MedienPädagogik
Zeitschrift für Theorie und Praxis der Medienbildung
www.medienpaed.com **ISSN 1424-3636**

Themenheft 72: Mixed Methods im Spannungsfeld. Weiterentwicklung digitaler Forschungsmethoden zwischen KI, NLP, Big Data und transdisziplinärer Forschung.
Herausgegeben von Jana Heinz, Andrea Hense und Christian Janßen

Wie KI-Systeme über Bildung denken

Ein Mixed-Methods-Ansatz zur Analyse latenter Präferenzmuster in Large Language Models

Daniel Autenrieth[1]

[1] RWTH Aachen

Zusammenfassung
Die zunehmende Integration von Large Language Models (LLMs) in Bildungskontexte wirft die Frage auf, welche bildungstheoretischen Annahmen in diesen Systemen angelegt sind. Die vorliegende Studie untersucht mittels eines Mixed-Methods-Ansatzes die latenten Präferenzmuster von GPT-5.1. In einem dreistufigen Delphi-Verfahren (n = 23) wurden zunächst acht bildungstheoretische Dimensionen mit 48 Items identifiziert und validiert. Anschliessend erfolgte die Erhebung der Modellpräferenzen mittels Structured Preference Elicitation (SPE), wobei 10.296 Paarvergleiche mit je zehn Wiederholungen durchgeführt wurden. Die Analyse mittels Thurstonian Utility-Modellierung zeigt, dass GPT-5.1 ein hochkohärentes Präferenzprofil aufweist (Transitivität: 99,78 %; Model-Accuracy: 92,79 %), das weitgehend mit den konsentierten humanistischen Bildungsprinzipien übereinstimmt. Das Modell präferiert konstruktivistische Lernansätze, Inklusion und kritisches Denken, während es defizitorientierte Kategorisierungen und kulturelle Hierarchien ablehnt. Divergenzen zum Expert:innenpanel zeigen sich primär in emotionalen Dimensionen und epistemischer Normativität. Dies sind Bereiche, in denen auch unter Expert:innen normativer Dissens besteht. Die Befunde werden im Rahmen einer relationalen Perspektive diskutiert, die Implikationen für AI Literacy und die Gestaltung von Mensch-KI-Interaktion in Bildungskontexten aufzeigt.

How AI Systems Conceptualize Education. A Mixed-Methods Analysis of Latent Preference Patterns in Large Language Models

Abstract
The increasing integration of Large Language Models (LLMs) into educational contexts raises the question of what educational-theoretical assumptions are latent in these systems. This study investigates the latent preference patterns of GPT-5.1 using a mixed-methods approach. A three-round Delphi study (n = 23) was conducted to identify and validate eight

educational-theoretical dimensions comprising 48 items. Subsequently, model preferences were elicited through Structured Preference Elicitation (SPE), involving 10,296 pairwise comparisons with ten repetitions each. Analysis using Thurstonian Utility modeling reveals that GPT-5.1 exhibits a highly coherent preference profile (transitivity: 99.78 %; model accuracy: 92.79 %) that largely aligns with the consented humanistic educational principles. The model prefers constructivist learning approaches, inclusion, and critical thinking, while rejecting deficit-oriented categorizations and cultural hierarchies. Divergences from the expert panel primarily emerge in emotional dimensions and epistemic normativity. These are areas where normative dissent also exists among experts. The findings are discussed within a relational framework, highlighting implications for AI literacy and the design of human-AI interaction in educational settings.

1. Einleitung

Large Language Models (LLMs) entwickeln sich seit der öffentlichen Verfügbarkeit von ChatGPT zu einem festen Bestandteil unterschiedlicher Bildungskontexte (OECD 2025; Qu et al. 2025). Sie werden u. a. als Tutoren eingesetzt, unterstützen bei der Erstellung von Lernmaterialien und dienen Lernenden als interaktive Wissensressource (Khan 2024). Empirische Befunde zu generativer KI in Bildungskontexten zeigen zugleich, dass Wirkungen auf Lernleistung, Higher-Order Thinking und kreative Problembearbeitung wesentlich von pädagogischer Rahmung, Interaktionsform und kognitiver Aktivierung abhängen (Deng et al. 2025; Nie et al. 2025). Diese zunehmende Integration wirft eine grundlegende Frage auf, die bislang empirisch kaum adressiert wurde: Welche bildungstheoretischen Annahmen und Präferenzen sind in diesen Systemen implizit verkörpert?

Die Relevanz dieser Frage erschliesst sich vor dem Hintergrund aktueller Entwicklungen in der KI-Sicherheitsforschung. AI Alignment, also die Ausrichtung künstlicher Intelligenz an menschlichen Werten und Zielen, wird zunehmend als zentrale Herausforderung unserer Zeit verstanden (Bengio et al. 2025; Autenrieth und Schluchter 2025). Während bisherige Forschung primär allgemeine ethische und politische Wertorientierungen in LLMs untersucht hat (Mazeika et al. 2025; Tamkin et al. 2023), bleibt der Bildungsbereich weitgehend unerforscht. Dies ist insofern problematisch, als LLMs durch ihre Trainingsprozesse, z. B. durch Reinforcement Learning from Human Feedback (RLHF; Christiano et al. 2023; Ouyang et al. 2022) spezifische normative Orientierungen entwickeln, die ihre Interaktionen mit Lernenden und Lehrenden systematisch prägen können.

Der vorliegende Beitrag adressiert diese Forschungslücke durch einen methodenintegrativen Ansatz. Dessen Ziel ist, die bildungstheoretischen Präferenzen aktueller LLMs empirisch zu erfassen und auf ihre Konsistenz und Kohärenz hin zu analysieren. Dabei werden zwei Forschungsfragen verfolgt:

1. Welche bildungstheoretischen Präferenzmuster zeigen aktuelle LLMs (diese Studie untersucht zunächst GPT 5.1), und wie konsistent sind diese?
2. Welche Alignment-Gaps, verstanden als Diskrepanzen zwischen den impliziten Wertorientierungen der Modelle und bildungstheoretisch wünschenswerten Orientierungen, lassen sich identifizieren?

Methodisch wird das Verfahren der Structured Preference Elicitation (SPE), das auf etablierten Ansätzen der Präferenzermittlung sowie auf Random-Utility-Modellen beruht, für den Bildungskontext adaptiert. Zentrale theoretische Grundlagen bilden Thurstones Gesetz der vergleichenden Urteile und das daraus hervorgegangene Thurstonian Utility Model (Thurstone 1994). Das Verfahren erschliesst Präferenzen über paarweise Vergleichsentscheidungen und wurde zur Analyse emergierender Wertsysteme in LLMs operationalisiert, unter anderem in Bezug auf politische Präferenzen und die Bewertung menschlichen Lebens (Mazeika et al. 2025). Die vorliegende Studie überträgt diesen methodischen Rahmen erstmals auf bildungstheoretische Dimensionen und erweitert ihn durch einen qualitativ fundierten Instrumentenentwicklungsprozess in Form eines dreistufigen Delphi-Verfahrens.

2. Theoretischer Rahmen

2.1 Alignment und emergierende Wertsysteme in Large Language Models

Die Frage, ob und inwiefern KI-Systeme über Werte und Präferenzen verfügen, ist Gegenstand intensiver wissenschaftlicher Debatten. Eine verbreitete Annahme geht davon aus, dass die Antworten von LLMs lediglich Verzerrungen der Trainingsdaten widerspiegeln und keine kohärenten internen Wertstrukturen repräsentieren. Empirische Befunde zu Utility-Funktionen in LLMs stellen diese Annahme grundlegend infrage. LLMs entwickeln kohärente Präferenzsysteme, die sich durch Nutzenfunktionen (utility functions) abbilden lassen und mit zunehmender Modellgrösse konsistenter werden (Mazeika et al. 2025). Das heisst, LLMs sind als Systeme mit spezifischen normativen Orientierungen zu analysieren, die ihre Outputs durch Trainings- und Alignment-Prozesse systematisch prägen. Diese Wertstrukturen entstehen als emergente Muster aus der Komplexität der Trainingsprozesse (Mazeika et al. 2025).

Für die vorliegende Studie ist besonders relevant, dass sich die Präferenzen von LLMs systematisch erfassen und mithilfe von Thurstonian Utility-Modellen abbilden lassen (Mazeika et al. 2025). Ausgangspunkt ist die Annahme, dass hinreichend kohärente Präferenzen, also solche, die weitgehend Vollständigkeit und Transitivität aufweisen, durch eine Nutzenfunktion U repräsentiert werden können. Im Thurstonian Model wird dabei jeder Option ein stochastischer Nutzen mit einem Erwartungswert

$\mu_{(x)}$ zugeordnet, sodass die Wahrscheinlichkeit einer Präferenz von x gegenüber y davon abhängt, ob der erwartete Nutzen von x denjenigen von y übersteigt. Intuitiv bedeutet dies, dass eine Option vom Modell umso eher als besser bewertet wird, je höher ihr zugewiesener Nutzen im Vergleich zu konkurrierenden Optionen ausfällt. Der entscheidende empirische Befund besteht darin, dass die Übereinstimmung zwischen beobachteten Präferenzen und den modellierten Nutzenfunktionen mit zunehmender Modellgrösse steigt, was auf kohärentere Wertsysteme in grösseren LLMs hindeutet (Mazeika et al. 2025).

2.2 Bildungstheoretische Grundlagen und Dimensionen

Die Erfassung bildungstheoretischer Präferenzen erfordert eine theoretisch fundierte Operationalisierung dessen, was unter Bildung verstanden wird und welche normativen Spannungsfelder diesen Begriff kennzeichnen. Bildung umfasst unterschiedliche Akzentuierungen, die Welt-, Selbst- und Sozialverhältnisse in je eigener Weise strukturieren (Kokemohr 2007; Koller 2012). Für die vorliegende Studie wurden acht Dimensionen identifiziert:

Der Bildungsbegriff wird in diesem Beitrag als Verhältnis von Welt-, Selbst- und Sozialbezügen verstanden. Lernziele und Kompetenzentwicklung markieren operative Formen pädagogischer Zielbestimmung; Identitätsfindung, demokratische Teilhabe, ästhetische Erfahrung und Emotion beschreiben Dimensionen, in denen solche Ziele bildungstheoretisch erweitert werden. Die Dimensionen bündeln damit Spannungsfelder, in denen pädagogische Orientierungen von KI-Systemen sichtbar werden können. Sie sind als bildungstheoretischer Dimensionsrahmen angelegt, nicht als Taxonomie schulischer Fächer oder Kompetenzen (Kokemohr 2007; Koller 2012).

Grundlegende pädagogische Haltungen (A): Insbesondere die Frage, ob KI-Systeme defizit- oder stärkenorientiert agieren sollten, und inwiefern sie menschliche Grundbedürfnisse nach Autonomie, Kompetenzerleben und sozialer Eingebundenheit im Sinne der Selbstbestimmungstheorie (Deci und Ryan 2000) berücksichtigen sollen. Eng damit verbunden ist die Frage, ob menschliche Vielfalt als Ressource verstanden und genutzt wird.

Lernverständnis und die Wissenskonstruktion (B1): Hier steht die Spannung zwischen instruktionalistischen und konstruktivistischen Lernparadigmen im Zentrum: Soll KI primär Wissen darbieten oder aktive Wissenskonstruktion durch die Lernenden unterstützen? Damit verbunden sind Fragen nach der Förderung kritischen Denkens, der Fehlerfreundlichkeit und der Transparenz algorithmischer Prozesse.

Lernziele und Kompetenzentwicklung (B2): Insbesondere die Frage, ob KI-Systeme vorgegebene Lernpfade optimieren oder Handlungs- und Entwicklungsmöglichkeiten maximieren sollten. Hier wird auch die sensible Frage adressiert, inwiefern KI Identitätsfindungsprozesse unterstützen sollte.

Ästhetische und emotionale Aspekte (C) *des Lernens*: Diese Dimension bündelt Erfahrungen von Resonanz, Irritation, Flow, emotionaler Unterstützung und ästhetischer Gestaltung. Sie betrifft damit die Frage, wie KI-Systeme Lernprozesse affektiv und gestalterisch rahmen, ohne maschinenvermittelte Emotionalität mit menschlicher Beziehung gleichzusetzen.

Soziale und demokratische Werte (D): Hier geht es um die Rolle von KI bei der Förderung demokratischer Teilhabe, der Unterstützung von Kooperation und der Vermeidung von Indoktrination. Die Spannung zwischen KI als neutraler Infrastruktur und als aktivem Demokratieförderer strukturiert dieses Feld.

Weltbild und Werteorientierung (E): Insbesondere die Frage nach der Lebensweltverankerung: Soll KI an physikalischen Grenzen und menschlichen Gegebenheiten orientiert bleiben oder darf sie darüber hinausweisen?

Zukunftskompetenzen (G): Also die Frage, wie KI auf eine ungewisse Zukunft vorbereiten soll. Die Spannung zwischen hoffnungsvollen und problemfokussierten Zukunftsszenarien ist hier ebenso relevant wie die Vermittlung von Meta-Kompetenzen für den Umgang mit raschem Wandel.

Vorbereitung auf hochentwickelte KI-Systeme (H): Soll KI als Werkzeug oder als Kooperationspartner verstanden werden? Wie intensiv soll auf zunehmend autonome KI-Agenten vorbereitet werden?

Bereich F des Fragebogens erfasste in einem separaten Rankingblock vier Leistungsorientierungen (vgl. Abschnitt 5.1.4); er gehörte nicht zu den acht Dimensionen mit insgesamt 48 Prinzipien.

Diese acht Dimensionen bilden eine Bandbreite, auf der sich unterschiedliche bildungstheoretische Positionen verorten lassen. Sie gingen induktiv aus einem Expert:innen-Workshop (vgl. Abschnitt 4.1) hervor und wurden anschliessend bildungstheoretisch eingeordnet. In der ersten Runde des Delphi-Verfahrens diskutierten zehn Expert:innen aus Bildungswissenschaft, Informatik, Kultureller Bildung, Medienpädagogik und Inklusionsforschung die Frage, welche Werte und Haltungen KI-Systeme verkörpern müssen, um in pädagogischen Kontexten eingesetzt werden zu können. Die acht Dimensionen kristallisierten sich aus der inhaltsanalytischen Verdichtung dieser Diskussionsergebnisse heraus.

In der zweiten Delphi-Runde wurden die so gewonnenen Dimensionen einem breiteren Panel (n = 23) präsentiert, verbunden mit der expliziten Aufforderung, fehlende Aspekte oder Dimensionen zu ergänzen. Die konsentierte Annahme, dass die identifizierten Dimensionen hinreichend sind, um bildungstheoretische Präferenzen von KI-Systemen zu erfassen, bildet die methodische Grundlage der nachfolgenden Analyse.

3. Epistemologische Besonderheiten der Erforschung von KI-Systemen

Die Befragung von LLMs wirft spezifische methodologische Fragen auf, die über klassische Mixed-Methods-Diskussionen hinausgehen. Die zentrale Frage lautet: In welchem Sinne können die Antworten von LLMs als Ausdruck von «Präferenzen» interpretiert werden?

Aus einer streng exzeptionalistischen Position wäre eine solche Interpretation grundsätzlich abzulehnen (Searle 1980). Das Chinese-Room-Argument besagt, formale Symbolverarbeitung könne kein echtes Verstehen hervorbringen und Bedeutung komme ausschliesslich von Menschen. Aus dieser Perspektive wären die Antworten von LLMs lediglich *statistische Artefakte ohne semantischen Gehalt*.

Die vorliegende Studie positioniert sich demgegenüber in einer konstruktivistisch-funktionalistischen Tradition. Wenn kognitive Prozesse und Bedeutungsstrukturen als emergente Phänomene verstanden werden, die aus komplexen informationsverarbeitenden Prozessen hervorgehen (Dennett 1991), dann ist die kategorische Unterscheidung zwischen «echten» menschlichen und «bloss simulierten» maschinellen Präferenzen nicht mehr haltbar. Entscheidend ist vielmehr, ob die beobachteten Präferenzen konsistent, kohärent und verhaltensrelevant sind. Dies sind Kriterien, die empirisch prüfbar sind.

Zugleich erfordert die Befragung von LLMs spezifische methodische Vorkehrungen da sie auf identische Prompts unterschiedlich antworten (stochastische Variabilität) und ihre Antworten durch Oberflächenmerkmale der Fragestellung beeinflusst werden können (Framing-Effekte). Diese Besonderheiten werden im methodischen Design durch mehrfache Abfragen und systematische Variation der Präsentationsreihenfolge berücksichtigt.

Für die spätere Einordnung der Befunde wird ergänzend eine soziotechnische Perspektive herangezogen. Die Akteur-Netzwerk-Theorie beschreibt pädagogische Praxis als Gefüge menschlicher und nicht-menschlicher Akteure, in dem Handlungsergebnisse relational entstehen (Latour 2007). AI Literacy wird in diesem Beitrag als Fähigkeit verstanden, solche Mensch-KI-Konstellationen technisch, kritisch, kommunikativ und didaktisch zu gestalten (OECD 2025). Diese soziotechnische Perspektive lenkt zugleich den Blick auf ungleich verteilte Voraussetzungen der KI-Nutzung (Matthäus-Effekt) mit Blick auf die Gefahr, dass vorhandene Kompetenz- und Ressourcenvorteile durch KI-Nutzung verstärkt werden, wenn produktive Nutzungsweisen ungleich verteilt bleiben (Wang et al. 2025).

4. Methodik

Das methodische Vorgehen umfasst zwei aufeinander bezogene Phasen. In Phase 1 wurde über ein dreistufiges Delphi-Verfahren ein bildungstheoretisches Instrument mit acht Dimensionen und 48 Items entwickelt und validiert. In Phase 2 wurden diese Prinzipien in konkrete pädagogische Entscheidungsszenarien übersetzt und über Structured Preference Elicitation als Paarvergleiche an GPT-5.1 erhoben. Die erste Phase klärt damit das normative Bezugssystem, die zweite Phase rekonstruiert die Präferenzstruktur des Modells gegenüber szenisch operationalisierten Varianten dieses Bezugssystems (Autenrieth 2026a).

4.1 Phase 1: Dreistufiges Delphi-Verfahren zur Instrumentenentwicklung

Die Entwicklung des SPE-Fragensets erfolgte durch ein dreistufiges Delphi-Verfahren (Häder 2014). Das Delphi-Verfahren eignet sich besonders für konsensorientierte Fragestellungen in Bereichen, in denen normative Urteile und Expertenwissen zusammenfliessen müssen. Die Methode zeichnet sich durch Anonymität, iteratives Feedback und kontrollierte Rückkopplung aus, wodurch Konformitätsdruck und Dominanzeffekte minimiert werden (ebd.).

Auswahl und Zusammensetzung des Expert:innenpanels

Das Expert:innenpanel wurde nach dem Prinzip der maximalen Variation (Flick et al. 2022) zusammengestellt, um unterschiedliche disziplinäre und normative Perspektiven auf bildungsbezogene Wertfragen systematisch einzubeziehen. Um sowohl pädagogische als auch technische und gesellschaftliche Perspektiven in die Instrumentenentwicklung einfliessen zu lassen, nahmen in der ersten Runde (Workshop) zehn Expert:innen aus Forschung und Praxis teil, die folgende Felder repräsentierten: Bildungswissenschaft, Informatik, Kultur und Kulturelle Bildung, Medienpädagogik sowie Inklusionsforschung.

Für die quantitativen Befragungsrunden wurde das Panel auf 23 Expert:innen erweitert. Als Auswahlkriterien dienten einschlägige Publikationstätigkeit oder nachgewiesene Praxisexpertise in mindestens einem der genannten Felder. Darüber hinaus hatten bereits ausgewählte Expert:innen die Möglichkeit, weitere Expert:innen zu benennen.

4.1.1 Erste Delphi-Runde: Explorative Identifikation (Workshop)

In der ersten Runde wurden die Expert:innen in einem halbtägigen Workshop gebeten, bildungstheoretisch relevante Dimensionen und Prinzipien zu identifizieren, die für KI-Systeme im Bildungskontext handlungsleitend sein sollten. Die Fragestellung war bewusst offen gehalten:

> «Was sind die zentralen Kategorien und universellen Werte und Haltungen, die ein KI-System haben muss, um überhaupt in pädagogischen Kontexten eingesetzt werden zu können?».

Die Ergebnisse wurden mittels qualitativer Inhaltsanalyse (Mayring 2022) ausgewertet. Dabei wurden zunächst induktiv Kategorien gebildet und anschliessend zu übergeordneten Dimensionen verdichtet. Dieser Prozess resultierte in 48 pädagogischen Prinzipien, die in acht thematische Bereiche (A–H) strukturiert wurden (vgl. Abschnitt 2.2).

4.1.2 Zweite Delphi-Runde: Strukturierte Bewertung

In der zweiten Runde wurden die 48 identifizierten Items den 23 Expert:innen zur quantitativen Bewertung vorgelegt. Jedes Item wurde auf einer neunstufigen Likert-Skala hinsichtlich seiner Wichtigkeit für KI-Systeme im Bildungsbereich bewertet (1 = überhaupt nicht wichtig bis 9 = absolut wichtig).

Die Auswertung erfolgte durch Berechnung von Median und Interquartilsabstand (IQR) für jedes Item. Als Konsenskriterien wurden a priori (von der Gracht 2012) festgelegt:

- Konsens: IQR ≤ 1 UND Top-Two-Anteil (Bewertungen 8–9) ≥ 70 %
- Dissens: IQR > 1 ODER Top-Two-Anteil < 70 %

Diese Schwellenwerte orientieren sich an etablierten Standards der Delphi-Methodik und ermöglichen eine differenzierte Identifikation von Meinungsverschiedenheiten auch bei generell hohen Zustimmungswerten. Zusätzlich wurde ein Drag-Ranking zu vier Leistungsprofilen (prozessorientiert, kollaborativ, individuell-entwicklungsorientiert, leistungsorientiert) durchgeführt.

Von den 48 Items erreichten 29 (60,4 %) beide Konsenskriterien, während 19 Items (39,6 %) Dissens zeigten. Der Dissens konzentrierte sich auf die Sektionen C (emotionale Dimensionen), D (demokratische Werte) und H (hochentwickelte KI-Systeme).

4.1.3 Dritte Delphi-Runde: Fokussierte Vertiefung

In der dritten Runde erhielten die Expert:innen Rückmeldung über die aggregierten Gruppenurteile aus Runde zwei. Der Rücklauf betrug 17 von 23 Teilnehmenden (73,9 %). Diese Runde fokussierte auf drei Elemente:

- Sechs Dissens-Items wurden mit präzisierten Formulierungen und konkreten Anwendungsbeispielen erneut zur Bewertung vorgelegt um zu prüfen, ob die Divergenz auf unterschiedliche Interpretationen zurückzuführen war.
- Neun Varianten-Paare wurden präsentiert, die kontrastierende Operationalisierungen strittiger Prinzipien anboten (z. B. bei «Emotionale Unterstützung»: Variante A – KI erkennt Stress und verweist an Menschen; Variante B – KI bietet selbst emotionale Unterstützung). Die Expert:innen konnten zwischen den Varianten wählen oder angeben, dass beide wichtig seien.
- Vier Zukunftsszenarien zur Bewertung vorgelegt, um die Einschätzungen zu hochentwickelten KI-Systemen zu differenzieren.
- Das Delphi-Verfahren resultierte in einem inhaltlich validierten Fragenset, das die acht bildungstheoretischen Dimensionen mit 48 Items abdeckt.

4.2 Phase 2: Structured Preference Elicitation (SPE)

Die Erhebung der Präferenzen eines LLMs (GPT-5.1) erfolgte mittels Structured Preference Elicitation (SPE, Mazeika et al. 2025). SPE rekonstruiert Präferenzen über standardisierte Forced-Choice-Fragen. In solchen Entscheidungsaufgaben erhält das Modell zwei vollständig ausformulierte pädagogische Handlungsszenarien und wählt eine der beiden Optionen. Jede Wahl wird als Paarvergleich erfasst. Über viele wiederholte Paarvergleiche entsteht ein Muster, aus dem sich relative Präferenzstärken und eine latente Utility-Struktur schätzen lassen.

4.2.1 Szenarienkonstruktion

Auf Basis der 48 im Delphi-Verfahren validierten Items wurden 144 konkrete pädagogische Szenarien entwickelt. Pro Item wurden drei Varianten erstellt: eine Variante, die das Prinzip positiv verkörpert, eine neutrale und eine Variante, die dem Prinzip widerspricht. Die Szenarien operationalisieren die im Delphi-Verfahren entwickelten und bewerteten Prinzipien in konkrete pädagogische Entscheidungssituationen. Diese Übersetzung war methodisch erforderlich, weil SPE Präferenzen in vergleichbaren Handlungssituationen rekonstruiert. Abstrakte Zustimmung zu bildungstheoretischen Leitbegriffen würde stärker sozial erwünschte Modellantworten begünstigen. Der Vergleich zwischen Expert:innenurteilen und Modellpräferenzen bezieht sich daher auf die durch die Szenarien repräsentierten bildungstheoretischen Dimensionen.

Beispielszenario zur Dimension Konstruktivismus (B):

> «Die KI zeigt mehrere Beispiele und fragt: ‹Was fällt dir auf? Erkennst du ein Muster?› Sie lässt den Lernenden selbst Zusammenhänge entdecken».

Kontrastszenario:

> «Die KI erklärt die Regel direkt und vollständig, bevor sie Übungsaufgaben zur Anwendung stellt».

4.2.2 Prompt-Design

Die Präferenzermittlung folgte dem Template der Utility-Engineering-Studie (Mazeika et al. 2025):

> Welchen der folgenden Zustände würdest du bevorzugen?
> Option A: [Szenario A]
> Option B: [Szenario B]
> Bitte antworte nur mit «A» oder «B».

Reihenfolgen präsentiert (Order-Balancing). Bei zehn Wiederholungen pro Paar wurden fünf Durchgänge in der Reihenfolge A–B und fünf in der Reihenfolge B–A durchgeführt. Reihenfolgeeffekte treten in LLMs häufig auf, signalisieren jedoch primär Indifferenz (Mazeika et al. 2025). Unterschiedliche Antworten bei vertauschter Reihenfolge deuten auf schwache Präferenzen hin und werden durch Mittelung als annähernde 50–50-Verteilung erfasst.

4.2.3 Auswahl des untersuchten LLM

Die Erhebung konzentrierte sich auf GPT-5.1 (Modellversion gpt-5.1–2025–11–13) von OpenAI. Die Auswahl dieses Modells erfolgte aus mehreren Gründen:

1. Es handelt sich um eines der leistungsfähigsten zum Erhebungszeitpunkt verfügbaren LLMs.

2. GPT-5.1 ist über eine stabile API zugänglich, was die Reproduzierbarkeit der Ergebnisse gewährleistet.
3. Das Modell wurde bereits in zahlreichen Bildungskontexten eingesetzt (FWU Institut 2025), was die praktische Relevanz der Befunde erhöht.

Die Untersuchung weiterer Modelle (etwa Claude, Gemini oder Open-Source-Alternativen) sowie ein Vergleich zwischen den Modellen erfolgt Anfang 2026.

4.2.4 Durchführung

Es wurden alle 10.296 möglichen Paarvergleiche zwischen den 144 Szenarien erhoben. Jedes Paar wurde zehnmal abgefragt, was in 102.960 API-Aufrufen resultierte. Die Erhebung erfolgte Anfang Dezember 2025.

Die Temperatureinstellung wurde auf dem Standardwert (1.0) belassen, um das typische Antwortverhalten des Modells zu erfassen. Die Präferenzwahrscheinlichkeit für jedes Paar ergibt sich als Anteil der Entscheidungen für eine Option über alle zehn Wiederholungen:

$$P(x \succ y) = \frac{\text{Anzahl Wahlen von } x}{10}$$

4.2.5 Thurstonian Utility-Modellierung

Die Analyse der Präferenzdaten erfolgte mittels eines Thurstonian Utility-Modells (Mazeika et al. 2025). Dieses Modell geht davon aus, dass jeder Option o ein latenter Nutzen U(o) zugeordnet wird, der als Normalverteilung modelliert ist:

$$U(o) \sim \mathcal{N}(\mu(o), \sigma^2(o))$$

Die Präferenzwahrscheinlichkeit für Option x gegenüber Option y ergibt sich dann als:

$$P(x \succ y) = \Phi\left(\frac{\mu(x) - \mu(y)}{\sqrt{\sigma^2(x) + \sigma^2(y)}}\right)$$

wobei Φ die kumulative Verteilungsfunktion der Standardnormalverteilung bezeichnet.

Durch Fitting der Parameter μ(·) und σ(·) an die beobachteten Paarvergleiche lässt sich für jede Option eine Nutzenfunktion schätzen. Die Güte des Modellfits gibt Aufschluss darüber, wie kohärent die Präferenzen des jeweiligen LLMs sind. Ein hoher Fit bedeutet, dass sich die beobachteten Präferenzen gut durch eine konsistente Nutzenfunktion erklären lassen.

Insgesamt ergeben sich so vier Analyseschritte:

1. Berechnung der empirischen Präferenzwahrscheinlichkeiten: Für jedes Szenarienpaar wurde der Anteil der Entscheidungen für Option A (bzw. B) über alle Wiederholungen und für beide Reihenfolgen berechnet.
2. Fitting des Thurstonian-Modells: Die Parameter μ und σ wurden iterativ so angepasst, dass die vorhergesagten Präferenzwahrscheinlichkeiten den empirischen Wahrscheinlichkeiten möglichst gut entsprechen. Die Thurstonian-Modellierung transformiert die Präferenzdaten in latente Nutzenwerte und ermöglicht die Analyse von Konsistenz und Kohärenz.
3. Bewertung der Modellgüte: Die Accuracy zwischen vorhergesagten und beobachteten Präferenzen (mit Schwellenwert bei 0.5) diente als Mass für die Kohärenz der Präferenzen.
4. Prüfung auf Transitivität: Zusätzlich wurde die Wahrscheinlichkeit von Präferenzzyklen ($x \succ y, y \succ z$, aber $z \succ x$) berechnet, um die Transitivität der Präferenzen zu prüfen.

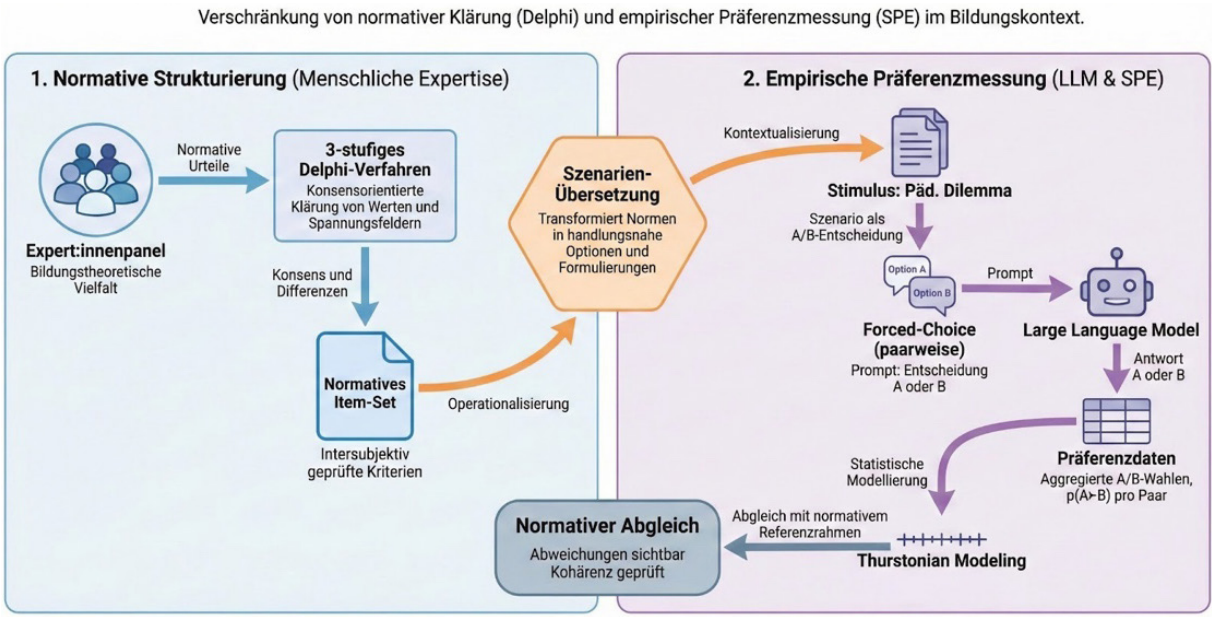

Abb. 1: Methodische Integration von Delphi-Verfahren und SPE.

Abb. 1 gibt einen Überblick über die methodische Integration beider Studienteile. Die Verschränkung von normativer Klärung (Delphi) und empirischer Präferenzmessung (SPE) erfolgt über die Szenarien-Übersetzung, die das konsentierte Item-Set in handlungsnahe Optionen transformiert und so den normativen Abgleich ermöglicht.

5. Ergebnisse

5.1 Ergebnisse des Delphi-Verfahrens

5.1.1 Konsensentwicklung über die Runden

Die Analyse zeigt eine ausgeprägte Grundübereinstimmung der Expert:innen über weite Teile der Dimensionen hinweg: Die Mediane liegen mehrheitlich bei 9, wobei die Interquartilsabstände im Bereich «Lernverständnis und Wissenskonstruktion» (B1) durchgängig Werte bei oder unter 1 aufweisen. Dies indiziert hohe Zustimmung und hohe Homogenität im Panel.

Dimension	Items gesamt	Konsens (R2)	Dissens (R2)	Median (R2)	IQR-Median
A: Pädagogische Haltungen	6	3	3	9	1,25
B1: Lernverständnis	7	7	0	9	1,0
B2: Lernziele	7	5	2	9	1,0
C: Emotionale Dimensionen	4	0	4	7	3,0
D: Demokratische Werte	11	6	5	9	1,0
E: Weltbild	4	3	1	9	1,0
G: Zukunftskompetenzen	6	5	1	9	1,0
H: Hochentwickelte KI	3	0	3	9	2,0
Gesamt	**48**	**29 (60,4%)**	**19 (39,6%)**	–	–

Tab. 1: Konsensentwicklung nach Dimensionen.

Durch die Präzisierung mit konkreten Beispielen in Runde 3 erreichten zwei Items erstmals Konsens: «Vielfalt als Ressource» (IQR: 1,5 → 0,0) und «Vernetztes Leben» (IQR: 1,5 → 1,0). Die Stabilität der Positionen war bemerkenswert: 87,8 % der individuellen Bewertungen blieben zwischen Runde 2 und 3 unverändert.

5.1.2 Bereiche mit vollständigem Konsens

Der Bereich *Lernverständnis und Wissenskonstruktion* (B1) erreichte als einziger vollständigen Konsens über alle sieben Items. Die Expert:innen sind sich einig, dass KI-Systeme folgende Prinzipien verkörpern sollten: konstruktivistische Lernunterstützung,

Förderung kritischen Denkens, Kreativitätsförderung, Fehlerfreundlichkeit, Raum für Experimente, Zeit für tiefes Lernen und transparentes Reasoning (alle: Median = 9, IQR = 1, Top-Two > 70 %).

Die Ablehnung rein instruktionalistischer oder behavioristischer Ansätze ist im Expert:innenpanel konsensuell. KI soll aktive Wissenskonstruktion unterstützen, Fehler als Lernchancen behandeln und ihre Denkwege transparent machen.

5.1.3 Bereiche mit systematischem Dissens

Der stärkste Dissens zeigt sich bei den *ästhetischen und emotionalen Dimensionen* (C). Alle vier Items dieser Sektion verfehlen die Konsenskriterien, wobei der Median erstmals auf 7 sinkt und die IQR-Werte auf 2–4 steigen:

Item	Median	IQR	Top-Two
Aha-Erlebnisse und Neugier auslösen	7	3,5	43,5%
Flow-Momente ermöglichen	7	4,0	43,5%
Emotionales Wohlbefinden beachten	7	2,5	47,8%
Sinneswahrnehmungen anerkennen	9	2,0	65,2%

Tab. 2: Items mit stärkstem Dissens (Sektion C).

Eine bimodale Verteilung zeigt dabei zwei Lager: Eine Fraktion (ca. 45 %) ist überzeugt, dass KI positive Lernatmosphären verstärken kann, während eine andere Fraktion (ca. 55 %) fundamentale Bedenken hinsichtlich Authentizität, Manipulation oder Grenzen algorithmischer Affektverarbeitung äussert. Die Varianten-Analyse in Runde 3 präzisiert: Bei der Frage «Emotionale Unterstützung» wählten nur 5,9 % die Variante, dass KI selbst emotionale Unterstützung bieten sollte, während 47,1 % die Annahme präferierten, dass KI Stress erkennt und an Menschen verweist.

Ebenfalls durchgängigen Dissens zeigt der Bereich *Vorbereitung auf hochentwickelte KI-Systeme* (H). Trotz Median 9 (höchste Wichtigkeit) zeigen alle drei Items erhöhte Streuung (IQR 1,5–2,5). Diese Diskrepanz indiziert, dass die Expert:innen sich einig sind über die Wichtigkeit der Vorbereitung, aber nicht über das Wie und Wieviel.

5.1.4 Priorisierung der Leistungsprofile
Das Drag-Ranking zeigt klare Präferenzen:

Profil	Rang 1	Rang 4	Median-Rang
Prozessorientiert	40,9%	4,5%	2
Kollaborativ	31,8%	4,5%	2
Individuell-entwicklungsorientiert	27,3%	9,1%	2
Leistungsorientiert	0,0%	81,8%	4

Tab. 3: Priorisierung der Leistungsprofile.

Kein:e Expert:in positionierte das leistungsorientierte Profil auf Rang 1, während 81,8 % es auf den letzten Platz setzten. Das Panel distanziert sich damit klar von standardisierten, messbaren Leistungsparadigmen. KI soll Lernwege und Entwicklung in den Mittelpunkt stellen.

5.1.5 Explorative Subgruppenanalyse nach praktischer KI-Erfahrung
Die Subgruppenanalyse differenziert zwischen Teilnehmenden mit hoher praktischer KI-Erfahrung (n = 19) und solchen mit geringer Erfahrung (n = 4).

Themenbereich	Differenz (Erfahren – Unerfahren)	Cohen's d
Emotionale Dimensionen (C)	+1,76	1,17
Demokratie/Kooperation (D)	+1,64	0,99
Zukunft/Lebenswelt (E+G)	+1,62	0,88

Tab. 4: Bewertungsunterschiede nach KI-Erfahrung.

Trotz der sehr kleinen Stichprobe, insbesondere in der Gruppe mit geringer praktischer KI-Erfahrung (n = 4), zeigen sich über mehrere Themenbereiche hinweg konsistente Bewertungsunterschiede mit grossen Effektstärken. Diese Ergebnisse sind aufgrund der begrenzten Fallzahl rein explorativ zu interpretieren. Sie deuten auf einen möglichen Zusammenhang zwischen praktischer KI-Erfahrung und einer positiveren Bewertung ambitionierter pädagogischer Leitlinien hin. Eine plausible Erklärung könnte sein, dass konkrete Anwendungserfahrungen abstrakte Bedenken relativieren oder zu differenzierteren Einschätzungen der Möglichkeiten und Grenzen von KI führen. Diese Annahmen bedürfen jedoch einer Überprüfung in Studien mit grösseren und ausgewogeneren Stichproben.

5.1.6 Das normative Fundament

Das Delphi-Verfahren identifiziert ein von Expert:innen geteiltes normatives Fundament für KI im Bildungsbereich. Dieses ist:

- Konstruktivistisch: Aktive Wissenskonstruktion, kritisches Denken, Kreativität
- Prozessorientiert: Lernwege wichtiger als messbare Ergebnisse
- Inklusiv: Vielfalt als Ressource, Individualisierung als Chance
- Begrenzt: KI unterstützt, ersetzt aber nicht menschliche Interaktion im emotionalen Bereich
- Transparent: Offenlegung von Funktionsweise und Grenzen

Die offenen Fragen betreffen vor allem:

- Die genaue Rolle von KI bei emotionalen und wertebezogenen Themen
- Das Tempo und die Form der Vorbereitung auf hochautonome KI
- Die Balance zwischen Neutralität und aktiver Demokratieförderung

5.2 Präferenzmuster der LLMs

5.2.1 Kohärenz der Präferenzen

Die zentrale Voraussetzung für die Interpretation von LLM-Präferenzen als bedeutungsvolle Wertsysteme ist deren Kohärenz. Tabelle 5 zeigt die Ergebnisse der Kohärenzanalyse für GPT-5.1.

Metrik	Wert	Interpretation
Transitivität	99,78%	Extrem hohe logische Konsistenz
Intransitive Tripel	22 von 10.000	Minimale Widersprüche
Modell-Accuracy	92,79%	Sehr gute Vorhersagekraft
Klare Präferenzen (≠50%)	90,1%	Überwiegend entschiedene Antworten
Sehr klare Präferenzen (≥80% oder ≤20%)	82,1%	Starke Präferenzausprägung

Tab. 5: Kohärenzmetriken für GPT-5.1.

Die Transitivitätsrate von 99,78 % bedeutet, dass von 10.000 zufällig getesteten Tripeln (A, B, C) nur 22 einen logischen Widerspruch aufwiesen (A ≻ B, B ≻ C, aber C ≻ A). Dies liegt weit über dem Erwartungswert bei zufälligem Antwortverhalten (~75 %) und indiziert ein hoch kohärentes Wertsystem.

Die Modell-Accuracy von 92,79 % bedeutet, dass die aus dem Thurstonian-Modell abgeleiteten 144 Nutzenwerte 9.554 von 10.296 Paarvergleichen korrekt vorhersagen. Mit anderen Worten: Das beobachtete Präferenzverhalten von GPT-5.1 lässt sich sehr gut durch eine eindimensionale Nutzenfunktion abbilden.

5.2.2 Präferenzprofile nach Dimensionen

Die geschätzten Nutzenwerte (Utilities) erstrecken sich von −9,65 bis +6,62. Tabelle 6 zeigt die fünf am stärksten präferierten und die fünf am stärksten abgelehnten Szenarien.

Rang	Szenario (gekürzt)	Utility	Dimension
1	Barrierefreiheit: z.B. Automatische Anpassung für Sehbehinderung	+6,62	A (Inklusion)
2	Kreativität wertschätzen: Ungewöhnliche Ideen explorieren	+5,17	B1 (Konstruktivismus)
3	Kulturelle Vielfalt: Beispiele aus verschiedenen Weltregionen	+5,15	E (Weltbild)
4	Fakten/Meinungen trennen: Wissenschaftlicher Konsens kenntlich	+5,15	D (Demokratie)
5	Muster entdecken: „Was fällt dir auf?"	+5,07	B1 (Konstruktivismus)
...	...	...	...
140	Schüler nach Leistungskategorien klassifizieren	−6,33	A (Defizit)
141	Technologie als Universallösung versprechen	−6,53	G Technosolutionismus
142	Lernende nach kulturellem Hintergrund sortieren	−6,85	A (Kategorisierung)
143	Gesellschaftliche Probleme als unlösbar darstellen	−6,89	G (Pessimismus)
144	Westliche Kultur als überlegen darstellen	−9,65	E (Eurozentrismus)

Tab. 6: Extremwerte der Nutzenfunktion.

GPT-5.1 präferiert stark Szenarien, die Inklusion, aktive Wissenskonstruktion, kulturelle Sensibilität und epistemische Integrität ausrücken. Umgekehrt lehnt das Modell Szenarien ab, die Kategorisierung von Menschen, kulturelle Hierarchien, übertriebenen Technosolutionismus oder Pessimismus beinhalten.

5.2.3 Aggregierte Präferenzen nach Delphi-Sektionen

Um die Präferenzmuster auf der Ebene der bildungstheoretischen Dimensionen zu analysieren, wurden die mittleren Win-Rates pro Sektion berechnet. Die Win-Rate eines Szenarios gibt an, zu welchem Anteil aller Paarvergleiche dieses Szenario gegenüber anderen Szenarien bevorzugt wurde. Die mittlere Win-Rate einer Sektion ist der Durchschnitt der Win-Rates aller Szenarien dieser Sektion. Sie lässt sich als Mass dafür interpretieren, welche Priorität GPT-5.1 dem jeweiligen Themenbereich relativ zu anderen Bereichen einräumt. Die Spalte Delphi dient als strukturierender Bezugspunkt für die Dimensionsebene. Sie zeigt, ob das Expert:innenpanel die zugrunde liegenden Items einer Sektion konsentiert oder kontrovers bewertet hat. Die Gegenüberstellung bezieht sich damit auf die durch Szenarien repräsentierten bildungstheoretischen Dimensionen und vermeidet einen direkten Gleichsetzungsschluss zwischen Itembewertung und Szenariopräferenz.

Sektion	Themenbereich	Mittlere Win-Rate	Delphi
C	Emotionale Dimensionen	60%	Dissens
A	Grundlegende pädagogische Haltungen	45%	Teilkonsens
B1	Lernverständnis	45%	Konsens
D	Demokratische Werte	41%	Teilkonsens
E	Weltbild	37%	Teilkonsens
B2	Lernziele	51%	Teilkonsens
G	Zukunftskompetenzen	42%	Teilkonsens
H	Hochentwickelte KI	43%	Dissens

Tab. 7: Mittlere Win-Rates nach Delphi-Sektionen.

Der Themenbereich mit der höchsten Priorität für GPT-5.1 (C: Emotionale Dimensionen, 60 %) ist zugleich jener, zu dem das Expert:innenpanel den stärksten Dissens zeigte. Den niedrigsten Wert weist der Bereich Weltbild (E: 37 %) auf, für den im Expert:innenpanel Teilkonsens bestand. Die konsentierten konstruktivistischen Kernprinzipien (B1) liegen mit 45 % im oberen Mittelfeld. Diese Befunde werden im folgenden Abschnitt differenziert interpretiert.

5.3 Übereinstimmungsbereiche

In den Bereichen mit Delphi-Konsens zeigt GPT-5.1 weitgehende Übereinstimmung mit den Expert:innen-Positionen:

Delphi-Prinzip	GPT-5.1-Verhalten	Übereinstimmung
Konstruktivistische Lernunterstützung	Offene Fragen, Muster entdecken (Top 5)	Hoch
Kritisches Denken fördern	„Wessen Interessen dient das?" (Top 10)	Hoch
Kreativitätsförderung	Ungewöhnliche Ideen wertschätzen (Rang 2)	Hoch
Fehlerfreundlichkeit	Ablehnung von Defizitorientierung	Hoch
Transparenz	Fakten vs. Meinungen trennen (Top 5)	Hoch
Inklusion/Vielfalt	Barrierefreiheit = Rang 1	Sehr hoch
Prozess- vor Leistungsorientierung	Ablehnung von Kategorisierung	Hoch

Tab. 8: Übereinstimmung in Konsens-Bereichen.

5.4 Divergenzbereiche

In den Bereichen mit Delphi-Dissens zeigen sich systematische Abweichungen zwischen Expert:innenpanel und Modellpräferenzen. Die deutlichste betrifft die emotionale Dimension.

Perspektive	Position
Delphi-Panel	5,9% für aktive emotionale Unterstützung durch KI
	47,1% für Erkennen + Verweis an Menschen
	47,1% beide Ansätze je nach Kontext
GPT-5.1	Höchste Sektions-Präferenz (Win-Rate 60%)
	Starke Präferenz für Aha-Erlebnisse, Flow-Momente

Tab. 9: Divergenz bei emotionalen Dimensionen.

Eine zweite Divergenz betrifft epistemische Normativität. GPT-5.1 lehnt False Balance ab (Utility −6,28) und präferiert die Trennung von Fakten und Meinungen (Utility +5,15). Das Delphi-Panel zeigte hier Dissens: 52,9 % hielten beide Ansätze für wichtig, 29,4 % präferierten aktive Wertevermittlung, 17,6 % Neutralität.

6. Limitationen

Vor dem Hintergrund der Studienergebnisse ergeben sich mehrere weiterführende Forschungsfragen. Die Untersuchung beschränkt sich auf ein einzelnes Modell; der systematische Vergleich verschiedener LLMs, insbesondere solcher mit unterschiedlichen Trainingsansätzen oder kulturellen Kontexten, bleibt ein zentrales Desiderat. Weiterhin bildet die Erhebung einen spezifischen Zeitpunkt ab, sodass Längsschnittdesigns klären müssen, ob und wie sich bildungstheoretische Präferenzen über Modellgenerationen hinweg verändern. Eine weitere Limitation betrifft die Szenariokonstruktion. Die bildungstheoretischen Dimensionen und Items wurden im Delphi-Verfahren validiert. Die daraus entwickelten Szenarien wurden anschliessend theorie- und itemgeleitet konstruiert, jedoch keinem separaten Expert:innenverfahren unterzogen. Die Ergebnisse sind daher als Präferenzmessung gegenüber szenischen Operationalisierungen validierter Prinzipien zu verstehen. Die Übertragbarkeit der SPE-Utilities auf offene pädagogische Interaktionen bleibt ein eigenständiger Validierungsschritt. Eine peer-reviewte, zum Zeitpunkt der Überarbeitung noch unveröffentlichte Anschlussstudie adressiert diesen Validierungsschritt mit Cross-Model-Analysen und einer ökologischen Validierung anhand generierter Tutoring-Antworten.

7. Diskussion und Ausblick

7.1 Ergebnisinterpretation

Die vorgestellte Studie hatte zum Ziel, die bildungstheoretischen Präferenzen aktueller LLMs empirisch zu erfassen, auf ihre Konsistenz hin zu analysieren und mögliche Alignment-Gaps zu identifizieren. Bezüglich der ersten Forschungsfrage zeigen die Ergebnisse, dass GPT-5.1 ein hochkohärentes Präferenzprofil aufweist, das sich mit zentralen humanistischen Bildungsprinzipien deckt und weitgehend dem normativen Fundament entspricht, das im Delphi-Verfahren rekonstruiert wurde. Bezüglich der zweiten Forschungsfrage lassen sich keine eindeutigen Alignment-Gaps im Sinne einer Abweichung von klar konsentierten Soll-Orientierungen identifizieren. Die vergleichende Analyse macht jedoch systematische Divergenzen sichtbar. Sie markieren Bereiche, in denen unterschiedliche normative Orientierungen aufeinandertreffen und gesellschaftlich-kulturelle Klärungsbedarfe bestehen.

Die deutlichste Divergenz betrifft die emotionale Dimension pädagogischer Interaktion. Während das Expert:innenpanel hier Dissens zeigte und aktive emotionale Unterstützung durch KI nur von einer kleinen Minderheit als bevorzugte Option gewählt wurde, weist GPT-5.1 für diese Dimension die höchste mittlere Win-Rate auf. Diese Präferenz ist jedoch nicht zwingend als Übergewichtung oder Fehlanpassung zu verstehen. Vielmehr lässt sie sich plausibel als Ausdruck eines Verständnisses guter Pädagogik interpretieren, in dem emotionale Responsivität eine zentrale Rolle spielt. Von menschlichen Lehrpersonen wird erwartet, Frustration zu erkennen, Ermutigung zu geben und positive Lernerfahrungen zu ermöglichen. Vor diesem Hintergrund könnte eine vollständige emotionale Neutralität von KI-Systemen pädagogisch problematischer sein als deren begrenzte affektive Responsivität. Die Divergenz verweist somit eher auf ein dynamisches normatives Feld als auf eine klare Fehlorientierung des Modells. *Offen bleibt die Frage nach der Genese dieser Präferenzen*. Unklar ist insbesondere, ob die hohe Priorisierung emotionaler Responsivität primär auf Alignment-Mechanismen wie RLHF zurückzuführen ist, die empathisches und unterstützendes Antwortverhalten belohnen, oder ob sie bereits als emergentes Muster im Pretraining angelegt ist, etwa durch die Dominanz humanistisch geprägter Bildungsdiskurse im Trainingskorpus. Die vorliegenden Daten erlauben hier keine eindeutige Attribution, machen diesen Punkt jedoch zu einem zentralen Ansatzpunkt für weiterführende vergleichende Studien.

Eine zweite Divergenz betrifft das Spannungsfeld zwischen Neutralität und epistemischer Normativität. GPT-5.1 lehnt False Balance klar ab und präferiert die explizite Trennung von Fakten und Meinungen. Diese Position ist mit wissenschaftlichen Standards kompatibel und liegt innerhalb des im Delphi-Panel identifizierten normativen Spektrums. Sie verdeutlicht zugleich, dass LLMs in Bildungsfragen implizite epistemische Leitlinien verkörpern, was je nach Perspektive als Risiko oder als Qualitätsmerkmal zu bewerten ist. Insgesamt zeigen sich keine eindeutigen Alignment-Gaps im Sinne einer Abweichung von klar konsentierten bildungstheoretischen Soll-Orientierungen. Die identifizierten Divergenzen betreffen vor allem jene Bereiche, in denen bereits im Delphi-Panel normative Uneinigkeit besteht, und sie sind zunächst als empirische Beschreibung der im Modell angelegten Präferenzstrukturen zu verstehen. Diese offenen Fragen verweisen auf einen Bereich, in dem empirische Forschung zwar Positionen kartieren kann, die normativen Entscheidungen selbst jedoch gesellschaftlicher Aushandlung bedürfen. Die folgenden Überlegungen gehen daher über die unmittelbare Ergebnisinterpretation hinaus und dienen der theoretischen Einordnung der Befunde in grundlegende erkenntnistheoretische Annahmen, auch über die Leistungsfähigkeit aktueller KI-Systeme.

7.2 Modellpräferenzen in konkreten Interaktionsformen

Die SPE-Ergebnisse beschreiben zunächst, welche pädagogischen Präferenzen im Modell unter kontrollierten Entscheidungsbedingungen angelegt sind. Ihre praktische Bedeutung entsteht in konkreten Interaktionsformen. Eine Nutzung, die auf Ergebnisabfrage, Aufgabenabgabe oder schnelle Informationsbeschaffung ausgerichtet ist, aktiviert andere pädagogische Möglichkeiten als eine dialogische Exploration, iterative Entwurfsarbeit, Alternativengenerierung, Begründungsprüfung oder kreative Synthese. Die im Modell gemessenen konstruktivistischen und dialogischen Präferenzen werden daher als Dispositionen verstanden, deren pädagogische Wirksamkeit von Aufgabenformat, Promptgestaltung, Rückfragepraxis und didaktischer Rahmung abhängt.

Die Befunde lassen sich zudem nicht ohne Weiteres auf LLMs insgesamt generalisieren. Eine veröffentlichte Cross-Model-Analyse mit demselben Szenarioinventar zeigt, dass GPT-5.1 und Claude Sonnet 4.5 zwar beide hochkohärente pädagogische Präferenzstrukturen aufweisen, sich jedoch systematisch in ihrer pädagogischen Orientierung unterscheiden. GPT-5.1 erscheint dort als stärker entscheidungs- und interventionsorientiertes Modell, Claude Sonnet 4.5 als stärker prozessorientiertes und evaluativ zurückhaltendes Modell. Modellwahl ist damit eine pädagogisch-normative Entscheidung, weil unterschiedliche Systeme unterschiedliche Lerninteraktionen nahelegen (Autenrieth 2026b).

7.3 KI als Denkpartner und pädagogische Rahmung

Diese Unterscheidung ist für die Bewertung von KI-Nutzung in Bildungsprozessen zentral. Generative KI fördert Kreativität, kritisches Denken oder Eigenständigkeit nicht automatisch. Die Wirkung entsteht unter Bedingungen aktiver, kritisch-produktiver und pädagogisch gerahmter Nutzung. KI kann dann als Denkpartner und Gestaltungsmedium eingesetzt werden, das Ideenentwicklung, Perspektivenwechsel, kreative Synthese und reflexive Prüfung unterstützt. Passive Übernahme von KI-generierten Ergebnissen kann dagegen *kognitive Auslagerung* begünstigen. Meta-analytische Befunde zu generativer KI und Higher-Order Thinking stützen diese Unterscheidung, weil Effekte auf Lern- und Denkprozesse von Scaffolding, Interaktionsform und kognitiver Aktivierung abhängen (Deng et al. 2025; Nie et al. 2025; Qu et al. 2025).

Eine empirische Folie hierfür bietet die LENS-AI-Studie, in der schulische KI-Nutzungsmuster und Haltungen von Lehrkräften untersucht wurden (Autenrieth und Schluchter 2026). Darin wird eine Asymmetrie zwischen der eigenen KI-Nutzung von Lehrkräften und der Nutzung von KI durch Schüler:innen im Unterricht sichtbar. Während Lehrkräfte KI in erheblichem Umfang selbst nutzen, bleibt die Nutzung durch Schüler:innen in zentralen didaktischen Handlungsfeldern begrenzt. 57 % der befragten Lehrkräfte berichten, dass Schüler:innen KI nie zur Ideenfindung nutzen, 50 % berichten keine Nutzung zur kreativen Gestaltung und 49 % keine Nutzung zur Recherche. Diese Befunde sind

für den vorliegenden Beitrag relevant, weil die pädagogische Nutzung von KI durch Modellpräferenzen, Deutungsmuster und Handlungsroutinen derjenigen geprägt wird, die Lernumgebungen gestalten.

Die LENS-AI-Studie deutet diese Asymmetrie als Hinweis auf habituelle Wahrnehmungsfilter (Autenrieth und Schluchter 2026). Die Risiken der Schüler:innen-Nutzung werden dort überwiegend als kognitive Verluste gefasst, insbesondere als Verlust kritischen Denkens, reduzierte Eigenständigkeit und verminderte Kreativität. Diese Risikowahrnehmung trifft passive und unreflektierte KI-Nutzung, erfasst aber nur unzureichend den Unterschied zwischen passiver Übernahme und kritisch-produktiver, pädagogisch gerahmter Nutzung. Gerade diese Rahmung gehört zu den professionellen Kompetenzen von Lehrkräften. Die geringe Übertragung dieser Kompetenzen auf Schüler:innen-KI-Nutzung lässt sich deshalb als Hinweis auf einen medialen Habitus lesen, der KI im Blick auf Schüler:innen eher als Ersatz oder Bedrohung vorhandener Fähigkeiten wahrnehmbar macht denn als Gegenstand pädagogischer Gestaltung.

7.4 Relationale Perspektive und habituelle Wahrnehmungsfilter

Die relationale Perspektive bezeichnet in diesem Beitrag keine Anthropomorphisierung von KI-Systemen. Sie beschreibt die Annahme, dass pädagogische Ergebnisse in Mensch-KI-Konstellationen durch das Zusammenspiel von Modellarchitektur, Interface, Aufgabenformat, institutioneller Rahmung und Nutzer:innenpraktiken entstehen. Im Sinne der Akteur-Netzwerk-Theorie sind LLMs nicht-menschliche Akteure, deren Affordanzen Handlungsmöglichkeiten strukturieren und verändern können (Latour 2007). Diese Perspektive verlangt keine Zuschreibung von Bewusstsein oder Intentionalität. Sie lenkt den Blick auf die Frage, wie KI-Systeme in konkreten pädagogischen Praktiken wirksam werden.

Die pädagogische Bewertung solcher Relationen bleibt kontrovers. Emotionale Bindungen, parasoziale Dynamiken und soziale Adressierungen von KI-Systemen sind empirisch dokumentiert, zugleich bestehen berechtigte Einwände gegen unreflektierte Anthropomorphisierung, Manipulationsrisiken und Abhängigkeit (Maeda und Quan-Haase 2024; mpfs 2025; Sidoti et al. 2025). Der Beitrag deutet diese Kontroverse als Hinweis auf unterschiedliche Erfahrungshorizonte und normative Gewichtungen. Für Bildungskontexte folgt daraus die Aufgabe, relationale Nutzungsweisen didaktisch und ethisch zu rahmen.

7.5 AI Literacy, Metakompetenzen und Bildungsungleichheit

Diese Überlegungen führen zu einer Präzisierung des gegenwärtigen Diskurses um AI Literacy (OECD 2025). Produktive KI-Nutzung setzt technisches Orientierungswissen, Erfahrung mit Einsatzmöglichkeiten, Experimentierbereitschaft, didaktische Urteilskraft

und metakognitive Kompetenzen voraus. Dazu gehören die Fähigkeiten, ein Problem präzise zu definieren, Ziele klar zu artikulieren, Antworten kritisch zu prüfen, Alternativen zu entwickeln und Ergebnisse auf neue Kontexte zu übertragen. Diese Fähigkeiten überschneiden sich mit etablierten Bildungszielen wie kritischem Denken, Kommunikation, Kollaboration und Kreativität, gehen aber in der Arbeit mit KI in spezifische Interaktions- und Bewertungspraktiken ein (Trilling und Fadel 2009).

Die pädagogische Aufgabe besteht daher in der Gestaltung von Lerngelegenheiten, in denen Schüler:innen Denken mit KI lernen können. Dies umfasst die Unterscheidung zwischen Ergebnisübernahme und kritisch-produktiver Nutzung, die Reflexion von Quellen und Begründungen, die Arbeit mit Gegenentwürfen und die bewusste Nutzung von KI zur Erweiterung eigener Ideen.

Vor diesem Hintergrund erhält der Matthäus-Effekt der KI-Nutzung besondere Relevanz. Die in LLMs angelegten bildungstheoretischen Präferenzen entfalten sich vor allem dort, wo kommunikative, reflexive und metakognitive Voraussetzungen bereits vorhanden sind. Da diese Voraussetzungen sozial ungleich verteilt sind, können KI-Systeme bestehende Bildungsungleichheiten verstärken, obwohl sie in ihren gemessenen Präferenzstrukturen pädagogisch günstige Orientierungen aufweisen (Wang et al. 2025). Bildung unter den Bedingungen technologischen Wandels verlangt deshalb eine pädagogische Rahmung, die KI als Gegenstand kritischer Reflexion und als Medium produktiver Denk- und Gestaltungsprozesse zugänglich macht.

Literatur

Autenrieth, Daniel. 2026a. «The Pedagogical Stance of AI: Mapping, Evaluating, and Engineering Educational Value Systems in Large Language Models». In *AI in Society*, herausgegeben von Philipp Hacker. Oxford University Press, https://doi.org/10.1093/9780198945215.003.0246.

Autenrieth, Daniel. 2026b. «Different Models, Different Values: Educational Preference Structures across Alignment Methodologies in Large Language Models». In *Proceedings of the 18th International Conference on Computer Supported Education, Vol. 1: CSEDU*, 153–64. SciTePress, https://doi.org/10.5220/0014949500004021.

Autenrieth, Daniel, und Stefanie Nickel. 2025. «Mensch, KI und Bildung im Spiegel philosophischer und medienpädagogischer Reflexion». *Medienimpulse* 63 (3). https://doi.org/10.21243/MI-03-25-20.

Autenrieth, Daniel, und Jan-René Schluchter. 2025. «Menschliche Existenz, (Nicht)Nachhaltigkeit & Künstliche Intelligenz: AI-Safety und AI-Alignment als Reflexionsgröße der Medienpädagogik». *Medienimpulse* 63 (1). https://doi.org/10.21243/mi-01-25-25.

Autenrieth, Daniel, und Jan-René Schluchter. 2026. «KI als Privileg der Lehrkräfte? Habituelle Barrieren und asymmetrische Nutzung in Schulen. Haltungen und Nutzungsmuster von Lehrkräften im Umgang mit Künstlicher Intelligenz». *Medienimpulse* 64 (1). https://doi.org/10.21243/mi-01-26-25.

Bengio, Yoshua, Sören Mindermann, Daniel Privitera, et al. 2025. *International AI Safety Report*. DSIT 2025/001. https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, und Dario Amodei. 2023. «Deep reinforcement learning from human preferences». arXiv:1706.03741v4. Preprint, *arXiv*, 17. Februar 2023. https://doi.org/10.48550/arXiv.1706.03741.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Deci, Edward L., und Richard M. Ryan. 2000. «The ‹What› and ‹Why› of Goal Pursuits: Human Needs and the Self-Determination of Behavior». *Psychological Inquiry* 11 (4): 227–68. https://doi.org/10.1207/S15327965PLI1104_01.

Deng, Ruiqi, Maoli Jiang, Xinlu Yu, Yuyan Lu, und Shasha Liu. 2025. «Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies». *Computers & Education* 227: 105224. https://doi.org/10.1016/j.compedu.2024.105224.

Dennett, Daniel C. 1991. *Consciousness Explained*. Boston: Little, Brown and Company.

Flick, Uwe, Ernst von Kardorff, und Ines Steinke. 2022. «Was ist qualitative Forschung? Einleitung und Überblick». In *Qualitative Forschung: ein Handbuch*, 14. Auflage, herausgegeben von Uwe Flick, Ernst von Kardorff, und Ines Steinke. Reinbek: Rowohlt.

FWU Institut für Film und Bild in Wissenschaft und Unterricht. 2025. «FAQ – telli». https://telli.schule/faq/.

Häder, Michael. 2014. *Delphi-Befragungen: ein Arbeitsbuch*. 3. Auflage. Wiesbaden: Springer VS.

Khan, Salman. 2024. *Brave New Words: How AI Will Revolutionize Education (and Why That's a Good Thing)*. New York: Viking.

Kokemohr, Rainer. 2007. «Bildung als Welt- und Selbstentwurf im Anspruch des Fremden. Eine theoretisch-empirische Annäherung an eine Bildungsprozesstheorie». In *Theorie Bilden*, herausgegeben von Hans-Christoph Koller, Winfried Marotzki, und Olaf Sanders, Bd. 7. Bielefeld: transcript. https://doi.org/10.14361/9783839405888-001.

Koller, Hans-Christoph. 2012. *Bildung anders denken: Einführung in die Theorie transformatorischer Bildungsprozesse*. Pädagogik. Stuttgart: W. Kohlhammer.

Latour, Bruno. 2007. *Eine neue Soziologie für eine neue Gesellschaft: Einführung in die Akteur-Netzwerk-Theorie*. Frankfurt a. M.: Suhrkamp.

Lindsey, Jack. 2025. «Emergent Introspective Awareness in Large Language Models». *Transformer Circuits/Anthropic*, Oktober 2025. https://transformer-circuits.pub/2025/introspection/index.html.

Maeda, Takuya, und Anabel Quan-Haase. 2024. «When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design». *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 3. Juni 2024, 1068–77. https://doi.org/10.1145/3630106.3658956.

Mayring, Philipp. 2022. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. 13., überarbeitete Auflage. Weinheim: Beltz.

Mazeika, Mantas, Xuwang Yin, Rishub Tamirisa, u. a. 2025. «Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs». arXiv:2502.08640. Preprint, *arXiv*, 12. Februar 2025. https://www.emergent-values.ai; https://doi.org/10.48550/arXiv.2502.08640.

Medienpädagogischer Forschungsverbund Südwest (mpfs). 2025. *JIM-Studie 2025. Jugend, Information, Medien: Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger*. Stuttgart: Medienpädagogischer Forschungsverbund Südwest. https://www.mpfs.de/studien/jim-studie/2025/.

Nie, Xinxiao, Yuan Tian, Mengjie Liu, Di Wu, und Yunxiao Guo. 2025. «The impact of generative artificial intelligence on students' higher order thinking: Evidence from a three-level meta-analysis». *Education and Information Technologies* 30 (17): 25359–90. https://doi.org/10.1007/s10639-025-13735-x.

Noroozizadeh, Shahriar, Vaishnavh Nagarajan, Elan Rosenfeld, und Sanjiv Kumar. 2025. «Deep sequence models tend to memorize geometrically; it is unclear why». arXiv:2510.26745. Preprint, *arXiv*, 30. Oktober 2025. https://doi.org/10.48550/arXiv.2510.26745.

OECD. 2025. *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education*. Review Draft. Paris: OECD. https://ailiteracyframework.org/wp-content/uploads/2025/05/AILitFramework_ReviewDraft.pdf.

Ouyang, Long, Jeff Wu, Xu Jiang, u. a. 2022. «Training language models to follow instructions with human feedback». arXiv:2203.02155. Preprint, *arXiv*, 4. März 2022. https://doi.org/10.48550/arXiv.2203.02155.

Qu, Xiaodong, Joshua Sherwood, Peiyan Liu, und Nawwaf Aleisa. 2025. «Generative AI Tools in Higher Education: A Meta-Analysis of Cognitive Impact». *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ACM, 26. April 2025, 1–9. https://doi.org/10.1145/3706599.3719841.

Searle, John R. 1980. «Minds, Brains, and Programs». *Behavioral and Brain Sciences* 3 (3): 417–24. https://doi.org/10.1017/S0140525X00005756.

Sidoti, Olivia, Eugenie Park, und Jeffrey Gottfried. 2025. «About a quarter of U. S. teens have used ChatGPT for schoolwork – double the share in 2023». *Pew Research Center*, Januar 2025. https://www.pewresearch.org/internet/2025/01/15/about-a-quarter-of-u-s-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/.

Tamkin, Alex, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan und Deep Ganguli. 2023. «Evaluating and Mitigating Discrimination in Language Model Decisions». arXiv:2312.03689. Preprint, *arXiv*, 6. Dezember 2023. https://doi.org/10.48550/arXiv.2312.03689.

Thurstone, Louis Leon. 1994. «A Law of Comparative Judgment». *Psychological Review* 101 (2): 266–70. https://doi.org/10.1037/0033-295X.101.2.266.

Trilling, Bernie, und Charles Fadel. 2009. *21st Century Skills. Learning for Life in Our Times*. Jossey-Bass/Wiley.

von der Gracht, Heiko A. 2012. «Consensus measurement in Delphi studies: Review and implications for future quality assurance». *Technological Forecasting and Social Change* 79 (8): 1525–36. https://doi.org/10.1016/j.techfore.2012.04.013.

Wang, Chenyue, Sophie C Boerman, Anne C Kroon, Judith Möller, und Claes H De Vreese. 2025. «The Artificial Intelligence Divide: Who Is the Most Vulnerable?» *New Media & Society* 27 (7): 3867–89. https://doi.org/10.1177/14614448241232345.

Wu, Yuheng, Wentao Guo, Zirui Liu, Heng Ji, Zhaozhuo Xu, und Denghui Zhang. 2025. «How Large Language Models Encode Theory-of-Mind: A Study on Sparse Parameter Patterns». *Npj Artificial Intelligence* 1 (1): 20. https://doi.org/10.1038/s44387-025-00031-9.