prüfen etc.[82] In solchen Fällen steigt das Risiko, dass eine Bildung von Paaren nicht möglich ist. Allerdings lässt sich dies nicht pauschal sagen: Weichen Studierende zwar im Ergebnis ab, greifen aber zentrale Begriffe und den Sinngehalt des Themas auf, kann die Zuordnung durch das Modell dennoch funktionieren – nur mit erhöhter Distanz und damit größerer Unsicherheit. Sicher nicht zugeordnet werden hingegen Absätze, die das Thema vollständig verfehlen oder keinerlei einschlägige Schlüsselbegriffe enthalten.

3. KI-gestütztes Feedback

Zu jedem gefundenen Vergleichspärchen werden anschließend Randbemerkungen generiert.

a) Formatives Feedback

Dabei wird jedes gefundene Paar mit einem Prompt an ein Large Language Model (LLM) übergeben, welches die Inhalte analysiert und eine Randbemerkung erzeugt. Basierend auf diesem Vergleich überprüft das Modell, inwieweit die studentische Fallbearbeitung die inhaltlichen Anforderungen der Lösungsskizze erfüllt, und gibt eine detaillierte Rückmeldung zu den jeweiligen Stärken und Schwächen der Teilleistung.

aa) Analyse und Rückmeldung durch die KI

In jedem Vergleichspaar dient der entsprechende Musterlösungsteil als Referenzrahmen für die anschließende Bewertung. Die Software analysiert, welche Inhalte aus der Musterlösung in der studentischen Antwort vollständig oder teilweise enthalten sind und wo Übereinstimmungen bestehen. Gleichzeitig werden Aspekte identifiziert, die fehlen oder fehlerhaft dargestellt wurden.

Auf Basis dieser Analyse generiert das LLM gezielt Verbesserungsvorschläge. Diese zeigen auf, wie die studentische Antwort inhaltlich besser wäre. Dabei kann es sich um Ergänzungen, Umformulierungen oder eine klarere Herleitung der Argumentation handeln. Durch diese Rückmeldung erhalten Studierende nicht nur eine Einschätzung ihrer Leistung, sondern auch konkrete Hinweise zur inhaltlichen Optimierung ihrer Klausurbearbeitung.

bb) Umgang mit Grenzen der Absatzzuordnung

Falls für einen Abschnitt der Musterlösung kein entsprechender Absatz im studentischen Gutachten gefunden wird, bedeutet dies, dass dieser Aspekt der Klausur nicht

81 Vgl. Fn. 63.

82 Zu den allgemeinen Herausforderungen juristischen Denkens bei der Anwendung von KI und speziell bei Prüfungen vgl. die Schweizer Studie aus dem Mai 2025 von Yu Fan *et al.*, LEXam: Benchmarking Legal Reasoning on 340 Law Exams, pre print: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5265144. Das dort untersuchte System LEXAM enthielt allerdings andere Aufgabentypen (Fragen mit richtiger oder falscher Antwort, Multiple-Choice-Fragen und offene Fragen, die den Anforderungen in Deutschland am nächsten kommen).

bearbeitet wurde. In diesen Fällen gibt das System lediglich einen entsprechenden Hinweis. Eine weitergehende inhaltliche Analyse findet nicht statt, da kein Textabschnitt für den Abgleich vorliegt.

Schwieriger stellt sich der umgekehrte Fall dar, wenn für einen Abschnitt einer studentischen Lösung kein entsprechender Absatz der Musterlösung gefunden werden kann. Für diese Fälle ist weitere Forschung erforderlich.[83]

Zur Lösung des Problems, dass Studierende zwar richtige Aspekte ansprechen, diese jedoch im falschen Zusammenhang darstellen, werden zwei Arten von Randbemerkungen unterschieden: inhaltliche und methodische. Bei der inhaltlichen Prüfung geht es darum, ob bestimmte Inhalte überhaupt vorhanden sind und korrekt dargestellt wurden. Die methodischen Randbemerkungen folgen einem anderen System: Einzelne Absätze, die gemeinsam eine Teilprüfung ergeben (z. B. einen Anspruch abbilden), werden zu Einheiten zusammengeführt. So kann bewertet werden, ob die studentische Strukturierung methodisch nachvollziehbar ist – also ob das Gutachten juristisch sauber gegliedert ist. Dieses Verfahren zur Strukturprüfung befindet sich aktuell noch in der Entwicklung.

b) Summatives Feedback

Nachdem die Randbemerkungen für die einzelnen Vergleichspaare erstellt wurden, nutzt die KI diese, um ein umfassendes Endvotum zu generieren. Dieses dient als abschließende Bewertung der Klausur und fasst die Gesamtleistung des Studierenden zusammen. Das Endvotum besteht aus Erwartungshorizont, inhaltlicher Bewertung, formeller Bewertung, Note und Ausblick.

Die KI generiert alle diese Bestandteile eigenständig, indem sie die bisher erstellten Randbemerkungen und die Musterlösung einschließlich eines Prompts als Grundlage nutzt. Die Note wird von der KI entweder frei bestimmt (Phase I) oder – sofern Teilpunkte und eine Notentabelle hinterlegt sind (Phase II) – rechnerisch ermittelt. Eine verlässliche, homogene Notengebung für die Gesamtleistung ist ohne Rohpunkteschema kaum möglich, die Nutzung von Teilpunkten und Notentabellen sorgt auch hier für ein Mehr an Objektivität und Nachvollziehbarkeit.

Michael B. Strecker ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Öffentliches Recht, insbesondere Verwaltungsrecht (Prof. Dr. Thomas Wischmeyer) an der Humboldt-Universität zu Berlin sowie Gründer des Online-Gesetzbuchs LexMea.

Prof. Dr. Susanne Hähnchen ist Inhaberin des Lehrstuhls für Bürgerliches Recht und Rechtsgeschichte an der Universität Potsdam.

Dr. Martin Heidebach ist Akademischer Oberrat und Habilitand an der Juristischen Fakultät der LMU München.

Prof. Dr. Marie Herberger, LL.M. ist Inhaberin des Lehrstuhls für Bürgerliches Recht, Zivilverfahrensrecht, Methodenlehre, Recht der Digitalisierung und Legal Tech an der Universität Bielefeld.

Simon Alexander Nonn ist Wissenschaftlicher Mitarbeiter an der Lehrprofessur für Öffentliches Recht (Prof. Dr. Urs Kramer) im Forschungsprojekt DeepWrite und Doktorand am Lehrstuhl für Deutsches und Europäisches Privatrecht, Zivilverfahrensrecht und Rechtstheorie (Prof. Dr. Thomas Riem) an der Universität Passau.

Sarah Großkopf war Wissenschaftliche Mitarbeiterin.. an der Lehrprofessur für Öffentliches Recht (Prof. Dr. Urs Kramer) im Forschungsprojekt DeepWrite an der Universität Passau.

Gökhan Erol ist Gründer des LegalTech-Start-ups KlausurenKIste und Diplom-Jurist.

Menaf Erol ist Gründer des LegalTech-Start-ups KlausurenKIste und Absolvent der Rechtswissenschaften an der Universität zu Köln.

Ilja Garber ist Gründer des LegalTech-Start-ups KlausurenKIste und Absolvent der Rechtswissenschaften an der Universität Münster.

Constantin Höhmann ist Vorstandsvorsitzender der Munich Legal Tech Student Association e. V. und Student der Rechtswissenschaft an der Ludwig-Maximilians-Universität München.

Enci Huang ist Vorstandsvorsitzender der Munich Legal Tech Student Association e. V. und Student der Rechtswissenschaft an der Ludwig-Maximilians-Universität München.

Clemens Hufeld ist Vice President Legal Data der Noxtua AG und promoviert in München zur Digitalisierung des Jurastudiums.

Louisa Zachmann ist Wissenschaftliche Mitarbeiterin und Doktorandin am Lehrstuhl für Staats- und Verwaltungsrecht von Prof. Dr. Ann-Katrin Kaufhold an der Ludwig-Maximilians-Universität München.

83 Vgl. dazu oben unter D. II. 5.