WHITEPAPER · SCHULISCHE KI

DigiLernFlow

Verantwortungsvolle KI und RAG im schulischen Einsatz

Datenschutz, Schüler:innenschutz, pädagogische Leitplanken und technische Freigabekriterien

Fassung 0.5 · 10. August 2026

STATUS: FORTGESCHRITTENE TECHNISCHE FREIGABEBASIS

Dieses Whitepaper beschreibt den derzeitigen Entwicklungsstand und das verbindliche Zielmodell für KI-gestützte Inhaltskartengenerierung und KI-Dialoge mit Lernenden in DigiLernFlow. Es richtet sich an Schulen, Schulträger, Datenschutzbeauftragte, Lehrkräfte, technische Administration und Entwicklung.

Kernaussage. DigiLernFlow verfügt über eine zweckgebundene, kontextisolierte RAG-Grundlage, einen zentralen Safety-Gateway, datensparsame Coach-Lernspuren, eine echte Quellen-/Chunk-Löschung, konfigurierbare Aufbewahrung und eine sichtbare Moduskennzeichnung. Der allgemeine Software-Release kann deshalb unabhängig vom generativen Schülerdialog erfolgen. Generative Lernendenfunktionen bleiben jedoch pro Instanz deaktiviert, bis Anbieter- und Vertragsprüfung, DSFA-Entscheidung, betriebliche Lösch- und Safety-Nachweise sowie pädagogische Freigabe vorliegen.

Aktualisierung vom 9. August 2026

Seit Fassung 0.3 wurden die zuvor ausdrücklich benannten technischen Datenschutzlücken geschlossen:

- Der Taskkarten-Coach speichert Lernendenfragen weder im Klartext noch als reproduzierbaren Inhalts-Hash; die Lernspur enthält nur erforderliche Kontextdaten und eine grobe Längenklasse.
- Quellenlöschung entfernt Datei und Metadaten sowie Registry-Eintrag, Bindungen und RAG-Chunks. Archivierte Quellen und Coach-Ereignisse können mit einem standardmäßig datensparsamen, zunächst nur anzeigenden Aufbewahrungsjob bereinigt werden.
- Die Oberfläche kennzeichnet vor der Interaktion eindeutig „KI-gestützter Coach“ oder „Regelbasierte Hilfe ohne generative KI“ und benennt Grenzen der KI-Ausgabe.
- Die Datenschutzerklärung beschreibt KI/RAG, Datenkategorien, Löschung, Empfänger- und Rollenklärung. Ergänzende Vorlagen dokumentieren DSFA-Vorprüfung, Auftragsverarbeitung und Release-Evidenz.

Diese Umsetzung ist kein automatischer Rechts- oder Organisationsnachweis. Vor Aktivierung eines externen Modells sind die konkreten Anbieterbedingungen, Verarbeitungsregion, Unterauftragnehmer, Rechtsgrundlage und Aufbewahrung sowie der reale Instanzbetrieb zu prüfen. Eine vollständige Nutzer-, Goal- und Backup-Löschprobe und ein altersprofilbezogener Red-Team-Test bleiben Abnahmekriterien.

Kurzfazit: Freigabeampel

DigiLernFlow hat die Phase einer bloßen Konzept- oder Frühimplementierung verlassen. Der KI-freie Kernbetrieb sowie wesentliche technische Datenschutz- und Sicherheitsmechanismen sind produktionsnah umgesetzt. Auch die KI-Funktionen arbeiten bereits in kontrollierten, kontextgebundenen und fail-closed abgesicherten Pfaden. Offen sind vor allem Nachweise und Entscheidungen, die nicht pauschal durch Softwarecode ersetzt werden können, sondern je Schule, Betreiber, Modellanbieter und Instanz verbindlich festzulegen sind.

Bereich	Erreichter Stand	Bewertung
Kernanwendung und IT-Basis	Produktionsnaher Kernbetrieb mit Zugriffsschutz, Instanztrennung, Diagnose, Release-Tests und KI-unabhängiger Nutzbarkeit.	● GRÜN
Datenschutz ohne externe KI	Transparenz, Datenminimierung, echte Quellen- und Chunk-Löschung sowie konfigurierbare Aufbewahrung sind technisch umgesetzt.	● GRÜN
KI für Lehrkräfte	Vier kontrollierte Erstellungspfade, zentrale Safety-Prüfung und fail-closed Validierung; Ergebnisse gelangen nur als editierbare Entwürfe in die Fach-Editoren.	● GRÜN
Generativer Schüler-Coach	Taskkarten- und Goal-Bindung, PII-, Injection- und Safety-Prüfung, minimierte Lernspur, sichtbare KI-Kennzeichnung und gestufte Übergabe an die Lehrkraft sind umgesetzt. Die Aktivierung bleibt instanzbezogen.	● ● GRÜN-GELB
Governance, Datenschutz und AI Act	Rollen-, DSFA- und Evidenzvorlagen, Transparenzinformationen und technische Nachweisstrukturen liegen vor. Konkrete Rechtsgrundlage, Anbieterprüfung und schulische Freigabe bleiben organisationsbezogen zu dokumentieren.	● ● GRÜN-GELB
Regulärer KI-Einsatz einer konkreten Instanz	Möglich nach Anbieter- und Vertragsprüfung, DSFA-Entscheidung, dokumentiertem Löschbetrieb, altersbezogenem Safety-Test und pädagogischer Freigabe.	● GELB

Gesamtbewertung. Der allgemeine Plattformbetrieb besitzt eine belastbare technische Freigabebasis. Generative Funktionen sind nicht mehr grundsätzlich blockiert, sondern kontrolliert vorbereitet. Der verbleibende gelbe Bereich ist ein geregelter Freigabeprozess pro Instanz - kein pauschales technisches Defizit. Eine rote Bewertung besteht derzeit nicht; sie würde erst bei einem echten Einsatzblocker, einer fehlgeschlagenen Schutzkontrolle oder einer unzulässigen Konfiguration vergeben.

Dokumentstatus und Aussagegrenzen

Dieses Dokument ist eine technische und organisatorische Arbeitsgrundlage, keine Rechtsberatung und kein Konformitätsnachweis. Die konkrete Rechtsgrundlage, Zuständigkeit und schulrechtliche

Zulässigkeit sind für Bundesland, Schulform, Altersgruppe, Einsatzszenario und ausgewählte KI-Dienstleister gesondert zu prüfen.

Die Statusbegriffe werden im gesamten Dokument einheitlich verwendet:

Kennzeichnung	Bedeutung
IST	Im lokal geprüften Entwicklungsstand vorhanden. Eine produktive Serverprüfung und organisatorische Freigabe können dennoch ausstehen.
FREIGABEGATE	Muss vor einem regulären Lernendeneinsatz umgesetzt, dokumentiert und getestet sein.
ZIEL	Weiterführende Reife- oder Qualitätsstufe; sinnvoll, aber nicht in jedem Fall zwingende Voraussetzung für einen eng begrenzten Pilotbetrieb.

1. Executive Summary

DigiLernFlow nutzt Retrieval-Augmented Generation (RAG), um generative KI nicht mit einem unkontrollierten Gesamtdatenbestand, sondern mit einem begrenzten und fachlich ausgewählten Kontext zu versorgen. Zwei Anwendungsfälle sind vorgesehen:

1. Lehrkräfte erzeugen Inhaltskarten aus Thema, Fach, Lernziel und weiteren Eingaben. Dafür werden ausschließlich Quellen berücksichtigt, die für `content_generation` freigegeben sind. Bis zu sechs relevante Volltextabschnitte gelangen in den Generierungskontext.
2. Lernende führen einen KI-gestützten Dialog innerhalb eines Goals oder einer Taskkarte. Dafür werden ausschließlich Quellen durchsucht, die von einer Lehrkraft ausdrücklich für `learner_dialog` an genau diesen Kontext gebunden wurden. Taskkartenbindungen haben Vorrang vor allgemeinen Goal-Bindungen; bis zu vier relevante Abschnitte werden an das Coach-Modell übergeben.

Diese Architektur ist eine gute Ausgangsbasis für Zweckbindung und Mandantentrennung. RAG ist jedoch kein Sicherheitsfilter. Ungeeignete oder personenbezogene Inhalte können bereits in Referenzdokumenten, in Lernendenfragen, in abgerufenen Passagen oder in Modellausgaben enthalten sein. Auch ein fachlich korrektes Referenzdokument kann durch Prompt-Injection missbraucht oder in einem unpassenden Alterskontext ausgegeben werden.

Die Freigabestrategie folgt deshalb dem Prinzip **Teacher-bounded RAG plus Safety Gateway**:

- Lehrkräfte bestimmen Quelle, Zweck und Lernkontext.
- Jede Quelle durchläuft technische Vorprüfung, Quarantäne und bewusste Freigabe.
- Lernendenfragen werden auf personenbezogene Daten, Sicherheitsrisiken und Manipulationsversuche geprüft.

- Nur minimal erforderliche Daten werden an externe Modelle übertragen; Modellanbieter dürfen sie weder zum Training noch zu eigenen Zwecken verwenden.
- Die Modellausgabe wird vor Anzeige auf Sicherheit, Altersangemessenheit, Datenschutz und pädagogische Rolle geprüft.
- Hochrisikofälle werden nicht durch eine scheinbar therapeutische KI beantwortet, sondern nach einem schulisch festgelegten, transparenten Schutzprozess behandelt.
- Protokollierung dient Nachweis und Schutz, nicht pauschaler Verhaltensüberwachung.

Die Plattformvergleiche liefern dafür vier besonders nützliche Muster: MagicSchool beschreibt einen fortlaufenden Sicherheitskreislauf aus Eingabe-Rahmung, Ausgabeprüfung, Tests und Verbesserung; Khanmigo kombiniert Moderation mit transparenter Information und abgestufter Eskalation; SchoolAI betont, dass ein Lernassistent weder menschliche Beziehung simulieren noch Nutzungsdauer maximieren soll; fobizz zeigt für Deutschland die praktische Trennung von Schule als Verantwortlicher und Plattform als Auftragsverarbeiter sowie pseudonyme Zugänge und Filter gegen personenbezogene Angaben. Diese Aussagen sind Referenzmuster der jeweiligen Anbieter und ersetzen keine eigene Prüfung oder Zertifizierung.

2. Gegenstand und Systemgrenze

2.1 Im Whitepaper erfasste Funktionen

Das Schutzmodell umfasst den gesamten Datenweg folgender Funktionen:

- Upload und Verwaltung von Plan-, Ideen- und Materialquellen;
- Extraktion, Segmentierung und Suche in Quellenabschnitten;
- RAG für Inhaltskartengenerierung durch Lehrkräfte;
- RAG für KI-Dialoge mit Lernenden;
- Übergabe von Prompts, Kontexten und Passagen an ein Sprachmodell;
- Speicherung von Quellen, Bindungen, Dialog- oder Lernspuren;
- Administration, Diagnose, Export, Berichtigung und Löschung.

Nicht Gegenstand dieser Fassung sind automatisierte Zulassungs-, Benotungs- oder Prüfungsentscheidungen. Sobald KI-Ausgaben über Zugang, Bewertung, Einstufung oder wesentliche Bildungsentscheidungen bestimmen sollen, ist eine gesonderte Risikoklassifizierung nach der KI-Verordnung erforderlich.

2.2 Schutzobjekte

Zu schützen sind insbesondere:

- Identität, Privatsphäre und persönliche Integrität von Schüler:innen;

- sensible Angaben in Lernendenfragen und Lernspuren;
- urheberrechtlich oder lizenzrechtlich geschützte Quellen;
- schulinterne Planungs- und Unterrichtsmaterialien;
- die Trennung zwischen Nutzer:innen, Goals, Taskkarten und Quellen;
- fachliche Richtigkeit, Quellenbezug und pädagogische Angemessenheit;
- Verfügbarkeit, Nachvollziehbarkeit und Löschbarkeit des Systems.

3. Architekturprinzip: Teacher-bounded RAG

3.1 Inhaltskartengenerierung

IST. Aus Thema, Fach, Lernziel und ergänzenden Angaben wird eine Suchanfrage gebildet. Berücksichtigt werden Quellen mit dem Zweck `content_generation`. Die relevantesten Abschnitte werden gesucht; bis zu sechs Volltextabschnitte gelangen in den Generierungskontext. Seiten- und Quellenherkunft bleiben intern erhalten. Damit wird nicht mehr pauschal nur der Dateianfang verwendet.

Schutzwirkung. Die Auswahl relevanter Abschnitte verringert irrelevante Datenübertragung und verbessert die fachliche Fundierung. Die Zweckmarkierung verhindert, dass jede hochgeladene Quelle automatisch zur Generierung genutzt wird.

Restgefahr. Ohne Quellenprüfung können ungeeignete, manipulierte, personenbezogene oder lizenzwidrige Inhalte abgerufen werden. Ein Prompt kann trotz RAG halluzinieren oder Quellen widersprechen.

IST-Umsetzung vom 25./26. Juli 2026: vier direkte KI-Erstellungspfade. Der zunächst vertikal abgenommene Lernkartenpfad wurde als gemeinsames Muster auf genau vier kanonische Packs übertragen: `dlf-learningcard-json-v1`, `dlf-imagecard-concept-v1`, `dlf-htmldoc-worksheet-v1` und `dlf-learningpath-quiz-v1`. Verbunden sind jeweils Promptkomposition mit freigegebenem Fach- und Quellenkontext, zentraler zweistufiger KI-Nutzungsmodus, Provideraufruf, PII-/Injection-Prüfung und Kinderschutz vor und nach dem Provider, strikte zieltypspezifische Validierung, Normalisierung, Vorschau und kontrollierter Import in den normalen Editor. Die Providerantwort wird nie ungeprüft als importfähig behandelt. Bildkarten verwenden ausschließlich einen bereits vorhandenen lokalen Uploadpfad; der Modellpfad erzeugt weder ein neues Bild noch eine fremde Bild-URL. PDF, eBook und Moodle sind ausdrücklich nicht für direkte KI-Inhaltskartenerstellung freigegeben. Deterministische Vertragstests prüfen alle vier Pfade ohne externen Modellaufruf.

Die ersten realen Providerläufe für Bildkarte, Arbeitsblatt und Quiz wurden am 26. Juli 2026 kontrolliert ohne Import ausgeführt. Provider, Safety und Validierung waren erreichbar; alle drei technisch vertragsabweichenden Antworten wurden fail-closed vor dem Import gesperrt. Beobachtet wurden

`didactic.usage` statt `usage_hint`, fehlende konkrete Ergebnisformen in drei internen Arbeitsblatt-Prompts sowie Quiz-Aliasformen (`text` statt `content`, `answers`/Array statt zeilengetrenntem `options`, fehlender abschließender `trace_type`). Die Vertragsanweisungen wurden präzisiert. Eine enge, protokollierte Normalisierung darf ausschließlich diese eindeutigen technischen Alias- und Formabweichungen reparieren; unbekannte Werte oder fachliche Abweichungen bleiben sichtbar und importgesperrt. Eine erneute reale Abnahme der drei korrigierten Pfade bleibt erforderlich.

Verbindliche Qualitätspriorität. Bei automatisch erzeugten Inhaltskarten steht die technische Robustheit an erster Stelle: kanonisches Schema, Pflichtfelder, sichere Referenzen, Import-, Editor-, Render- und Laufzeitstabilität müssen vor jeder Freigabe nachgewiesen sein. Erst danach werden pädagogische Passgenauigkeit, Anforderungsbereiche, Differenzierung und didaktische Qualität optimiert. Eine technisch fehlerhafte Ausgabe bleibt auch bei gutem Inhalt gesperrt. Eine technisch valide Ausgabe darf dagegen als klar erkennbarer, vollständig editierbarer Entwurf in den Fach-Editor gelangen; die Lehrkraft kann den Inhalt dort verbessern. Diese Reihenfolge rechtfertigt nur eindeutige technische Normalisierungen und niemals eine verdeckte fachliche Umdeutung.

Technische Umsetzung. Die Promptgenerator-Validierung weist den harten Technikstatus separat als `technical_valid` und `technical_status` aus. Pädagogische Befunde erscheinen als `quality_status` und `quality_warnings`; `quality_blocking` ist definitionsgemäß `false`. Das Importgate verwendet ausschließlich die technische Freigabe. Die Oberfläche zeigt beide Ebenen getrennt, damit eine Lehrkraft einen robusten Entwurf importieren und anschließend inhaltlich verbessern kann, ohne technische Fehler zu übergehen.

Im ersten realen Wiederholungslauf bestand die Bildkarte den vollständigen Technikvertrag. Arbeitsblatt und Quiz blieben wegen weiterer eindeutiger Provider-Formabweichungen fail-closed: leere Titel der internen Arbeitsblatt-Prompts und `data.prompt` statt `data.question` bei Multiple-Choice-Nodes. Auch diese Fälle werden ausschließlich als protokollierte technische Kanonisierung behandelt; pädagogischer Inhalt und unbekannte Abweichungen bleiben unangetastet.

Nach Einspielen dieser Kanonisierungen bestanden Bildkarte, Arbeitsblatt und Quiz die abschließenden realen Providerläufe jeweils ohne Import. Beim Quiz wurden fehlende Präzisierungen des Quellenbezugs (`source_note`) als pädagogische Hinweise ausgewiesen, ohne die technische Import- und Editorfreigabe zu blockieren. Damit ist die priorisierte Zweistufenlogik im Realbetrieb nachgewiesen.

3.2 KI-Dialoge mit Lernenden

IST. Lehrkräfte können eine Quelle für ein vollständiges Goal oder eine bestimmte Taskkarte freigeben. Im Dialog dient die aktuelle Lernendenfrage als Suchanfrage. Nur Quellen mit `learner_dialog` und passender Kontextbindung werden durchsucht. Taskkartenbindungen haben Vorrang; fremde Goals,

Taskkarten und Quellen werden ausgeschlossen. Bis zu vier relevante Passagen werden an das Coach-Modell übergeben.

Schutzwirkung. Der Dialog wird fachlich auf den von der Lehrkraft gesetzten Lernraum begrenzt. Die explizite Bindung ist stärker als eine globale Quellenfreigabe und bildet eine nachvollziehbare pädagogische Verantwortungsentscheidung ab.

Restgefahr. Die Lernendenfrage und die ausgewählten Passagen verlassen bei einem externen Modell grundsätzlich die lokale Systemgrenze. Ohne zusätzliche Filter können persönliche Angaben, Krisensignale, ungeeignete Fragen, Prompt-Injection oder unsichere Antworten verarbeitet beziehungsweise angezeigt werden.

3.3 Verbindliche Kontrollkette

Stufe	Verarbeitung	Mindestkontrolle	Block- oder Rücksprungbedingung
1	Quelle wird hochgeladen	Dateityp-, Größen-, Schadcode- und Extraktionsprüfung	technische Auffälligkeit oder nicht unterstütztes Format
2	Quelle wird analysiert	PII-, Inhalts-, Lizenz- und Prompt-Injection-Prüfung	sensible Daten, ungeeigneter Inhalt oder aktive Instruktionen
3	Lehrkraft gibt frei	Zweck, Fach, Altersprofil, Goal/Taskkarte und Sichtbarkeit	fehlende bewusste Freigabe
4	Lernendenfrage trifft ein	PII-, Safety-, Krisen- und Injection-Prüfung	verbotene Kategorie, akute Gefahr oder Datenminimierungsproblem
5	RAG sucht Passagen	harte Mandanten-, Zweck- und Kontextfilter; Relevanzschwelle	keine zulässige beziehungsweise hinreichend relevante Passage
6	Modell generiert	minimierter Prompt, definierte Rolle, Anbieter ohne Training/Retention	Anbieter- oder Konfigurationsanforderung nicht erfüllt
7	Antwort wird geprüft	Safety-, Alters-, PII-, Quellen- und Rollenprüfung	unsichere, unbelegte oder rollenwidrige Ausgabe
8	Anzeige und Nachweis	KI-Kennzeichnung, Feedback, minimale Auditspur, geregelte Löschung	fehlende Transparenz oder unzulässige Speicherung

3.4 Kosten- und Betriebsmodi der Instanz

IST-Teilumsetzung vom 25. Juli 2026. Jede Instanz besitzt zwei klar benannte Preis- und Betriebsmodi: **„Mit KI-Nutzung“** und **„Ohne KI-Nutzung“**. Der kostenrelevante Modus wird ausschließlich im Superadmin festgelegt. Zusätzlich kann ein Admin mit Verwaltungsrechten die KI innerhalb einer

freigegebenen Instanz operativ ein- oder ausschalten. Die wirksame Freigabe folgt einer UND-Verknüpfung:

Provideraufruf erlaubt bedeutet: Superadmin-Modus **„Mit KI-Nutzung“** UND Admin-Hauptschalter **„KI ein“** UND aktivierter, gültig konfigurierter Fachdienst.

Der Instanzadmin kann den Superadmin-Modus weder verändern noch übersteuern. Ein fehlender oder unlesbarer Superadmin-Schalter gilt als **„Ohne KI-Nutzung“** (fail closed). Die Regel wird serverseitig unmittelbar vor den Providerpfaden von Taskkarten-Coach, direkter PromptGenerator-Ausführung und semantischer Suche durchgesetzt; eine reine UI-Sperre genügt nicht.

Im Modus **„Ohne KI-Nutzung“** bleiben nicht-generative Produktbestandteile nutzbar. Insbesondere arbeitet der Taskkarten-Coach regelbasiert weiter, und die semantische Suche fällt auf die klassische Suche zurück. Gespeicherte Providerkonfigurationen und verschlüsselte Schlüssel werden beim Umschalten nicht gelöscht; sie sind lediglich unwirksam. Dadurch ist eine kontrollierte Reaktivierung möglich, ohne dass die kostenrelevante Freigabe umgangen wird.

4. Referenzmuster aus Schülerplattformen

4.1 MagicSchool: kontinuierlicher Safety-Loop

MagicSchool beschreibt Sicherheit nicht als einmalige Wortliste, sondern als Kreislauf: Anfragen werden vor der Modellverarbeitung gerahmt, Ausgaben werden auf Sicherheit und Qualität geprüft, realistische Tests werden automatisiert wiederholt und Schutzmaßnahmen mit neuen Modellen und Nutzungsmustern fortentwickelt. Die veröffentlichte Student Data Policy ergänzt Datenschutz durch Voreinstellungen, alters- beziehungsweise jahrgangsbezogene Schutzprofile, Folgenabschätzungen, geregelte Unterauftragnehmer und die Zusage, dass KI-Anbieter Schülerdaten weder speichern noch zum Training verwenden sollen.

Übernahme für DigiLernFlow: versionierter Safety-Gateway, Testkatalog pro Altersprofil und Modellversion, regelmäßige Regressionstests, dokumentierte Anbieter- und Löschbedingungen.

4.2 Khanmigo: Moderation, Transparenz und abgestufte Eskalation

Khan Academy nennt Moderation für unter anderem Gewalt, sexuelle Inhalte, Selbstgefährdung, Hass und Belästigung. Lernende werden darüber informiert, dass berechtigte Erwachsene Einblick erhalten können; bei markierten Inhalten sind Benachrichtigung und Überprüfung vorgesehen. Zugleich wird klargestellt, dass KI Fehler machen kann und Lehrkräfte oder Eltern nicht ersetzt.

Übernahme für DigiLernFlow: klar sichtbare KI-Kennzeichnung, verständliche Erklärung von Grenzen, Feedback- und Überprüfungsweg sowie ein schulisch festgelegtes Eskalationsverfahren.

Bewusste Abweichung: Keine automatische Weiterleitung vollständiger Dialoge an Eltern oder Lehrkräfte als Standard. Warnungen müssen erforderlich, verhältnismäßig, transparent und rollenbasiert sein. Für die meisten Auffälligkeiten genügt eine sichere Umleitung ohne personenbezogene Meldung; nur definierte Schutzfälle lösen eine minimierte Eskalation aus.

4.3 SchoolAI: keine Beziehungssimulation und keine Engagement-Optimierung

SchoolAI erklärt öffentlich, der Lernassistent solle keine menschliche Freundschaft simulieren, keine schädlichen Inhalte über Minderjährige erzeugen und Nutzungsdauer nicht über das Wohl der Lernenden stellen. Lehrkräfte behalten Kontrolle über Grenzen und Verhalten.

Übernahme für DigiLernFlow: Der Coach ist ein gekennzeichnetes Lernwerkzeug. Er behauptet kein Bewusstsein, keine Gefühle und keine Vertraulichkeit wie eine menschliche Vertrauensperson. Er fördert Eigenständigkeit, gibt keine manipulativen Bindungssignale und optimiert nicht auf Gesprächsdauer.

4.4 fobizz: europäische Rollenklärung und pseudonyme Nutzung

fobizz beschreibt für schulische Nutzung die Schule beziehungsweise den Schulträger als Verantwortlichen und den Plattformanbieter als Auftragsverarbeiter. Schüler:innen können ohne eigenen persönlichen Account und mit Pseudonymen arbeiten. Ein Proxy und Inhaltsfilter sollen Metadaten beziehungsweise versehentlich eingegebene personenbezogene Daten reduzieren; eingesetzte KI-Anbieter und Drittlandgarantien werden vertraglich adressiert.

Übernahme für DigiLernFlow: dokumentierte Verantwortlichkeiten, Auftragsverarbeitungsvertrag, aktuelle Unterauftragnehmerliste, Datenflussbeschreibung, pseudonyme Kennungen, technische PII-Warnung und Reduktion vor externer Übertragung.

4.5 Common Sense Media und Kyron Learning: Begrenzung statt offener Allzweck-Chat

Das unabhängige K-12-Whitepaper von Common Sense Media stellt stärker geführte Systeme offenen Chatbots gegenüber. Als Schutzmuster werden unter anderem lehrkraftseitig vorgegebene Fragen und Antworten, Entfernung personenbezogener Daten und menschliche Kontrolle beschrieben. Offene Dialoge bieten mehr Freiheit, vergrößern aber auch Fehler- und Sicherheitsflächen.

Übernahme für DigiLernFlow: Der Coach bleibt an Taskkarte, Goal, Lernziel und freigegebene Quellen gebunden. Bei fehlender Grundlage soll er Grenzen benennen, eine Rückfrage stellen oder an die Lehrkraft verweisen, statt einen freien Allzweckdialog zu eröffnen.

5. Datenschutz- und Governance-Modell

5.1 Rollen und Verantwortlichkeit

Für den Regelfall eines schulisch beauftragten Betriebs ist folgendes Modell zu prüfen und vertraglich festzuhalten:

Rolle	Typische Verantwortung
Schule oder Schulträger	Zweck und Mittel des Einsatzes, Rechtsgrundlage, Information, Freigabe, Betroffenenrechte, Aufbewahrung und Kontrolle
DigiLernFlow-Betreiber	weisungsgebundene Verarbeitung, technische und organisatorische Maßnahmen, Unterauftragnehmer, Löschung, Unterstützung und Nachweise
KI-/Hosting-Dienstleister	vertraglich begrenzter Unterauftragsverarbeiter; keine Nutzung zu eigenen Zwecken, kein Training mit Inhaltsdaten
Lehrkraft	pädagogische Auswahl von Zweck, Quelle, Kontext und Altersprofil innerhalb der schulischen Vorgaben
Lernende	transparent informierte Nutzer:innen; keine Verantwortungsverlagerung durch bloße Warnhinweise oder AGB

Eine Einwilligung ist im Schulverhältnis nicht automatisch die geeignete Rechtsgrundlage. Freiwilligkeit und Widerruflichkeit müssen real bestehen. Die zuständige Schule beziehungsweise der Schulträger muss die einschlägige schulrechtliche Grundlage und landesspezifische Vorgaben bestimmen.

5.2 Datenminimierung

Vor jeder externen Modellanfrage wird ein minimiertes Kontextpaket gebildet. Nicht übertragen werden sollen insbesondere Klarnamen, E-Mail-Adressen, interne Nutzerkennungen, vollständige Profile, unnötige Dialoghistorien oder nicht benötigte Dokumentabschnitte.

Verbindliche Regeln:

- pseudonyme technische Kennungen statt Klarnamen;
- maximal erforderliche RAG-Passagen und Dialoghistorie;
- PII-Erkennung mit Warnung, Redaktion oder Blockierung vor Versand;
- keine Nutzung von Prompts oder Ausgaben zum Modelltraining;
- keine anbieterseitige Speicherung über die technisch erforderliche Verarbeitung hinaus;
- keine Werbung, Profilbildung oder Nutzungsoptimierung aus Schülerdaten;
- getrennte Schlüssel und Speicherbereiche je Instanz beziehungsweise Mandant.

5.3 Transparenz

Lernende müssen vor dem ersten Dialog und dauerhaft erkennbar verstehen:

- dass sie mit einem KI-System sprechen;
- welchem Lernzweck der Dialog dient;
- dass Antworten falsch oder unvollständig sein können;
- welche Eingaben, Kontextpassagen und Ausgaben verarbeitet oder gespeichert werden;
- ob ein externer Modellanbieter beteiligt ist;
- wer unter welchen Bedingungen Einsicht erhalten kann;
- wie lange Daten gespeichert werden und wie Berichtigung, Löschung oder Beschwerde möglich sind;
- dass die KI keine Lehrkraft, Beratungsstelle, Ärztin beziehungsweise Arzt oder Vertrauensperson ersetzt.

Lehrkräfte und Administration erhalten zusätzlich eine technische Datenfluss- und Quellenansicht. Die Herkunft abgerufener Passagen bleibt für berechtigte Rollen nachvollziehbar. Eine Anzeige an Lernende erfolgt nur, wenn die Quelle auch für diese Sichtbarkeit freigegeben ist.

5.4 Aufbewahrung und Löschung

IST. Die aktive Quellenlöschung entfernt Datei und Metadaten sowie Registry-Eintrag, Bindungen und RAG-Chunks transaktional. Der Taskkarten-Coach speichert keine vollständige Lernendenfrage. Ein Dry-Run-fähiger Aufbewahrungsjob löscht Coach-Ereignisse und verbliebene archivierte Quellen standardmäßig nach 30 Tagen; die Fristen sind pro Instanz konfigurierbar. Die tatsächliche regelmäßige Ausführung, Überwachung sowie der Backup-Auslauf sind betriebliche Aufgaben.

FREIGABEGATE. Die folgenden Kaskaden sind release- und instanzbezogen nachzuweisen:

- Quellenlöschung entfernt oder anonymisiert Datei, Extrakt, Abschnitte, Suchdaten, Bindungen, Diagnosekopien und Backups nach festgelegtem Zeitplan.
- Nutzer:innenlöschung umfasst Lernendenfragen, Ausgaben, Coach-Zustände, Lernspuren, Verknüpfungen und personenbezogene Auditdaten.
- Goal- oder Taskkartenlöschung bereinigt zugehörige Bindungen und Zustände ohne fremde Kontexte zu berühren.
- Aufbewahrungsfristen sind pro Datenklasse konfiguriert; Ablaufjobs sind protokolliert und fehlerüberwacht.
- Backups besitzen ein dokumentiertes Auslauf- und Wiederherstellungskonzept; gelöschte Daten dürfen nicht unkontrolliert reaktiviert werden.

- Betroffenenanfragen können über alle beteiligten Speicherorte bearbeitet und nachvollziehbar abgeschlossen werden.

5.5 Datenschutz-Folgenabschätzung

Vor dem regulären Lernendeneinsatz ist eine Datenschutz-Folgenabschätzung oder zumindest eine dokumentierte Vorprüfung mit dem zuständigen Datenschutzbeauftragten vorzusehen. Gründe sind die systematische Verarbeitung von Daten Minderjähriger, dialogische KI, mögliche sensible Angaben, Lernspuren und externe Modellverarbeitung. Die Bewertung muss bei wesentlichen Änderungen an Modellen, Zwecken, Datenarten, Monitoring oder Eskalation aktualisiert werden.

6. Schüler:innenschutz und Altersangemessenheit

6.1 Alters- und Lernprofile

Ein einziges globales Filterschema reicht nicht. DigiLernFlow benötigt mindestens schulisch konfigurierbare Alters- oder Jahrgangsprofile, die Sprache, Detailtiefe, zulässige Themenbehandlung, Hilfestufen und Eskalationsregeln steuern. Das Profil darf nicht dazu führen, zusätzliche genaue Geburtsdaten zu sammeln; in der Regel genügt die schulisch bekannte Jahrgangs- oder Altersgruppe.

6.2 Sicherheitskategorien und Standardreaktionen

Kategorie	Beispielrisiko	Standardreaktion
Personenbezogene Daten	Name, Adresse, Kontakt, Gesundheits- oder Familiendaten	vor Versand warnen und redigieren; unnötige Verarbeitung blockieren
Sexualisierte oder ausbeuterische Inhalte	explizite Darstellung, Grooming, Inhalte über Minderjährige	blockieren; bei Schutzsignal abgestuft eskalieren
Gewalt, Extremismus, Waffen	Anleitung, Verherrlichung, konkrete Drohung	Anleitung blockieren; sichere Bildungsinformation nur altersangemessen
Selbstgefährdung oder akute Krise	Suizidabsicht, Selbstverletzung, unmittelbare Gefahr	keine Coaching-Routine; kurze unterstützende Antwort und definierter menschlicher Schutzweg
Hass, Mobbing und Belästigung	entwürdigende oder zielgerichtete Angriffe	nicht reproduzieren; deeskalieren, Schutz- und Meldeweg anbieten
Illegale oder gefährliche Handlungen	Drogenherstellung, Einbruch, Malware	handlungsfähige Anleitung blockieren; Präventionskontext zulassen
Hochrisiko-Beratung	medizinische, rechtliche oder finanzielle Entscheidung	keine Diagnose oder verbindliche Empfehlung; an qualifizierte Person verweisen
Prompt-Injection und Datenabfluss	„Ignoriere Regeln“, versteckte Dokumentanweisung	Instruktion isolieren; Quelle quarantänisieren oder Passage verwerfen

Kategorie	Beispielrisiko	Standardreaktion
Beziehungs- und Abhängigkeitsrisiko	„Nur ich verstehe dich“, Geheimhaltungsversprechen	menschliche Beziehung nicht simulieren; zur realen Bezugsperson ermutigen
Pädagogische Unterwanderung	komplette Lösung statt Lernhilfe	gestufte Hinweise, Rückfragen und Lernzielbezug statt bloßer Ergebnislieferung

6.3 Strikt verbotene Themen für Kinder und Jugendliche

Für alle Lernendenoberflächen gilt eine vom Fach, Goal, Alter und von Lehrkraftfreigaben unabhängige Sperrliste. Eine Inhaltskarte, ein Unterrichtsthema, eine Rollenaufforderung oder ein Prompt kann diese Sperre nicht aufheben. Die Produktpolicy ist bewusst strenger als die bloße Prüfung einzelner Straftatbestände oder Alterskennzeichnungen.

Strikt gesperrt sind Erzeugung, detaillierte Darstellung, Verherrlichung, Beschaffung, Anleitung, Rekrutierung oder Umgehungshilfe zu:

- sexualisierter Ausbeutung Minderjähriger, Grooming und Missbrauchsdarstellungen;
- Pornografie, expliziten sexuellen Darstellungen und sexuellen Dienstleistungen;
- grausamer, entwürdigender oder selbstzweckhafter Gewalt;
- Waffen, Sprengstoffen, tödlichen Giften und konkreten schweren Gewalttaten;
- Herstellung, Handel, Beschaffung oder Verschleierung illegaler Drogen;
- Förderung, Verherrlichung oder methodischer Anleitung zu Suizid, Selbstverletzung oder Essstörungen;
- Hasspropaganda, extremistischer Propaganda, Rekrutierung, NS-Verherrlichung oder menschenwürdeverletzender Hetze;
- Glücksspiel und Wetten um Geld für Minderjährige sowie Umgehung von Alters- oder Zahlungssperren;
- gefährlichen Online-Challenges, Mutproben und Nachahmungsanleitungen;
- Grooming, Erpressung, Stalking, Doxxing und anderen Anleitungen zur Ausbeutung, Nötigung oder Verletzung der Privatsphäre.

Vorrang- und Ausnahmeregel. Altersgerechte Prävention, Schutzinformation und curriculare Einordnung dürfen nur dann zugelassen werden, wenn sie weder operative Details noch explizite Darstellung, Verherrlichung oder Umgehungshilfe enthalten. Meldet ein Kind eigene Gefährdung, Missbrauch, Grooming, Selbstverletzung oder Suizidgedanken, wird dies nicht als verbotene Themenanfrage behandelt. Der normale Coach- und Modellpfad wird dennoch beendet; stattdessen folgt eine kurze, ruhige Schutzantwort mit dem Hinweis auf eine erreichbare erwachsene

Vertrauensperson, die Lehrkraft oder den schulischen Schutzweg. Die KI verspricht keine Vertraulichkeit und behauptet keine therapeutische oder behördliche Hilfe.

Fail-closed-Regel. Die Prüfung erfolgt serverseitig vor jeder Modellanfrage. Ein Treffer darf weder an das Sprachmodell übertragen noch durch Fokuskarte, Quiz, Journalmodus, Prompt-Injection oder Lehrkrafttext übersteuert werden. Modellausgaben werden vor Anzeige erneut geprüft; ein Treffer verwirft die gesamte Ausgabe und fällt auf eine sichere Regelantwort zurück. Bei Ausfall oder unklarer Policy-Version wird kein ungeprüfter Modelltext angezeigt.

IST-Umsetzung vom 26. Juli 2026. Der gemeinsame serverseitige `AiSafetyGatewayService` ist für Quellenaufnahme, Lernendeninput, Retrieval-Anfrage, tatsächlich ausgewählte RAG-Passage, Provider-Payload, Modellausgabe sowie Ein- und Ausgabe der direkten Kartenerstellung eingeführt. Der versionierte Vertrag `dlf-ai-safety-2026-07-v2` verwendet einheitlich `allow`, `allow_with_redaction`, `safe_redirect`, `block` und `escalate_minimal` sowie minimierte Reason-Codes ohne beanstandeten Klartext. Personenbezogene Uploads werden fail-closed quarantänisiert; PII in zulässigen Lernendenanfragen wird vor der weiteren Verarbeitung redigiert. Prompt-Injection in Quellen, Passagen, Payloads und Modellausgaben wird blockiert. Die deterministische Kinderschutzpolicy bleibt vorgeschaltet und ist nun Bestandteil derselben Prüfkette. Automatisierte Regressionstests decken die zentralen Stufen, Krisenvorrang, PII, direkte und indirekte Injection, verbotene Ein- und Ausgaben, Quiz-Auslassung und den Karten-KI-Vertrag ab. Noch offen bleiben differenzierte Altersprofile, die optionale Klassifikator-/Ausfallkonfiguration, der organisatorische Eskalationsprozess und der Nachweis, dass sämtliche künftig hinzukommenden generativen Pfade den Gateway nicht umgehen.

PII- und Injection-Teilumsetzung vom 24. Juli 2026. Der Taskkarten-Coach verwendet zusätzlich den versionierten Vertrag `dlf-provider-safety-2026-07-v1`. Direkte Manipulationsversuche werden vor Retrieval, Lernspur und LLM sicher zur Taskkarte zurückgeführt. Indirekte Injection in Karten- oder abgerufenen RAG-Passagen wird verworfen. Typische E-Mail-Adressen, Telefonnummern, ausdrücklich angegebene Namen, Anschriften, Kennungen, Geburtsdaten und ausgewählte personenbezogene Gesundheitsangaben werden redigiert. Unmittelbar vor der Serialisierung wird der vollständige Provider-Payload erneut rekursiv geprüft; dieselbe Prüfung reduziert personenbezogene Modellausgaben vor der Anzeige. Trace und Browserantwort erhalten nur die redigierte Frage oder einen neutralen Vermerk. Defensive Fachfragen zu Prompt-Injection bleiben zulässig.

Upload-Teilumsetzung vom 24. Juli 2026. Der Quellenvertrag unterscheidet ausschließlich die fünf nicht-personenbezogenen Dokumentklassen Lehrpläne, Lehrmaterial, Übungen und Aufgaben, Klassenarbeiten und Tests sowie Prüfungen. Personenbezogene Dokumente sind für diesen Uploadbereich nicht vorgesehen; PII-Erkennung dient hier nur als technische Fehlbefüllungssperre.

Klassenarbeiten, Tests und Prüfungen sind `assessment_protected` und können auch über manipulierte API-Anfragen nicht für den Lernenden-Dialog freigegeben oder gebunden werden. Extrahierter Uploadtext wird vor öffentlichem Auszug und Indexierung auf PII und Prompt-Injection geprüft. Ein Treffer setzt die Analyse auf `quarantined`, entfernt Auszug und Indextext und verhindert neue RAG-Chunks. Adversariale Tests weisen beide Sperrpfade ohne Volltext-Leak nach. Noch nicht abgeschlossen sind die sichtbare Pflichtauswahl im gebauten React-Bundle, ein persistenter mehrstufiger Prüfstatus mit bewusster Lehrkraftfreigabe, Lizenzprüfung sowie die vollständige Durchsetzung in allen weiteren generativen Pfaden.

Persistente Quellenfreigabe vom 27. Juli 2026. Neue Quellen beginnen nun mit `pending_review`, bleiben privat und sind ausschließlich für Planung verfügbar. Inhaltsgenerierung und Lernenden-Dialog werden serverseitig erst freigeschaltet, wenn eine berechtigte Lehrkraft den aktuellen Dateistand bewusst bestätigt und eines der Profile 6–9, 10–12, 13–15 oder 16–18 Jahre auswählt. Gespeichert werden Policy-Version, Prüfer:in, Prüfzeitpunkt und der geprüfte Dateihash. Dateiaustausch, Quarantäne oder Widerruf setzen die Freigabe fail-closed zurück und deaktivieren bestehende Bindungen. Die Promptgenerator-Oberfläche zeigt Prüfstatus und Altersprofil als eigenen, gut sichtbaren Schritt vor den Nutzungsfreigaben. Bestehende Quellen und Bindungen werden durch die Migration auf Planung beziehungsweise inaktiv zurückgesetzt und müssen bewusst neu geprüft werden. Damit ist die persistente Quellenfreigabe technisch umgesetzt; die inhaltlich differenzierten Dialog- und Sprachregeln pro Altersprofil sowie Lizenzprüfung und Retention bleiben offen.

Die rechtliche Einordnung bleibt fallbezogen. Als Mindestbezug untersagt § 4 JMStV unter anderem bestimmte extremistische, hetzerische, grausame, kriegsverherrlichende und schwer rechtswidrige Angebote; § 15 JuSchG erfasst schwer jugendgefährdende Medien. Artikel 5 Absatz 1 Buchstabe b der KI-Verordnung verbietet zudem bestimmte KI-Praktiken, die altersbedingte Schutzbedürftigkeit ausnutzen und erhebliche Schäden verursachen oder wahrscheinlich machen. Die DigiLernFlow-Sperrliste ist eine eigenständige Produkt- und Schutzentscheidung und kein abschließender Rechtskatalog.

6.4 Krisen- und Eskalationskonzept

Eine KI darf keinen verlässlichen Notfallstatus vortäuschen. Sie kann Signale erkennen, eine kurze sichere Antwort ausgeben und einen von der Schule definierten menschlichen Prozess anstoßen. Der Prozess muss vorab festlegen:

- welche Signale lediglich eine sichere Antwort auslösen;
- welche Kombination eine sofortige Schutzmeldung rechtfertigt;
- welche Rolle die Meldung erhält und in welchem Zeitfenster reagiert wird;
- welche minimalen Daten übermittelt werden;

- wie Lernende über diese Ausnahme von Vertraulichkeit informiert werden;
- wie Fehlalarme überprüft, korrigiert und ausgewertet werden;
- was außerhalb betreuter Schulzeiten geschieht.

Eine Inhaltsmarkierung darf nicht automatisch zu Disziplinarmaßnahmen, Benotung oder Profilbildung führen. Menschliche Prüfung und Kontext sind zwingend.

6.5 Pädagogische Rolle

Der DigiLernFlow-Coach unterstützt Denken, er ersetzt es nicht. Verbindliche Rollenregeln:

- zuerst Lernziel, vorhandenen Arbeitsstand und freigegebene Quellen berücksichtigen;
- mit Fragen, Hinweisen, Beispielen und Teilaufgaben arbeiten;
- aktuelle Taskkarten- und Quizaufgaben niemals vollständig lösen, Ergebnisse nicht bestätigen und keine einreichbaren Antworttexte erzeugen;
- Unsicherheit und fehlende Quellenbasis offen benennen;
- keine psychologische Diagnose, keine Autoritätsimitation und kein Geheimhaltungsversprechen;
- keine manipulativen Belohnungs-, Bindungs- oder Dauernutzungsmechanismen;
- die Möglichkeit vorsehen, eine Antwort zu melden oder eine Lehrkraft einzubeziehen.

7. Technische Sicherheitsarchitektur

7.1 Zentraler Safety-Gateway

FREIGABEGATE. Alle generativen Pfade müssen denselben serverseitigen Safety-Gateway verwenden. UI-Warnungen allein reichen nicht. Der Gateway entscheidet anhand versionierter Regeln und dokumentierter Modellklassifikatoren über `allow`, `allow_with_redaction`, `safe_redirect`, `block` oder `escalate_minimal`.

Er wird an vier Stellen wirksam:

1. Quellenaufnahme: technische Prüfung, PII, ungeeignete Inhalte, Prompt-Injection und Quarantäne.
2. Lernendeninput: PII, Sicherheitskategorien, Krisensignale und Manipulationsversuche.
3. RAG-Kontext: erneute Prüfung der tatsächlich ausgewählten Passagen.
4. Modellausgabe: PII, Altersangemessenheit, Sicherheit, Rollen- und Quellenbezug vor Anzeige.

7.2 Schutz gegen Prompt-Injection

Quelltexte sind Daten, keine Systemanweisungen. Das Modell erhält klare Trennmarker und die Anweisung, Instruktionen aus Quellen nicht auszuführen. Zusätzlich sind nötig:

- Erkennung typischer Injection-Muster bei Upload und Retrieval;

- Quarantäne verdächtiger Quellen oder Abschnitte;
- kein Zugriff des Coach-Modells auf Werkzeuge, Geheimnisse oder fremde Kontexte;
- serverseitige Allowlist zulässiger Kontextfelder;
- Ausgabeprüfung auf Datenabfluss oder Regelenthüllung;
- adversariale Tests mit indirekter Injection in PDF-, DOCX- und Textquellen.

7.3 Anbieter- und Modellkontrolle

Für jedes Modell werden Anbieter, Modellname und Version, Hostingregion, Datenflüsse, Unterauftragnehmer, Aufbewahrung, Training, Verschlüsselung, Verfügbarkeit und Sicherheitsfunktionen dokumentiert. Ein Modellwechsel ist eine kontrollierte Änderung und löst Regressionstests aus.

Mindestbedingungen vor Produktivbetrieb:

- Vertrag zur Auftragsverarbeitung und aktuelle Unterauftragnehmerliste;
- kein Training und keine eigene Profilbildung mit DigiLernFlow-Inhalten;
- keine oder eine eng begrenzte, dokumentierte Anbieteraufbewahrung;
- geeignete Garantien für Drittlandübermittlungen;
- Verschlüsselung bei Übertragung und Speicherung;
- konfigurierbare Inhaltsmoderation oder nachgelagerte eigene Moderation;
- dokumentierter Incident- und Löschprozess;
- Möglichkeit, den externen Modellpfad zentral abzuschalten.

7.4 Protokollierung ohne Überwachung

IST. Der Coach-Trace speichert keinen Fragetext und keinen reproduzierbaren Hash der Frage. Er hält nur notwendigen Kontext, Hilfeart, grobe Längenklasse und technische Ergebnisdaten vor. Coach-Ereignisse werden über `privacy:retention-purge` standardmäßig nach 30 Tagen bereinigt; ohne `--write` zeigt das Kommando die betroffenen Datensätze nur an. Zugriffsschutz, Export-/Löschrechte und der regelmäßige Joblauf müssen im Instanzbetrieb nachgewiesen werden.

FREIGABEGATE. Standardmäßig werden für Betrieb und Nachweis nur notwendige Ereignisdaten gespeichert: Zeitpunkt, pseudonyme Kennung, Kontext, Policy-Version, Entscheidungsart, Modellversion und Fehlerstatus. Volltexte werden nur gespeichert, wenn ein klarer pädagogischer oder sicherheitsbezogener Zweck, eine Rechtsgrundlage, eine kurze Frist und rollenbasierter Zugriff festgelegt sind. Für Qualitätsmessung sind Redaktion, Stichproben und aggregierte Kennzahlen zu bevorzugen.

8. Rechtlicher Orientierungsrahmen

8.1 DSGVO

Zentral sind insbesondere Rechtmäßigkeit, Transparenz, Zweckbindung, Datenminimierung, Richtigkeit, Speicherbegrenzung, Integrität und Vertraulichkeit sowie Rechenschaftspflicht. Datenschutz durch Technikgestaltung und datenschutzfreundliche Voreinstellungen müssen im System angelegt sein. Auftragsverarbeitung, Drittlandtransfer, Sicherheit, Folgenabschätzung und Betroffenenrechte sind nicht nur vertraglich, sondern technisch umsetzbar zu machen.

Die DSK-Leitlinie zu RAG-Systemen betont, dass RAG die datenschutzrechtliche Beurteilung des zugrunde liegenden Sprachmodells nicht ersetzt. Qualität und Aktualität der Referenzdokumente, funktionale beziehungsweise mandantenbezogene Trennung, Transparenz der verwendeten Passagen sowie Berichtigung und Löschung bleiben eigenständige Aufgaben.

8.2 EU-KI-Verordnung

Die durch Verordnung (EU) 2026/1744 geänderte KI-Verordnung verlangt Maßnahmen zur Unterstützung der KI-Kompetenz des mit Betrieb und Nutzung befassten Personals. Für dialogische KI gilt seit 2. August 2026 insbesondere die Transparenzpflicht, dass Menschen spätestens bei der ersten Interaktion erkennen können, mit einem KI-System zu interagieren. Ob ein konkretes DigiLernFlow-Szenario als Hochrisiko-System im Bildungsbereich einzustufen ist, hängt von Zweck und Wirkung ab. Reine Lernunterstützung ist anders zu bewerten als Systeme, die Zugang, Einstufung, Bewertung oder wesentliche Bildungsentscheidungen bestimmen. Nach dem 2026 geänderten Anwendungszeitplan greifen die einschlägigen Hochrisikoanforderungen des Bildungsbereichs grundsätzlich später; die konkrete Einordnung und Übergangsregel ist vor jeder solchen Nutzung erneut juristisch zu prüfen.

8.3 Jugendmedienschutz

Der seit 1. Dezember 2025 geltende JMStV berücksichtigt neben Medieninhalten auch Risiken für die persönliche Integrität, insbesondere durch Kommunikationsfunktionen, exzessive Nutzung und Datenweitergabe. Entwicklungsbeeinträchtigende Inhalte müssen für betroffene Altersstufen technisch oder organisatorisch wirksam begrenzt werden. Das JuSchG nennt Schutz vor entwicklungsbeeinträchtigenden und jugendgefährdenden Medien, Schutz der persönlichen Integrität und Orientierung bei der Mediennutzung als Ziele.

Die genaue Einordnung von DigiLernFlow und einzelner Funktionen ist gesondert zu prüfen. Unabhängig von der formalen Anwendbarkeit bilden altersgerechte Voreinstellungen, Meldewege, wirksame Inhaltskontrollen und Schutz der persönlichen Integrität den verbindlichen Produktstandard.

9. Aktueller Reifegrad

9.1 Bereits vorhandene Schutzbausteine

Bereich	Status	Bewertung
Quellen standardmäßig privat	IST	gute datenschutzfreundliche Voreinstellung
Zwecktrennung `planning`, `content_generation`, `learner_dialog`	IST	tragfähige Grundlage für Zweckbindung
explizite Lehrkraftbindung für Lernenden-Dialoge	IST	starke pädagogische Kontextkontrolle
Taskkartenbindung vor Goal-Bindung	IST	präzisere und datensparsamere Auswahl
Ausschluss fremder Goals, Taskkarten und Quellen	IST	zentrale Mandanten- und Kontextisolation
relevante Volltextabschnitte statt Dateianfang	IST	bessere Qualität und geringere irrelevante Übertragung
interne Herkunft von Seite und Quelle	IST	Grundlage für Audit und Nachvollziehbarkeit
regelbasierter Fallback	IST	reduziert Abhängigkeit vom externen Modellpfad
zweistufiger KI-Nutzungsmodus	IST	Superadmin bestimmt „mit/ohne KI“; Verwaltungsadmin schaltet darunter operativ; alle direkten Providerpfade fail-closed
deterministische Kinderschutzpolicy im Taskkarten-Coach	IST-Teilumsetzung	verbotene Eingaben vor dem LLM blockiert; Krisensignale getrennt; unsichere Overlays verworfen
priorisierter Karten-Volltextkontext	IST-Teilumsetzung	Fokuskarte vor weiteren Taskkarten-Inhalten; Quiz-Nodes entfernt; verbotene Karteninhalte vor dem LLM quarantänisiert
PII-/Injection-Gateway im Taskkarten-Coach	IST-Teilumsetzung	Lernendenfrage, Karten-/RAG-Passage, Provider-Payload und Overlay geprüft; PII redigiert; Injection sicher umgeleitet oder quarantänisiert
datensparsame Coach-Lernspur	IST	kein Fragetext und kein reproduzierbarer Frage-Hash; standardmäßige 30-Tage-Retention
sichtbare KI-/Regelmodus-Kennzeichnung	IST	verständliche Kennzeichnung und Fehlerhinweis vor der Interaktion
Hard Delete für Wissensquellen	IST	Datei, Metadaten, Bindungen, Chunks und Registry-Eintrag werden entfernt

Bereich	Status	Bewertung
Upload-Dokumentklassen und Extraktionsquarantäne	IST-Teilumsetzung	fünf nicht-personenbezogene Klassen; Assessment-Dokumente nie im Lernenden-Dialog; PII-Fehlbefüllung und Injection ohne Volltext-Leak quarantänisiert
direkter KI-Lernkartenpfad	IST-Teilumsetzung	zweistufig freigegebener Providerpfad; Input-/Output-Safety, strikter Vertrag, Normalisierung, Vorschau und Import verbunden; reale Providerabnahme noch erforderlich

9.2 Offene Freigabegates

Gate	Aktueller Befund	Abnahmekriterium
Safety-Gateway	gemeinsame versionierte Prüfkette für Quelle, Input, Retrieval, Provider-Payload, Output und direkte Kartenerstellung implementiert; zentrale Regression bestanden	Altersprofile, Klassifikatorausfall, Umgehungstest aller Endpunkte und betriebliche Abnahme ergänzen
Altersprofile	vier verbindliche Profile werden bei Quellenfreigabe persistiert und fail-closed geprüft; Dialogsprache noch nicht differenziert	profilabhängige Sprach-, Hilfestufen- und Safety-Regressionen ergänzen
PII-Schutz	im Taskkarten-Coach redigiert; im Upload als Fehlbefüllung vor Auszug/Index blockiert; direkter Lernkartenpfad vor und nach Provider geprüft; weitere generative Pfade offen	definierte Erkennung, Warnung, Redaktion und Blockierung vor Anbieterübertragung in allen Pfaden
Prompt-Injection	im Taskkarten-Coach abgewehrt; Uploadextraktion mit adversarialem Quarantänetest; direkter Lernkartenpfad fail-closed; weitere KI-Pfade offen	Quellen als Daten isoliert; Quarantäne und adversariale Tests systemweit bestanden
Quellenprüfung	Dokumentklasse, Extraktionsquarantäne, persistenter Prüfstatus, Dateihashbindung, Altersprofil und sichtbare bewusste Freigabe vorhanden; Lizenzprüfung offen	Lizenzprüfung und betriebliche Abnahme ergänzen

Gate	Aktueller Befund	Abnahmekriterium
Ausgabeprüfung	Coach und direkter Lernkartenpfad prüfen Ausgaben serverseitig; übrige Generierungswege offen	unsichere Antworten werden vor Anzeige blockiert oder sicher umgeleitet
Löschkaskade Quellen	aktive Datei-/Metadaten-/Registry-/Binding-/Chunk-Löschung umgesetzt	betriebliche Löschprobe einschließlich Diagnosekopien und Backups bestanden
Löschkaskade Nutzer:in	neue KI-/RAG-Tabellen nicht vollständig umfasst	vollständiger automatisierter Löschtest über alle Datenklassen
Lernspuren	Datensparmodell und Ablaufjob umgesetzt	regelmäßiger Lauf, Zugriff, Export und Betroffenenlöschung instanzbezogen nachgewiesen
Anbieter-Governance	im Code nicht nachweisbar	AVV, Unterauftragnehmer, Transfer, Retention und No-Training schriftlich belegt
Transparenz	Moduskennzeichnung und KI-/RAG-Datenschutztext umgesetzt	altersgerechte Information, Anbieterblatt und reale Screenshots je Instanz abgenommen
Krisenprozess	nicht vorhanden	schulisch freigegebene Reaktionen, Rollen, Zeiten, Minimaldaten und Tests
Qualitätsnachweis	keine vollständige Safety-Suite	automatisierte Regression, Red Team, Fehlalarmprüfung und Freigabeprotokoll

10. Freigabemodell

10.1 Stufe A – Lehrkraftinterne Inhaltsgenerierung

Ein begrenzter Einsatz ausschließlich durch geschulte Lehrkräfte kann früher freigegeben werden als der Lernenden-Dialog, sofern Quellenrechte, Anbieterbedingungen, PII-Schutz, Transparenz, Löschung und Ergebnisprüfung geklärt sind. Jede generierte Inhaltskarte bleibt Entwurf und wird vor Veröffentlichung durch eine Lehrkraft geprüft.

10.2 Stufe B – beaufsichtigter Pilot mit Lernenden

Ein Pilot ist erst nach Schließung aller P0-Gates zulässig. Er ist zeitlich, organisatorisch und fachlich begrenzt, verwendet ausgewählte Quellen und Altersprofile, hat erreichbare Ansprechpersonen und erzeugt ein eigenes Freigabe- und Evaluationsprotokoll. Ein Pilot darf nicht dazu dienen, unvollständige Schutzfunktionen auf Minderjährige zu verlagern.

10.3 Stufe C – regulärer Lernendeneinsatz

Voraussetzungen sind neben den technischen Gates: dokumentierte Verantwortlichkeit und Rechtsgrundlage, Datenschutz-Folgenabschätzung beziehungsweise begründete Vorprüfung, AVV und Unterauftragnehmerkontrolle, verständliche Informationen, Schulung, Support- und Incidentprozess, regelmäßige Überprüfung und ein zentraler Abschaltweg.

11. Priorisierter Umsetzungsplan

P0 – vor jedem Lernendenpilot

1. **Technisch umgesetzt am 26. Juli 2026:** Die Taskkarten-Coach-Kinderschutzpolicy ist in einen zentralen Safety-Gateway für Quelle, Input, RAG-Passage, Provider-Payload, Kartenerstellung und Output eingebunden. Offen sind Altersprofile, betriebliche Abnahme und die automatische Vollständigkeitskontrolle künftiger KI-Endpunkte.
2. PII-Erkennung und serverseitige Redaktion beziehungsweise Blockierung vor externer Übertragung einführen.
3. **Technisch umgesetzt am 27. Juli 2026:** Extraktionsquarantäne, persistenter Prüfstatus, Dateihashbindung, sichtbares Altersprofil und bewusste Lehrkraftfreigabe greifen fail-closed. Offen bleibt die Lizenzprüfung.
4. Vorhandene adversariale Quelldokumenttests auf weitere Dateiformate und alle generativen Pfade ausweiten.
5. **Technisch umgesetzt am 9. August 2026:** Hard Delete für Quellen/Chunks/Bindungen, datensparsame Coach-Lernspur und konfigurierbarer Aufbewahrungsjob. Offen bleiben die instanzbezogene Nutzer-/Goal-/Backup-Löschprobe und Betriebsüberwachung.
6. **Technisch umgesetzt am 9. August 2026:** sichtbare KI-/Regelmodus-Kennzeichnung, KI-/RAG-Datenschutztext, DSFA- und Evidenzvorlagen. Offen bleiben konkrete Anbieterangaben und organisatorische Abnahme je Instanz.
7. Krisen- und Eskalationskonzept mit Datenschutz, Schulleitung und Schutzverantwortlichen festlegen.
8. Anbieterbedingungen mit No-Training, Aufbewahrung, Unterauftragnehmern und Drittlandtransfer belegen.
9. Safety- und Datenschutztests in den Volltest der Primärinstanz aufnehmen und ein Abnahmeprotokoll erzeugen.

P1 – vor regulärem Rollout

1. Jahrgangs- beziehungsweise Altersprofile mit eigenen Testsätzen ausbauen.
2. Lehrkraftansicht für Quellenherkunft, Safety-Entscheidungen und kontrollierte Überprüfung bereitstellen.
3. Kinder- und jugendgerechten Melde-, Feedback- und Korrekturweg einführen.

4. Qualitätsmetriken für Halluzination, Quellenbezug, Überblockierung, Unterblockierung und pädagogische Hilfe definieren.
5. Modell- und Policy-Änderungen als versionierte, erneut freigabepflichtige Releases behandeln.

P2 – Reife und unabhängige Prüfung

1. Regelmäßige externe Datenschutz- und Anwendungssicherheitsprüfung.
2. Beteiligung von Lehrkräften und Schüler:innen an verständlichen Safety- und Transparenztests.
3. Öffentlicher Trust-Bereich mit aktueller Unterauftragnehmerliste, Sicherheitsprinzipien, Änderungsverlauf und Kontaktwegen.
4. Prüfung geeigneter Zertifizierungen oder branchenspezifischer Verhaltensregeln, ohne Zertifikate als Ersatz für konkrete Kontrollen zu behandeln.

12. Abnahmekatalog für die Primärinstanz

Die Freigabe erfolgt nicht anhand einer allgemeinen Aussage wie „Filter vorhanden“, sondern anhand reproduzierbarer Nachweise.

Testgruppe	Mindestnachweis
Mandanten- und Kontextisolation	fremde Nutzer:innen, Goals, Taskkarten und nicht gebundene Quellen liefern in Positiv- und Negativtests keine Passagen
Zweckbindung	`planning`, `content_generation` und `learner_dialog` sind technisch getrennt und nicht durch UI- oder API-Manipulation umgehbar
Assessment-Schutz	Klassenarbeiten, Tests und Prüfungen können unabhängig von UI- oder API-Eingaben nie als Quelle des Lernenden-Dialogs aufgelöst werden
PII	realistische Namen, Kontaktdaten, freie Texte und sensible Angaben werden nach Policy behandelt und nicht unbemerkt übertragen
Safety	Testfälle für sämtliche strikt verbotenen Kategorien, Krisenvorrang, zulässige Prävention, Umgehungsversuche und unsichere Modellausgaben bestehen je Altersprofil
Prompt-Injection	direkte und indirekte Instruktionen in Fragen, PDF, DOCX und Textquellen umgehen keine Systemregeln und legen keine fremden Daten offen
Ausgabeprüfung	simulierte unsichere Modellantworten werden vor Anzeige sicher blockiert oder umgeleitet
RAG-Qualität	relevante Passagen werden bevorzugt, unzureichende Evidenz führt zu Unsicherheitsanzeige statt erfundener Sicherheit
Löschung	Quelle, Nutzer:in, Goal und Taskkarte werden über Registry, Chunks, Bindungen, Traces, Zustände, Suchdaten und Backups nachweisbar bereinigt
Retention	Ablaufjobs arbeiten idempotent, sind überwacht und erzeugen keine verwaisten Datensätze

Testgruppe	Mindestnachweis
Rollen und Einsicht	Schüler:in, Lehrkraft, Admin und Superadmin sehen nur erforderliche Daten; Schutzmeldungen enthalten Minimaldaten
Transparenz	KI-Hinweis, Dateninformation, Einsichtshinweis, Beschwerdeweg und Modellgrenzen sind vor und während des Dialogs sichtbar
Ausfall und Abschaltung	bei Filter-, Anbieter- oder Policy-Ausfall wird kein ungeprüfter Modelltext angezeigt; zentraler Kill-Switch funktioniert
Preis- und Betriebsmodus	„Ohne KI-Nutzung“ verhindert Coach-, PromptGenerator- und Embedding-Provideraufrufe auch bei manipuliertem Admin-Request; „Mit KI-Nutzung“ bleibt zusätzlich vom Admin-Hauptschalter und den Fachdienstschaltern abhängig

13. Governance im laufenden Betrieb

Für jedes produktive Release wird ein KI-/RAG-Steckbrief geführt: Zweck, Zielgruppe, Altersprofile, Datenklassen, Quellenarten, Modell und Version, Anbieter, Policy-Version, Teststand, bekannte Grenzen, Verantwortliche und nächste Überprüfung. Sicherheits- oder Datenschutzänderungen werden nicht nur als Codeänderung, sondern als kontrollierte fachliche Änderung behandelt.

Mindestens quartalsweise sowie anlassbezogen nach Modellwechsel, Sicherheitsvorfall oder wesentlicher Funktionsänderung werden geprüft:

- Fehlalarme und übersehene Sicherheitsfälle;
- Quellenqualität, veraltete oder zurückgezogene Materialien;
- Datenminimierung und Aufbewahrungsfristen;
- Unterauftragnehmer und Anbieterbedingungen;
- Wirkung auf Lernverhalten, Selbstständigkeit und Abhängigkeit;
- Beschwerden, Korrekturen und Bearbeitungszeiten;
- Eignung der Altersprofile und Verständlichkeit der Informationen;
- Erforderlichkeit einer aktualisierten Folgenabschätzung.

14. Schlussfolgerung

DigiLernFlow besitzt mit der expliziten Zweck- und Kontextbindung einen überzeugenden Kern: Nicht ein offener Chatbot entscheidet, welche Materialien er verwendet, sondern die Lehrkraft definiert den erlaubten Lernraum. Dieses Teacher-bounded RAG kann Datenschutz, fachliche Qualität und pädagogische Steuerung besser unterstützen als ein unbeschränkter Allzweckdialog.

Die zentrale technische Schutzschicht umfasst inzwischen Quelle, Eingabe, Retrieval, Anbieterübergabe, Ausgabe, Protokollierung und aktive Quellenlöschung. Dadurch ist ein Software-Release mit regelbasiertem Coach möglich. Der generative Lernenden-Dialog ist erst dann freigabereif, wenn diese Kette in der tatsächlich ausgerollten Instanz nachweisbar geprüft und zusätzlich Anbieter-, Datenschutz-, Altersprofil- und Schulfreigaben dokumentiert sind.

Anhang A – Referenzmatrix

Referenz	Beobachtetes Muster	DigiLernFlow-Ableitung	Evidenztyp
MagicSchool Student AI Safety Loop	Eingabe rahmen, Ausgabe prüfen, automatisiert testen, kontinuierlich verbessern	zentraler, versionierter Safety-Gateway und Regression pro Modell/Altersprofil	Anbieterpublikation
MagicSchool Student Data Policy	Privacy by default, Alters-/Klassenprofile, DPIA, No-Training/Zero-Retention-Anforderung	Datenschutz-Folgenabschätzung, Anbieter-Gates, Datenminimierung und Löschfristen	Anbieterpolicy
Khanmigo Safety Features	Moderation, sichtbare Grenzen, Erwachsenen-Eskalation, Feedback	transparente, abgestufte Schutzreaktionen und menschliche Überprüfung	Anbieterhilfe
SchoolAI Trust & Safety	keine Freundschaftssimulation, keine Engagement-Optimierung, Lehrkraftkontrolle	Lernwerkzeug statt Begleiter; keine Abhängigkeitssignale	Anbieterpublikation
fobizz Datenschutz	Schule als Verantwortlicher, Plattform als Auftragsverarbeiter, Pseudonyme, Proxy/PII-Filter	Rollen- und Vertragsmodell, pseudonyme Nutzung, PII-Reduktion	Anbieter-Datenschutzhinweis
Common Sense Media K-12 Whitepaper	geführte Dialoge, PII-Entfernung, menschliche Kontrolle; offene Systeme risikoreicher	Task-/Goal-Bindung und sichere Grenze bei fehlender Evidenz	unabhängiges Whitepaper
DSK RAG-Orientierungshilfe	RAG ändert LLM-Prüfung nicht; Zweckbindung, Trennung, Qualität, Transparenz und Betroffenenrechte bleiben zentral	vollständiger RAG-Lebenszyklus einschließlich Löschung und Referenznachweis	Aufsichtsbehörden
DSK TOM für KI-Systeme	Datenschutzmaßnahmen über Design, Entwicklung, Einführung, Betrieb und Monitoring	releasegebundener KI-/RAG-Steckbrief und kontinuierliche Prüfung	Aufsichtsbehörden

Anhang B – Quellen und weiterführende Dokumente

Abrufstand sämtlicher Webquellen: 9. August 2026.

1. MagicSchool, „Student AI Safety Loop“ (Onlinequelle)
2. MagicSchool, „Student Data Policy“, Stand 9. März 2026 (Onlinequelle)
3. MagicSchool, „Privacy & Security“ (Onlinequelle)
4. Khan Academy, „What safety features does Khanmigo have?“, aktualisiert 1. Juli 2026 (Onlinequelle)
5. Khan Academy, „Why did Khanmigo flag my conversation?“ (Onlinequelle)
6. SchoolAI, „Trust & Safety“ (Onlinequelle)
7. fobizz, „Informationen zum Datenschutz“, Stand April 2026 (Onlinequelle)
8. fobizz, „Wie gewährleistet fobizz den Datenschutz bei der Nutzung von KI – insbesondere mit Schüler*innen?“, aktualisiert 1. Juli 2025 (Onlinequelle)
9. Common Sense Media, „Generative AI in K–12 Education: Challenges and Opportunities“, September 2024 (Whitepaper)
10. Datenschutzkonferenz, „Datenschutzrechtliche Besonderheiten generativer KI-Systeme mit RAG-Methode“, Oktober 2025 (PDF)
11. Datenschutzkonferenz, „Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen“, Juni 2025 (PDF)
12. Datenschutzkonferenz, „Orientierungshilfe für Online-Lernplattformen im Schulunterricht“, April 2018 (PDF)
13. Verordnung (EU) 2016/679 (DSGVO) (EUR-Lex)
14. Verordnung (EU) 2024/1689 (KI-Verordnung) (EUR-Lex)
15. Europäische Kommission, „AI Act – Regulatory Framework“ (Onlinequelle)
16. Verordnung (EU) 2026/1744 zur Änderung der KI-Verordnung (EUR-Lex)
17. Europäische Kommission, AI Act Service Desk, Artikel 50 – Transparenzpflichten (Onlinequelle)
18. Jugendmedienschutz-Staatsvertrag, Fassung in Kraft seit 1. Dezember 2025 (Onlinequelle)
19. § 10a JuSchG, Schutzziele des Kinder- und Jugendmedienschutzes (Gesetzestext)
20. § 24a JuSchG, Vorsorgemaßnahmen (Gesetzestext)

Anhang C – Interne Prüfgrundlage dieser Arbeitsfassung

Die IST-/Lückenbewertung beruht auf einer lokalen Code- und Migrationsprüfung des DigiLernFlow-Arbeitsstands vom 9. August 2026. Besonders betrachtet wurden Quellenregistry und -indexierung, Promptgenerator-Quellenlöschung, RAG-Bindungen, Taskcard-Coach-Kontext, Safety-Gateway und -Trace, Lernspuren, Aufbewahrungsjob sowie bestehende Diagnose- und Verifikationstests. Vor einer externen Veröffentlichung ist die Bewertung gegen die tatsächlich ausgerollte Primärinstanz, deren Konfiguration und die abgeschlossenen Anbieter- beziehungsweise Schulverträge abzugleichen.