

# WhiteExam

## Eine didaktisch kontrollierte Dual-Loop-Architektur für KI-gestütztes Lernen und kompetenzorientiertes digitales Prüfen Design Paper und Forschungsagenda

Reinhard Weiß  
Gymnasium Kirschgarten Basel

Referenzfassung, Juli 2026

### Zusammenfassung

Der Beitrag entwickelt WHITEEXAM als wissenschaftlich begründete Referenzarchitektur für eine integrierte Lern- und Prüfungsplattform. Ausgangspunkt ist ein Strukturproblem digitaler Bildung: Lernmanagement, generative KI, Aufgabenbanken und digitale Prüfungssysteme werden häufig als getrennte Werkzeuge betrieben. Dadurch bleiben Lernziele, Lernaktivitäten, Kompetenznachweise, Rückmeldung und Qualitätsentwicklung nur lose gekoppelt. Zugleich gefährdet eine undifferenzierte Integration generativer KI die Eigenaktivität im Lernen und die Validität summativer Leistungsentscheidungen.

Das Design Paper rekonstruiert den implementierten Systemstand von WHITEEXAM anhand der Codebasis, des relationalen Datenmodells, der Benutzeroberflächen und der Entwicklungsdokumentation. Es verbindet diese Artefaktanalyse mit Constructive Alignment, Design-Based Research, Forschung zu intelligenten Tutorsystemen, formativer Rückmeldung, Learning Analytics und Assessment-Validität. Als zentraler Entwurfsbeitrag wird eine *didaktisch kontrollierte Dual-Loop-Architektur* formuliert: Eine formative Lernschleife kombiniert lehrpersonenkuratierte Lernpfade, klassische Aufgaben und begrenzte KI-Dialoge; eine davon technisch, normativ und datenseitig getrennte Prüfungsschleife erzeugt belastbare Kompetenznachweise. Verbunden werden beide Schleifen durch gemeinsame fachliche Artefakte, Kompetenzziele und kontrollierte Rückkopplung, nicht durch eine Vermischung von Tutor- und Prüfungslogik.

Der Beitrag beschreibt zehn Designanforderungen, sechs Gestaltungsprinzipien, die System- und Datenarchitektur, das Rollenmodell, den Aufgabenlebenszyklus sowie Governance- und Datenschutzanforderungen. Implementierte Funktionen werden konsequent von Designzielen und Forschungsoptionen unterschieden. Es werden keine Wirksamkeitsansprüche erhoben. Stattdessen resultieren prüfbare Designhypothesen und eine mehrstufige Forschungsagenda für spätere Studien zu KI-Dialogen, Prüfungsvalidität, Learning Analytics und fachlicher Übertragbarkeit. Damit dient der Text als Grundsatzartikel, auf den spezialisierte empirische Veröffentlichungen aufbauen können.

**Schlüsselwörter:** KI in der Bildung; digitale Prüfungen; KI-Tutor; Constructive Alignment; Design-Based Research; Learning Analytics; Assessment-Validität; Human-in-the-Loop

# Inhaltsverzeichnis

|          |                                                                   |           |
|----------|-------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Einleitung</b>                                                 | <b>4</b>  |
| <b>2</b> | <b>Methodischer Status und Materialbasis</b>                      | <b>5</b>  |
| 2.1      | Design Paper im Horizont von Design-Based Research . . . . .      | 5         |
| 2.2      | Analysierte Artefakte . . . . .                                   | 5         |
| 2.3      | Evidenz- und Statuslogik . . . . .                                | 5         |
| <b>3</b> | <b>Stand der Forschung</b>                                        | <b>6</b>  |
| 3.1      | Von intelligenten Tutorsystemen zu generativer KI . . . . .       | 6         |
| 3.2      | Formative Rückmeldung und Selbstregulation . . . . .              | 6         |
| 3.3      | Assessment, Kompetenz und Validität . . . . .                     | 6         |
| 3.4      | Learning Analytics als pädagogischer Regelkreis . . . . .         | 7         |
| 3.5      | Forschungslücke . . . . .                                         | 7         |
| <b>4</b> | <b>Anforderungen an Plattformen im KI-Zeitalter</b>               | <b>7</b>  |
| <b>5</b> | <b>Zentraler Entwurf: die Dual-Loop-Architektur</b>               | <b>8</b>  |
| 5.1      | Trennung ohne Entkopplung . . . . .                               | 8         |
| 5.2      | Sechs Gestaltungsprinzipien . . . . .                             | 9         |
| <b>6</b> | <b>Systemarchitektur</b>                                          | <b>9</b>  |
| 6.1      | Architekturstil . . . . .                                         | 9         |
| 6.2      | Domänenmodell . . . . .                                           | 9         |
| 6.3      | Rollenmodell . . . . .                                            | 11        |
| <b>7</b> | <b>Die formative Lernschleife</b>                                 | <b>13</b> |
| 7.1      | Lernumgebungen als kuratierte Pfade . . . . .                     | 13        |
| 7.2      | Klassische Aufgaben und KI-Dialoge . . . . .                      | 13        |
| 7.3      | Dialogischer KI-Tutor . . . . .                                   | 16        |
| 7.4      | Learning Analytics und adaptive Lernpfade . . . . .               | 16        |
| <b>8</b> | <b>Die summative Prüfungsschleife</b>                             | <b>17</b> |
| 8.1      | Prüfung als Evidenzarchitektur . . . . .                          | 17        |
| 8.2      | Kompetenzorientierte Prüfungsarchitektur . . . . .                | 17        |
| 8.3      | Durchführung und Sicherheit . . . . .                             | 19        |
| 8.4      | KI im Prüfungslebenszyklus . . . . .                              | 19        |
| <b>9</b> | <b>Aufgabenlebenszyklus und Materialimport</b>                    | <b>19</b> |
| 9.1      | Von vorhandenen Materialien zu versionierten Artefakten . . . . . | 19        |
| 9.2      | Sammlungen und fachliche Erweiterbarkeit . . . . .                | 20        |

|                                                             |           |
|-------------------------------------------------------------|-----------|
| <b>10 Kompetenzmodell und Constructive Alignment</b>        | <b>20</b> |
| 10.1 Konzeptionelles Kompetenzmodell . . . . .              | 20        |
| 10.2 Alignment als prüfbare Relation . . . . .              | 22        |
| <b>11 Datenschutz, KI-Governance und Verantwortlichkeit</b> | <b>22</b> |
| 11.1 Datenfluss . . . . .                                   | 22        |
| 11.2 Governance-Anforderungen . . . . .                     | 23        |
| <b>12 Design-Based-Research-Agenda</b>                      | <b>24</b> |
| 12.1 Von Designthesen zu prüfbaren Hypothesen . . . . .     | 24        |
| 12.2 Vorgeschlagene Forschungsphasen . . . . .              | 25        |
| 12.3 Messmodell für Learning Analytics . . . . .            | 26        |
| <b>13 Diskussion</b>                                        | <b>26</b> |
| 13.1 Wissenschaftlicher Beitrag . . . . .                   | 26        |
| 13.2 Spannungsfelder . . . . .                              | 26        |
| 13.3 Grenzen . . . . .                                      | 27        |
| <b>14 Ausblick</b>                                          | <b>27</b> |
| <b>15 Fazit</b>                                             | <b>28</b> |
| <b>A Abbildungsinventar und Reproduzierbarkeit</b>          | <b>30</b> |
| <b>B Traceability-Matrix der Designprinzipien</b>           | <b>30</b> |

# 1 Einleitung

Generative Künstliche Intelligenz verändert die Bedingungen, unter denen Lernen unterstützt und Leistung nachgewiesen wird. Lernende können Erklärungen, Beispiele, Programmcode oder Lösungsvorschläge in Echtzeit erzeugen lassen. Lehrpersonen können Materialien variieren, Rückmeldungen entwerfen und große Antwortmengen strukturieren. Damit wächst jedoch nicht automatisch die Qualität des Lernens oder Prüfens. Der zentrale Gestaltungsgegenstand ist nicht das Sprachmodell, sondern die pädagogische, technische und institutionelle Architektur, in die es eingebettet wird.

Diese Verschiebung ist für digitale Bildungsplattformen folgenreich. Klassische Lernmanagementsysteme organisieren Inhalte und Abgaben; Aufgabensysteme verwalten Items; Prüfungsbrowser sichern Endgeräte; Chatbots erzeugen Dialoge; Analysewerkzeuge verdichten Nutzungsdaten. Wenn diese Komponenten lediglich nebeneinanderstehen, entstehen Medienbrüche, widersprüchliche Rollenmodelle und unklare Verantwortlichkeiten. Vor allem bleibt offen, wie formative Unterstützung und summativer Nachweis auf dieselben Kompetenzziele bezogen werden können, ohne ihre unterschiedlichen Funktionen zu vermischen.

Die aktuelle Forschung liefert hierfür keine einfache Lösung. Intelligente Tutorsysteme können Lernen wirksam unterstützen (Kulik & Fletcher, 2016; VanLehn, 2011). Für generative KI zeigen neuere Studien sowohl bedeutsame Lerngewinne unter eng gestalteten Bedingungen (Kestin et al., 2025) als auch negative Transfereffekte bei unbeschränkter Hilfestellung (Bastani et al., 2025). Die Differenz verweist auf eine Designfrage: Entscheidend ist, ob ein System Antworten liefert oder Denkprozesse anstößt, ob fachliche Grenzen gelten, ob Hinweise abgestuft werden und ob selbstständige Kompetenz anschließend unabhängig nachgewiesen wird.

WHITEEXAM setzt an dieser Designfrage an. Das System entstand zunächst aus Anforderungen kompetenzorientierter digitaler Prüfungen und wurde zu einer Plattform erweitert, die Aufgabenentwicklung, Materialimport, Lernaktivitäten, KI-Lerndialoge, Lernumgebungen, Durchführung, Korrektur und Analyse verbindet. Sein wissenschaftliches Potenzial liegt nicht in einer einzelnen Funktion, sondern in der Trennung und kontrollierten Kopplung zweier Regelkreise:

1. einer *formativen Lernschleife*, in der Fehler erwünscht, Hilfen adaptiv und Wiederholungen möglich sind;
2. einer *summativen Prüfungsschleife*, in der Standardisierung, Nachvollziehbarkeit, Sicherheit und Validität Vorrang haben.

**Designthese.** KI-gestütztes Lernen und kompetenzorientiertes Prüfen benötigen eine gemeinsame fachliche Infrastruktur, aber getrennte Interaktions-, Daten- und Entscheidungslogiken.

Der Beitrag beantwortet drei Fragen:

- F1:** Welche Anforderungen ergeben sich an eine integrierte Lern- und Prüfungsplattform im KI-Zeitalter?
- F2:** Wie operationalisiert WHITEEXAM diese Anforderungen in einer Dual-Loop-Architektur, einem Aufgabenlebenszyklus und einem rollenbasierten Governance-Modell?
- F3:** Welche überprüfbaren Designhypothesen und Forschungsprogramme folgen aus der Architektur?

Der Text ist ein Design Paper, keine Produktbroschüre und keine Wirksamkeitsstudie. Er beschreibt ein digitales Artefakt, begründet seine Entwurfsentscheidungen, markiert Grenzen und formuliert prüfbare Anschlüsse. Damit schafft er eine gemeinsame Referenz für spätere Einzelstudien, ohne deren Ergebnisse vorwegzunehmen.

## 2 Methodischer Status und Materialbasis

### 2.1 Design Paper im Horizont von Design-Based Research

Design-Based Research (DBR) verbindet die Entwicklung von Interventionen in authentischen Kontexten mit Theoriearbeit und systematischer Untersuchung (Brown, 1992; Cobb et al., 2003; Design-Based Research Collective, 2003). Das Ziel ist weder eine rein lokale Optimierung noch eine von der Praxis abgelöste Theorie, sondern übertragbares Gestaltungswissen. Typisch sind iterative Zyklen aus Analyse, Entwurf, Erprobung und Revision (McKenney & Reeves, 2019).

Der vorliegende Beitrag dokumentiert die *Design- und Rekonstruktionsphase* eines solchen Programms. Er erfüllt noch nicht die Anforderungen einer vollständigen DBR-Studie: Es liegt hier kein kontrollierter Wirksamkeitsvergleich, kein abgeschlossenes Mixed-Methods-Design und keine formale Auswertung mehrerer dokumentierter Iterationszyklen vor. Wissenschaftlich belastbar sind dagegen die Artefaktbeschreibung, die explizite Herleitung von Anforderungen, die Zuordnung von Mechanismen zu theoretischen Annahmen und die Formulierung falsifizierbarer Designhypothesen.

### 2.2 Analyisierte Artefakte

Die Rekonstruktion basiert auf dem Systemstand vom 29. Juli 2026. Untersucht wurden:

- die PHP-Anwendung mit Präsentations-, Anwendungs- und Konfigurationsschicht;
- 66 relationale Datenbankmigrationen sowie die produktnahe MariaDB-Instanz;
- die Module für Aufgaben, Sammlungen, Prüfungen, Lernaktivitäten, Lernumgebungen, KI-Dialoge, Klassen/Kurse, Korrektur und Analyse;
- automatisierte Tests und technische Entwicklungsdokumente;
- die aktuellen Lehrpersonen- und Lernendenoberflächen;
- reale, aber für die Publikation datensparsam ausgewählte Test- und Demonstrationsartefakte.

Zum Stichtag umfasste die lokale Instanz unter anderem 760 Aufgaben, 3 Prüfungsvorlagen, 29 Lernaktivitäten und 2 Lernumgebungen. Diese Zahlen belegen Systemumfang, nicht Nutzungserfolg. Screenshots wurden mit einem Autoren-/Testkonto erstellt; individuelle Lernendenlisten und personenbezogene Antwortdaten werden nicht abgebildet.

### 2.3 Evidenz- und Statuslogik

Um Entwicklungsstand und wissenschaftlichen Anspruch nicht zu vermengen, verwendet der Beitrag drei Statusklassen:

**Tabelle 1:** Statusklassen für Aussagen über WHITEEXAM

|                              |                                                                                 |
|------------------------------|---------------------------------------------------------------------------------|
| <b>implementiert</b>         | In Code, Datenmodell und Benutzeroberfläche nachweisbare Funktion zum Stichtag. |
| <b>Designziel</b>            | Aus der Architektur ableitbares, aber noch nicht vollständig realisiertes Ziel. |
| <b>Forschungsperspektive</b> | Hypothese oder Evaluationsoption, deren empirische Prüfung aussteht.            |

Diese Unterscheidung verhindert zwei typische Fehlschlüsse: Aus einer implementierten Funktion folgt noch keine pädagogische Wirkung; aus einer plausiblen Architektur folgt noch keine institutionelle Skalierbarkeit.

## 3 Stand der Forschung

### 3.1 Von intelligenten Tutorsystemen zu generativer KI

KI in der Bildung ist älter als generative Sprachmodelle. Forschungsstränge zu intelligenten Tutorsystemen, adaptiven Systemen, automatisierter Bewertung sowie Profilierung und Vorhersage bestehen seit Jahrzehnten. Meta-analytische Befunde zeigen, dass intelligente Tutorsysteme unter geeigneten Bedingungen lernwirksam sein können (Kulik & Fletcher, 2016; VanLehn, 2011). Zugleich kritisieren systematische Reviews eine schwache pädagogische Theoriebildung, geringe Beteiligung von Lehrpersonen und unzureichende Reflexion ethischer Fragen (Zawacki-Richter et al., 2019). Historische Analysen warnen deshalb davor, Neuheit allein aus technologischen Fähigkeiten abzuleiten (Williamson & Eynon, 2020).

Generative KI erweitert den Möglichkeitsraum: Dialoge müssen nicht mehr vollständig vorprogrammiert werden; Rückmeldungen können sprachlich variiert und auf offene Eingaben bezogen werden. Gleichzeitig sind Ausgaben probabilistisch, können sachlich falsch, verzerrt oder pädagogisch unzweckmäßig sein. Kasneci et al. (2023) folgern daher, dass LLM-basierte Systeme klare pädagogische Strategien, kritische Prüfung und menschliche Aufsicht erfordern.

Die aktuelle Evidenz spricht gegen pauschale Wirkungsannahmen. Kestin et al. (2025) berichten in einem Crossover-RCT mit 194 Studierenden einer Physikeinführung größere kurzfristige Lerngewinne und geringere Lernzeit für einen fachlich sorgfältig gestalteten KI-Tutor gegenüber einer aktiven Präsenzlektion. Die Autoren betonen selbst die Bedeutung von Expert:innen-Prompts, fachlichen Referenzantworten und Scaffolding. Bastani et al. (2025) zeigen dagegen in einem Feldexperiment mit nahezu 1.000 Schülerinnen und Schülern, dass unbeschränkte GPT-Hilfe zwar die Übungsleistung erhöht, aber die anschließende selbstständige Prüfungsleistung verschlechtern kann; lernorientierte Leitplanken verminderten diesen Effekt. Der gemeinsame Befund lautet daher nicht “KI-Tutoren wirken”, sondern: *Interaktionsregeln und Übergänge zur selbstständigen Leistung sind kausal relevante Designmerkmale.*

### 3.2 Formative Rückmeldung und Selbstregulation

Formative Bewertung nutzt Informationen über den aktuellen Lernstand, um nächste Lernhandlungen zu verbessern (Black & Wiliam, 1998). Wirksame Rückmeldung soll verständlich, zielbezogen, handlungsnah und zeitgerecht sein (Hattie & Timperley, 2007; Shute, 2008). Sie darf die kognitive Tätigkeit nicht ersetzen. Nicol und Macfarlane-Dick (2006) verknüpfen Feedback deshalb mit Selbstregulation: Lernende benötigen Kriterien, Möglichkeiten zur Selbsteinschätzung, Dialog und Gelegenheit zur Revision. Zimmerman (2002) beschreibt Selbstregulation als zyklischen Prozess aus Zielsetzung, Handlungskontrolle und Reflexion.

Für WHITEEXAM folgen daraus vier Konsequenzen. Erstens braucht jeder Lernbaustein ein explizites Lernziel. Zweitens müssen Hinweise gestuft sein und eigene Erklärungen voraussetzen. Drittens sollen Fehler nicht unmittelbar in Noten überführt werden. Viertens muss der Wechsel vom unterstützten Lernen zum eigenständigen Nachweis sichtbar sein.

### 3.3 Assessment, Kompetenz und Validität

Constructive Alignment fordert die Abstimmung intendierter Lernergebnisse, Lernaktivitäten und Leistungsbewertung (Biggs, 1996; Biggs & Tang, 2011). Kompetenzorientierung bedeutet dabei mehr als die Verwendung anspruchsvoller Aufgabentypen. Entscheidend ist, welche Schlussfolgerung

aus beobachtbarem Handeln gezogen werden darf. Validität bezieht sich entsprechend auf die Interpretation und Nutzung von Ergebnissen, nicht bloß auf das Instrument (American Educational Research Association et al., 2014; Kane, 2013).

Generative KI verschärft diese Frage. Wenn ein Produkt vollständig oder teilweise durch ein Assistenzsystem erzeugt werden kann, ist unklar, ob es die Kompetenz der geprüften Person repräsentiert. Aktuelle Reviews fordern daher stärker prozessbezogene, höherstufige und kontextualisierte Prüfungen sowie explizite Regeln für erlaubte KI-Nutzung (Bower et al., 2024; Xia et al., 2024). Evaluative Urteilkraft – das begründete Einschätzen der Qualität eigener und fremder Arbeit – gewinnt an Bedeutung (Bearman et al., 2024). Eine aktuelle kritische Übersicht rückt Assessment-Validität ausdrücklich ins Zentrum der GenAI-Debatte (Ali & Maroulis, 2026).

Digitale Prüfungssysteme sollten somit nicht nur Aufgaben ausliefern. Sie müssen Nachweissituationen modellieren: Aufgabenblöcke, Abhängigkeiten, Navigation, Zeit, erlaubte Ressourcen, Bewertungsraster, Entscheidungspfade und Revisionen. WHITEEXAM interpretiert Prüfung daher als gestaltete Evidenzarchitektur.

### 3.4 Learning Analytics als pädagogischer Regelkreis

Learning Analytics verspricht, Lernprozesse sichtbar und Entscheidungen dateninformiert zu machen (Ferguson, 2012; Siemens, 2013). Ohne Lerntheorie können Dashboards jedoch zu bloßer Aktivitätsmessung werden (Wise & Shaffer, 2015). Klickzahlen, Verweildauer oder Einfügeanteile sind keine Kompetenzen. Sie erhalten Bedeutung erst durch ein theoretisch begründetes Interpretationsmodell.

Hinzu kommen Datenschutz, Machtasymmetrien und Zweckänderungsrisiken (Drachsler & Greller, 2016; Slade & Prinsloo, 2013). Für schulische Kontexte sind Datenminimierung, Verständlichkeit, Zugriffsbeschränkung und Widerspruchsmöglichkeiten besonders relevant. WHITEEXAM behandelt Analysedaten deshalb als diagnostische Hinweise, nicht als automatische Zuschreibungen.

### 3.5 Forschungslücke

Die Literatur behandelt KI-Tutoren, digitale Assessments, Learning Analytics und Governance meist in getrennten Diskursen. Offen bleibt, wie eine einzelne Plattform

1. dieselben fachlichen Artefakte für Lernen und Prüfen wiederverwendet,
2. formative Hilfe und summative Entscheidung technisch trennt,
3. Lehrpersonen didaktische Autorenschaft belässt,
4. KI über den Aufgabenlebenszyklus einsetzt, ohne Entscheidungsverantwortung zu automatisieren, und
5. Daten aus beiden Schleifen für nachvollziehbare, theoriegebundene Verbesserung nutzt.

Diese Integrationslücke bildet den Gegenstand des vorliegenden Designs.

## 4 Anforderungen an Plattformen im KI-Zeitalter

Aus Theorie, Systemanalyse und schulischem Anwendungskontext werden zehn Designanforderungen abgeleitet.

**Tabelle 2:** Designanforderungen an eine integrierte Lern- und Prüfungsplattform

| ID  | Anforderung                         | Operationalisierung                                                                                                                                     |
|-----|-------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| A1  | Funktionale Trennung                | Lernversuche, Tutorinteraktionen und formative Quizze müssen von Prüfungsversuchen, Bewertungen und Freigaben getrennt gespeichert und geregelt werden. |
| A2  | Alignment                           | Lernziel, Material, Aktivität, Aufgabe, Bewertungsraster und Ergebnisinterpretation müssen explizit aufeinander beziehbar sein.                         |
| A3  | Lehrpersonen-Autorenschaft          | Lehrpersonen kuratieren Material, Lernweg, Hilfen, Aufgabenkontext, Kriterien und Freigaben; KI bleibt Assistenz.                                       |
| A4  | Begrenzte KI                        | Dialoge benötigen Themenbindung, Lösungsschutz, abgestufte Hinweise, Zuglimits, überprüfbare Zustandsübergänge und Fehlerbehandlung.                    |
| A5  | Aufgabenlebenszyklus                | Import, Entwurf, Review, Versionierung, Sammlung, Lernnutzung, Prüfungsnutzung, Korrektur und Revision sollen zusammenhängend unterstützt werden.       |
| A6  | Kompetenzorientierte Prüfung        | Mehrteilige Aufgaben, verschiedene Antwortformate, Navigation, Ressourcen und Bewertungsraster müssen passend zum Kompetenzziel kombinierbar sein.      |
| A7  | Nachvollziehbarkeit                 | Snapshots, Bewertungsrevisionen, KI-Aufrufe, menschliche Entscheidungen und Freigaben müssen auditierbar bleiben.                                       |
| A8  | Rollen und institutioneller Kontext | Schule, Schuljahr, Klasse, Kurs, Fach, Lehrperson, Lernende und Administration benötigen klar begrenzte Zugriffe.                                       |
| A9  | Datenschutz und Governance          | Zweckbindung, Datenminimierung, externe Providerkontrolle, Aufbewahrung, Löschung und menschliche Letztentscheidung sind Systemanforderungen.           |
| A10 | Fachliche Erweiterbarkeit           | Fachspezifische Aufgabenformate und Tutorprofile sollen auf einem gemeinsamen, fachneutralen Kern aufbauen.                                             |

## 5 Zentraler Entwurf: die Dual-Loop-Architektur

### 5.1 Trennung ohne Entkopplung

Lernen und Prüfen ähneln sich oberflächlich: In beiden Fällen bearbeiten Lernende Aufgaben. Ihre normativen Funktionen unterscheiden sich jedoch grundlegend. Im Lernen soll Unsicherheit sichtbar werden; im Prüfen soll aus einer standardisierten Beobachtung eine begründbare Leistungsinterpretation entstehen. Im Lernen sind Hinweise, Wiederholung und Dialog erwünscht; im Prüfen können dieselben Hilfen die beabsichtigte Aussage des Ergebnisses zerstören.

Abbildung 1 zeigt die Kernidee. Die Schleifen teilen nicht denselben Versuch, sondern referenzieren versionierte fachliche Artefakte. Dadurch kann eine Aufgabe als Ausgangspunkt für ein Quiz, einen KI-Dialog oder eine Prüfung dienen, ohne dass formative Fehlerdaten automatisch Teil einer summativen Bewertung werden.

**Tabelle 3:** Didaktische und technische Differenz von Lern- und Prüfungsmodus

| Dimension     | Lernmodus                                                     | Prüfungsmodus                                                                                      |
|---------------|---------------------------------------------------------------|----------------------------------------------------------------------------------------------------|
| Zweck         | Verständnis entwickeln, diagnostizieren, revidieren           | Kompetenz unter definierten Bedingungen nachweisen                                                 |
| Fehler        | produktive Information                                        | Bestandteil der Evidenz                                                                            |
| Unterstützung | Hinweise, Dialog, Wiederholung                                | nur explizit erlaubte Ressourcen                                                                   |
| KI-Rolle      | tutorielle und formative Assistenz                            | keine verdeckte Hilfe während des Nachweises; Assistenz bei Authoring und nachgelagerter Korrektur |
| Daten         | Lernversuch, Schritte, Nachrichten, formative Zusammenfassung | Prüfungsversuch, Antwortsnapshot, Punkte, Feedback, Revision                                       |
| Entscheidung  | lernenden- und lehrpersonengesteuerte nächste Schritte        | institutionell verantwortete Leistungsentscheidung                                                 |

## 5.2 Sechs Gestaltungsprinzipien

**P1 – Didaktik vor Automatisierung.** Technologie darf didaktische Entscheidungen operationalisieren, aber nicht stillschweigend ersetzen.

**P2 – Kompetenz vor Aufgabentyp.** Multiple Choice, offene Antwort, Code oder Dialog sind Mittel zur Sichtbarmachung unterschiedlicher Leistungen.

**P3 – Geteilte Artefakte, getrennte Versuche.** Wiederverwendung erfolgt über versionierte Aufgaben und Materialien, nicht über Vermischung von Lern- und Prüfungsdaten.

**P4 – KI als begrenzter Akteur.** Modelle diagnostizieren, formulieren, variieren und clustern. Serverseitige Regeln bestimmen Zustandswechsel, Freigaben und Abschluss.

**P5 – Menschliche Geltungsentscheidung.** Lehrpersonen verantworten Lernziele, Aufgabenauswahl, Kriterien, Freigaben und finale Bewertung.

**P6 – Nachvollziehbare Evolution.** Importierte Materialien, Lernaktivitäten, Prüfungen und Bewertungen bleiben versioniert und revidierbar.

# 6 Systemarchitektur

## 6.1 Architekturstil

WHITEEXAM ist **implementiert** als serverseitig gerendeter modularer Monolith auf Basis von PHP und MariaDB/MySQL realisiert. Präsentationsseiten liegen in getrennten Bereichen für Administration, Lehrpersonen, Lernende und Prüfungsdurchführung. Domänenlogik ist in PHP-Modulen gekapselt; JavaScript ergänzt interaktive Lern-, Prüfungs- und Analysefunktionen. Dieser Stil reduziert verteilte Systemkomplexität und erleichtert die Nachvollziehbarkeit von Fachregeln. Grenzen bestehen bei horizontaler Skalierung, Hintergrundjob-Orchestrierung, standardisierten Fremdschnittstellen und mandantenfähiger Rechteverwaltung.

## 6.2 Domänenmodell

Das Datenmodell organisiert fünf miteinander verbundene Bereiche:

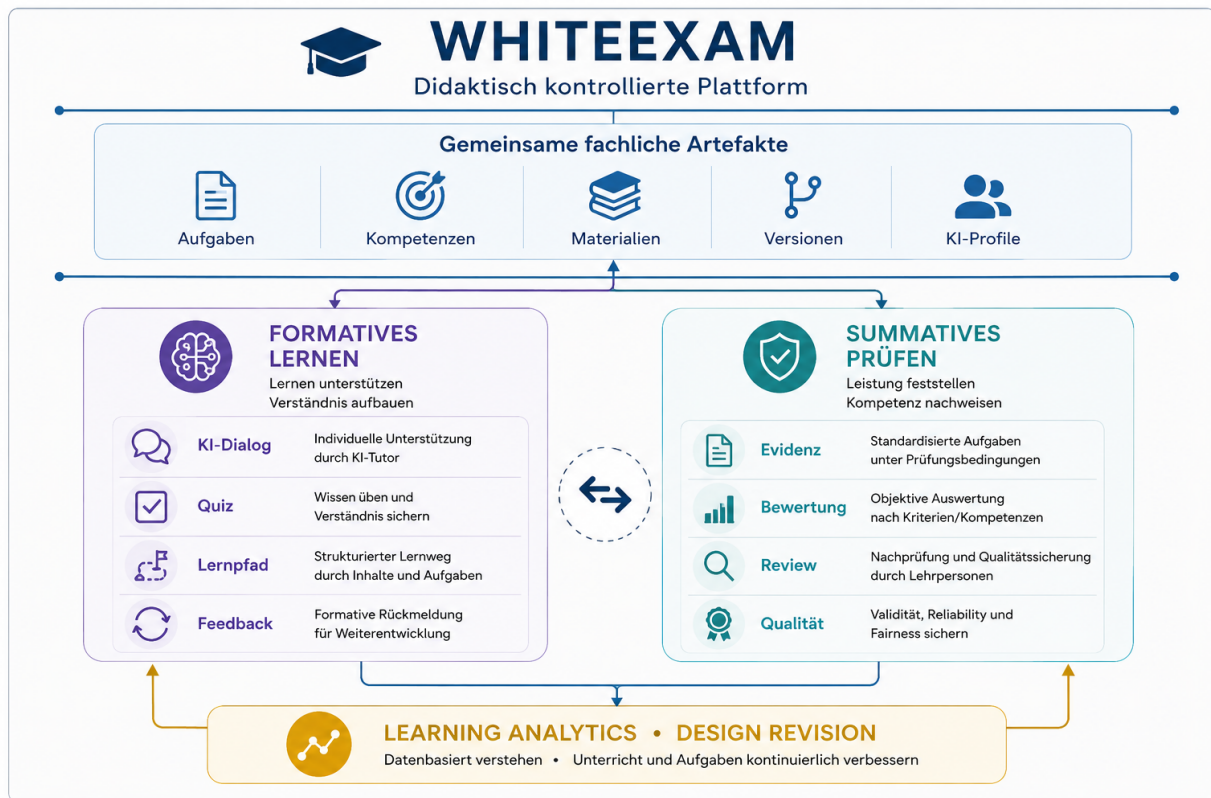


Abbildung 1: Dual-Loop-Architektur: getrennte Regelkreise mit gemeinsamen fachlichen Artefakten

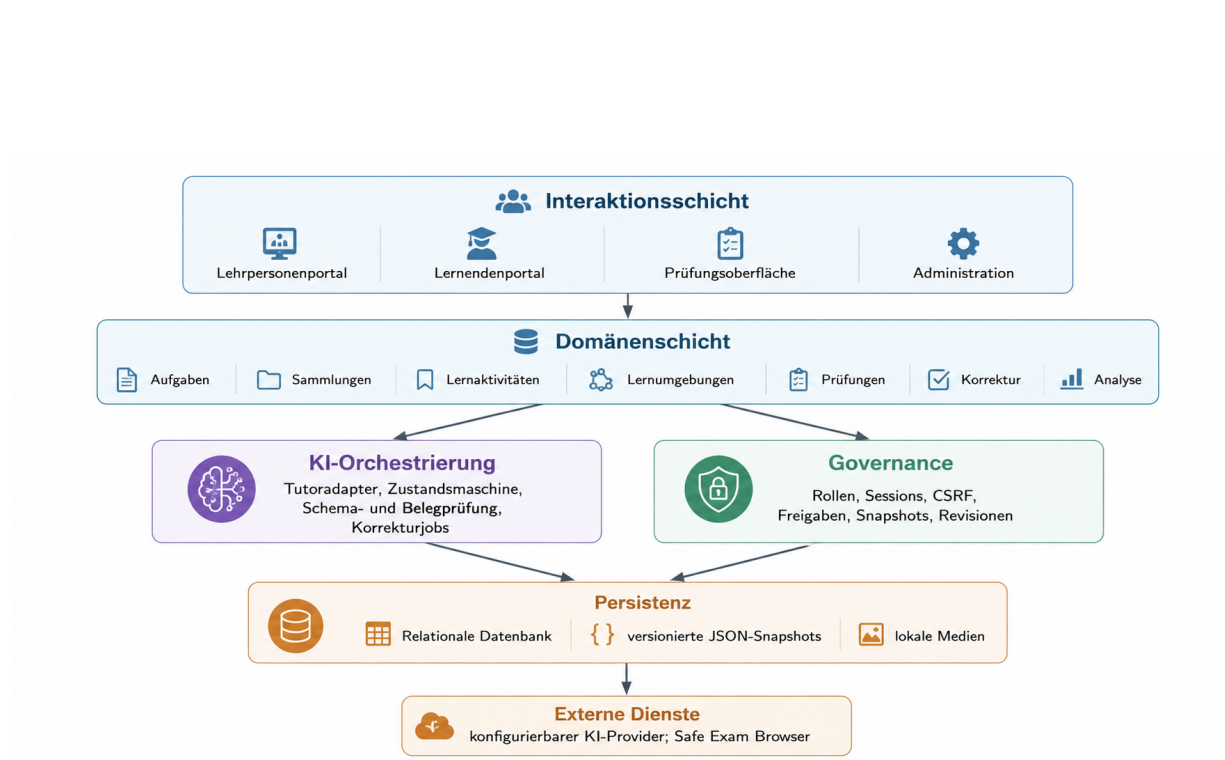


Abbildung 2: Logische Gesamtarchitektur von WHITEEXAM

1. **Organisation:** Benutzer, Schulen, Schuljahre, Stammklassen, Kurse und Zuordnungen;
2. **Wissens- und Aufgabenbasis:** Fächer, Stufen, Themen, Sammlungen, Aufgaben, Assets und Aufgabenblöcke;
3. **Lernen:** versionierte Lernaktivitäten, Lernumgebungen, Freigaben, Quizversuche, Dialogsitzungen und Lernschrittstatus;
4. **Prüfen:** Prüfungen, Sektionen, Aufgabenblöcke, Zugangscodes, Versuche, Antworten und SEB-Konfiguration;
5. **Bewerten und Analysieren:** manuelle Bewertungen, KI-Korrekturjobs, Revisionen, Lern- und Prüfungsstatistiken.

Die gemeinsame Aufgabenentität ist ein Integrationspunkt, aber kein globaler Wahrheitscontainer. Beim Start einer Lern- oder Prüfungssitzung werden relevante Konfigurationen als Snapshot gebunden. Nachträgliche Änderungen am Entwurf verändern damit nicht rückwirkend die Bearbeitungsbedingungen eines laufenden Versuchs.

### 6.3 Rollenmodell

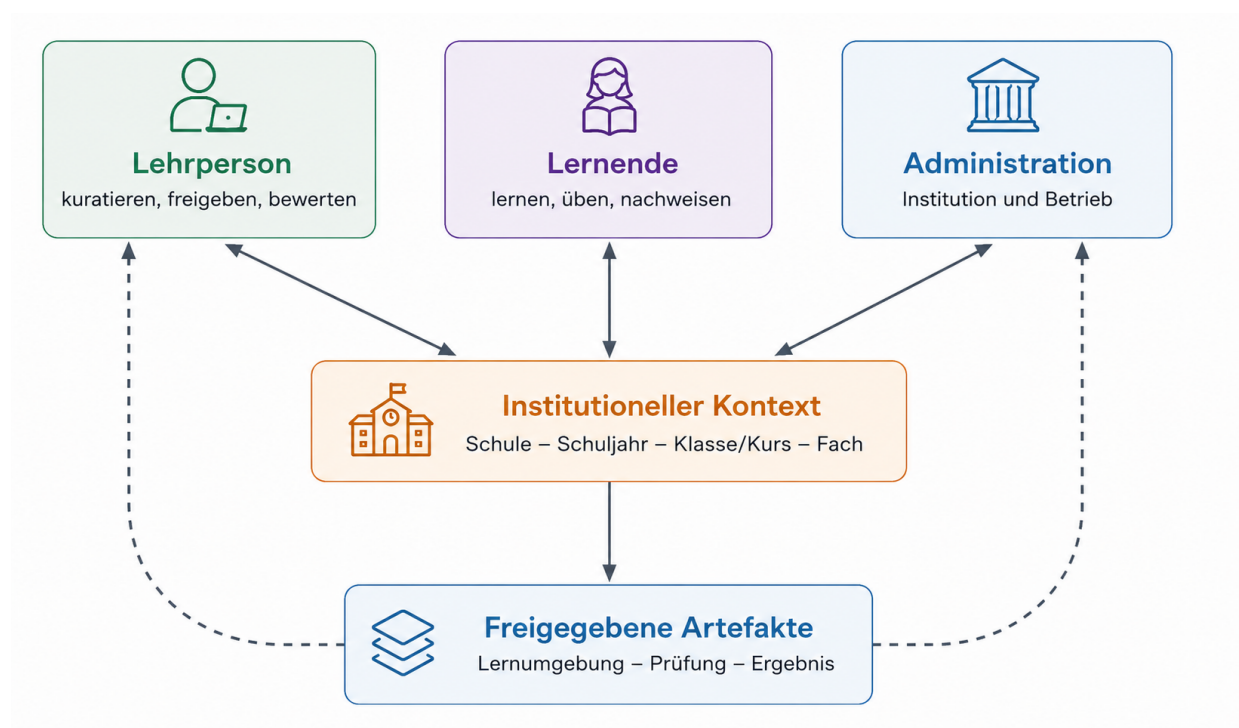


Abbildung 3: Rollenmodell und institutioneller Zugriffskontext

Die **implementiert** Rollen **student**, **teacher** und **admin** werden durch Klassen- und Schulzuordnungen kontextualisiert. Lernende sehen ausschließlich für ihre Kurse freigegebene Lernumgebungen und eigene Ergebnisse. Lehrpersonen bearbeiten eigene oder für sie sichtbare Artefakte und sehen Lernverläufe nur im legitimierten Unterrichtskontext. Für größere Verbünde ist ein feingranulares Policy- und Mandantenmodell **Designziel**.

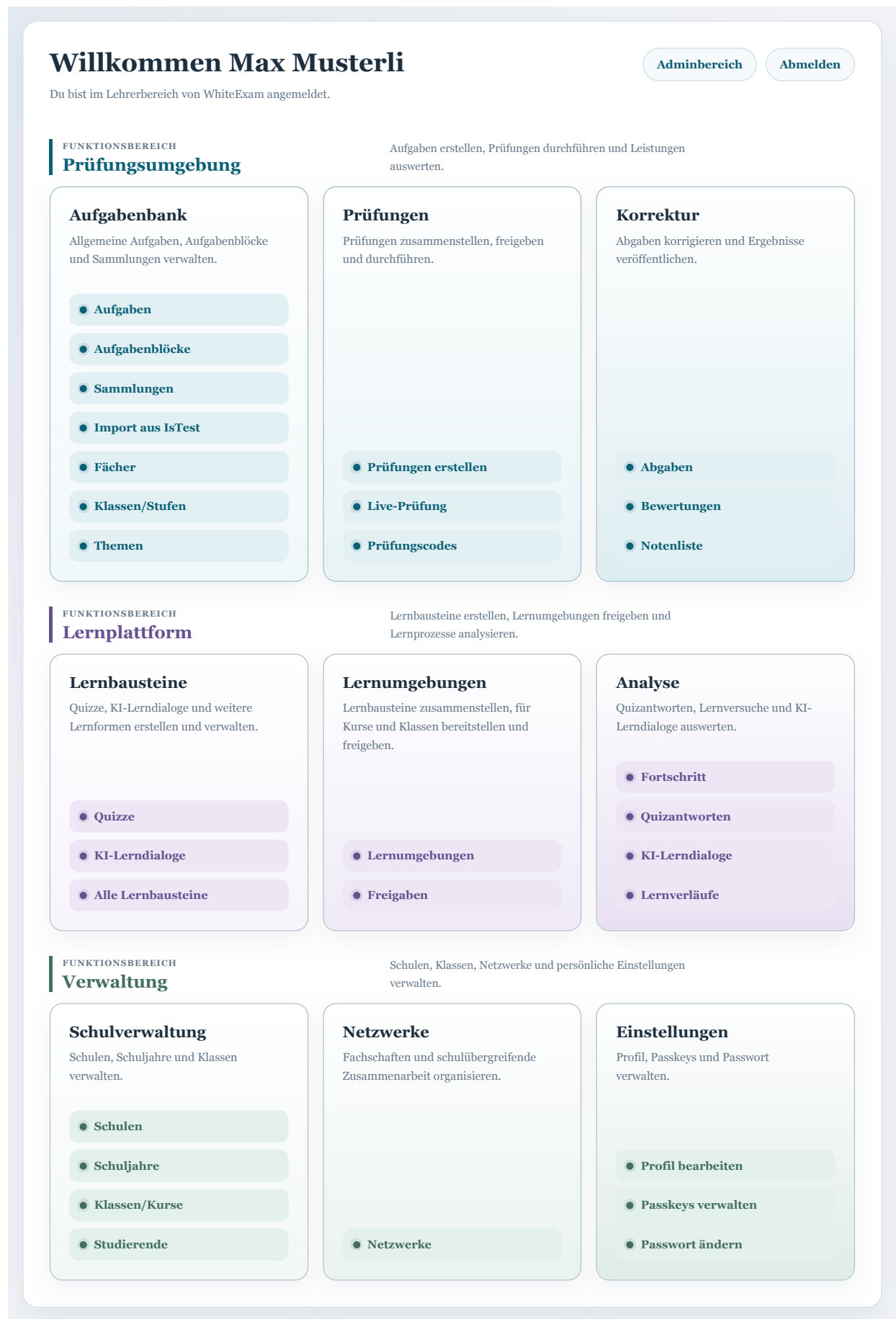


**Abbildung 4:** Lehrpersonen-Dashboard: funktionale Trennung von Prüfungsumgebung, Lernplattform und Verwaltung (Systemstand 29. Juli 2026)

## 7 Die formative Lernschleife

### 7.1 Lernumgebungen als kuratierte Pfade

Eine Lernumgebung ordnet versionierte Lernaktivitäten in eine von der Lehrperson verantwortete Reihenfolge und gibt sie für Klassen, Kurse oder Gruppen frei. **implementiert** sind Quizaktivitäten, offene Aufgaben und KI-Lerndialoge kombinierbar. Die Reihenfolge ist im aktuellen Stand eine Empfehlung, keine algorithmisch erzwungene adaptive Route. Diese Grenze ist didaktisch relevant: WHITEEXAM behauptet derzeit keine autonome Personalisierung.

Adaptive Unterstützung liegt auf drei Ebenen:

1. **Mikroadaption:** Hinweise und Tutorreaktionen beziehen sich auf die aktuelle Antwort und den gespeicherten Lernschritt;
2. **Lehrpersonenadaption:** Lehrpersonen verändern Reihenfolge, Freigabe und Materialien auf Basis beobachteter Lernstände;
3. **Pfadadaption:** datenbasierte Verzweigung oder Freischaltung zukünftiger Lernaktivitäten ist **Designziel** und muss vor Einsatz empirisch und ethisch validiert werden.

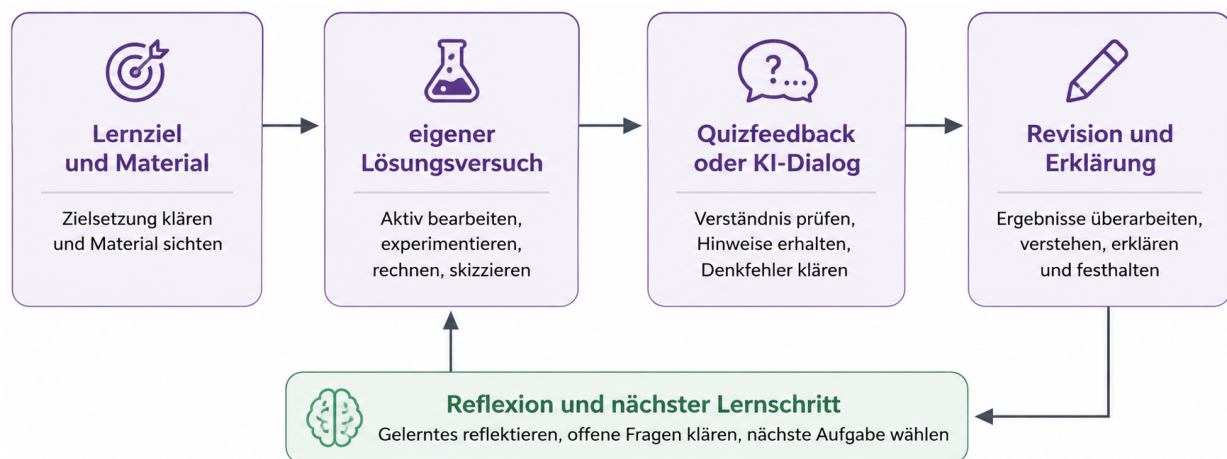


Abbildung 5: Lernzyklus aus eigenem Versuch, begrenzter Unterstützung und Revision

### 7.2 Klassische Aufgaben und KI-Dialoge

Die Kombination klassischer Aufgaben mit Dialogen erfüllt unterschiedliche Funktionen. Auswahlfragen können Begriffsunterscheidungen und typische Fehlvorstellungen effizient diagnostizieren. Offene Aufgaben erzwingen eine eigene sprachliche oder mathematische Externalisierung. Ein anschließender KI-Dialog kann genau diese Antwort aufgreifen, Nachfragen stellen und zur Revision anregen.

Diese Abfolge vermeidet, den Dialog zum universellen Lernformat zu erklären. Lernende sollen nicht bloß mit einem Modell sprechen, sondern zwischen Erkennen, Produzieren, Begründen, Prüfen und Überarbeiten wechseln. Das entspricht der Annahme, dass geeignete Repräsentationswechsel und dosierte Unterstützung die kognitive Verarbeitung fördern können (Mayer, 2009; Sweller et al., 2011).

Meine Kurse → GKG Klasse 1c (Physik) → Mechanik 1 – Einführung, Trägheit und Kräfte

LERNWEG

Mechanik 1 – Einführung, Trägheit und Kräfte

Zurück zum Kurs

Kurs: GKG Klasse 1c (Physik) Schuljahr 2026/27

In Bearbeitung

Ein geführter Lernweg zur Entwicklung der Mechanik, zum Trägheitsprinzip, zum Fadenexperiment und zu den Wirkungen von Kräften. Fachliche Grundlage: „Phy1, 1. Mechanik (Script).pdf“.

LERNBAUSTEINE

Dein Lernweg

1

QUIZ · VERPFLICHTEND

Abgeschlossen

2

QUIZ · VERPFLICHTEND

Abgeschlossen

3

QUIZ · VERPFLICHTEND

Abgeschlossen

4

QUIZ · VERPFLICHTEND

Abgeschlossen

5

QUIZ · VERPFLICHTEND

Abgeschlossen

6

QUIZ · VERPFLICHTEND

Abgeschlossen

7

OFFENE FRAGE · VERPFLICHTEND

Abgeschlossen

8

KI-LERNDIALOG · VERPFLICHTEND

In Bearbeitung

Abbildung 6: Lernendenansicht eines kuratierten Lernpfads mit Quiz, offener Aufgabe und KI-Lerndialog

14

LEHRPERSONENVORSCHAU · KEINE STATISTIK

Bearbeiten

# Mechanik 1 – Quizfrage 7: Bewegungszustand ohne resultierende Kraft

[Zurück zu den Quizen](#)

Version 1 · Veröffentlicht

## QUIZFRAGE 1 VON 1

### Mechanik 1 – Quizfrage 7: Bewegungszustand ohne resultierende Kraft

Auf zwei Körper wirkt jeweils keine resultierende äußere Kraft. Welche Aussagen folgen aus dem Trägheitsprinzip? **Es können mehrere**

**Antworten richtig sein.**

Antwortmöglichkeiten

- ☒ A. Ein ruhender Körper bleibt in Ruhe.
- ☒ B. Ein geradlinig gleichförmig bewegter Körper bewegt sich geradlinig gleichförmig weiter.
- ☐ C. Ein ruhender Körper setzt sich ohne weitere Einwirkung von selbst in Bewegung.
- ☐ D. Ein bewegter Körper benötigt ständig eine resultierende Kraft, um seine Geschwindigkeit beizubehalten.

Richtige Auswahl in der Lehrpersonenvorschau: A, B

Prüfversuche: 2 · Hinweise: aktiv · Lösung nach Grenze: ja

### Erklärung nach Abschluss

Das Trägheitsprinzip umfasst beide Bewegungszustände: Ruhe und geradlinig gleichförmige Bewegung. Ohne resultierende äußere Kraft ändert sich der jeweilige Zustand nicht.

(a) Formatives Quiz mit erklärter Lösung

LEHRPERSONENVORSCHAU · TESTDATEN · KEINE STATISTIK

[Zur Konfiguration](#)

## KI-Lerndialog 2 – Das Fadenexperiment verstehen

### Mechanik 1 – Offene Frage 2: Das Fadenexperiment erklären

Ein Massenstück hängt an einem oberen Faden. An seiner Unterseite ist ein zweiter Faden befestigt. Beim langsamen Ziehen am unteren Faden reißt der obere Faden. Beim schnellen, ruckartigen Ziehen reißt dagegen der untere Faden.

Erkläre beide Beobachtungen zusammenhängend. Berücksichtige Gewichtskraft, Zugkraft, die Belastung beider Fäden, die Trägheit des Massenstücks und den Unterschied zwischen langsamem und ruckartigem Ziehen.

#### Lernziel

Du erklärst beide Fälle des Fadenexperiments mit Gewichtskraft, Zugkraft, Fadenbelastung, Trägheit und kurzer Einwirkungszeit und vergleichst sie fachlich korrekt.

TEST

orientation

0/6

KI-Tutor

orientation

Willkommen!

Nutze deine eben gespeicherte Erklärung zum Fadenexperiment als Ausgangspunkt. Betrachte langsames und ruckartiges Ziehen getrennt und verbessere anschließend deinen Vergleich.

Antwort prüfen

Hinweis

Ich komme nicht weiter

Dialog vorzeitig beenden

Testsitzung zurücksetzen und löschen

(b) Begrenzte Lehrpersonenvorschau eines KI-Dialogs

**Abbildung 7:** Komplementäre Lernaktivitäten in WHITEEXAM

### 7.3 Dialogischer KI-Tutor

Der **implementiert** Tutorprozess folgt einem providerneutralen Vertrag. Das Modell erhält einen Aktivitätssnapshot, den serverseitigen Zustand und das Transkript des aktuellen Versuchs. Es liefert eine strukturierte Antwort. WHITEEXAM prüft anschließend Schema, Lernschritt-IDs und wortgetreue Belege gegen gespeicherte Lernendennachrichten. Erst danach werden Lernschritt, Phase, Hinweisstufe oder Abschluss verändert.

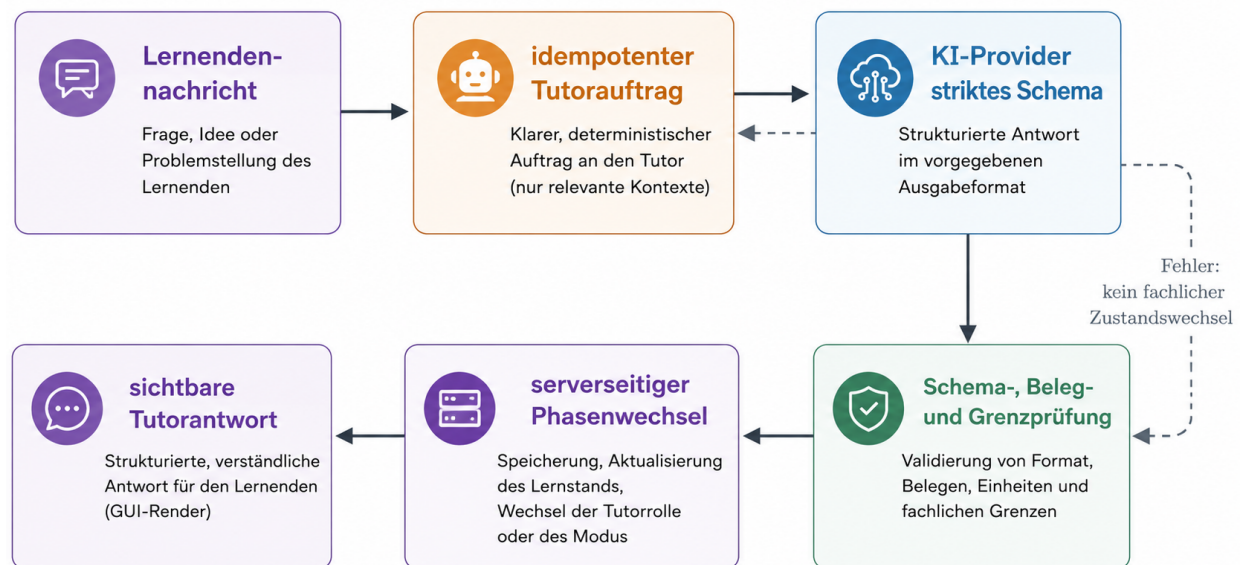


Abbildung 8: Kontrollierter Ablauf eines KI-Lerndialogs

Die Zustandsmaschine strukturiert den Dialog von Orientierung und erster Vermutung über Bedingungsprüfung, schrittweise Spur, Erklärung und Reflexion bis zum Abschluss. Weitere Leitplanken sind:

- keine Ausgabe einer vollständigen Musterlösung (`solution_policy = never`);
- maximale Zahl von Tutorantworten und Zeichenlimits;
- abgestufte Hinweise statt sofortiger Lösung;
- Rückführung bei Themenabweichung oder Manipulationsversuch;
- keine Benotung im Lerndialog;
- Providerfehler verändern den fachlichen Zustand nicht;
- technische Metadaten werden ohne Schlüssel oder vollständige Providerpayloads gespeichert.

Diese Regeln sind keine Garantie fachlicher Korrektheit. Sie reduzieren jedoch den Freiheitsgrad des Modells und machen pädagogisch relevante Zustände überprüfbar. Das Design folgt damit der Evidenz, dass nicht bloßer KI-Zugang, sondern Scaffolding und Guardrails die Lernwirkung prägen (Bastani et al., 2025; Kestin et al., 2025; OECD, 2026).

### 7.4 Learning Analytics und adaptive Lernpfade

WHITEEXAM speichert bei Lernaktivitäten Fortschritt, Quizversuche, Dialognachrichten, Lernschrittstatus und begrenzte Eingabemetriken. Einzelne Tasten oder Zwischenablageinhalte werden nicht erfasst. Ein hoher Einfügeanteil wird ausdrücklich als Beobachtung, nicht als Täuschungsbeweis behandelt.

Aus diesen Daten können drei analytische Ebenen entstehen:

1. **deskriptiv:** Welche Aktivitäten wurden begonnen oder abgeschlossen?
2. **diagnostisch:** Welche Fehlvorstellungen oder Lernschritte treten wiederholt auf?
3. **präskriptiv:** Welcher nächste Lernbaustein wäre fachlich begründbar?

Nur die erste Ebene ist weitgehend **implementiert**. Die zweite wird durch Dialogstatus und Quizanalysen vorbereitet. Die dritte ist **Forschungsperspektive**. Ein adaptiver Pfad darf nicht allein aus Aktivitätsdaten berechnet werden; er benötigt ein explizites Kompetenzmodell, Unsicherheitsangaben, Einsichtsmöglichkeiten und die Möglichkeit lehrpersoneller Übersteuerung.

## 8 Die summative Prüfungsschleife

### 8.1 Prüfung als Evidenzarchitektur

Eine Prüfung ist in WHITEEXAM hierarchisch als Prüfung, Sektion, Aufgabenblock und Teilaufgabe modelliert. Jede Ebene kann Reihenfolge, Beschreibung, Navigation und teilweise Zeit- oder Bewertungslogik bestimmen. Dadurch lassen sich zusammenhängende Szenarien abbilden, in denen Teilfragen aufeinander aufbauen oder bewusst voneinander unabhängig bleiben.

Kompetenzorientierung wird nicht durch einen Aufgabentyp garantiert. Ein belastbarer Prüfungsplan muss mindestens spezifizieren:

$$\begin{aligned} &\text{Kompetenzziel} \rightarrow \text{beobachtbare Handlung} \rightarrow \text{Aufgabenformat} \\ &\quad \rightarrow \text{Bewertungsregel} \rightarrow \text{Interpretation.} \end{aligned} \quad (1)$$

Die Validitätsfrage lautet, ob die beobachtete Antwort die beabsichtigte Kompetenz unter den gewählten Bedingungen hinreichend repräsentiert (American Educational Research Association et al., 2014; Kane, 2013).

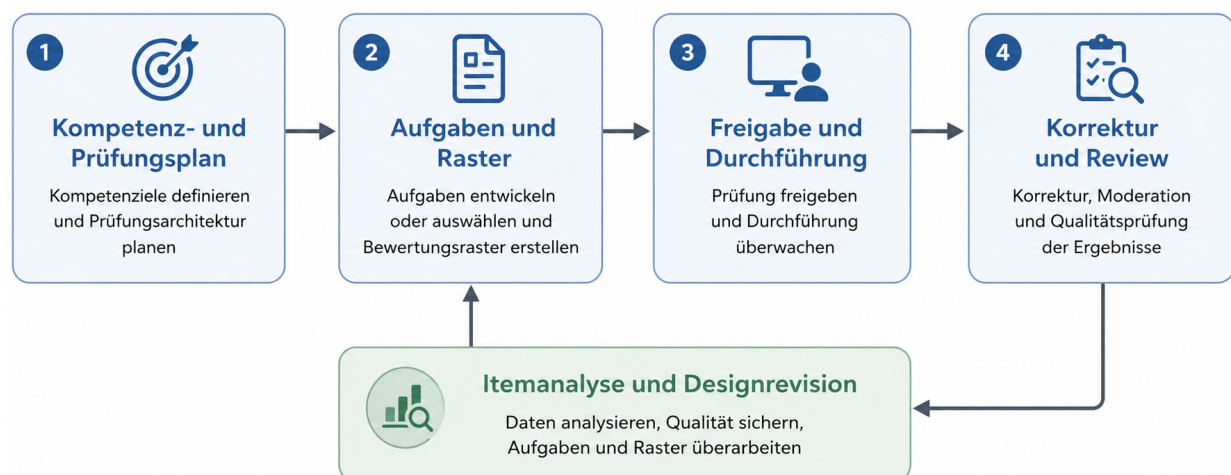


Abbildung 10: Prüfungszyklus von der Spezifikation zur qualitätsgesicherten Revision

### 8.2 Kompetenzorientierte Prüfungsarchitektur

Der **implementiert** Editor verbindet freie und lineare Navigation, Sektionen, Aufgabenblöcke, Punktgewichte, Zeitlimits und Safe-Exam-Browser-Konfiguration. Klassische Auswahlaufgaben können mit offenen Erklärungen, mehrschrittigen Rechenaufgaben, experimentbezogenen Transfers und Python-Code kombiniert werden. Dadurch wird eine Prüfung nicht auf automatisch

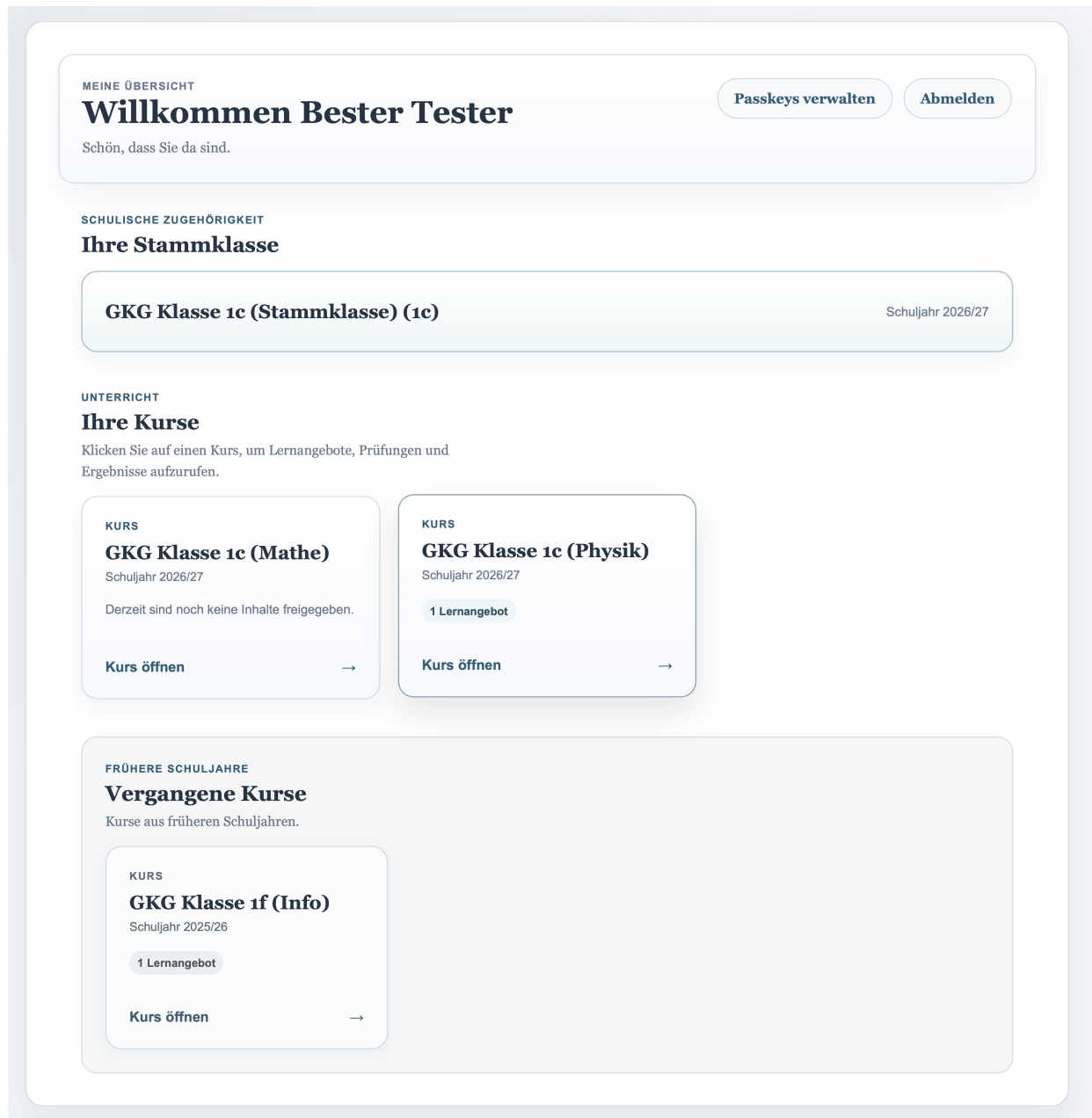


Abbildung 9: Studierenden-Dashboard mit institutioneller Zugehörigkeit und kursbezogenen Lernangeboten

korrigierbare Items reduziert.

Ein exemplarisches Prüfungsdesign unterscheidet:

1. **Konzeptverständnis:** Auswahl, Zuordnung oder kurze Begründung typischer Modelle und Fehlvorstellungen;
2. **Anwendung:** strukturierte, mehrschrittige Bearbeitung bekannter Problemklassen;
3. **Transfer:** begründete Anwendung auf neue Kontexte, Daten, Experimente oder offene Szenarien.

Diese Dreiteilung ist kein universelles Taxonomieschema. Sie zeigt, wie unterschiedliche Evidenzarten innerhalb einer Prüfung bewusst kombiniert werden können.

### 8.3 Durchführung und Sicherheit

PrüfungsCodes verbinden Prüfung, Zielgruppe, Zeitfenster und Modus. Live-Ansichten zeigen Status und Fortschritt; Unterbrechungen können kontrolliert behandelt werden. Safe Exam Browser ist **implementiert** als optionaler oder verpflichtender Baustein integrierbar. Erlaubte externe URLs ermöglichen ressourcenbezogene Prüfungen, beispielsweise mit Simulationen oder Fachwerkzeugen. Sicherheit ist dabei kein Selbstzweck. Eine stark gesicherte Umgebung kann die Vergleichbarkeit erhöhen, aber auch Authentizität und Zugänglichkeit einschränken. Umgekehrt kann eine offene Ressourcenprüfung valide sein, wenn gerade Recherche, Werkzeugnutzung und kritische Bewertung geprüft werden. Die Sicherungsstufe muss deshalb aus dem Kompetenzziel folgen.

### 8.4 KI im Prüfungslebenszyklus

Die konsequente Trennung von Lern- und Prüfungsmodus bedeutet kein generelles KI-Verbot. KI kann an verschiedenen Stellen unterschiedliche Rollen übernehmen:

- **Authoring:** Varianten, Distraktoren oder Formulierungsvorschläge als Entwurf;
- **Review:** Hinweise auf Mehrdeutigkeit, fehlende Kriterien oder unplausible Punkteverteilung;
- **Durchführung:** nur wenn KI-Nutzung selbst expliziter Teil des Kompetenznachweises ist;
- **Korrektur:** Clustering und Vorschläge auf Basis von Aufgabe, Musterlösung und Bewertungsraster;
- **Analyse:** Identifikation auffälliger Items oder wiederkehrender Fehlermuster.

WHITEEXAM realisiert gegenwärtig vor allem die KI-gestützte Nachbearbeitung offener Antworten. Antworten werden aufgabenweise gesammelt, in Qualitätsgruppen strukturiert und mit Vorschlägen für Punkte, lernendenorientiertes Feedback und interne Begründung versehen. Lehrpersonen prüfen und verändern die Vorschläge. Revisionen speichern alte und neue Werte sowie den Lehrpersonenhinweis. Diese Human-in-the-Loop-Struktur ist notwendig, aber nicht hinreichend: Für summative Nutzung müssen zusätzlich Güte, Fairness, Interrater-Übereinstimmung und Fehlerprofile empirisch untersucht werden.

## 9 Aufgabenlebenszyklus und Materialimport

### 9.1 Von vorhandenen Materialien zu versionierten Artefakten

Innovationen scheitern häufig, wenn sie vollständige Neuerstellung voraussetzen. WHITEEXAM behandelt bestehende Materialien deshalb als Ausgangspunkt eines kontrollierten Transformationsprozesses. **implementiert** sind der Import aus IsTest-/GIFT-nahen Formaten sowie spezialisierte

Importpfade für Quiz- und Dialogmaterialien. Importierte Inhalte werden nicht automatisch zu freigegebenen Prüfungsisems, sondern als prüfbare Entwürfe in die Aufgabenbank überführt.

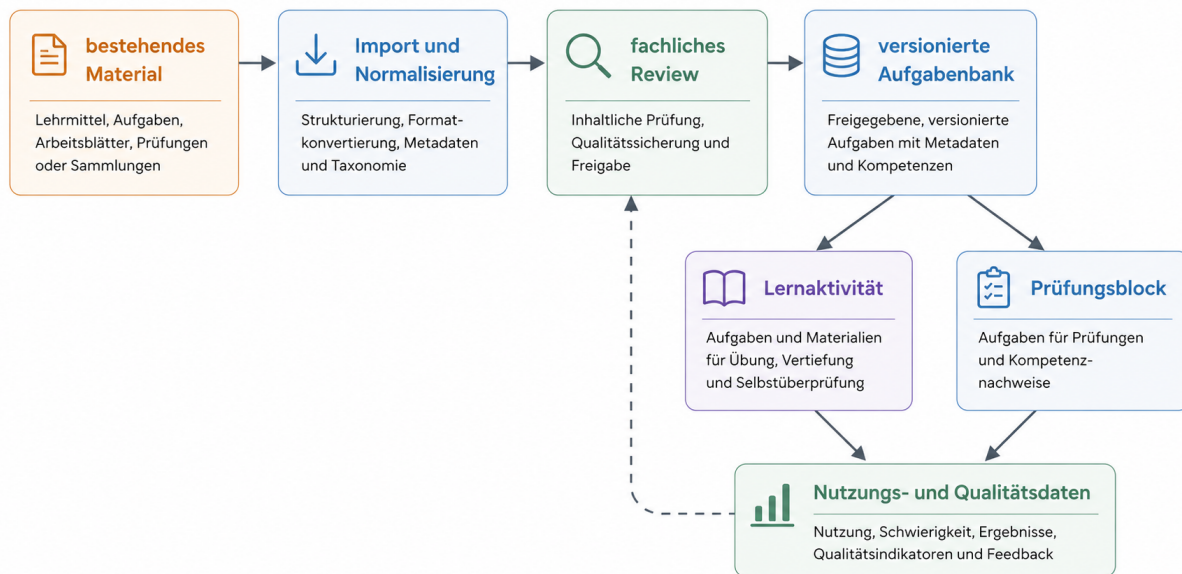


Abbildung 12: Unterstützung des Aufgabenlebenszyklus

Der Lebenszyklus umfasst Provenienz, fachliche Zuordnung, Sichtbarkeit, Status, Version, Lösung, Scoring, Assets und Nutzungskontext. Er ermöglicht schrittweise Qualitätsentwicklung: Ein vorhandenes Item kann importiert, überarbeitet, in einer formativen Aktivität erprobt, in einem Aufgabenblock arrangiert und später in einer Prüfung genutzt werden. Umgekehrt können Analyseergebnisse eine Revision auslösen, ohne historische Versuche zu verändern.

## 9.2 Sammlungen und fachliche Erweiterbarkeit

Sammlungen strukturieren Aufgaben nach Fach, Stufe und Thema und besitzen abgestufte Sichtbarkeit. Der gemeinsame Kern ist fachneutral; Fachspezifik entsteht durch Inhalte, Aufgabenformate, Bewertungsraster und Tutorprofile. Der aktuelle Bestand umfasst Physik, Informatik, Mathematik und PAM. Naturwissenschaftliche Experimente, mathematische Herleitungen, sprachliche Dialoge oder Programmieraufgaben stellen dennoch unterschiedliche Anforderungen. Erweiterbarkeit bedeutet daher nicht, jedes Fach in dasselbe Interaktionsmuster zu pressen, sondern fachspezifische Erweiterungen an gemeinsame Lebenszyklus-, Rollen- und Governance-Regeln anzuschließen.

# 10 Kompetenzmodell und Constructive Alignment

## 10.1 Konzeptionelles Kompetenzmodell

WHITEEXAM verfügt **implementiert** über Fach-, Stufen- und Themenzuordnungen sowie über Lösungen und Bewertungsraster. Ein eigenständiges, hierarchisches Kompetenzmodell mit Kompetenzdimensionen, Niveaustufen und Evidenzregeln ist zum Stichtag jedoch nicht als vollständiges Modul implementiert. Es ist **Designziel** und eine Voraussetzung für belastbare adaptive Lernpfade und längsschnittliche Kompetenzprofile.

PRÜFUNGSUMGEBUNG

Prüfung bearbeiten

Vorschau

Zurück zu Prüfungen

Pflege die Prüfungsdaten und stelle die Abschnitte aus Aufgabenblöcken zusammen.

Gesamtpunktzahl: 54 Punkte

Prüfungsdaten

Titel

Sj. 25-26, Physik, Klasse 1a, 3. Prüfung (Thermodynamik)

Beschreibung

Die Prüfung besteht aus drei Teilen:

\*\*Teil 1: Multiple-Choice-Aufgaben\*\*

12 Multiple-Choice-Fragen zu den Inhalten der Unterrichtseinheit und den Brian-Aufgaben.

\*\*Teil 2: Rechen- und Verständnisaufgaben\*\*

Fach

Physik

Stufe

Klasse 1

Status

Entwurf

Zeitlimit (optional)

90

Sichtbarkeit

Privat

Wirkungsweise der Sichtbarkeit

Von enthaltenen Aufgaben übernehmen

Kontextsichtbarkeit wird verfügbar, sobald die Migration 044\_add\_context\_visibility\_overrides.sql eingespielt ist.

Abbildung 11: Prüfungseditor mit Prüfungsdaten, SEB-Rahmen und mehrteiliger Aufgabenarchitektur

PRÜFUNGSUMGEBUNG

Sammlungen verwalten

Zurück zu Sammlungen

Bestehende Sammlungen filtern, prüfen und bearbeiten.

Alle Sammlungen

Fach

Stufe

Thema

Status

Sichtbarkeit

Informatik

Alle Stufen

Algorithmen

Alle Status

Alle Sichtbarkeiten

Filtern

Filter zurücksetzen

| TITEL                                       | FACH       | STUFE    | THEMA       | SICHTBARKEIT |
|---------------------------------------------|------------|----------|-------------|--------------|
| Info - Algorithmen, Heronverfahren          | Informatik | Klasse 1 | Algorithmen | Privat       |
| Info - Algorithmen, Rekursive und Iterative | Informatik | Klasse 1 | Algorithmen | Privat       |

Klassen & Kurse

Zum Dashboard

Stammklasse oder Kurs anlegen

Eine Stammklasse filtert SchülerInnen, die kein überlappend bestehende Schüler innen aus gewählten Stammklassen.

Stammklassenliste importieren

Aktives Schuljahr: 2026/27

Schule

Typ

Name

Kürzel

Gymnasium Kirchgarten

Stammklasse

z.B. 2a oder Physik LK

optional

Anlegen und Mitglieder verwalten

Klassenübersicht

Alle Klassen und Kurse Ihrer Schulen mit Status und Belegung.

Schule

Schuljahr

Status

Typ

Namen, Kürzel, Schule

2026/27

Aktive

Alle

Anwenden

Zurücksetzen

Klassen/Kurse

Schulen

Lehrpersonen-Zuordnungen

Studierenden-Möglichkeiten

8

1

8

165

Neues Gymnasium Kirchgarten

Schuljahr 2026/27

| Name/Kürzel            | Typ  | Studierende | Lehrpersonen | Status | Abgleich | Aktion |
|------------------------|------|-------------|--------------|--------|----------|--------|
| GRG Klasse 1c (Mathe)  | Kurs | 21          | 1            | Aktiv  | Aktuell  | Öffnen |
| GRG Klasse 1c (Physik) | Kurs | 21          | 1            | Aktiv  | Aktuell  | Öffnen |

(a) Sammlungen nach Fach, Stufe, Thema und Sichtbarkeit

(b) Klassen- und Kursverwaltung

Abbildung 13: Wiederverwendung von Materialien im institutionellen Kontext

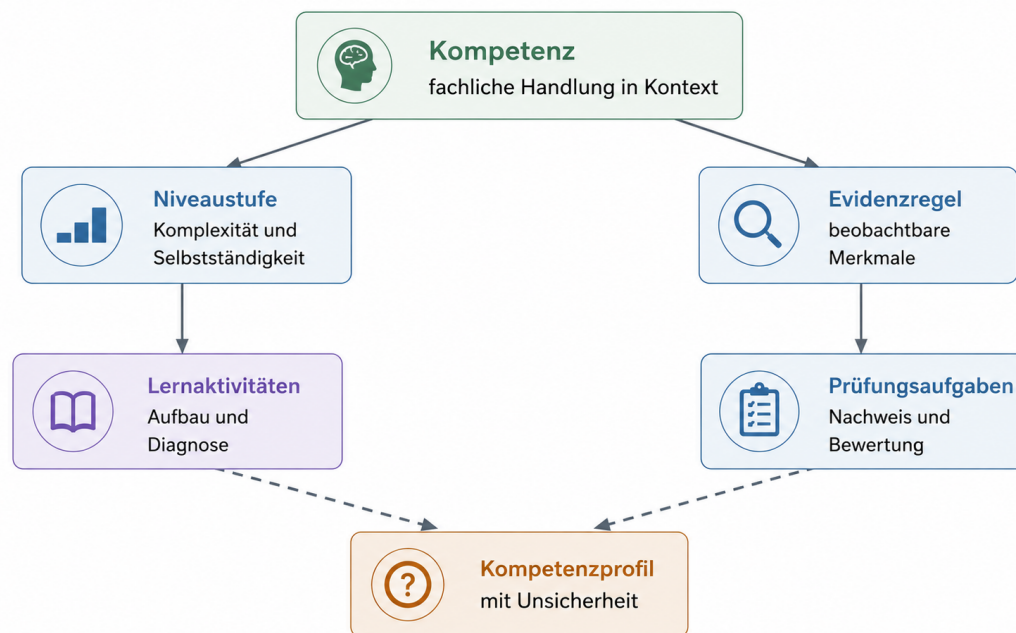


**Abbildung 14:** Konzeptionelles Kompetenzmodell für künftige Alignment- und Adaptionfunktionen

*Benötigte künftige UI-Abbildung:* Kompetenzmodell-Editor mit Kompetenzdimension, Niveaustufen, verknüpften Lernaktivitäten, Prüfungsaufgaben, Evidenzregeln und sichtbarer Unsicherheit. Dieser Platzhalter ist absichtlich konzeptionell; ein entsprechender aktueller Screenshot wäre irreführend.

## 10.2 Alignment als prüfbare Relation

Constructive Alignment wird in der Plattform nicht durch ein Schlagwort, sondern durch Beziehungen operationalisiert:

$$\mathcal{A} = \{(c, m, l, a, r) \mid c = \text{Kompetenz}, m = \text{Material}, \\ l = \text{Lernaktivität}, a = \text{Assessment}, r = \text{Rubrik}\}. \quad (2)$$

Eine Alignment-Prüfung kann dann fragen, ob jede intendierte Kompetenz durch Lerngelegenheiten aufgebaut und durch mindestens eine passende Evidenzregel geprüft wird. Umgekehrt sollen Aufgaben keine hohe Gewichtung besitzen, wenn sie nicht auf deklarierte Lernziele zurückführbar sind.

Diese Formalisierung eröffnet eine spätere **Forschungsperspektive** Analyse von curricularer Abdeckung, Überprüfung und Übergewichtung. Sie erlaubt auch, KI im Authoring nicht bloß Texte generieren zu lassen, sondern Lücken oder Inkonsistenzen in einem expliziten Alignment-Graphen vorzuschlagen.

# 11 Datenschutz, KI-Governance und Verantwortlichkeit

## 11.1 Datenfluss

WHITEEXAM verarbeitet Stammdaten, institutionelle Zuordnungen, Lern- und Prüfungsversuche, Antworten, Bewertungen, Feedback, technische Sitzungsdaten und KI-bezogene Metadaten. Nicht jede Verarbeitung benötigt dieselbe Rechtsgrundlage oder Aufbewahrungsfrist. Besonders sensibel

sind Leistungsdaten und freie Texte, die an externe KI-Provider übermittelt werden können.

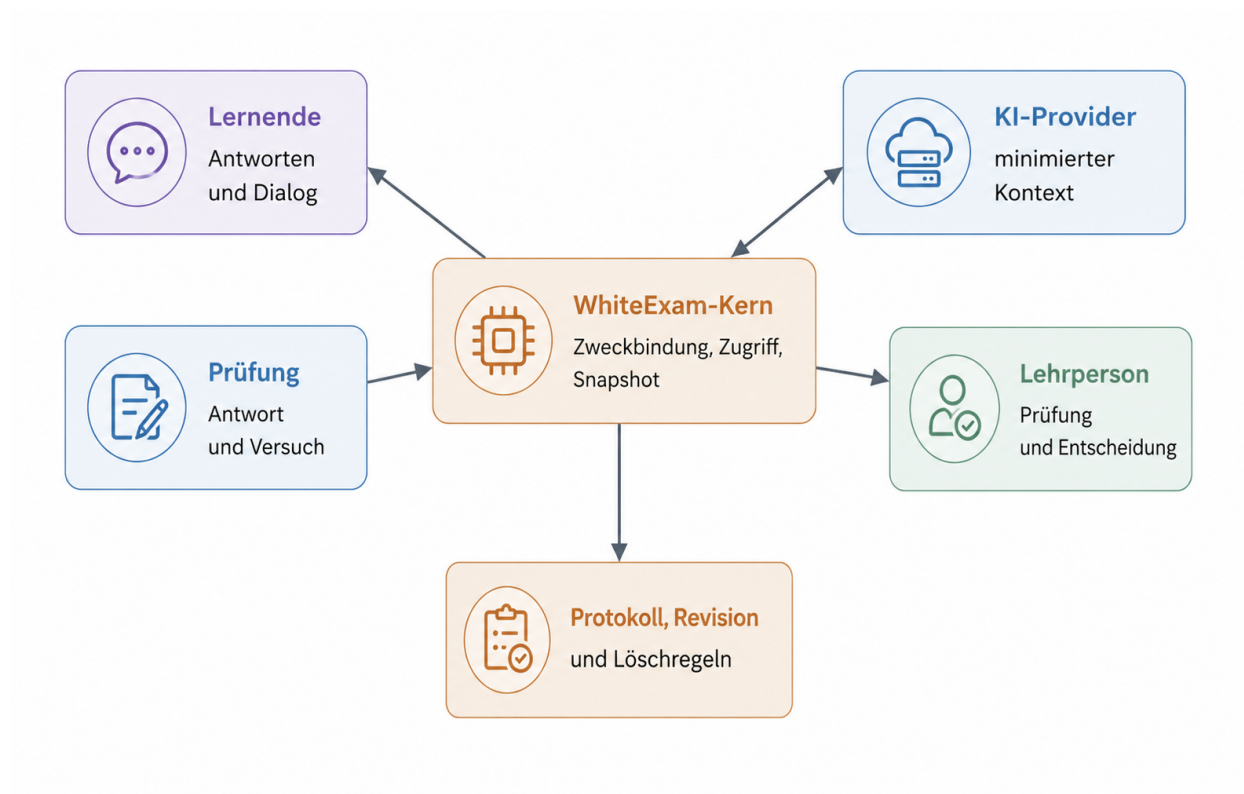


**Abbildung 15:** Datenfluss mit kontrollierter Providergrenze und menschlicher Entscheidung

## 11.2 Governance-Anforderungen

UNESCO und OECD betonen einen menschenzentrierten, zweckgebundenen und evidenzorientierten Einsatz generativer KI (Miao & Holmes, 2023; OECD, 2023, 2026). Die europäischen Leitlinien fordern AI- und Data-Literacy sowie kontextbezogene ethische Entscheidungen (European Commission, 2026). Der EU AI Act ordnet bestimmte Systeme zur Bewertung von Lernergebnissen oder zur Überwachung in Prüfungen als Hochrisikooanwendungen ein (European Union, 2024). Welche konkrete Klassifizierung gilt, hängt von Funktion, Anbieterrolle und Einsatzkontext ab; dieser Beitrag ersetzt keine Rechtsprüfung.

**Tabelle 4:** Governance-Kontrollen und Systemstatus

| Kontrolle                     | Konkretisierung                                                                                         | Status            |
|-------------------------------|---------------------------------------------------------------------------------------------------------|-------------------|
| Zwecktrennung                 | Lern-, Prüfungs- und Korrekturprozesse besitzen getrennte Tabellen und Berechtigungen.                  | <b>realisiert</b> |
| Menschliche Letztentscheidung | KI-Vorschläge werden nicht verdeckt als finale summative Bewertung gespeichert.                         | <b>realisiert</b> |
| Providerkontrolle             | Adapter, Modellwahl, <code>store:false</code> , Schema- und Fehlerstatus; keine Schlüssel in Lerndaten. | <b>realisiert</b> |
| Provenienz und Revision       | Snapshots, Modellmetadaten und Bewertungsrevisionen.                                                    | <b>realisiert</b> |
| Aufbewahrung und Löschung     | institutionell definierbare Fristen je Datenart und Export-/Löschworkflow.                              | <b>Designziel</b> |
| Fairness-Monitoring           | Vergleich von Vorschlägen, Lehrpersonenentscheidungen und Fehlern nach relevanten Gruppen.              | <b>Forschung</b>  |
| Betroffenenrechte             | verständliche Information, Einsicht, Korrektur und Widerspruch.                                         | <b>Designziel</b> |

Die DSGVO verlangt unter anderem Zweckbindung, Datenminimierung, Transparenz und angemessene Sicherheit (European Union, 2016). Für WHITEEXAM folgt daraus: Rohdaten sollen nicht vorsorglich für unbestimmte Analysen gesammelt werden; Providerkontexte sollen minimiert und dokumentiert sein; Lernanalysen dürfen nicht in verdeckte Verhaltensbewertungen umgedeutet werden; und automatisierte Vorschläge müssen durch kompetente Menschen überprüfbar bleiben.

## 12 Design-Based-Research-Agenda

### 12.1 Von Designthesen zu prüfbaren Hypothesen

Die Architektur erzeugt keine Wirksamkeit aus sich selbst. Sie formuliert Mechanismen, die empirisch geprüft werden können.

**Tabelle 5:** Forschungsprogramm für WHITEEXAM

| ID        | Designhypothese                                                                                                  | Geeignete Daten                                             | Primäres Kriterium               |
|-----------|------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------|----------------------------------|
| <b>H1</b> | Begrenzte KI-Dialoge fördern Revisionstiefe stärker als generische Chatbot-Hilfe.                                | randomisierte Vergleichsstudie, Dialogcoding, Vor-/Nachtest | selbstständiger Transfer ohne KI |
| <b>H2</b> | Die Sequenz eigene Antwort – Dialog – Revision reduziert Lösungskopien und erhöht fachliche Begründungsqualität. | Prozessdaten, Textrevisionen, Blindrating                   | Qualität der Begründung          |

| ID | Designhypothese                                                                                                                            | Geeignete Daten                                                     |        | Primäres Kriterium                        |
|----|--------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|--------|-------------------------------------------|
| H3 | Dual-Loop-Trennung erhöht die Interpretierbarkeit summarischer Ergebnisse gegenüber vermischten Lern-/Prüfungssettings.                    | Validitätsargument, Interviews, Aufgabenanalyse                     | Inter- | Vertretbarkeit der Ergebnisinterpretation |
| H4 | Aufgabenweise KI-Korrektur mit Raster und Review erhöht Konsistenz, ohne systematische Benachteiligung zu erzeugen.                        | Doppelkorrektur, Abweichungsanalyse, Fairness-Audit                 | Abwei- | Übereinstimmung und Fehlerasymmetrie      |
| H5 | Lebenszyklusdaten verbessern die Aufgabenqualität über Iterationen.                                                                        | Itemstatistik, Versionsvergleich, Lehrpersonenprotokolle            | Ver-   | Revision problematischer Items            |
| H6 | Lehrpersonenkuratierte Pfade mit Mikroadaption sind über Fächer transferierbar, benötigen aber fachspezifische Dialog- und Evidenzmodelle. | multiple Fallstudien in Physik, Informatik, Mathematik und Sprachen |        | Designinvarianz und Fachvariation         |

## 12.2 Vorgeschlagene Forschungsphasen

**Phase 1 – Usability und Implementierungsfidelität.** Beobachtungen, Think-Aloud-Studien und Loganalysen prüfen, ob Rollen, Übergänge und Hinweise wie beabsichtigt verstanden werden. Zentral ist nicht Zufriedenheit allein, sondern ob Lernende eigene Antworten produzieren und Lehrpersonen die Systemgrenzen korrekt einschätzen.

**Phase 2 – Lokale Lernmechanismen.** Kleine randomisierte oder quasiexperimentelle Studien untersuchen einzelne Dialogmuster, Hinweisstufen und Aufgabenfolgen. Primäre Endpunkte sind verzögerter Transfer und selbstständige Leistung, nicht bloß erfolgreiche Bearbeitung mit Hilfe.

**Phase 3 – Prüfungsvalidität und Korrekturgüte.** Für ausgewählte Prüfungen werden Kompetenzmodell, Aufgabenplan, Bewertungsraster, Interrater-Reliabilität und KI-Mensch-Abweichungen analysiert. Entscheidungen mit hoher Tragweite erfordern explizite Validitätsargumente.

**Phase 4 – DBR über Fächer und Institutionen.** Mehrere Lehrpersonen adaptieren dieselben Designprinzipien. Dokumentierte Iterationen zeigen, welche Komponenten stabil bleiben und welche fach- oder institutionsspezifisch verändert werden müssen.

**Phase 5 – Langzeit- und Governancefolgen.** Untersucht werden Selbstregulation, Abhängigkeit von Hilfen, Chancengerechtigkeit, Arbeitsbelastung, Datenschutzpraxis und institutionelle Verantwortlichkeit.

### 12.3 Messmodell für Learning Analytics

Ein zukünftiges Analysemodell sollte vier Ebenen strikt unterscheiden:

1. **Ereignis:** beobachtete Aktion, etwa Antwort, Hinweisaufruf oder Revision;
2. **Indikator:** regelgeleitete Verdichtung, etwa abgeschlossener Lernschritt;
3. **Interpretation:** theoriebezogene Annahme, etwa anhaltende Fehlvorstellung;
4. **Entscheidung:** pädagogische Handlung, etwa zusätzliches Material oder Gespräch.

Zwischen diesen Ebenen bestehen Unsicherheiten. Das System soll sie sichtbar machen, statt aus einem Ereignis unmittelbar eine Kompetenz oder Absicht abzuleiten. Dies folgt der Forderung, Learning Analytics theoretisch zu fundieren (Wise & Shaffer, 2015) und ethisch zu begrenzen (Slade & Prinsloo, 2013).

## 13 Diskussion

### 13.1 Wissenschaftlicher Beitrag

Der Beitrag von WHITEEXAM liegt in vier miteinander verbundenen Konstrukten.

Erstens präzisiert die *Dual-Loop-Architektur* das Verhältnis von Lernen und Prüfen. Die Schleifen sind curricular gekoppelt, aber hinsichtlich Versuchen, Hilfen, Daten und Entscheidungen getrennt. Dies operationalisiert Constructive Alignment, ohne formative und summative Funktionen gleichzusetzen.

Zweitens erweitert der *Aufgabenlebenszyklus* die Betrachtung einzelner Items. Aufgaben entstehen aus Materialien, werden importiert, kuratiert, versioniert, in Lern- und Prüfungskontexten verwendet, bewertet und anhand von Evidenz revidiert. KI kann an mehreren Stellen assistieren, bleibt aber an Status-, Review- und Freigaberegeln gebunden.

Drittens wird der *dialogische KI-Tutor* nicht als freier Chatbot, sondern als hybrides System modelliert: Das Sprachmodell formuliert und diagnostiziert; der Plattformkern kontrolliert Zustand, Belege, Grenzen und Abschluss. Diese Arbeitsteilung verbindet sprachliche Flexibilität mit überprüfbaren pädagogischen Regeln.

Viertens verknüpft das *Governance-Modell* didaktische Freiheit mit institutioneller Verantwortung. Lehrpersonen behalten Autorschaft und Urteil, handeln jedoch innerhalb nachvollziehbarer Rollen, Freigaben, Snapshots und Revisionspfade.

### 13.2 Spannungsfelder

**Freiheit versus Standardisierung.** Lehrpersonen benötigen Gestaltungsspielraum; Prüfungsentscheidungen benötigen zugleich Vergleichbarkeit und dokumentierte Kriterien. WHITEEXAM löst den Konflikt nicht abstrakt, sondern macht konfigurierbare Entscheidungen sichtbar.

**Personalisierung versus Eigenaktivität.** Hilfe kann Überforderung reduzieren, aber auch notwendige Anstrengung ersetzen. Deshalb wird Adaptivität als abgestufte Unterstützung und nicht als möglichst schnelle Lösungsproduktion verstanden.

**Datenreichtum versus Datenschutz.** Prozessdaten können Diagnosen verbessern, erzeugen aber Überwachungsrisiken. Das Design bevorzugt minimale, zweckgebundene Indikatoren und menschliche Interpretation.

**KI-Effizienz versus Bewertungsvalidität.** Automatisierte Vorschläge können Korrekturarbeit

strukturieren. Je höher die Konsequenz einer Entscheidung, desto stärker müssen Referenzgrundlage, menschliche Prüfung, Fairnessanalyse und Revisionsmöglichkeit sein.

**Integration versus Systemkomplexität.** Eine durchgängige Plattform reduziert Medienbrüche, vergrößert aber den Verantwortungs- und Sicherheitsumfang. Der modulare Monolith ist für Pilotierung nachvollziehbar; institutionelle Skalierung erfordert zusätzliche Betriebs-, Schnittstellen- und Rechtearchitektur.

### 13.3 Grenzen

Erstens ist WHITEEXAM ein entwickeltes Artefakt, aber noch kein empirisch validiertes Gesamtmodell. Die vorhandenen Daten erlauben keine kausalen Aussagen über Lernzuwachs, Prüfungsqualität oder Entlastung.

Zweitens stammt die Architektur aus einem konkreten schulischen Entwicklungskontext und ist stark durch naturwissenschaftliche und informatische Anforderungen geprägt. Fachübertragbarkeit ist plausibel, aber nicht belegt.

Drittens ist dynamische Pfadadaption noch nicht vollständig implementiert. Der gegenwärtige Lernpfad ist kuratiert; Adaptivität liegt vor allem im Dialog und in lehrpersoneller Steuerung.

Viertens fehlt ein ausgebautes Kompetenzmodell mit expliziten Niveaustufen, Evidenzregeln und Unsicherheit. Ohne dieses Modell dürfen Aktivitäts- und Prüfungsdaten nicht zu umfassenden Kompetenzprofilen überinterpretiert werden.

Fünftens ist der KI-Betrieb providerabhängig. Modelländerungen können Verhalten, Kosten und Fehlerprofile verändern. Erforderlich sind versionierte Evaluationssätze, Monitoring und Fallbacks.

Sechstens ersetzt die technische Human-in-the-Loop-Struktur keine rechtliche oder ethische Zulässigkeitsprüfung. Institutionen müssen Verarbeitungszwecke, Verträge, Aufbewahrung, Betroffenenrechte und Verantwortlichkeiten selbst klären.

## 14 Ausblick

Die nächste Entwicklungs- und Forschungsetappe sollte nicht möglichst viele KI-Funktionen hinzufügen, sondern die begründeten Beziehungen des Designs vervollständigen.

Priorität besitzt ein explizites Kompetenzmodell. Es sollte Kompetenzen, Niveaustufen, Evidenzregeln, Lernaktivitäten und Prüfungsisems verbinden und Unsicherheit sichtbar halten. Darauf können adaptive Pfade aufbauen, die Vorschläge machen, aber lehrpersonell übersteuerbar bleiben.

Zweitens benötigt die KI-Korrektur eine formale Evaluationsumgebung. Ein versionierter Goldstandard mit Doppelratings, Fehlerkategorien und Fairnessprüfungen sollte Modell- und Promptänderungen vor produktivem Einsatz testen.

Drittens sollte der Import bestehender Materialien um nachvollziehbare Dokumentsegmente, Formelerkennung, Asset-Provenienz und Freigabesperren erweitert werden. KI-generierte Aufgaben bleiben Entwürfe, bis fachliche und didaktische Prüfung abgeschlossen ist.

Viertens sind offene Standards und Schnittstellen wichtig, damit WHITEEXAM nicht zum isolierten Datensilo wird. Exporte, institutionelle Identitäten, Aufgabenstandards und Forschungsdatenpakete benötigen kontrollierte, datensparsame Schnittstellen.

Fünftens sollte ein DBR-Verbund mit mehreren Lehrpersonen, Fächern und Schulen aufgebaut werden. Der Grundsatzartikel liefert dafür gemeinsames Vokabular, Designprinzipien und Hypothesen;

spätere Publikationen können einzelne Mechanismen empirisch untersuchen.

## 15 Fazit

WHITEEXAM wird in diesem Beitrag als didaktische Infrastruktur und Forschungsartefakt beschrieben. Die zentrale Innovation ist nicht ein einzelner Chatbot, ein Prüfungseditor oder eine Korrekturfunktion. Sie liegt in der kontrollierten Verbindung des gesamten Aufgaben-, Lern- und Prüfungslebenszyklus.

Die Dual-Loop-Architektur hält formative Unterstützung und summativen Nachweis auseinander, ohne sie curricular zu entkoppeln. Gemeinsame, versionierte fachliche Artefakte ermöglichen Wiederverwendung; getrennte Versuche, Hilfen, Daten und Entscheidungen schützen die jeweilige Funktion. Der dialogische Tutor wird durch serverseitige Zustände und Grenzen gerahmt. Kompetenzorientierte Prüfungen werden als Evidenzarchitekturen modelliert. KI unterstützt Authoring, Dialog, Korrektur und Analyse, während die Lehrperson die fachliche, didaktische und normative Verantwortung behält.

Der wissenschaftliche Anspruch des Designs besteht gerade darin, seine Reichweite zu begrenzen. Implementierung ist kein Wirksamkeitsnachweis, Aktivitätsdaten sind keine Kompetenz, und Human-in-the-Loop ist keine automatische Garantie fairer Entscheidungen. Aus diesen Grenzen folgt ein konkretes Forschungsprogramm. Wenn die formulierten Mechanismen in unterschiedlichen Fächern und Institutionen systematisch untersucht werden, kann WHITEEXAM zur Grundlage kumulativer Forschung über KI-Dialoge, Assessment-Validität, Learning Analytics und schulische KI-Governance werden.

LERNPLATTFORM · LERNUMGEBUNGEN

Neue Lernumgebung erstellen

Zu den Freigaben

Zur Analyse

Zurück zum Dashboard

# Lernumgebungen

Ordne veröffentlichte Lernbausteine und gib sie gezielt für Klassen, Gruppen oder Einzelpersonen frei.

Lernumgebungen

Freigaben

LERNUMGEBUNG

Aktiv

## Mechanik 1 – Einführung, Trägheit und Kräfte

Ein geführter Lernweg zur Entwicklung der Mechanik, zum Trägheitsprinzip, zum Fadenexperiment und zu den Wirkungen von Kräften. Fachliche Grundlage: „Phy1, 1. Mechanik (Script).pdf“.

|                            |                                                |
|----------------------------|------------------------------------------------|
| Lernbausteine<br><b>16</b> | Freigaben<br><b>1</b>                          |
| Lernversuche<br><b>18</b>  | Zuletzt geändert<br><b>2026-07-28 20:04:24</b> |

Zeitraum: sofort – offen

Letzte Lernaktivität: 2026-07-28 21:24:22

Fortschritt ansehen

Vorschau

Freigaben

Bearbeiten

LERNUMGEBUNG

Aktiv

## Kapitel 1 - Python: Listen und NumPy-Arrays

Pilot/Test-Lernpfad aus dem importierten Kurs „Informatik Grundlagenfach (Ws)“. Einstieg in NumPy, Listen und Arrays mit drei Quizfragen und zwei sokratischen KI-Dialogen.

|                           |                                                |
|---------------------------|------------------------------------------------|
| Lernbausteine<br><b>6</b> | Freigaben<br><b>1</b>                          |
| Lernversuche<br><b>22</b> | Zuletzt geändert<br><b>2026-07-28 11:30:38</b> |

Zeitraum: sofort – offen

Letzte Lernaktivität: 2026-07-28 14:25:21

Fortschritt ansehen

Vorschau

Freigaben

Bearbeiten

**Abbildung 16:** Lehrpersonenansicht veröffentlichter Lernumgebungen als Ausgangspunkt für fächerübergreifende DBR-Zyklen

29

## A Abbildungsinventar und Reproduzierbarkeit

Die Screenshots wurden am 29. Juli 2026 aus der lokalen Entwicklungsinstanz erzeugt. Sie dokumentieren den Systemstand und sind keine Nutzungsdatenstudie.

**Tabelle 6:** Verfügbare UI-Artefakte der Referenzfassung

| Datei                       | Inhalt                                           | Verwendung |
|-----------------------------|--------------------------------------------------|------------|
| teacher-dashboard.png       | Funktionsbereiche Prüfung, Lernen und Verwaltung | Abb. 4     |
| student-dashboard.png       | Stammklasse, Kurse und Lernangebote              | Abb. 9     |
| student-learning-path.png   | Sequenz aus Quiz, offener Aufgabe und KI-Dialog  | Abb. 6     |
| quiz-preview.png            | Auswahlfrage, Lösung und Erklärung               | Abb. 7     |
| ai-dialog-preview.png       | Tutoraufgabe, Lernziel und Dialogsteuerung       | Abb. 7     |
| exam-editor.png             | Prüfungsdaten, SEB und Prüfungsteile             | Abb. 11    |
| collections.png             | Sammlungsverwaltung                              | Abb. 13    |
| class-course-management.png | Klassen- und Kursverwaltung                      | Abb. 13    |
| learning-environments.png   | veröffentlichte Lernumgebungen                   | Abb. 16    |

## B Traceability-Matrix der Designprinzipien

**Tabelle 7:** Bezug zwischen Theorie, Design und Evaluationsbedarf

| Theoretischer Bezug        | Designentscheidung                                                               | Offene Prüfung                                              |
|----------------------------|----------------------------------------------------------------------------------|-------------------------------------------------------------|
| Constructive Alignment     | gemeinsame Kompetenz- und Aufgabenartefakte; getrennte Lern- und Prüfungsnutzung | curriculare Abdeckung und Validitätsargument                |
| Formative Rückmeldung      | gestufte Hinweise, Revision, keine Note im Dialog                                | Qualität, Timing und Transfer der Rückmeldung               |
| Selbstregulation           | eigener Versuch vor Hilfe; Reflexionsphase                                       | Abhängigkeit von Hilfe und langfristige Selbstständigkeit   |
| Intelligente Tutorsysteme  | strukturierter Dialog und Mikroadaptation                                        | Vergleich mit generischem Chat und menschlicher Rückmeldung |
| Learning Analytics         | Ereignis – Indikator – Interpretation – Entscheidung                             | diagnostische Güte, Fairness und Verständlichkeit           |
| Assessment-Validität       | Prüfungsplan, Ressourcenregeln, Raster und Review                                | Vertretbarkeit der Ergebnisinterpretation                   |
| Human-in-the-Loop          | KI-Vorschlag mit menschlicher Revision                                           | Automationsbias, Interrater-Güte und Arbeitsbelastung       |
| Datenschutz und Governance | Zwecktrennung, minimale Providerdaten, Rollen und Protokolle                     | institutionelle Rechts- und Folgenabschätzung               |

## Literatur

- Ali, R., & Maroulis, J. (2026). Assessment Validity in the Age of Generative Artificial Intelligence: A Critical Review [Online first]. *Assessment & Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2026.2683046>
- American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences*, 122(14), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjaw, R. (2024). Developing Evaluative Judgment for a Time of Generative Artificial Intelligence. *Assessment & Evaluation in Higher Education*, 49(6), 893–905. <https://doi.org/10.1080/02602938.2024.2335321>
- Biggs, J. (1996). Enhancing Teaching through Constructive Alignment. *Higher Education*, 32(3), 347–364. <https://doi.org/10.1007/BF00138871>
- Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University* (4. Aufl.). Open University Press.
- Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Bower, M., Torrington, J., Lai, J. W. M., Petocz, P., & Alfano, M. (2024). How Should We Change Teaching and Assessment in Response to Increasingly Powerful Generative Artificial Intelligence? Outcomes of the ChatGPT Teacher Survey. *Education and Information Technologies*, 29, 15403–15439. <https://doi.org/10.1007/s10639-023-12405-0>
- Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal of the Learning Sciences*, 2(2), 141–178. [https://doi.org/10.1207/s15327809jls0202\\_2](https://doi.org/10.1207/s15327809jls0202_2)
- Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher*, 32(1), 9–13. <https://doi.org/10.3102/0013189X032001009>
- Design-Based Research Collective. (2003). Design-Based Research: An Emerging Paradigm for Educational Inquiry. *Educational Researcher*, 32(1), 5–8. <https://doi.org/10.3102/0013189X032001005>
- Drachsler, H., & Greller, W. (2016). Privacy and Analytics: It's a DELICATE Issue. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 89–98. <https://doi.org/10.1145/2883851.2883893>
- European Commission. (2026). *Guidelines on the Ethical Use of Artificial Intelligence and Data in Teaching and Learning for Educators* (Guidelines). European Commission. Brussels. <https://education.ec.europa.eu/focus-topics/digital-education/actions/plan/ethical-guidelines-for-educators-on-using-artificial-intelligence>
- European Union. (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation)* (Regulation). European Union. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- European Union. (2024). *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence* (Regulation). European Union. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

- Ferguson, R. (2012). Learning Analytics: Drivers, Developments and Challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304–317. <https://doi.org/10.1504/IJTEL.2012.051816>
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing a Novel Research-Based Design in an Authentic Educational Setting. *Scientific Reports*, 15, 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, 86(1), 42–78. <https://doi.org/10.3102/0034654315581420>
- Mayer, R. E. (2009). *Multimedia Learning* (2. Aufl.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511811678>
- McKenney, S., & Reeves, T. C. (2019). *Conducting Educational Design Research* (2. Aufl.). Routledge. <https://doi.org/10.4324/9781315105642>
- Miao, F., & Holmes, W. (2023). *Guidance for Generative AI in Education and Research* (Guidance). UNESCO. Paris. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- OECD. (2023). *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem* (Report). OECD Publishing. Paris. <https://doi.org/10.1787/c74f03de-en>
- OECD. (2026). *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education* (Report). OECD Publishing. Paris. <https://doi.org/10.1787/062a7394-en>
- Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist*, 57(10), 1380–1400. <https://doi.org/10.1177/0002764213498851>
- Slade, S., & Prinsloo, P. (2013). Learning Analytics: Ethical Issues and Dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529. <https://doi.org/10.1177/0002764213479366>
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- Williamson, B., & Eynon, R. (2020). Historical Threads, Missing Links, and Future Directions in AI in Education. *Learning, Media and Technology*, 45(3), 223–235. <https://doi.org/10.1080/17439884.2020.1798995>

- Wise, A. F., & Shaffer, D. W. (2015). Why Theory Matters More than Ever in the Age of Big Data. *Journal of Learning Analytics*, 2(2), 5–13. <https://doi.org/10.18608/jla.2015.22.2>
- Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A Scoping Review on How Generative Artificial Intelligence Transforms Assessment in Higher Education. *International Journal of Educational Technology in Higher Education*, 21, 40. <https://doi.org/10.1186/s41239-024-00468-z>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, 41(2), 64–70. [https://doi.org/10.1207/s15430421tip4102\\_2](https://doi.org/10.1207/s15430421tip4102_2)