OneTutor

Outsourcing Learning to Generative AI?

Why AI in Higher Education Is a Design Challenge

ONETUTOR PERSPECTIVE | SEPTEMBER 2026

Written by Prof. Dr. Alexander Pretschner, Jann Winter, Max Eisner, Nicolas Ebner, and Philipp Csistian

A white paper for university leaders, from presidents to deans of studies, and for everyone who designs teaching in the age of generative AI.

What This Paper Is About

AUDIENCE

This paper is addressed to universities. However, its underlying principles (where digital tools should support learning and where they should deliberately hold back) are not limited to universities. They apply to every education provider and every organization with an interest in continuing education that takes teaching and certification seriously.

Contents

Summary

Students have long been using AI tools, with or without their university's involvement. This not only changes the role universities still play in delivering content; it also raises a question that is existential for them: must education and certification always go together? This paper argues that each university's own competitive response to AI is an essential part of how universities answer this question. It is not one option among several but an existential design task. A four-layer model (phases, motivation, types of effort and cross-cutting factors) shows where this response should start.

What often gets lost in the AI debate is that "assessment" refers to two different things. The summative function, the assessment outcome with legal consequences, is the foundation that AI erodes. The formative function, feedback that is useful to no one but the learner and the instructors, is by contrast the one assessment function that AI does not threaten but, for the first time, makes possible across the board. Formative assessment was never a pedagogical problem but always a problem of scale. And that is the constraint that falls away.

There are no shortcuts to learning. The bottleneck is not the time students spend learning but how they use it. A tool alone does not change that, and it can even harm learning if it replaces effort instead of prompting it. Only universities have the leverage to use AI so that all students learn more effectively: curated material, control over assessment, and the learning behavior of an entire cohort. OneTutor is our attempt to put this leverage to use. The question is whether students will choose their own university's material.

CHAPTER 1

The Universities' Loss of Control

> **For more than twenty years, students have increasingly been learning from material produced by external providers. AI chatbots did not start this development; they are its latest step so far. The exodus from campus is a visible consequence.**

For some twenty years, universities have been losing authority over where students actually learn, and from what material. Lecture recordings make the lecture hall seem optional. Lecture notes and past exams reconstructed from memory circulate between cohorts and even between universities. In students' eyes, YouTube educational videos, MOOCs and Khan Academy often explain the same topic more clearly than their own lecture does. ChatGPT and Gemini Notebook now offer personalized explanations at any time. 88% of young people in education have already used generative AI (Stürz et al. 2025).

From the students' perspective, this is not a workaround. At least as they see it, alternative materials are often on a par with the university's own course materials in content and pedagogy, especially in undergraduate courses. For twenty years, digital technologies have made it possible to decouple what used to be an essential part of the learning process, the lecture, from the physical university. Generative AI now allows students to engage interactively with materials from their own university and from other providers, independently of traditional teaching formats, time and place.

One clear sign of this loss of control is the growing trend of students staying away from campus: attendance is declining, sometimes despite a formal attendance requirement. This may reflect the everyday reality of a new generation of students. But it is certainly also a rational response to the fact that, compared with digital alternatives, attending in person no longer clearly looks like the best way to understand the material. Learning interactions increasingly take place neither on campus nor in the course's established digital forums. Students increasingly use their own tools. More and more, substantive discussion happens in isolation rather than collectively.

At the same time, a meta-analysis of 69 studies covering some 21,000 students identifies attendance as the strongest known single predictor of academic performance: stronger than standardized admission tests, school grades or study strategies, and largely independent of conscientiousness and motivation (Credé et al. 2010). In a synthesis of more than a hundred factors related to academic success, attendance ranks fairly high, though not unrivaled at the top (Schneider & Preckel 2017). From the students' perspective, attending less may be a rational gain in efficiency. And although the relationship is correlational and does not prove causality, learning may suffer when students stay away from campus.

One plausible partial explanation for this attendance effect is that attending in person creates a regular, consequence-free rhythm of opportunities to check one's understanding: the question asked in passing, the show of hands, a glance at the notes of the person sitting next to you, the conversation over lunch. If so, formative assessment (Section 2.2) may be the most promising way to partly reproduce this rhythm off campus.

To be explicit, we do not conclude from this that lectures are simply the wrong format (too long, too one-way, not personalized, not interactive), or that formats such as project work, seminars or flipped classrooms could simply replace them. That depends on the subject, on the instructors' personalities, and on the students' motivations, abilities, preferences and cultural backgrounds.

CHAPTER 2

The Dual Mission and Its Quiet Decoupling

> **Universities couple education and assessment within a single institution. In education as a whole, this is the exception, not the rule. AI threatens this coupling from several directions. Formative assessment, which can scale for the first time, turns the threat into an opportunity.**

Alongside research, universities traditionally fulfill two tasks at once: they build knowledge and competence, and they certify that this competence has been acquired. In the public mind, the two are inseparable. A university degree stands for both.

2.1 Decoupling as a Real Vulnerability

In other areas of education, this coupling is absent or weaker. Examinations by Chambers of Industry and Commerce certify regardless of where and how candidates acquired the necessary knowledge. A Scrum Master or IT security certificate assesses an outcome, not the path to it; candidates can prepare with a book, a course, YouTube or a chatbot.

For universities, decoupling is not a hypothetical scenario but a potentially existential vulnerability. In law at German universities, commercial exam-preparation courses run by private providers (Repetitorien) have long been a clearly visible form of decoupling. If instruction is now potentially available everywhere, and often just as good in every subject, then what structurally remains of the university is above all certification, at least wherever the added value of on-site education cannot be convincingly justified. Certifying bodies compete differently from full degree programs: they are cheaper, faster and more specialized. Calling the university's role into question in this way is the classic setting for a possible disintermediation, as observed in digital transformation projects of every kind.

We are not alone in believing that AI fundamentally calls the certification function into question in any case. In a covert field test at a British university, submissions written entirely by AI were fed into the live examination system in five regular psychology modules. 94% went undetected, and on average they received grades half a grade boundary higher than those of real students (Scarfe et al. 2024). So for unproctored assessment formats, falling back on certification means relying on a foundation that is already shaky. AI in education is therefore not a marginal issue; it affects both sides of the university's mission.

A second factor contributing to this erosion has nothing to do with AI: it is unclear whether degrees actually reflect what they are meant to signify. A degree works as a signal to the outside world because it is assumed to reflect ability. The validity of this indicator is debated in the research literature; findings on the relationship between academic performance and later career success range from "useful but limited" to "significantly overestimated" and have shifted over the decades (Roth et al. 1996; Van Iddekinge et al. 2024). In addition, the German Science and Humanities Council notes critically that differences between subjects and institutions weaken the informative value of individual grades, a problem compounded by a persistent trend toward better grades (Wissenschaftsrat 2012, based on the 2010 examination year). A signal that does not differentiate does not inform.

Contested predictive power does not necessarily mean a worthless degree. How much a degree actually tells us, however, is an open empirical question. Against this background, the finding by Scarfe et al. is not a new threat but the sharpening of an old ambiguity. That is the real vulnerability: not that the degree is worthless, but that its informative value is, not least, an assumption.

2.2 Two Kinds of Assessment: Formative Assessment as an Opportunity

The debate about AI and assessment suffers from the fact that "assessment" refers to two fundamentally different things: summative and formative assessment. Summative assessment faces outward. Formative assessment, the continual retrieval of knowledge as part of learning, faces inward and serves to make learning more effective. Generative AI affects these two functions in exactly opposite ways: it calls summative assessment into question and makes formative assessment possible. The reason is obvious. Formative assessment was probably never as much a question of pedagogy as a question of scale. An instructor with three hundred students could never give frequent, timely feedback. The limit was availability, and that is the limit that falls away.

TWO FUNCTIONS OF ASSESSMENT AND THE EFFECT OF GENERATIVE AI

	Summative assessment	Formative assessment
Purpose	To determine and certify	To determine in order to guide
Result	Grade, credits, degree, with legal consequences	Feedback to the learner and to the instructors
Consequences outside the learning process	Considerable and legally binding	None, which is probably a precondition for its effectiveness
Historical limit	Proctoring and proof of authorship	Instructors' working time, i.e., purely a matter of scale
Effect of generative AI	Erodes. Wherever assessment is unproctored, certification loses its basis.	Becomes possible for the first time. The scaling limit disappears.

Since the review by Black & Wiliam (1998), it has been considered established that frequent, timely feedback produces learning gains. How large these gains are, however, is disputed. Later studies find considerably smaller effects and criticize the quality of the underlying research (Kingston & Nash 2011; Bennett 2011; Xuan et al. 2022), while others in turn challenge these reanalyses on methodological grounds (Briggs et al. 2012). So there is robust evidence that formative assessment works, but no robust figure for how well. This is not peculiar to this topic but the normal state of much of education research, and it is why each university needs its own evaluation research (Chapter 7).

One aspect of this debate is particularly noteworthy to us: among the forms of implementation that Kingston & Nash distinguish, computer-based formative systems performed above average. However, this sub-analysis rests on very few studies, and other work places computer-assisted feedback at the lower end. The finding cannot serve as proof, but we can take it seriously as an indication.

The question is therefore no longer only how much authority over assessment a university can defend against AI, but also which assessment function it can exercise for the first time with AI. Thinking about AI in education only defensively means overlooking the area where the benefits

are relatively clear, and where the university has a structural advantage. Formative feedback is valuable when it relates to what is actually taught and assessed, that is, to curated material and to the course's learning objectives. A general-purpose assistant can process this material as soon as someone uploads it, and students have long been doing so. What it lacks is the decision about which parts count, and the cohort that reveals where students get stuck. The university's advantage therefore lies not in access to the material but in curation and aggregation.

Feedback is not automatically effective, though. Shute (2008) shows that formative feedback must be non-evaluative, supportive, timely and specific, and that its effect interacts with characteristics of the learner and the task. Poorly designed feedback can also make learning worse. The opportunity therefore lies not in the quantity of feedback but in its design, and it depends on a condition that is easily overlooked: careful thought must be given to whether the feedback carries consequences, for example in the form of bonus points.

2.3 Looking Back: Lessons from the MOOC Wave

Around 2012, MOOCs were expected to reshape higher education. Instead, the field consolidated around a much older business model: outsourced online master's programs for working professionals. The vast majority of learners did not return after their first year, growth was concentrated almost entirely in the wealthiest countries, and the low completion rates did not improve over six years (Reich & Ruipérez-Valiente 2019). Germany also has considerable regulatory protection: academic degrees are protected by law, and degree programs are accredited. A private certifying body does not simply take their place.

Predictions of decoupling therefore need to explain why this time should be different. MOOCs replaced the lecture with another fixed, one-way artifact, a video that is the same for everyone. But unidirectional videos are weaker in exactly the dimension that matters: the follow-up question, individualization, the response to a specific error in reasoning. Generative AI replaces not the artifact but the interaction. That does not make decoupling inevitable, but it does make it considerably more plausible than in 2012.

CHAPTER 3

The Design Task

> **AI in degree programs is a typical digital transformation project. The mere existence of a tool makes no process more effective or efficient. What matters is solely how it improves learning.**

One can take the view, as during the MOOC hype of 2012, that the university as an institution is simply obsolete. In doing so, it is worth looking at what is at stake. Digital, individualized learning can lead to isolation. Students who learn alone with a video or a chatbot do not learn with and from their peers. They miss the question from the person next to them, the explanation they give in conversation, the look at someone else's flawed solution, which often teaches more than their own correct answer. Across 225 studies in STEM fields, collaborative, active engagement with the material has been shown to be measurably more effective than passive listening (Freeman et al. 2014). Most university graduates will say they learned at least as much from their peers as from their instructors. (Whether these categories apply to newer generations with different patterns of communication and social behavior in the same way, is a different question.)

And it is not only about content. Part of what students take away from university is personal development in the classical sense: a stance of one's own, the capacity for debate, calibrating one's judgment against others, contact with people who challenge them rather than merely confirm them. This is hard to measure and easily overlooked if education is seen purely as knowledge transfer.

For a university that (a) does not want its own role fundamentally called into question, (b) assumes that, depending on the subject, something essential is lost in fully digital, unguided learning, and (c) recognizes that students use AI anyway, the consequence is a design task: to offer its own appropriate use of AI within its own teaching.

"Appropriate" here explicitly means competitive with ChatGPT, Gemini Notebook or similar technology. The advantages a university is up against are clear: availability around the clock, no trip to campus, no waiting for Thursday's office hour, no question too trivial for an email reply. A university that offers nothing here meets these advantages with nothing, and loses its teaching role not to a better pedagogical concept but simply to availability.

This inevitably calls for managing expectations. AI, too, will not automatically turn a university into a place where suddenly more learning, or more committed learning, happens. As with any digital transformation project, simply digitizing a process usually changes nothing at all. The evaluation research on our own tool, OneTutor, confirms this. On average, no differences in self-reported learning outcomes were found between frequent users, infrequent users and non-users (Stumpf et al. 2026; see Chapter 7). Conversely, those who do use the tool are very satisfied with it.

What a good tool can realistically achieve at a university is that learning which takes place online anyway happens on the university's own infrastructure rather than with an interchangeable competitor, without losing what only the university can offer: curation, meaning the expert decision about what is relevant and in what depth; the cohort, in which many students work on the same material; and control over assessment, which determines in the first place what feedback can meaningfully relate to.

THREE PATHS OPEN TO A UNIVERSITY

Path	What it solves	What remains open
1. Offer nothing	Nothing. No effort, no costs, no governance questions.	The teaching role is lost to availability. Doing nothing is itself a decision.
2. Campus license for a general-purpose assistant	The availability problem, fully. Quick to decide, supposedly cheaper per student.	Students choose the material. The cohort's questions stay locked in individual accounts. The substitution path stays open. The university never learns what it has set in motion.
3. Tool tied to the course	Availability, curation, cohort signal and measurability, all at once.	The most demanding path. It requires decisions by instructors and governance rules for formative data.

CHAPTER 4

Two Design Principles

> **Learning knows no shortcuts. A tool may therefore increase efficiency only to a limited extent; its task is to make learning more effective. And only universities have the decisive leverage for this: the learning behavior of an entire cohort, which can be used to continuously improve teaching and learning.**

4.1 No Shortcuts

The time an individual needs to learn cannot be substantially reduced, because learning is time-intensive engagement with the subject matter. It is not a by-product that could be shortcut without defeating its purpose. What can be shortened is everything surrounding this engagement, which can be called friction: searching for the right passage in the lecture notes, waiting for an answer, being stuck with no one to ask, not being alerted that one has misunderstood something, or despairing over learning materials that are simply flawed.

This distinction is not new. The psychology of learning describes the counterpart of friction as desirable difficulties: conditions that impair short-term performance and, precisely because of that, improve long-term retention and transfer (Bjork & Bjork 2011). The point is not whether any learning happens under more difficult conditions, but that, among several ways of learning, the one that feels more laborious is usually also the more effective one. The only new element here is applying this as a design principle for a tool rather than as advice to learners.

Formative feedback is the model case on the permitted side of this rule. The delay between a student's mistake and the instructor's feedback is friction, not effort. A student who solves a problem incorrectly on Sunday evening and finds out on Thursday in the tutorial has spent four days consolidating a wrong model. This waiting time contributes nothing to learning; it does harm. The attempt to retrieve what has been learned, by contrast, stays with the learner. That is the effort that cannot be delegated.

If there are no shortcuts to learning, what is a tool for at all? Reducing friction is the obvious answer, but it is secondary, because friction is probably not the biggest obstacle to learning. The real purpose lies elsewhere: to use the available study time in a more targeted and more effective way. More targeted means that engagement starts where understanding actually breaks down. More effective means that the better activity takes place in the same amount of time.

The amount of time invested in studying is a poor indicator of actual performance (Plant et al. 2005). What truly matters is how that time is spent. When asked about their primary study strategy, the vast majority of students cite rereading, while only a small fraction mentions self-testing (Karpicke et al. 2009). Furthermore, studying is typically driven by deadlines rather than proactive planning; in other words, it is massed rather than spaced (Hartwig & Dunlosky 2012). The current state of research is remarkably clear: highlighting, rereading, and summarizing are among the least effective techniques, whereas practice testing and spaced practice are highly effective (Dunlosky et al. 2013; Roediger & Karpicke 2006; Cepeda et al. 2006). The issue is not a lack of study time, but rather how poorly that time is structured.

That this persists is no criticism of students. Fluent rereading feels like understanding; a failed retrieval attempt feels like failure. Learners therefore systematically choose the method that feels better and works worse. It is the same inversion of perceived and actual learning that Deslauriers

et al. (2019) measured for teaching formats. What is missing, then, is neither time nor insight, but something that prompts the retrieval attempt and reports back the result while there is still time to change course.

4.2 The Underrated Value of the Cohort

The cohort in which a student learns matters for at least two reasons. The first is social and was discussed in Chapter 3. The second is, for the university, the economically more important one: when many students work with the same material, their engagement can be used to improve it. Where a question comes up repeatedly in the chat, where a quiz question is systematically answered incorrectly, or where a chapter of the lecture notes prompts an unusual number of questions, a previously invisible gap becomes apparent. Materials can then be supplemented on the basis of evidence from an entire cohort. This is the side of formative assessment from which instructors also benefit, and through them, indirectly, the learners.

Cohort-wide, non-individual learning analytics of this kind provide value that ChatGPT or Gemini Notebook, used individually, structurally cannot generate. There, each conversation remains isolated and benefits no one but the individual user. A tool tied to an institution can aggregate these signals and feed them back into better curation. The AI provides the signal; humans decide what to do with it. Aggregation is also the best reason for keeping curation in human hands. And the same body of data is what makes evaluation research beyond the individual possible in the first place.

The value of these signals depends, however, on the data being collected without consequences. A quiz that counts toward the grade no longer measures understanding alone and can also encourage students to optimize for the grade. A chat question that students suspect might one day end up on record may not be asked. Ideally, formative instruments remain formative. Whether they are then used at all is another question, and bonus points can help motivation. This trade-off has to be decided case by case.

CHAPTER 5

A Layered Model

> **Four layers separate what the AI debate usually conflates: when and whether learning happens at all, what learning time consists of, and what cuts across everything. This separation shows where a tool can have an effect and where it cannot. Without it, every expectation of AI remains untestable.**

Solving the design task from Chapter 3 requires a vocabulary that separates four things that are usually conflated: when learning happens, whether it happens at all, what learning time consists of, and what cannot be assigned to any single phase.

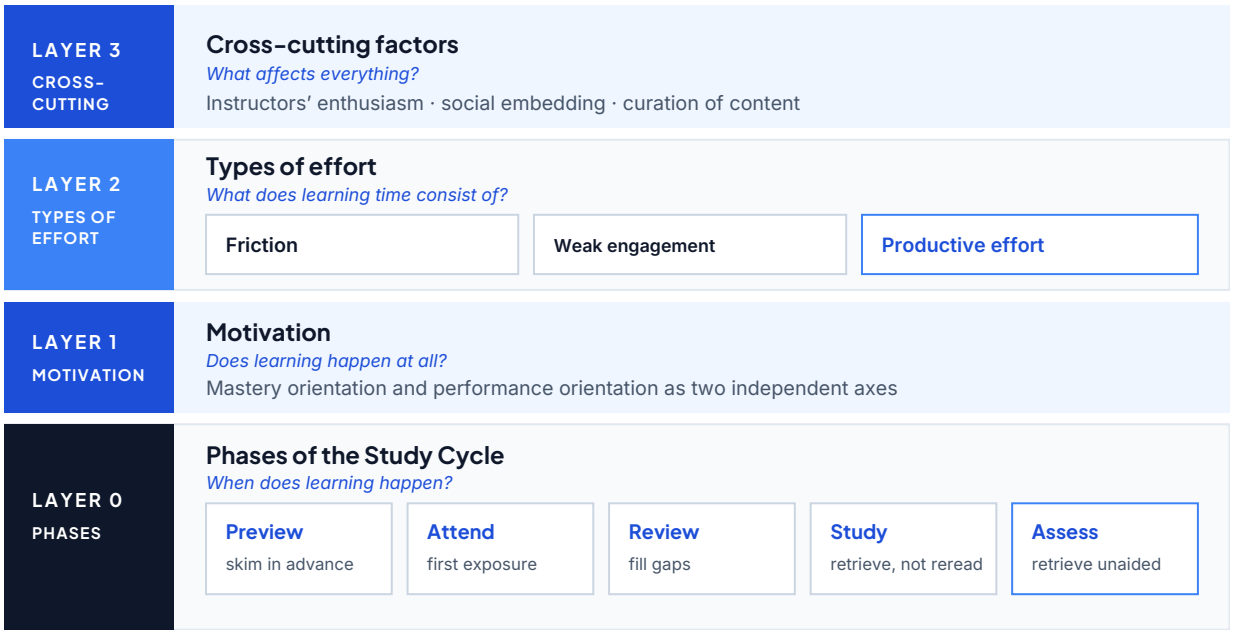

Figure 1 The layered model. Layer 0 follows the Study Cycle of the UNC Learning Center, which builds on Christ's (1997) PLRS system. A tool can intervene meaningfully only at Layer 2, where it converts weak engagement into productive effort.

Layer 0: Phases of the Study Cycle

The temporal framework is the Study Cycle as used by the UNC Learning Center, which has five phases and builds on Christ's (1997) PLRS system: Preview (skimming old and new material before the session), Attend (encountering the material for the first time), Review (organizing notes, answering quizzes and filling gaps within about a day), Study (practicing, explaining things to oneself, retrieving instead of rereading) and Assess (retrieving what has been learned without help and identifying gaps, i.e., formative assessment). The cycle is not linear. Assess leads back to Study, and the same material passes through the phases several times. We chose this model because it is activity-based: every phase involves observable actions, and tools can only be mapped meaningfully onto actions. It is a heuristic for learning strategies, not a validated theory of learning, and we use it as an organizing framework, not as evidence of effect. A tool alone cannot ensure that students follow this cycle appropriately. That requires further pedagogical and structural measures, which are beyond the scope of this paper.

Layer 1: Motivation

Whether a phase takes place at all is decided not by the tool but by the learners' motivation. A focus on understanding is the desire to truly master a subject, even beyond what the exam covers. A focus on performance is the desire to pass as efficiently as possible or to get a good grade. The two are independent axes, not opposite poles; the same person may go through the full cycle in one subject and study only to pass in another. The distinction corresponds to the mastery and performance axes of goal orientation research, whose independence is supported by factor analysis (Elliot & McGregor 2001). Goal orientations are not a purely stable personality trait; they also change within a single semester (Fryer & Elliot 2007).

It follows that a tool can amplify activity that is already taking place but usually cannot initiate it. No tool on its own gets someone to sit down and study. How that quarter of an hour is then used, however, can very much be designed, and that is what Layer 2 is about.

Layer 2: Friction, Weak Engagement, Productive Effort

Every learning phase breaks down into components that differ in how much they contribute to learning. Two are easy to name: accidental friction, which contributes nothing to learning, and productive effort, which is learning itself. But this two-way split is incomplete. Someone reading a chapter for the third time is not searching for a passage in the lecture notes or waiting for an answer. That is not friction but engagement with the subject matter. Yet it is not the kind of effort learning depends on either. Between the two, then, lies a third category, weak engagement: real but barely effective work on the content. It is not a marginal phenomenon but the empirical norm (Karpicke et al. 2009; Dunlosky et al. 2013).

THE DESIGN PRINCIPLE AT LAYER 2

Component of learning time	Example	What a tool may do with it
Friction	Searching for the passage in the lecture notes, waiting until Thursday for an answer	Reduce it freely. The benefit is undisputed, but not exclusive.
Weak engagement	Reading the same chapter for the third time, highlighting, summarizing	Convert it into productive effort. That is the real purpose.
Productive effort	Retrieving without help, putting things in one's own words, finding one's own reasoning errors	Structure it, trigger it, give feedback on it, but never replace it.

A tool that turns productive effort back into weak engagement by supplying the answer instead of demanding it defeats its purpose. And it does so unnoticed, because to the learner the result feels like progress.

This axis alone is not enough, however. It classifies by activity. A second axis is needed: what consequences may arise outside the individual learning process. How much support a tool may give depends on the external consequences of the learning outcome. Where there are none, the tool may help extensively, because students gain nothing by fooling themselves. Where the consequence is a degree or a grade bonus, however, the tool invites misuse. When learning efforts are rewarded externally, this second axis is why formative assessment is both the greatest opportunity and the most sensitive point of the whole model.

Layer 3: Cross-Cutting Factors

Three factors cannot be assigned to any single phase, yet they influence all of them. They are the same ones identified in Chapter 3 as genuinely at risk. Instructors' enthusiasm acts as a multiplier on the mastery axis; it survives a video stream in muted form and a set of lecture notes not at all. Social embedding comes from in-person attendance: study groups, hallway conversations, incidental exposure to how others think. Curation of learning content is the institution's responsibility. Today, conventional AI tools have practically no effect at this layer. Layer 3 is the part of the model that a purely reactive chat tool fundamentally cannot change, and where even a proactive tool has to be complemented by further pedagogical measures.

CHAPTER 6

Application: OneTutor

> **OneTutor is built for students but designed around instructors. Having instructors curate quizzes and guided conversations ensures quality, a close link to the course, and a focus on what matters. The cohort's learning behavior thus becomes the basis for better teaching and for evaluation research.**

OneTutor is an AI tutor built for students and designed around instructors. Instructors upload their course materials and curate quiz questions and AI-guided conversations. This design decision provides insight into the anonymized learning behavior of an entire cohort, above all into which parts were not understood. From this, instructors can develop better lecture materials and new quizzes with AI support.

40,000+
students use OneTutor.

40+
institutions use the tool.

800+
courses run on it.

As of September 2026

Chat: Known Unknowns

When students know what they do not know, they can ask questions, which OneTutor answers based on the course material, in the Preview, Attend, Review and Study phases. In the lecture hall, every question has a social price: admitting in front of three hundred people that you did not understand slide 12 costs something. The Chat, a RAG system based on the course materials and transcripts, removes exactly this price: anonymous, immediate, as often as you like, even in the middle of the lecture. From the students' perspective, ChatGPT or Gemini Notebook are equivalent here; from the university's perspective, the value of the cohort is lost. If searching for an answer in the material counts as friction in the sense described above, that friction is clearly reduced.

Quiz: Unknown Unknowns

Quiz questions curated by instructors pose a question without answering it. The cognitive work of retrieving the answer from memory stays with the learner, a basic principle of formative assessment. One of the most robust findings in the psychology of learning is that this kind of retrieval improves retention more than rereading the same material (Roediger & Karpicke 2006).

How far this effect extends beyond the practiced content, however, depends on how the questions are designed. Transfer occurs above all when the task requires learners to formulate an answer themselves rather than merely recognize it (Pan & Rickard 2018). Both points argue against freely generated questions and for curated ones aligned with the learning objectives.

This matches our experience: AI-generated quiz questions need to be curated by instructors. When an AI generates questions from teaching materials without any constraints, we see questions and answers that are sometimes subtly wrong, sometimes misleading, sometimes fundamentally wrong and sometimes simply irrelevant.

Quizzes are used mainly before exams. But we also see students working through them right before and right after the lecture, in the Preview and Review phases. A quiz before the session then acts as a pretest rather than an overview: the failed attempt to retrieve what was supposedly learned opens exactly the gap that the lecture then fills. This effect is documented. In five experiments, posttest performance after a pretest was better than after extended reading time, even when the analysis included only content that had not been correctly retrieved in the pretest (Richland et al. 2009).

Converse: Learning Through Guided Interaction

A third feature, currently in pilot, reverses the AI tutor's role, mainly in the Study and Assess phases. Instead of waiting for a question, the system leads and moderates the conversation, with an individual or a group, in units of about fifteen minutes, on content specified by the instructors. The prerequisite is that students have largely understood the material. In these conversations they learn to answer precisely and completely, deepening their understanding and identifying gaps along the way. Used in a group, ideally in person, the feature addresses the social embedding from Layer 3 that is lost as in-person attendance declines.

This feature comes closest to Shute's (2008) ideal of formative feedback: non-evaluative, immediate, specific to the answer just given, and repeatable as often as desired. It also requires exactly the kind of self-formulated answers that, according to Pan & Rickard (2018), promote transfer. Until now, this combination has not been achievable in a course with hundreds of participants. Looking ahead, and subject to legal and ethical considerations, it remains to be examined whether such conversations could in future also be used separately for summative assessment, as oral exams that scale without limit.

CHAPTER 7

Why Evaluation Research Is Not Optional

> **In what sense AI tutors are "useful" is a complex question. Those who do not ask it cannot know under what conditions AI really helps students and where it harms them, and are ultimately acting irresponsibly.**

At present, no one knows reliably how well AI tutors work in university teaching, or even what that means. By far most of what is said about AI in education today consists of plausibility arguments: theoretically grounded, internally coherent, empirically open. The mechanisms described in Chapter 6 are well founded too, but generally not yet proven for this product in this setting.

Even the question of whether an AI support system "works" has many layers. Correlations between grades and tool use can be misleading, because learners differ in type and motivation. Subjective ratings of usefulness from questionnaires must also be treated with caution: to students, reducing friction and replacing effort feel the same. In a randomized comparison, students in active learning settings demonstrably learned more but felt they were learning less. The authors conclude that teaching evaluations based on student perception can unintentionally reward the inferior, passive method (Deslauriers et al. 2019).

These shortcomings must not, however, become an excuse for not evaluating experience with AI tutors objectively. That is why we have had the use of OneTutor studied from the very beginning. The Bavarian Research Institute for Digital Transformation (bidt) of the Bavarian Academy of Sciences and Humanities is investigating its use at ten Bavarian universities in the project "AIffect-

iveness". The first study covers the 2025 summer semester: 84 courses, surveys of up to about 1,000 students and 31 instructors, and usage data from the system (Stumpf et al. 2026).

80%
of the students surveyed used the tutor actively.

90%
of the instructors see it as a useful addition.

84%
of the instructors would use it again.

Those who use the tool also value it. However, no differences in self-reported learning outcomes were found between frequent users, infrequent users and non-users. Non-users evidently achieve comparable results through other routes. The provisional conclusion is that OneTutor is broadly accepted, used voluntarily, and rated as helpful by instructors and users alike. What cannot be concluded is that all students learn more or better with it than without it. The claim, however, is not that students learn more, but that the composition of their learning time shifts. If that claim holds, then according to Deslauriers et al. (2019), self-reports would be more likely to underestimate the effect than to confirm it, because the more effective activity feels worse.

Relevant studies are no longer rare; reliable ones still are. Since 2024, a whole series of meta-analyses on the learning effects of generative AI has appeared, reporting mostly positive and sometimes large effects (e.g., Liu et al. 2025). Their value is limited: one widely cited paper has since been retracted (Wang & Fan 2025), and independent reanalyses of the same data show that the reported effects shrink sharply once corrected for publication bias. Across the field, most published work relies on self-reports and short-term measurements. Reliable data on transfer and retention beyond a single semester are almost entirely lacking.

That is the argument for conducting one's own evaluation research, and a stronger one than a mere gap in the literature would be. Relying on the published record means relying on effect sizes that range from "large" to "indistinguishable from zero", depending on who redoes the calculation. A university that wants to know what a tool achieves in its own context, in its own courses, with its own students and in its own subjects cannot get that answer from the literature. It has to gather that evidence itself.

At the same time, the few methodologically strong individual studies show how limited their reach is. In a randomized crossover comparison, Kestin et al. (2025) found more than twice the learning gain of an active learning session, but over two lessons, in one course at one institution, and with a purpose-built tutor developed by the instructor who also taught the comparison session; learning was measured immediately afterwards, not weeks later. That is a strong signal for well-built tools, not evidence for AI tutors in general. In a randomized study with nearly a thousand high school mathematics students, Bastani et al. (2024) show that plain ChatGPT without pedagogical guardrails leads to markedly worse exam results than learning without ChatGPT or with a version of ChatGPT that has guardrails. The pattern is notable: during practice these students performed better, in the exam worse.

None of this is an argument for holding back on evaluation research, because doing nothing is itself a decision with consequences. Three reasons make the obligation to measure compelling. The intervention affects the learning of an entire cohort. Its effects can be negative, and can be so without anyone noticing (Bastani et al. 2024). And the perception of those affected is systematically unreliable as a check (Deslauriers et al. 2019). Rolling out AI tutors without evaluation research means never finding out whether they help or harm, with no way of reconstructing it later.

CHAPTER 8

Six Questions for Evaluating a Tool

Whether students learn with AI is no longer an open question. What is open is whose AI they use, and what remains of their own university in the process: the cohort whose members learn from one another, the curation that grows out of its questions, and a degree that is ultimately worth something.

The common reading of this situation is defensive: AI threatens the certifying exam, that is, summative assessment. This reading is incomplete. The same technology makes the other half of assessment, the formative half, broadly available for the first time. That is an opportunity universities have and that an individually used chatbot structurally lacks, because it has no curated material, no learning objectives and no cohort. Seizing it takes two decisions: to deploy formative instruments, and to measure whether they work. Six questions help identify a tool that qualifies for this at all.

1. Does the tool draw on the curated material and learning objectives of the course, and is that selection made by the instructors rather than by the learners?
2. Does its use generate a signal about the cohort that feeds back into better curation, instead of remaining in individual student accounts?
3. Is the more effective activity the path of least resistance, rather than a recommended option alongside it?
4. Is the easy substitution path structurally blocked within the tool, rather than merely not recommended?
5. Is it intended that the formative data be structurally unusable for grading, and has this been communicated to students?
6. Does the university actually hold the data from which it could later determine what the tool has achieved?

A general-purpose chat assistant cannot answer four of these six questions with “yes”, not because of the quality of its answers but because of how it is built: it is tied to no course, knows no cohort, keeps every path open and leaves no analyzable traces at the institution.

There are no shortcuts to learning. Tools alone can therefore even harm learning. Only universities have the leverage to use AI in a way that lets all students learn more effectively. OneTutor can support that.

References

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 No. 4895486. doi.org/10.2139/ssrn.4895486. Preprint, not peer-reviewed.

Bennett, R. E. (2011): Formative Assessment: A Critical Review. Assessment in Education: Principles, Policy & Practice 18(1), pp. 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. (eds.): Psychology and the Real World. Worth Publishers, New York, pp. 56–64. Page numbers follow the first edition; the 2nd edition (2014) paginates pp. 59–68.

Black, P. & Wiliam, D. (1998): Assessment and Classroom Learning. Assessment in Education: Principles, Policy & Practice 5(1), pp. 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), pp. 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), pp. 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. Primary source of the PLRS system.

Credé, M., Roch, S. G. & Kieszczynka, U. M. (2010): Class Attendance in College. Review of Educational Research 80(2), pp. 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), pp. 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), pp. 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), pp. 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), pp. 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), pp. 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), pp. 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), pp. 471–479. doi.org/10.1080/09658210802647009

Kestin, G., Miller, K., Klales, A., Milbourne, T. & Ponti, G. (2025): AI tutoring outperforms in-class active learning. Scientific Reports 15, Article 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), pp. 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), pp. 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), pp. 96–116. doi.org/10.1016/j.cedpsych.2004.06.001

Reich, J. & Ruipérez-Valiente, J. A. (2019): The MOOC pivot. Science 363(6423), pp. 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), pp. 243–257. doi.org/10.1037/a0016496

Roediger, H. L. & Karpicke, J. D. (2006): Test-Enhanced Learning. Psychological Science 17(3), pp. 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), pp. 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), pp. 565–600. doi.org/10.1037/bul0000098

Shute, V. J. (2008): Focus on Formative Feedback. Review of Educational Research 78(1), pp. 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 No. 20, Munich, March 24, 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 No. 17, Munich. doi.org/10.35067/xypq-kn75

UNC Learning Center (n.d.): The Study Cycle. University of North Carolina at Chapel Hill. learningcenter.unc.edu/tips-and-tools/the-study-cycle. Adapted from Christ's PLRS system by the LSU Center for Academic Success, presented in 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), pp. 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, Article 621. doi.org/10.1057/s41599-025-04787-y. The article has since been retracted.

Wissenschaftsrat (2012): Prüfungsnoten an Hochschulen im Prüfungsjahr 2010. Arbeitsbericht mit einem wissenschaftspolitischen Kommentar, Drs. 2627-12, Cologne. No DOI; data from the Federal Statistical Office.

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

OneTutor GmbH | White paper, as of September 2026.

Suggested citation: OneTutor GmbH (2026): Outsourcing Learning to Generative AI? Why AI in Higher Education Is a Design Challenge. Munich.