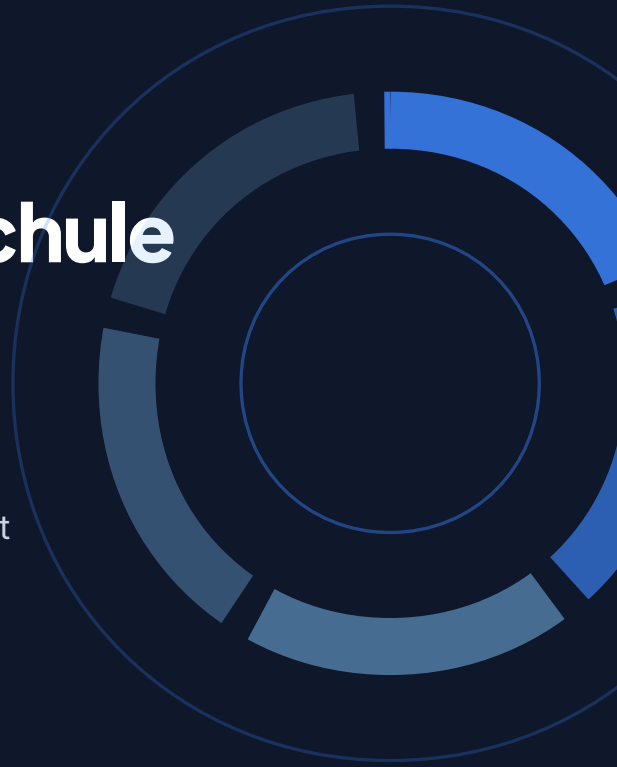

Lernen an der Hochschule im Wettbewerb mit generativer KI

Warum KI in der Lehre eine Gestaltungsaufgabe ist



ONETUTOR PERSPEKTIVE | AUGUST 2026

Verfasst von der OneTutor GmbH

Ein Whitepaper für Hochschulleitungen, Vizepräsidien für Lehre und Digitalisierung, Studiendekaninnen und Studiendekane sowie alle, die Lehre unter den Bedingungen generativer KI gestalten.

Worum es geht

ZIELGRUPPE

Dieses Papier richtet sich an Hochschulen (als Kurzform für Universitäten und Hochschulen der Angewandten Wissenschaften). Die zugrundeliegenden Prinzipien, an welcher Stelle digitale Werkzeuge unterstützen und wo sie sich bewusst zurückhalten sollen, sind aber nicht auf die Hochschule beschränkt. Sie gelten für jeden Bildungsanbieter und jede Organisation mit Interesse an Weiterbildung, die Vermittlung und Zertifizierung ernst nimmt.

Inhalt

Auf einen Blick	3
Zusammenfassung	4
1. Der Kontrollverlust der Hochschulen	5
2. Die doppelte Mission und ihre stille Entkopplung	6
2.1 Entkopplung als reale Verwundbarkeit	6
2.2 Zwei Arten des Prüfens	7
2.3 Rückschau: was die MOOC-Welle lehrt	8
3. Die Gestaltungsaufgabe	9
4. Zwei Grundsätze für die Gestaltung	11
4.1 Keine Abkürzungen	11
4.2 Der unterschätzte Wert der Kohorte	12
5. Ein Schichtenmodell	13
6. Anwendung: OneTutor	16
7. Warum Begleitforschung nicht optional ist	18
8. Sechs Fragen an ein Werkzeug	20
Quellen	21

Auf einen Blick

BEOBACHTUNG

- Hochschulen haben die Kontrolle über den Lernprozess schon vor KI verloren: Aufzeichnungen, Skripte, Erklärvideos. KI ist nicht der Anfang dieser Entwicklung, sondern ihre bislang letzte Stufe.
- Damit steht die Kopplung zur Debatte, die den Hochschulabschluss trägt. Wenn die Vermittlung von Bildungsinhalten überall gleich gut verfügbar ist, bleibt von der Hochschule strukturell die Zertifizierung übrig.
- Ausgerechnet dieses Fundament wackelt: Wo unbeaufsichtigt geprüft wird, ist Urheberschaft nicht mehr nachweisbar.

DIAGNOSE

- „Prüfen“ heißt zweierlei. KI erodiert das summative Prüfen und ermöglicht das formative erstmals in der Breite. Formatives Assessment war nie ein didaktisches, sondern vor allem ein Skalierungsproblem.
- Lernen kennt keine Abkürzungen. Der Engpass ist aber nicht die Lernzeit, sondern ihre Zusammensetzung: Erneutes Lesen ist der Normalfall, aktives Abrufen von Gelerntem die Ausnahme, wirksam ist Letzteres. Genau das ist gestaltbar.
- Drei Dinge hat nur die Hochschule: kuratiertes Material, die Nähe zur Prüfung und eine Kohorte, deren Fragen sich aggregieren und in bessere Lehre zurückführen lassen. Ein einzeln genutzter Chatbot kann das nicht.

KONSEQUENZ

- Studierende nutzen KI ohnehin. Die Wahl ist nicht „KI oder keine KI“, sondern gestaltetes oder ungestaltetes Werkzeug. Nichts anzubieten ist ebenfalls eine Entscheidung der Hochschule.
- Kein Werkzeug macht Studierende fleißiger. Es entscheidet nur, wo die ohnehin stattfindende Lernaktivität stattfindet und woraus sie besteht.
- Wir wissen heute zu wenig. Wer KI-Tutoren ohne Begleitforschung ausrollt, erfährt nie, ob sie helfen oder schaden.
- Was ein taugliches Werkzeug ausmacht, lässt sich in sechs Fragen fassen (Kapitel 8). OneTutor ist unser Versuch, sie alle mit Ja zu beantworten.

Zusammenfassung

Studierende nutzen KI-Werkzeuge längst, mit oder ohne ihre Hochschule. Das verändert nicht nur, welche Rolle die Hochschule bei der Vermittlung von Inhalten noch spielt, sondern rührt an einer für Hochschulen existentiellen Frage: ob Vermittlung und Zertifizierung von Ausbildung überhaupt noch zusammengehören müssen. Dieses Papier argumentiert, dass eine eigene, konkurrenzfähige Antwort auf KI wesentlicher Teil einer Antwort der Hochschulen auf diese Frage ist. Dieser Teil ist keine Option unter mehreren, sondern eine existentielle Gestaltungsaufgabe. Ein Modell aus vier Ebenen, nämlich Phasen, Motivation, Aufwandsarten und Querschnittsfaktoren, zeigt, wo diese Antwort ansetzen sollte.

In der KI-Debatte geht häufig verloren, dass „Prüfen“ zwei verschiedene Dinge bezeichnet. Die summative Funktion, also das Prüfungsergebnis mit Rechtsfolgen, ist das Fundament, das durch KI erodiert. Die formative Funktion, also Rückmeldungen, die niemandem etwas nützen außer der lernenden Person und den Dozierenden, ist umgekehrt die eine Assessment-Funktion, die durch KI nicht bedroht wird, sondern die KI erstmals in der Breite ermöglicht. Formatives Assessment war nie ein didaktisches, sondern immer ein Skalierungsproblem. Genau diese Beschränkung fällt.

Beim Lernen gibt es keine Abkürzungen. Der Engpass ist nicht die Zeit, die Studierende fürs Lernen aufwenden, sondern wie sie genutzt wird. Ein Werkzeug allein verschiebt daran nichts und kann dem Lernprozess sogar abträglich sein, wenn es Anstrengung ersetzt statt sie auszulösen. Nur Hochschulen haben den Hebel, KI so einzusetzen, dass alle Studierenden effektiver lernen: kuratiertes Material, die Nähe zur Prüfung und das Lernverhalten einer ganzen Kohorte. OneTutor ist unser Versuch, genau diesen Hebel nutzbar zu machen. Die Frage ist, ob die Wahl der Studierenden auf die eigene Hochschule fällt.

KAPITEL 1

Der Kontrollverlust der Hochschulen

Studierende lernen seit über zwanzig Jahren zunehmend mit Material externer Anbieter. KI-Chatbots sind nicht der Anfang dieser Entwicklung, sondern ihr derzeit letzter Schritt. Die Campusflucht ist die sichtbare Folge.

Seit rund zwanzig Jahren verlieren Hochschulen die Kontrolle darüber, wo und mit welchem Material Studierende tatsächlich lernen. Vorlesungsaufzeichnungen machen den Hörsaal scheinbar optional. Skripte und Prüfungsprotokolle kursieren zwischen Jahrgängen und sogar zwischen Hochschulen. YouTube-Erklärvideos, MOOCs und die Khan Academy erklären aus Studierendenperspektive dasselbe Thema oft klarer als die eigene Vorlesung. ChatGPT und NotebookLM erklären inzwischen personalisiert und jederzeit. 88 Prozent der jungen Menschen in Ausbildung haben generative KI bereits genutzt (Stürz et al. 2025).

Aus Sicht der Studierenden ist das keine Notlösung. Alternative Materialien sind zumindest in deren Wahrnehmung inhaltlich und didaktisch häufig mindestens ebenbürtig zum eigenen Kursmaterial, vor allem in den grundständigen Fächern. Digitale Technologien ermöglichen seit zwanzig Jahren die Entkopplung eines vormals wesentlichen Teils des Lernprozesses, nämlich der Vorlesung, vom physischen Ort der Hochschule. Generative KI erlaubt nun die interaktive Auseinandersetzung mit den Materialien der eigenen Hochschule und anderer Anbieter zeitlich und räumlich unabhängig von den tradierten Lehrformaten.

Eine sichtbare Ausprägung dieses Kontrollverlusts ist die zunehmende Campusflucht: sinkende Anwesenheit, sogar trotz manchmal formal bestehender Präsenzpflcht. Das mag die Lebensrealität einer neuen Studierendengeneration spiegeln. Es ist aber sicherlich auch eine rationale Reaktion darauf, dass Anwesenheit im Vergleich zu den Alternativen nicht mehr eindeutig als der beste Weg zum Verständnis erscheint. Lerninteraktionen finden zunehmend nicht mehr auf dem Campus statt, aber auch nicht mehr in den etablierten digitalen Foren des Kurses. Studierende nutzen vermehrt ihre eigenen Tools. Der inhaltliche Diskurs findet verstärkt nicht mehr gemeinsam, sondern isoliert statt.

Gleichzeitig weist eine Metaanalyse über 69 Studien mit rund 21.000 Studierenden Anwesenheit als den stärksten bekannten Einzelprädiktor für Studienleistung aus, stärker als standardisierte Aufnahmetests, Schulnoten oder Lernstrategien, und weitgehend unabhängig von Gewissenhaftigkeit und Motivation (Credé et al. 2010). In einer Synthese von über hundert Einflussgrößen auf Studienerfolg rangiert Anwesenheit weiter oben, aber nicht konkurrenzlos an der Spitze (Schneider & Preckel 2017). Sinkende Präsenz mag aus Studierendenperspektive einen rationalen Effizienzgewinn darstellen. Auch wenn der Zusammenhang korrelativ ist und keine Kausalität beweist, kann die Effektivität des Lernens bei Abwesenheit vom Campus leiden.

Eine plausible Teilerklärung für diesen Anwesenheitseffekt ist, dass Präsenz einen regelmäßigen, konsequenzfreien Rhythmus von Überprüfungsgelegenheiten erzeugt: die Zwischenfrage, das Handzeichen, der Blick auf das Blatt der Sitznachbarin, das Gespräch beim Mittagessen. Trifft das zu, dann ist formatives Assessment (Kapitel 2.2) vielleicht die vielversprechendste Strategie, diesen Rhythmus auch abseits des Campus teilweise zu reproduzieren.

Daraus folgt für uns ausdrücklich nicht, dass Vorlesungen schlicht das falsche Format sind, also zu lang, zu unidirektional, nicht personalisiert, nicht interaktiv, und dass Formate wie Projektarbeit, Seminare oder Flipped Classrooms sie einfach ersetzen könnten. Das hängt vom Fach ab, von der Persönlichkeit der Dozierenden sowie von den Motivationslagen, Fähigkeiten, Vorlieben und kulturellen Prägungen der Studierenden.

KAPITEL 2

Die doppelte Mission und ihre stille Entkopplung

Hochschulen koppeln Ausbildung und Prüfung in einer Institution. Im Bildungsbereich ist das die Ausnahme, nicht die Regel. KI greift diese Kopplung aus mehreren Richtungen an. Formatives Assessment, das erstmals skaliert, dreht die Bedrohung in eine Chance.

Hochschulen erfüllen neben der Forschung traditionell zwei Aufgaben gleichzeitig: Sie vermitteln Wissen und Kompetenz, und sie zertifizieren, dass diese Kompetenz erworben wurde. In der öffentlichen Wahrnehmung gehört beides untrennbar zusammen. Ein Hochschulabschluss steht für beides gleichzeitig.

2.1 Entkopplung als reale Verwundbarkeit

In anderen Bildungsbereichen ist diese Kopplung nicht vorhanden oder nicht so stark. IHK-Prüfungen zertifizieren unabhängig davon, wo und wie jemand sich das nötige Wissen angeeignet hat. Ein Scrum-Master-Zertifikat oder ein IT-Sicherheitszertifikat prüft ein Ergebnis, nicht den Weg dorthin; vorbereiten kann man sich mit einem Buch, einem Kurs, YouTube oder einem Chatbot.

Für Hochschulen ist eine Entkopplung kein hypothetisches Szenario, sondern eine potentiell existentielle Verwundbarkeit. In der Rechtswissenschaft stellen die von Externen kostenpflichtig angebotenen Repetitorien seit Langem eine solche deutlich sichtbare Entkopplung dar. Wenn nun potentiell in allen Fächern Vermittlung überall und oft ebenso gut verfügbar ist, bleibt von der Hochschule strukturell vor allem die Zertifizierung übrig, jedenfalls dort, wo sich der Mehrwert der Ausbildung vor Ort nicht überzeugend begründen lässt. Zertifizierungsinstanzen konkurrieren anders als vollständige Studiengänge: günstiger, schneller, spezialisierter. Die Infragestellung der Rolle der Hochschule in diesem Sinn ist die klassische Situation einer möglichen Disintermediation, wie sie in allen Digitalisierungsvorhaben beobachtbar ist.

Nicht nur wir sind der Ansicht, dass KI die Zertifizierungsfunktion ohnehin fundamental in Frage stellt. In einem verdeckten Feldtest an einer britischen Universität wurden in fünf regulären Psychologie-Modulen vollständig KI-geschriebene Abgaben in das laufende Prüfungssystem eingespeist. 94 Prozent blieben unentdeckt, und sie erhielten im Mittel eine halbe Notenstufe bessere Bewertungen als echte Studierende (Scarfe et al. 2024). Wer sich auf die Zertifizierung zurückzieht, zieht sich bei unbeaufsichtigten Prüfungsformen also auf ein bereits wackliges Fundament zurück. Die Frage nach KI in der Lehre ist damit keine Randfrage, sondern eine, die an beide Seiten der Mission von Hochschulen rührt.

Zur Erosion trägt ein zweiter, von KI unabhängiger Faktor bei: Es ist unklar, ob Zeugnisse abbilden, was sie abzubilden vorgeben. Ein Abschluss wirkt als Signal nach außen, weil unterstellt wird, dass er mit Fähigkeiten zusammenhängt. Wie tragfähig dieses Signal ist, ist in der Forschung umstritten; die Befundlage zum Zusammenhang zwischen Studienleistung und späterem Berufserfolg reicht von „nützlich, aber begrenzt“ bis „deutlich überschätzt“ und hat sich über die Jahrzehnte verändert (Roth et al. 1996; Van Iddekinge et al. 2024). Hinzu kommt, dass der Wissenschaftsrat kritisiert, dass Unterschiede zwischen Fächern und Standorten die Aussagekraft einzelner Noten schwächen, bei fortgesetzter Tendenz zu besseren Noten (Wissenschaftsrat 2012, auf Basis des Prüfungsjahres 2010). Ein Signal, das nicht unterscheidet, informiert nicht.

Aus einer umstrittenen Prognosekraft folgt nicht zwingend ein wertloser Abschluss. Die Frage aber, inwieweit ein Abschluss trägt, ist empirisch offen. Vor diesem Hintergrund ist der Befund

von Scarfe et al. keine neue Bedrohung, sondern die Zuspitzung einer alten Unklarheit. Das ist die eigentliche Verwundbarkeit: nicht dass das Zeugnis wertlos wäre, sondern dass seine Aussagekraft vor allem auch eine Annahme ist.

2.2 Zwei Arten des Prüfens: Formatives Assessment als Chance

Die Debatte über KI und Prüfungen leidet daran, dass „Prüfen“ zwei grundverschiedene Dinge bezeichnet, das summative und das formative Assessment. Das summative Assessment wirkt nach außen, das formative Assessment, also das ständige Abrufen von Wissen als Teil der Lernerfahrung, wirkt nach innen und dient einem effektiveren Lernprozess. Für diese beiden Funktionen wirkt generative KI in genau entgegengesetzte Richtungen: Das summative Assessment stellt sie in Frage, das formative Assessment ermöglicht sie. Der Grund ist naheliegend. Formatives Assessment war wohl nie eine Frage der Didaktik, sondern stets eine Frage der Skalierung. Ein Dozent mit dreihundert Hörern konnte häufige, zeitnahe Rückmeldung nie geben. Die Grenze war die Verfügbarkeit, und genau diese Grenze fällt.

ZWEI FUNKTIONEN DES PRÜFENS UND DIE WIRKUNG GENERATIVER KI

	Summatives Assessment	Formatives Assessment
Zweck	Feststellen und beurkunden	Feststellen, um zu steuern
Ergebnis	Note, Studienleistung, Zeugnis mit Rechtsfolgen	Rückmeldung an die lernende Person und an die Lehrenden
Folgen außerhalb des Lernprozesses	Erheblich und rechtlich bindend	Keine, und genau das ist die Bedingung seiner Wirksamkeit
Historische Grenze	Aufsicht und Nachweis der Urheberschaft	Arbeitszeit der Lehrenden, also reine Skalierung
Wirkung generativer KI	Erodiert. Wo unbeaufsichtigt geprüft wird, verliert die Beurkundung ihre Grundlage.	Wird erstmals möglich. Die Skalierungsgrenze fällt.

Dass häufige, zeitnahe Rückmeldung Lerngewinne erzeugt, gilt seit der Übersicht von Black & Wiliam (1998) als belegt. Wie groß diese Gewinne sind, ist allerdings umstritten. Spätere Arbeiten kommen zu deutlich kleineren Effekten und kritisieren die Qualität der zugrundeliegenden Studien (Kingston & Nash 2011; Bennett 2011; Xuan et al. 2022), während andere diese Neuberechnungen ihrerseits methodisch angreifen (Briggs et al. 2012). Es gibt also belastbare Indizien dafür, dass formatives Assessment wirkt, aber keine belastbare Zahl dafür, wie stark. Das ist kein Sonderfall dieses Themas, sondern der Normalzustand großer Teile der Bildungsforschung, was genau den Grund der Notwendigkeit eigener Begleitforschung in Kapitel 7 darstellt.

Ein Detail dieser Debatte ist für uns bemerkenswert: Unter den Umsetzungsformen, die Kingston & Nash unterscheiden, schnitten computerbasierte formative Systeme überdurchschnittlich ab. Diese Teilauswertung beruht allerdings auf sehr wenigen Studien, und andere Arbeiten ordnen computergestütztes Feedback umgekehrt am unteren Ende ein. Der Befund mag nicht als Beleg dienen, als Hinweis dürfen wir ihn ernst nehmen.

Die Frage ist damit nicht mehr nur, wie viel Prüfungshoheit eine Hochschule gegen KI verteidigen kann, sondern welche Prüfungsfunktion sie mit KI erstmals ausüben kann. Wer über KI in der Lehre nur defensiv nachdenkt, überspringt den Teil, an dem der Zugewinn vergleichsweise klar ist. Und es ist eine Stelle, an der die Hochschule strukturell im Vorteil ist. Formatives Feedback ist dann wertvoll, wenn es sich auf das bezieht, was tatsächlich geprüft und gelehrt wird, also auf

kuratiertes Material und die Lernziele der Veranstaltung. Ein allgemeiner Assistent kann dieses Material verarbeiten, sobald es jemand hochlädt, und Studierende tun das längst. Was er nicht hat, ist die Entscheidung darüber, was davon zählt, und die Kohorte, an der sichtbar wird, wo es klemmt. Der Vorteil der Hochschule liegt also nicht im Zugang zum Material, sondern in der Kuratierung und in der Aggregation.

Feedback ist dabei nicht automatisch wirksam. Shute (2008) arbeitet heraus, dass formatives Feedback nicht bewertend, unterstützend, zeitnah und spezifisch sein muss und dass seine Wirkung mit Merkmalen des Lernenden und der Aufgabe wechselwirkt. Falsch gebautes Feedback kann Lernen auch verschlechtern. Die Chance liegt also nicht in der Menge des Feedbacks, sondern in seiner Bauart und hängt an einer Bedingung, die leicht übersehen wird: dass sorgfältig überlegt sein will, ob das Feedback Folgen zeitigt, etwa in Form von Bonuspunkten.

2.3 Rückschau: was die MOOC-Welle lehrt

Um 2012 sollten MOOCs die Hochschulbildung neu ordnen. Tatsächlich konsolidierte sich das Feld um ein deutlich älteres Geschäftsmodell, das Outsourcing von Online-Masterprogrammen für Berufstätige. Die große Mehrheit der Lernenden kehrte nach dem ersten Jahr nicht zurück, das Wachstum konzentrierte sich fast vollständig auf die wohlhabendsten Länder, und die niedrigen Abschlussquoten verbesserten sich über sechs Jahre hinweg nicht (Reich & Ruipérez-Valiente 2019). Hinzu kommt in Deutschland ein erheblicher regulatorischer Schutz: Hochschulgrade sind rechtlich geschützt, Studiengänge akkreditiert. Eine private Zertifizierungsinstanz tritt nicht einfach an ihre Stelle.

Wer die Entkopplung heute prognostiziert, schuldet also eine Begründung, warum es diesmal anders sein sollte. MOOCs ersetzen die Vorlesung durch ein anderes festes, unidirektionales Artefakt, nämlich ein Video, das für alle gleich ist. Videos sind aber genau in der Dimension schwächer, auf die es ankommt: Rückfrage, Individualisierung, Reaktion auf den konkreten Denkfehler. Generative KI ersetzt nicht das Artefakt, sondern die Interaktion. Das macht die Entkopplung nicht unvermeidlich. Es macht sie aber wesentlich plausibler als 2012.

KAPITEL 3

Die Gestaltungsaufgabe

KI im Studium ist ein typisches Digitalisierungsvorhaben. Dass ein Werkzeug existiert, macht keinen Prozess effektiver oder effizienter. Entscheidend ist allein, wie es den Lernprozess verbessert.

Man kann wie im MOOC-Hype 2012 der Ansicht sein, die Institution der Hochschule sei einfach obsolet. Dabei lohnt der Blick auf das, was auf dem Spiel steht. Digitales, individualisiertes Lernen kann zu Vereinzelung führen. Wer allein mit einem Video oder einem Chatbot lernt, lernt nicht mit und von einer Peergroup. Er oder sie verpasst die Rückfrage der Sitznachbarin, die eigene Erklärung im Gespräch, den Blick auf einen fremden, falschen Lösungsweg, der oft mehr lehrt als die eigene richtige Antwort. Dass gemeinsames, aktives Arbeiten am Stoff messbar besser wirkt als passives Zuhören, ist für die MINT-Fächer über 225 Einzelstudien hinweg belegt (Freeman et al. 2014). Die meisten Akademiker werden angeben, von ihren Kommilitonen mindestens genau so viel gelernt zu haben wie von ihren Dozenten.

Und es geht nicht nur um Inhalte. Ein Teil dessen, was Studierende an einer Hochschule mitnehmen, ist Persönlichkeitsbildung im klassischen Sinne: Haltung, Diskursfähigkeit, Kalibrierung, der Kontakt mit Menschen, die einen herausfordern statt nur zu bestätigen. Das lässt sich schwer messen und wird leicht übersehen, wenn man Ausbildung ausschließlich als Wissenstransfer betrachtet.

Wer als Hochschule (a) die eigene Rolle nicht grundsätzlich infrage gestellt sehen will, (b) davon ausgeht, dass beim volldigitalen, unbegleiteten Lernen je nach Fach etwas Wesentliches verloren geht, und (c) zur Kenntnis nimmt, dass Studierende KI ohnehin nutzen, für den ist die Konsequenz eine Gestaltungsaufgabe: eine eigene, angemessene Nutzung von KI in der eigenen Lehre anzubieten.

Angemessen heißt hier ausdrücklich konkurrenzfähig zu ChatGPT, NotebookLM oder vergleichbarer Technologie. Die Vorteile, gegen die eine Hochschule antritt, sind klar: Verfügbarkeit 24 Stunden am Tag, keine Anfahrt zur Uni, kein Warten auf die Sprechstunde am Donnerstag, keine Frage, die zu banal für eine Rückmeldung per E-Mail wäre. Wer hier nichts anbietet, tritt gegen diese Vorteile mit nichts an und verliert die Vermittlungsrolle nicht an ein besseres pädagogisches Konzept, sondern schlicht an Verfügbarkeit.

Das führt zwingend zum Management von Erwartungen. Auch KI macht aus einer Hochschule nicht automatisch eine, an der plötzlich mehr oder engagierter gelernt wird. Es ist wie in allen Digitalisierungsvorhaben: Allein einen Prozess zu digitalisieren, ändert in der Regel gar nichts. Das zeigt auch die Begleitforschung zu unserem eigenen Werkzeug OneTutor. Zwischen Häufignutzenden, Wenignutzenden und Nichtnutzenden ließen sich im Schnitt keine Unterschiede im subjektiv wahrgenommenen Lernerfolg feststellen (Stumpf et al. 2026; siehe Kapitel 7). Umgekehrt gilt allerdings: Wer dieses Werkzeug nutzt, ist damit sehr zufrieden.

Was ein gutes Werkzeug an der Hochschule realistisch leisten kann, ist, dass die Lernaktivität, die ohnehin virtuell stattfindet, auf der eigenen Infrastruktur passiert statt bei einem austauschbaren Wettbewerber, und dass dabei nicht verloren geht, was nur die Hochschule geben kann: das Kuratieren, also die kompetente Entscheidung darüber, was relevant ist und in welcher Tiefe; die Kohorte, in der viele Studierende dasselbe Material bearbeiten; und die Nähe zur Prüfung, die überhaupt erst bestimmt, worauf sich eine Rückmeldung sinnvoll bezieht.

DREI WEGE, DIE EINER HOCHSCHULE OFFENSTEHEN

Weg	Was er löst	Was offen bleibt
1. Nichts anbieten	Nichts. Kein Aufwand, keine Kosten, keine Governance-Fragen.	Die Vermittlungsrolle geht an die Verfügbarkeit verloren. Nichtstun ist selbst eine Entscheidung.
2. Campus-Lizenz für einen allgemeinen Assistenten	Das Verfügbarkeitsproblem vollständig. Schnell beschlossen, günstig pro Kopf.	Das Material wählen die Studierenden. Die Fragen der Kohorte bleiben in einzelnen Konten liegen. Der Ersatzpfad bleibt offen. Die Hochschule erfährt nicht, was sie ausgelöst hat.
3. An die Veranstaltung gebundenes Werkzeug	Verfügbarkeit, Kuration, Kohortensignal und Messbarkeit gemeinsam.	Der aufwendigste Weg. Er verlangt Entscheidungen von Lehrenden und eine Governance für formative Daten.

KAPITEL 4

Zwei Grundsätze für die Gestaltung

Lernen kennt keine Abkürzungen. Ein Werkzeug darf die Effizienz deshalb nur begrenzt erhöhen; seine Aufgabe ist, den Lernprozess effektiver zu machen. Und nur Hochschulen haben dafür den entscheidenden Hebel: das Lernverhalten einer ganzen Kohorte, aus dem sich Lehre und Lernen fortlaufend verbessern lassen.

4.1 Keine Abkürzungen

Die individuell notwendige Zeit fürs Lernen lässt sich nicht wesentlich verkürzen, denn Lernen ist die zeitintensive Auseinandersetzung mit der Materie. Lernen ist kein Nebenprodukt, das man abkürzen könnte, ohne das Ziel zu verfehlen. Was sich verkürzen lässt, ist alles, was um diese Auseinandersetzung herum liegt und was man als Reibung bezeichnen kann: die Stelle im Skript suchen, auf eine Antwort warten, feststecken ohne Ansprechpartner, nicht darauf hingewiesen zu werden, dass man etwas missverstanden hat, oder daran verzweifeln, dass Lernmaterialien schlicht fehlerhaft sind.

Diese Trennung ist nicht neu. Die Lernpsychologie kennt sie als wünschenswerte Erschwernisse: Bedingungen, die die kurzfristige Leistung verschlechtern und gerade dadurch Behalten und Transfer langfristig verbessern (Bjork & Bjork 2011). Es geht dabei nicht darum, ob unter erschwerten Bedingungen überhaupt gelernt wird, sondern darum, dass unter mehreren Lernweisen in der Regel die subjektiv mühsamere auch die wirksamere ist. Neu ist hier nur die Anwendung als Konstruktionsprinzip für ein Werkzeug statt als Ratschlag an Lernende.

Formatives Feedback ist der Musterfall auf der erlaubten Seite dieser Regel. Die Latenz zwischen Fehler und Rückmeldung ist Reibung, nicht Anstrengung. Wer am Sonntagabend eine Aufgabe falsch löst und das am Donnerstag in der Übung erfährt, hat vier Tage lang ein falsches Modell konsolidiert. Diese Wartezeit trägt nichts zum Lernen bei, sie schadet. Der Versuch selbst, Gelerntes abzurufen, bleibt dagegen bei der lernenden Person. Das ist die Anstrengung, die nicht delegierbar ist.

Wenn es beim Lernen keine Abkürzungen gibt, wozu dann überhaupt ein Werkzeug? Reibung zu senken ist die naheliegende Antwort, aber sie ist nachrangig, denn sie stellt vermutlich nicht das größte Hindernis beim Lernen dar. Der Zweck liegt woanders: Die vorhandene Lernzeit soll treffgenauer und wirksamer genutzt werden. Treffgenauer heißt, dass die Auseinandersetzung dort ansetzt, wo das Verständnis tatsächlich abbricht. Wirksamer heißt, dass in derselben Zeit die bessere Tätigkeit stattfindet.

Wie viel Zeit jemand ins Lernen steckt, sagt über die erreichte Leistung wenig aus (Plant et al. 2005). Entscheidend ist, woraus diese Zeit besteht. Befragt man Studierende nach ihrer Hauptstrategie, nennt die große Mehrheit wiederholtes Lesen und nur eine kleine Minderheit das Selbsttesten (Karpicke et al. 2009). Gelernt wird außerdem terminabhängig statt geplant, also massiert statt verteilt (Hartwig & Dunlosky 2012). Der Forschungsstand ist ungewöhnlich klar: Markieren, Wiederlesen und Zusammenfassen gehören zu den schwachen Techniken, Praxistests und verteiltes Lernen zu den starken (Dunlosky et al. 2013; Roediger & Karpicke 2006; Cepeda et al. 2006). Die Lernzeit ist nicht zu knapp. Sie ist falsch zusammengesetzt.

Warum das so bleibt, ist kein Vorwurf an die Studierenden. Flüssiges Wiederlesen fühlt sich nach Verstehen an, ein misslungener Abrufversuch nach Scheitern. Lernende wählen deshalb systematisch die Methode, die sich besser anfühlt und schlechter wirkt. Es ist dieselbe Umkehrung von empfundenem und tatsächlichem Lernerfolg, die Deslauriers et al. (2019) für Lehrformate

gemessen haben. Es fehlt also weder an Zeit noch an Einsicht, sondern an einer Instanz, die den Abrufversuch anstößt und sein Ergebnis zurückmeldet, solange noch Zeit bleibt, etwas zu ändern.

4.2 Der unterschätzte Wert der Kohorte

Die Kohorte, in der ein Studierender lernt, ist aus mindestens zwei Gründen wichtig. Der erste ist sozial und in Kapitel 3 benannt. Der zweite ist für die Hochschule der wirtschaftlich wichtigere: Die Auseinandersetzung vieler Studierender mit demselben Material kann dazu verwendet werden, dieses Material zu verbessern. Wo eine Frage im Chat wiederholt auftaucht, eine Quizfrage systematisch scheitert oder ein Kapitel im Skript auffällig oft nachgefragt wird, zeigt sich eine Lücke, die vorher unsichtbar war. Aus dieser Häufung lassen sich Materialien anhand eines Befunds über eine ganze Kohorte hinweg ergänzen. Das ist die Seite des formativen Assessment, bei der auch die Lehrenden gewinnen und indirekt dadurch die Lernenden.

Diese kohortenumfassende, nicht individuelle Learning Analytics ist ein Wert, den ein einzeln genutztes ChatGPT oder NotebookLM strukturell nicht erzeugen kann. Dort bleibt jede Konversation isoliert und kommt niemandem außer der einen Person zugute. Ein Werkzeug, das an eine Institution gebunden ist, kann diese Signale aggregieren und in bessere Kuratierung zurückführen. Die KI liefert das Signal, der Mensch entscheidet, was daraus wird. Das ist zugleich der beste Grund, warum das Kuratieren beim Menschen bleiben sollte. Und es ist derselbe Datenbestand, der überindividuelle Begleitforschung überhaupt erst möglich macht.

Der Wert dieser Signale hängt allerdings davon ab, ob die Erhebung der Daten folgenlos ist. Ein Quiz, das zur Note beiträgt, misst nicht mehr nur Verständnis, sondern kann auch Optimierungsverhalten zeitigen. Eine Chatfrage, von der Studierende vermuten, dass sie irgendwann aktenkundig wird, wird vielleicht nicht gestellt. In einer idealen Welt bleiben formative Instrumente formativ. Ob sie dann aber überhaupt genutzt werden, ist eine andere Frage, und Bonuspunkte mögen der Motivation dienlich sein. Dieser Zielkonflikt muss im Einzelfall entschieden werden.

KAPITEL 5

Ein Schichtenmodell

Vier Ebenen trennen, was in der KI-Debatte meist vermischt wird: wann und ob überhaupt gelernt wird, woraus die Lernzeit besteht, und was quer zu allem liegt. Diese Trennung zeigt, wo ein Werkzeug wirken kann und wo nicht. Ohne sie bleibt jede Erwartung an KI unüberprüfbar.

Um die Gestaltungsaufgabe aus Kapitel 3 zu lösen, braucht es ein Vokabular, das vier Dinge trennt, die üblicherweise vermischt werden: wann gelernt wird, ob überhaupt gelernt wird, woraus die Lernzeit besteht, und was keiner einzelnen Phase zuzuordnen ist.

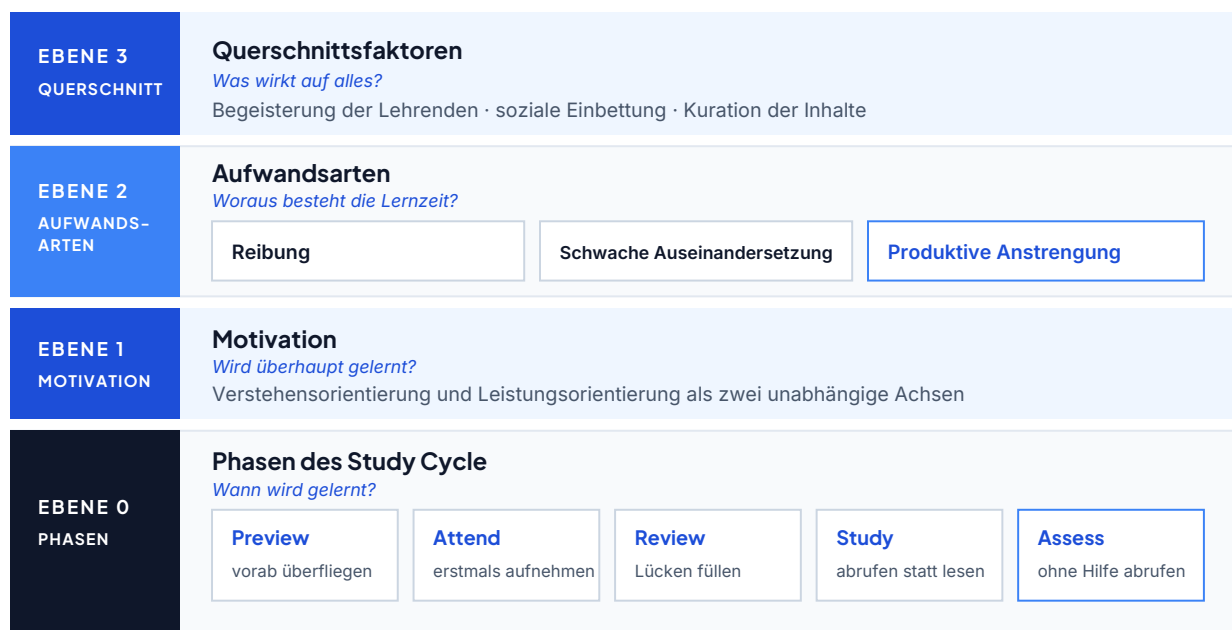


Abbildung 1 Das Schichtenmodell. Ebene 0 nach dem Study Cycle des UNC Learning Center, aufbauend auf dem PLRS-System von Christ (1997). Ein Werkzeug kann sinnvoll nur auf Ebene 2 ansetzen und dort die schwache Auseinandersetzung in produktive Anstrengung überführen.

Ebene 0: Phasen des Study Cycle

Als zeitliches Grundgerüst dient der Study Cycle in der Fassung des UNC Learning Center mit fünf Phasen, aufbauend auf dem PLRS-System von Christ (1997): Preview (altes und neues Material vor der Sitzung überfliegen), Attend (Stoff zum ersten Mal aufnehmen), Review (innerhalb etwa eines Tages Notizen ordnen, Quizzes beantworten, Lücken füllen), Study (üben, sich selbst erklären, abrufen statt wiederlesen) und Assess (ohne Hilfe abrufen, Lücken erkennen: formatives Assessment). Der Zyklus ist nicht linear. Assess führt zurück zu Study, und derselbe Stoff durchläuft die Phasen mehrfach. Wir wählen dieses Modell, weil es aktivitätsbasiert ist: Jede Phase hat beobachtbare Handlungen, und nur auf Handlungen lassen sich Werkzeuge sinnvoll abbilden. Es ist eine Lernstrategie-Heuristik, keine validierte Lerntheorie, und wir verwenden es als Ordnungsraster, nicht als Wirknachweis. Dass Studierende diesen Zyklus angemessen gestalten, kann ein Werkzeug allein nicht sicherstellen. Dazu sind weitere didaktische und strukturelle Maßnahmen nötig, die nicht im Fokus dieses Papiers liegen.

Ebene 1: Motivation

Ob eine Phase überhaupt stattfindet, entscheidet nicht das Werkzeug, sondern die Motivation der Lernenden. Verstehensorientierung ist der Wunsch, ein Fach wirklich zu durchdringen, auch über den Prüfungsstoff hinaus. Leistungsorientierung ist der Wunsch, möglichst effizient zu bestehen oder eine gute Note zu erreichen. Beide Achsen sind unabhängig und keine Gegenpole; dieselbe Person kann in einem Fach den vollen Zyklus durchlaufen und in einem anderen nur auf Bestehen hin lernen. Die Unterscheidung entspricht der Mastery- und Performance-Achse der Zielorientierungsforschung, deren Unabhängigkeit faktorenanalytisch gestützt ist (Elliot & McGregor 2001). Zielorientierungen sind dabei kein rein stabiles Personenmerkmal, sondern verändern sich auch innerhalb eines Semesters (Fryer & Elliot 2007).

Daraus folgt, dass ein Werkzeug ein Verstärker bereits stattfindender Aktivität sein kann, aber in der Regel kein Initiator. Kein Werkzeug allein bringt jemanden dazu, sich hinzusetzen und zu lernen. Wie die Viertelstunde dann genutzt wird, ist dagegen sehr wohl gestaltbar, und darum geht es in Ebene 2.

Ebene 2: Reibung, schwache Auseinandersetzung, produktive Anstrengung

Jede Lernphase zerfällt in Anteile, die sich nach ihrem Beitrag zum Lernen unterscheiden lassen. Zwei davon sind schnell benannt: Reibung, die nichts zum Lernen beiträgt, und produktive Anstrengung, die das Lernen selbst konstituiert. Diese Zweiteilung ist allerdings unvollständig. Wer ein Kapitel zum dritten Mal liest, sucht nicht die Stelle im Skript und wartet auf keine Antwort. Das ist keine Reibung, sondern Beschäftigung mit der Sache. Es ist aber auch nicht die Anstrengung, an der das Lernen hängt. Zwischen beiden liegt eine dritte Kategorie, die schwache Auseinandersetzung: echte, aber wenig wirksame Arbeit am Inhalt. Sie ist keine Randerscheinung, sondern der empirische Normalfall (Karpicke et al. 2009; Dunlosky et al. 2013).

DAS KONSTRUKTIONSPRINZIP AUF EBENE 2

Anteil der Lernzeit	Beispiel	Was ein Werkzeug damit tun darf
Reibung	Die Stelle im Skript suchen, auf eine Antwort bis Donnerstag warten	Beliebig senken. Hier ist der Nutzen unstrittig, aber nicht exklusiv.
Schwache Auseinandersetzung	Dasselbe Kapitel zum dritten Mal lesen, markieren, zusammenfassen	In produktive Anstrengung überführen. Das ist der eigentliche Zweck.
Produktive Anstrengung	Ohne Hilfe abrufen, selbst formulieren, den eigenen Denkfehler finden	Strukturieren, auslösen, rückmelden, aber niemals ersetzen.

Ein Werkzeug, das produktive Anstrengung in schwache Auseinandersetzung zurückverwandelt, indem es die Antwort mitliefert, statt sie zu verlangen, verfehlt seinen Zweck. Und es tut das unauffällig, weil sich das Ergebnis für die lernende Person wie Fortschritt anfühlt.

Diese Achse genügt allerdings nicht. Sie unterscheidet nach der Handlung. Es braucht eine zweite Achse, nämlich welche Konsequenzen ggf. außerhalb des individuellen Lernprozesses entstehen. Der zulässige Unterstützungsgrad eines Werkzeugs hängt von den externen Konsequenzen der Lernleistung ab. Wo es keine gibt, darf das Werkzeug weitgehend helfen, denn die Studierenden haben nichts davon, sich selbst zu täuschen. Wo ein Zeugnis oder ein Notenbonus die Konsequenz ist, lädt das Werkzeug zum Missbrauch ein. Diese zweite Achse ist der Grund, warum formatives Assessment die größte Chance und gleichzeitig die empfindlichste Stelle des ganzen Modells ist, wenn Lernanstrengungen extern belohnt werden.

Ebene 3: Querschnittsfaktoren

Drei Faktoren lassen sich keiner Phase zuordnen und beeinflussen doch alle. Es sind dieselben, die in Kapitel 3 als das eigentlich Gefährdete benannt wurden. Begeisterung entsteht bei den Dozierenden und wirkt als Multiplikator auf die Verstehensachse; sie überlebt einen Videostream gedämpft und ein Skript gar nicht. Soziale Einbettung entsteht aus Präsenz: Lerngruppen, Flurgespräche, das beiläufige Mitbekommen fremder Denkwege. Die Kuratierung von Lerninhalten liegt bei der Institution. Klassische KI-Werkzeuge wirken auf dieser Ebene heute praktisch nicht. Ebene 3 ist der Teil des Modells, an dem ein rein reaktives Chat-Werkzeug grundsätzlich nichts ändern kann, da bei dem auch ein proaktives Werkzeug durch weitere didaktische Maßnahmen komplementiert werden muss.

KAPITEL 6

Anwendung: OneTutor

OneTutor ist für Studierende gebaut, aber dozentenorientiert aufgesetzt. Das Kuratieren von Quizzes und geleiteten Gesprächen durch Dozenten sichert die Qualität, die Bindung an die Veranstaltung und die Konzentration auf das Wesentliche. Das Lernverhalten der Kohorte wird so zur Grundlage für bessere Lehre und für Begleitforschung.

OneTutor ist ein KI-Tutor, der für Studierende gedacht und dozentenorientiert entworfen ist. Dozierende laden die Materialien ihres Kurses hoch und kuratieren Quizfragen und KI-geleitete Diskussionen. Diese Designentscheidung ermöglicht den Zugriff auf das anonymisierte Lernverhalten einer ganzen Kohorte, vor allem darauf, welche Teile nicht verstanden wurden. Dozierende können daraus KI-gestützt bessere Vorlesungsmaterialien und neue Quizzes ableiten.

50.000+

Studierende nutzen OneTutor.

50+

Institutionen setzen das Werkzeug ein.

800+

Kurse arbeiten damit.

Stand: August 2026

Chat: Known Unknowns

Wissen Studierende, was sie nicht wissen, können sie Fragen stellen, die OneTutor relativ zum Veranstaltungsmaterial beantwortet, und zwar in den Phasen Preview, Attend, Review und Study. Im Hörsaal hat jede Frage einen sozialen Preis: Vor dreihundert Leuten zuzugeben, dass man Folie 12 nicht verstanden hat, kostet etwas. Der Chat, ein RAG-System auf Basis der Kursmaterialien und Transkripte, entfernt genau diesen Preis: anonym, sofort, beliebig oft, mitten in der Vorlesung. Aus Studierendenperspektive sind ChatGPT oder NotebookLM hier gleichwertig; aus Hochschulperspektive geht der Wert der Kohorte verloren.

Quiz: Unknown Unknowns

Von Dozierenden kuratierte Quizfragen stellen die Frage, beantworten sie aber nicht. Die Denkarbeit, die Antwort aus dem eigenen Kopf zu holen, bleibt bei der lernenden Person, ein Grundprinzip des formativen Assessment. Dass dieser Abrufversuch das Behalten stärker verbessert als erneutes Lesen desselben Materials, gehört zu den robustesten Ergebnissen der Lernpsychologie (Roediger & Karpicke 2006).

Wie weit dieser Effekt über die geübten Inhalte hinaus trägt, hängt allerdings von der Bauart der Fragen ab. Transfer tritt vor allem dann auf, wenn Übung und Prüfung inhaltlich zusammenpassen und die Aufgabe eine eigene Formulierung statt bloßen Wiedererkennens verlangt (Pan & Rickard 2018). Beides spricht gegen frei erzeugte Fragen und für kuratierte, an den Lernzielen ausgerichtete.

Das deckt sich mit unserer Erfahrung: KI-generierte Quizfragen müssen von Dozierenden kuratiert werden. Wo eine KI aus Unterrichtsmaterialien völlig frei Fragen erzeugt, beobachten wir Fragen und Antworten, die manchmal subtil falsch, manchmal irreführend, manchmal fundamental falsch und manchmal schlicht irrelevant sind.

Quizzes werden vor allem vor Prüfungen verwendet. Wir beobachten aber auch, dass ein Teil der Studierenden sie direkt vor und direkt nach der Vorlesung bearbeitet, also in den Preview- und

Review-Phasen. Ein Quiz vor der Sitzung ist dann ein Pretest, kein Überblick: Der gescheiterte Abrufversuch öffnet genau die Lücke, die die Vorlesung anschließend füllt. Der Effekt ist belegt. In fünf Experimenten war die Nachtestleistung nach einem Vortest besser als nach verlängerter Lesezeit, selbst wenn ausschließlich jene Inhalte ausgewertet wurden, die im Vortest nicht korrekt abgerufen worden waren (Richland et al. 2009).

Converse: Lernen durch geführte Interaktion

Eine dritte, derzeit in Erprobung befindliche Funktionalität dreht die Rolle des KI-Tutors um, vor allem in der Study- und Assess-Phase. Statt auf eine Frage zu warten, führt und moderiert das System das Gespräch, individuell oder mit einer Gruppe, in Einheiten von rund fünfzehn Minuten, zu Inhalten, die die Dozierenden vorgeben. Voraussetzung ist, dass die Studierenden den Stoff im Wesentlichen verstanden haben. In diesen Gesprächen lernen sie, präzise und vollständig zu antworten, vertiefen so ihr Verständnis und erkennen dabei Lücken. In einer Gruppe eingesetzt, idealerweise physisch vor Ort, adressiert die Funktion die soziale Einbettung aus Ebene 3, die mit sinkender Präsenz verschwindet.

Diese Funktionalität kommt dem Ideal formativen Feedbacks nach Shute (2008) am nächsten: nicht bewertend, unmittelbar, spezifisch auf die gerade gegebene Antwort bezogen und beliebig wiederholbar. Sie verlangt zudem genau die eigene Formulierung, die nach Pan & Rickard (2018) den Transfer begünstigt. Genau diese Kombination war in einer Veranstaltung mit Hunderten Teilnehmenden bisher nicht herstellbar. Perspektivisch ist unter Berücksichtigung rechtlicher und ethischer Rahmenbedingungen zu untersuchen, ob diese Gespräche nicht auch im summativen Assessment verwendet werden können, als beliebig skalierende mündliche Prüfungen.

KAPITEL 7

Warum Begleitforschung nicht optional ist

Inwiefern KI-Tutoren „nützen“, ist eine komplexe Frage. Wer sie nicht stellt, weiß nicht, unter welchen Bedingungen KI Studierenden wirklich hilft und wo es ihnen schadet, und handelt damit letztlich verantwortungslos.

Niemand weiß derzeit belastbar, wie gut KI-Tutoren in der Hochschullehre funktionieren und was das überhaupt heißt. Der weit überwiegende Teil dessen, was über KI in der Lehre gesagt wird, sind heute Plausibilitätsargumente: theoretisch fundiert, in sich schlüssig, empirisch offen. Auch die Mechanismen aus Kapitel 6 sind gut begründet, aber für dieses Produkt in diesem Setting in der Regel noch nicht bewiesen.

Schon die Frage, ob ein KI-Assistenzsystem „funktioniert“, ist vielschichtig. Korrelationen von Noten mit Werkzeugnutzung können in die Irre führen, weil es unterschiedliche Lerntypen und Motivationslagen gibt. Subjektive Nützlichkeitsurteile aus Fragebögen sind ebenfalls mit Vorsicht zu genießen: Der Abbau von Reibung und das Ersetzen von Anstrengung fühlen sich für Studierende identisch an. In einem randomisierten Vergleich an der Harvard University lernten Studierende in aktiven Lernsettings nachweislich mehr, hatten aber das Gefühl, weniger zu lernen. Die Autoren folgern, dass Lehrevaluation auf Basis studentischer Wahrnehmung unbeabsichtigt die schlechtere, passive Methode belohnen kann (Deslauriers et al. 2019).

Diese Unzulänglichkeiten dürfen aber keine Ausrede sein, Erfahrungen mit KI-Tutoren nicht zu objektivieren. Wir haben den Einsatz von OneTutor deshalb von Anfang an wissenschaftlich begleiten lassen. Das Bayerische Forschungsinstitut für Digitale Transformation (bidt) der Bayerischen Akademie der Wissenschaften untersucht im Projekt „Affectiveness“ den Einsatz an zehn bayerischen Hochschulen. Grundlage der ersten Studie sind das Sommersemester 2025, 84 Lehrveranstaltungen, Befragungen von bis zu rund 1.000 Studierenden und 31 Dozierenden sowie Nutzungsdaten aus dem System (Stumpf et al. 2026).

80 %

der befragten Studierenden nutzten den Tutor aktiv.

90 %

der Dozierenden sehen ihn als sinnvolle Ergänzung.

84 %

der Dozierenden würden ihn erneut einsetzen.

Zwischen Häufignutzenden, Wenignutzenden und Nichtnutzenden ließen sich dagegen keine Unterschiede im subjektiv wahrgenommenen Lernerfolg feststellen. Nichtnutzende erreichen mit anderen Lernwegen offenbar Vergleichbares. Umgekehrt gilt aber, dass wer das Werkzeug nutzt, es auch schätzt. Vorläufig lässt sich daraus schließen, dass OneTutor breit angenommen, freiwillig genutzt und von Lehrenden wie Nutzenden als hilfreich bewertet wird. Nicht ableiten lässt sich, dass alle Studierenden damit mehr oder besser lernen als ohne. Der Anspruch ist allerdings auch nicht, dass mehr gelernt wird, sondern dass sich die Zusammensetzung der Lernzeit verschiebt. Wäre er eingelöst, müsste eine Selbstauskunft ihn nach Deslauriers et al. (2019) eher unterschätzen als bestätigen, denn die wirksamere Tätigkeit fühlt sich schlechter an.

Relevante Studien sind inzwischen nicht mehr rar; belastbare schon. Seit 2024 ist eine ganze Reihe von Metaanalysen zum Lerneffekt generativer KI erschienen, die überwiegend positive und teils große Effekte berichten (etwa Liu et al. 2025). Ihr Wert ist begrenzt: Eine vielzitierte Arbeit wurde inzwischen zurückgezogen (Wang & Fan 2025), und unabhängige Neuanalysen desselben Datenbestands zeigen, dass die berichteten Effekte nach Korrektur um den Publikationsbias stark schrumpfen. Für das Feld insgesamt gilt, dass sich der Großteil der publizierten Arbeiten

auf Selbstauskünfte und kurzfristige Messungen stützt. Belastbare Daten zu Transfer und Behalten über ein Semester hinaus fehlen praktisch vollständig.

Genau das ist das Argument für eigene Begleitforschung, und zwar ein stärkeres als ein bloßer Mangel an Literatur es wäre. Wer sich auf den publizierten Bestand verlassen will, verlässt sich auf Effektstärken, die zwischen „groß“ und „nicht von null unterscheidbar“ schwanken, je nachdem, wer nachrechnet. Eine Hochschule, die wissen will, was ein Werkzeug in ihren Veranstaltungen, bei ihren Studierenden und in ihren Fächern bewirkt, kann diese Antwort nicht aus der Literatur beziehen. Sie muss sie selbst erheben.

Die wenigen methodisch starken Einzelstudien zeigen zugleich, wie eng ihre Reichweite ist. Kestin et al. (2025) fanden in einem randomisierten Crossover-Vergleich an der Harvard University mehr als doppelten Lernzuwachs gegenüber einer Active-Learning-Sitzung, allerdings über zwei Lektionen, einen Kurs, eine Institution und mit einem eigens gebauten Tutor, entwickelt von dem Lehrenden, der auch die Vergleichssitzung hielt; gemessen wurde unmittelbar danach, nicht Wochen später. Das ist ein starkes Signal für gut gebaute Werkzeuge, kein Beleg für KI-Tutoren im Allgemeinen. Bastani et al. (2024) zeigen im randomisierten Design mit knapp tausend Mathematik-Lernenden in der Schule: Ein reines ChatGPT ohne didaktische Guardrails führt zu deutlich schlechteren Prüfungsergebnissen als Lernen ohne ChatGPT oder mit einem ChatGPT mit Guardrails. Bemerkenswert ist dabei das Muster: Während der Übung waren diese Lernenden besser, in der Prüfung schlechter. Genau so sieht ersetzte Anstrengung von innen aus.

Daraus folgt kein Plädoyer für Zurückhaltung bei der Begleitforschung, denn Nichtstun ist selbst eine Entscheidung mit Folgen. Drei Gründe machen eine Pflicht zur Messung zwingend: Die Intervention greift in das Lernen einer ganzen Kohorte ein. Ihre Effekte können negativ sein, und zwar unbemerkt (Bastani et al. 2024). Und die Wahrnehmung der Betroffenen ist als Kontrollinstanz systematisch unzuverlässig (Deslauriers et al. 2019). Wer KI-Tutoren ohne Begleitforschung ausrollt, erfährt nicht, ob sie helfen oder schaden, und kann es später nicht mehr rekonstruieren.

KAPITEL 8

Sechs Fragen an ein Werkzeug

Ob Studierende mit KI lernen, ist keine offene Frage mehr. Offen ist, mit wessen KI und was von der eigenen Hochschule dabei übrig bleibt: die Kohorte, die sich gegenseitig etwas beibringt, das Kuratieren, das aus deren Fragen entsteht, und das Zeugnis, das am Ende etwas wert ist.

Die defensive Lesart dieser Lage ist die verbreitete: KI bedroht die beurkundende Prüfung, also das summative Assessment. Diese Lesart ist unvollständig. Denn dieselbe Technologie macht die andere Hälfte des Prüfens, die formative, erstmals in der Breite verfügbar. Das ist die Chance, die Hochschulen haben und die ein einzeln genutzter Chatbot strukturell nicht hat, weil ihm das kuratierte Material, die Lernziele und die Kohorte fehlen. Sie zu nutzen kostet zwei Entscheidungen: die formativen Instrumente einzusetzen, und zu messen, ob sie wirken. Woran ein Werkzeug erkennbar ist, das dafür überhaupt in Frage kommt, lässt sich an sechs Fragen festmachen.

- 1 Bezieht sich das Werkzeug auf das kuratierte Material und die Lernziele der Veranstaltung, und entscheiden über diese Auswahl die Lehrenden und nicht die Lernenden?
- 2 Entsteht aus der Nutzung ein Signal über die Kohorte, das in bessere Kuration zurückfließt, statt in einzelnen Studierendenkonten zu verbleiben?
- 3 Ist die wirksamere Tätigkeit der Weg des geringsten Widerstands und nicht die empfohlene Option daneben?
- 4 Ist der einfache Ersatzpfad im Werkzeug strukturell verschlossen und nicht nur nicht empfohlen?
- 5 Ist beabsichtigt, dass die formativen Daten für die Notengebung strukturell unverwertbar sind, und ist das den Studierenden kommuniziert?
- 6 Liegen bei der Hochschule überhaupt die Daten, an denen sich später ablesen ließe, was das Werkzeug bewirkt hat?

Vier dieser sechs Fragen kann ein allgemeiner Chat-Assistent nicht mit Ja beantworten, und zwar nicht wegen seiner Antwortqualität, sondern wegen seiner Bauform: Er ist an keine Veranstaltung gebunden, kennt keine Kohorte, hält jeden Pfad offen und hinterlässt bei der Institution keine auswertbaren Spuren.

Beim Lernen gibt es keine Abkürzungen. Werkzeuge allein können dem Lernprozess deswegen sogar abträglich sein. Nur Hochschulen haben den Hebel, KI so einzusetzen, dass alle Studierenden effektiver lernen können. OneTutor kann dabei helfen.

Quellen

- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö. & Mariman, R. (2024):** Generative AI Can Harm Learning. The Wharton School Research Paper, SSRN Working Paper Nr. 4895486. doi.org/10.2139/ssrn.4895486. Preprint, nicht peer-reviewed.
- Bennett, R. E. (2011):** Formative Assessment: A Critical Review. *Assessment in Education: Principles, Policy & Practice* 18(1), S. 5–25. doi.org/10.1080/0969594X.2010.513678
- Bjork, E. L. & Bjork, R. A. (2011):** Making Things Hard on Yourself, But in a Good Way. In: Gernsbacher, M. A. et al. (Hrsg.): *Psychology and the Real World*. Worth Publishers, New York, S. 56–64. Seitenangabe nach der Erstauflage; die 2. Auflage (2014) paginiert S. 59–68.
- Black, P. & Wiliam, D. (1998):** Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice* 5(1), S. 7–74. doi.org/10.1080/0969595980050102
- Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L. & Yin, Y. (2012):** Meta-Analytic Methodology and Inferences About the Efficacy of Formative Assessment. *Educational Measurement: Issues and Practice* 31(4), S. 13–17. doi.org/10.1111/j.1745-3992.2012.00251.x
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T. & Rohrer, D. (2006):** Distributed Practice in Verbal Recall Tasks. *Psychological Bulletin* 132(3), S. 354–380. doi.org/10.1037/0033-2909.132.3.354
- Christ, F. L. (1997):** Seven Steps to Better Management of Your Study Time. H&H Publishing, Clearwater, FL. Primärquelle des PLRS-Systems.
- Credé, M., Roch, S. G. & Kieszczynka, U. M. (2010):** Class Attendance in College. *Review of Educational Research* 80(2), S. 272–295. doi.org/10.3102/0034654310362998
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K. & Kestin, G. (2019):** Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *PNAS* 116(39), S. 19251–19257. doi.org/10.1073/pnas.1821936116
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J. & Willingham, D. T. (2013):** Improving Students' Learning With Effective Learning Techniques. *Psychological Science in the Public Interest* 14(1), S. 4–58. doi.org/10.1177/1529100612453266
- Elliot, A. J. & McGregor, H. A. (2001):** A 2 × 2 Achievement Goal Framework. *Journal of Personality and Social Psychology* 80(3), S. 501–519. doi.org/10.1037/0022-3514.80.3.501
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H. & Wenderoth, M. P. (2014):** Active learning increases student performance in science, engineering, and mathematics. *PNAS* 111(23), S. 8410–8415. doi.org/10.1073/pnas.1319030111
- Fryer, J. W. & Elliot, A. J. (2007):** Stability and Change in Achievement Goals. *Journal of Educational Psychology* 99(4), S. 700–714. doi.org/10.1037/0022-0663.99.4.700
- Hartwig, M. K. & Dunlosky, J. (2012):** Study strategies of college students. *Psychonomic Bulletin & Review* 19(1), S. 126–134. doi.org/10.3758/s13423-011-0181-y
- Karpicke, J. D., Butler, A. C. & Roediger, H. L. (2009):** Metacognitive strategies in student learning. *Memory* 17(4), S. 471–479. doi.org/10.1080/09658210802647009
- Kestin, G., Miller, K., Klaes, A., Milbourne, T. & Ponti, G. (2025):** AI tutoring outperforms in-class active learning. *Scientific Reports* 15, Artikel 17458. doi.org/10.1038/s41598-025-97652-6
- Kingston, N. & Nash, B. (2011):** Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice* 30(4), S. 28–37. doi.org/10.1111/j.1745-3992.2011.00220.x
- Liu et al. (2025):** The Impact of ChatGPT on Students' Academic Achievement: A Meta-Analysis. *Journal of Computer Assisted Learning*. doi.org/10.1111/jcal.70096
- Pan, S. C. & Rickard, T. C. (2018):** Transfer of Test-Enhanced Learning. *Psychological Bulletin* 144(7), S. 710–756. doi.org/10.1037/bul0000151
- Plant, E. A., Ericsson, K. A., Hill, L. & Asberg, K. (2005):** Why study time does not predict grade point average across college students. *Contemporary Educational Psychology* 30(1), S. 96–116. doi.org/10.1016/j.cedpsych.2004.06.001
- Reich, J. & Ruipérez-Valiente, J. A. (2019):** The MOOC pivot. *Science* 363(6423), S. 130–131. doi.org/10.1126/science.aav7958
- Richland, L. E., Kornell, N. & Kao, L. S. (2009):** The pretesting effect. *Journal of Experimental Psychology: Applied* 15(3), S. 243–257. doi.org/10.1037/a0016496
- Roediger, H. L. & Karpicke, J. D. (2006):** Test-Enhanced Learning. *Psychological Science* 17(3), S. 249–255. doi.org/10.1111/j.1467-9280.2006.01693.x
- Roth, P. L., BeVier, C. A., Switzer, F. S. & Schippmann, J. S. (1996):** Meta-Analyzing the Relationship Between Grades and Job Performance. *Journal of Applied Psychology* 81(5), S. 548–556. doi.org/10.1037/0021-9010.81.5.548
- Scarfe, P., Watcham, K., Clarke, A. & Roesch, E. (2024):** A real-world test of artificial intelligence infiltration of a university examinations system. *PLOS ONE* 19(6), e0305354. doi.org/10.1371/journal.pone.0305354
- Schneider, M. & Preckel, F. (2017):** Variables associated with achievement in higher education. *Psychological Bulletin* 143(6), S. 565–600. doi.org/10.1037/bul0000098
- Shute, V. J. (2008):** Focus on Formative Feedback. *Review of Educational Research* 78(1), S. 153–189. doi.org/10.3102/0034654307313795
- Stumpf, C., Löwel, V., Schlude, A. & Stürz, R. A. (2026):** Ein KI-Tutor geht in die Lehre. *bidt Analysen & Studien* Nr. 20, München, 24. März 2026. doi.org/10.35067/xypq-kn78
- Stürz, R. A., Schlude, A., Harles, D., Mendel, U. & Stumpf, C. (2025):** Das bidt-Digitalbarometer 2025. *bidt Analysen & Studien* Nr. 17, München. doi.org/10.35067/xypq-kn75
- UNC Learning Center (o. J.):** The Study Cycle. University of North Carolina at Chapel Hill. learningcenter.unc.edu/tips-and-tools/the-study-cycle. Adaptiert nach Christs PLRS-System durch das LSU Center for Academic Success, dargestellt bei McGuire, S. Y. (2018): *Teach Yourself How to Learn*. Stylus, Sterling, VA.
- Van Iddekinge, C. H., Arnold, J. D., Krivacek, S. J., Frieder, R. E. & Roth, P. L. (2024):** Making the Grade? A Meta-Analysis of Academic Performance as a Predictor of Work Performance and Turnover. *Journal of Applied Psychology* 109(12), S. 1972–1993. doi.org/10.1037/apl0001212
- Wang, J. & Fan, W. (2025):** The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking. *Humanities and Social Sciences Communications* 12, Artikel 621. doi.org/10.1057/s41599-025-04787-y. Der Artikel wurde inzwischen zurückgezogen.

Wissenschaftsrat (2012): Prüfungsnoten an Hochschulen im Prüfungsjahr 2010. Arbeitsbericht mit einem wissenschaftspolitischen Kommentar, Drs. 2627-12, Köln. Kein DOI; Daten des Statistischen Bundesamtes.

Xuan, Q., Cheung, A. & Sun, D. (2022): The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms. *Frontiers in Psychology* 13, 990196. doi.org/10.3389/fpsyg.2022.990196



Über OneTutor

OneTutor ist ein KI-Tutor für die Hochschullehre: dozentenorientiert entworfen, an die Veranstaltung und ihre kuratierten Materialien gebunden, mit Kohortenganalytik für Lehrende und wissenschaftlicher Begleitforschung von Anfang an. Gehostet in Deutschland, Privacy by Design.

Mehr unter onetutor.ai

OneTutor GmbH | Whitepaper, Stand August 2026.

Zitiervorschlag: OneTutor GmbH (2026): Lernen an der Hochschule im Wettbewerb mit generativer KI. Warum KI in der Lehre eine Gestaltungsaufgabe ist. München.