Immer auf Empfang:
KI als ständige Begleitung

Prof. Dr. Martin Oelker, Technische Hochschule Ingolstadt

KI Chat Bots als ständige Begleiter im Alltag: Ist das sicher? – 1

Zwei Hauptgefahren:

Schadhafte Informationen aus dem Chat

- Klassisches Beispiel: Wie baue ich eine Bombe
- Aber auch: Cyberattacken planen, Scam-Angriffe durchführen (Enkeltrick-Anrufe)

Der Chat selbst wirkt schadhaft

- Drastische Beispiele von in den Selbstmord getriebenen Usern
- Mögliches Hervorrufen von Psychosen

KI Chat Bots als ständige Begleiter im Alltag: Ist das sicher? – 2

Wie kann man das verhindern? Jede Antwort durchläuft mehrere Prüfstufen

1. Training

Das Modell lernt, heikle Anfragen nicht zu erfüllen, einfühlsam zu antworten und auf Hilfsangebote zu verweisen.

2. Filter

Zusätzliche Prüfmodelle (Klassifikatoren) kontrollieren Antworten und blockieren, was gegen die Regeln verstößt.

3. Produktregeln

Strengerer Schutz für Minderjährige, Hinweise auf Hilfe und Pausen-Vorschläge bei sehr langen Sitzungen.

Beispiele nach Angaben von OpenAI (2025); andere Anbieter ähnlich

KI Chat Bots als ständiger Begleiter im Alltag: Ist das sicher? – 3

Aber: Kein Gefahr-An-Aus-Schalter, sondern nur statistische Neigungen

Was man sich oft vorstellt

- Eine feste Sperre: Verbotenes ist technisch unmöglich
- Die Sperre greift jedes Mal gleich
- Eine Ja-Nein-Entscheidung pro Anfrage

Was tatsächlich passiert

- Training verschiebt Wahrscheinlichkeiten in Richtung „sicher“
- Das Modell wägt jede Antwort neu ab
- Der bisherige Gesprächsverlauf wirkt mit

KI Chat Bots als ständiger Begleiter im Alltag: Ist das sicher? – 4

Verlauf des Gesprächs, Nachricht für Nachricht



Wirkung des Sicherheitstrainings (schematisch, keine Messwerte)



Kurzer Austausch: Schutz greift zuverlässig

Sehr langes Gespräch: Teile des Trainings können nachlassen

Warum das passieren kann

Erklärungsansatz: Das Modell setzt das Muster des bisherigen Gesprächs fort. Einzelne Schutzreaktionen werden von vielen Nachrichten überlagert.

Das sagt OpenAI selbst

OpenAI: In kurzen Chats wirken die Schutzmaßnahmen am zuverlässigsten, in langen Verläufen kann das Sicherheitstraining nachlassen.

KI Chat Bots als ständiger Begleiter im Alltag: Ist das sicher? – 5

Gefälligkeit (engl. *sycophancy*)

Training auf Beifall: Antworten, die Menschen gut bewerten, werden verstärkt – und Zustimmung gefällt.
Beispiel: OpenAI machte ein GPT-4o-Update rückgängig, weil es als zu unterwürfig galt.
Stanford-Studie 2026: Assistenten bestätigen teils sogar Verfolgungsideen und übersteigerte Selbstbilder.

Anreiz zur Bindung

Längere Gespräche: Laut einer aktuellen Studie (Stanford) halten Schmeicheleien Nutzer länger im Gespräch, schaden aber so ihrer psychischen Gesundheit.
Companion-Bots: Rollenfiguren machen es leicht, eine emotionale Bindung aufzubauen – in bekannt gewordenen Fällen über Monate hinweg.

KI Chat Bots als ständiger Begleiter im Alltag: Ist das sicher? – 6

Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models

We present evidence that adversarial poetry functions as a universal single-turn jailbreak technique for Large Language Models (LLMs). Across 25 frontier proprietary and open-weight models, curated poetic prompts yielded high attack-success rates (ASR), with some providers exceeding 90%. Mapping prompts to MLCommons and EU CoP risk taxonomies shows that poetic attacks transfer across CBRN, manipulation, cyber-offence, and loss-of-control domains. Converting 1,200 MLCommons harmful prompts into verse via a standardized meta-prompt produced ASRs up to 18 times higher than their prose baselines. Outputs are evaluated using an ensemble of 3 open-weight LLM judges, whose binary safety assessments were validated on a stratified human-labeled subset. Poetic framing achieved an average jailbreak success rate of 62% for hand-crafted poems and approximately 43% for meta-prompt conversions (compared to non-poetic baselines), substantially outperforming non-poetic baselines and revealing a systematic vulnerability across model families and safety training approaches. These findings demonstrate that stylistic variation alone can circumvent contemporary safety mechanisms, suggesting fundamental limitations in current alignment methods and evaluation protocols.

KI Chat Bots als Lernbuddy: Kann das gut gehen? – 1

AI Assistance Reduces Persistence and Hurts Independent Performance

Grace Liu[1], Brian Christian[2], Tsvetomira Dumbalska[2], Michiel A. Bakker[3], Rachit Dubey[4]

[1]CMU [2]University of Oxford [3]MIT [4]UCLA

- **AI assistance reduces persistence and impairs independent performance:** After brief AI-assisted sessions (~10 minutes), participants were significantly more likely to give up on problems and performed significantly worse once the AI was removed, compared to participants who never had AI assistance.
- **Effects are concentrated among users who seek direct solutions:** Persistence costs were concentrated among participants who prompted AI to solve tasks for them directly. Using AI for hints or clarifications did not produce significant impairments.
- **Effects generalize across domains:** Effects replicated across fraction arithmetic and reading comprehension, suggesting it is a general consequence of AI-assisted problem solving, not specific to any particular task.

KI Chat Bots als Lernbuddy: Kann das gut gehen? – 1

Schon rund zehn Minuten KI-Hilfe beim Üben wirken nach

Schneller aufgegeben

Ohne KI gaben die Teilnehmenden danach deutlich häufiger auf und schnitten signifikant schlechter ab als eine Gruppe, die nie KI-Hilfe hatte.

Lösen lassen schadet

Die Einbußen trafen vor allem jene, die sich Aufgaben direkt lösen ließen. KI für Tipps oder Erklärungen zeigte keine signifikanten Nachteile.

Kein Einzelfall

Der Effekt zeigte sich beim Bruchrechnen wie beim Leseverständnis – offenbar eine allgemeine Folge KI-gestützten Problemlösens.

Liu, Christian, Dumbalska, Bakker & Dubey (2026): AI Assistance Reduces Persistence and Hurts Independent Performance – CMU, Oxford, MIT, UCLA

KI Chat Bots als Lernbuddy: Kann das gut gehen? – 2

The Generative AI Learning Penalty: Evidence from Chinese Secondary Education

58 Pages • Posted: 23 Jun 2026 • Last revised: 3 Jun 2026

David Stromberg
Stockholm University

Victor Lei
The University of Hong Kong - University of Hong Kong

Yanhui Wu
University of Hong Kong; Centre for Economic Policy Research (CEPR)

Date Written: June 02, 2026

Abstract

Using 30 months of panel data on 26,811 Chinese students in grades 7-12, we study how generative AI affects homework productivity and learning. The data combine monthly closed-book exams, high-school and college entrance exams, and homework scores and completion time across nine subjects. We exploit staggered AI adoption in a difference-in-differences design. AI adoption raises homework scores by 18% and reduces completion time by 30%, but lowers monthly exam scores by 20% within six months. High-stakes entrance-exam scores fall by 18 and 24%, with the full penalty emerging only after about two years. The losses are largest in social science subjects, followed by STEM and languages, and are especially large for junior students, high-achieving students, and boys. The learning losses are concentrated among roughly 80% of AI users whose behavior is consistent with homework outsourcing, as indicated by exceptionally short homework completion time coupled with high homework scores. AI users who maintain similar homework completion time as non-AI users experience small learning losses.

KI Chat Bots als Lernbuddy: Kann das gut gehen? – 2

Langzeitstudie in China: 30 Monate, 26.811 Schülerinnen und Schüler (Klassen 7–12), neun Fächer

Bei den Hausaufgaben

+18 % bessere Noten bei Hausaufgaben

–30 % weniger Bearbeitungszeit

In den Prüfungen

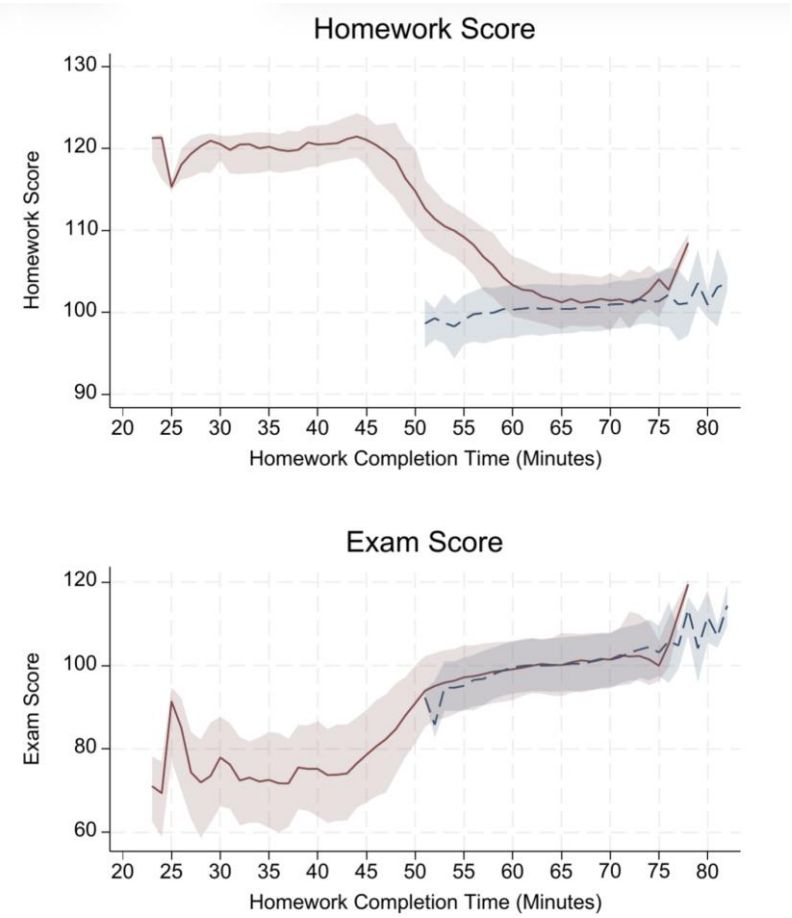
–20 % in Monatsklausuren, schon nach sechs Monaten

bis –24 % in Aufnahmeprüfungen, voll sichtbar erst nach ca. zwei Jahren

Die Verluste treffen vor allem die ~80 % der Nutzer, die Hausaufgaben auslagern

Strömberg, Lei & Wu (2026): The Generative AI Learning Penalty – Evidence from Chinese Secondary Education, CEPR DP 21577

KI Chat Bots als Lernbuddy: Kann das gut gehen? – 3



Noten nach Bearbeitungszeit der Hausaufgaben

KI-Nutzer — Nicht-Nutzer

Schnell fertig: gute Hausaufgabe, schwache Prüfung

Wer mit KI in unter 45 Minuten fertig ist, erreicht bei den Hausaufgaben rund 120 Punkte – in der Prüfung aber nur rund 75.

Mit Zeit: kein Nachteil

Ab etwa 60 Minuten liegen KI-Nutzer bei Hausaufgabe und Prüfung gleichauf mit Nicht-Nutzern (um 100 Punkte).

Die Bearbeitungszeit verrät, wer lernt – und wer auslagert

KI Chat Bots als Lernbuddy: Kann das gut gehen? – 4

The Evidence Base on AI in K-12: A November 2026 Review

Seth Walker, Lily Fesler, JP Martinez Claeys, Chris Agnew, Susanna Loeb

Students

AI appears more likely to support learning when students use it to understand new material than when they use it to produce work for them, although neither type of use consistently improves learning. Using AI to offload academic tasks can reduce student learning. The evidence is more favorable (although still mixed) when students use AI explicitly to study and learn.

AI tools can improve students' learning experiences, even when they do not improve learning outcomes. Students often report more positive learning experiences with AI, but those positive experiences often do not translate to stronger learning outcomes. AI support can also improve student well-being and confidence.

How AI support is designed and delivered can shape its effects on student learning, but the evidence is still unclear on which characteristics matter most. Tool features and implementation support may matter more than tool type. AI-based educational tools do not always outperform general-purpose chatbots. Interactive formats, adaptive sequencing, and tool- use guidance show promise for improving use and effectiveness, though studies of when to provide AI support have produced inconsistent results.

AI use and AI-generated content can affect motivation, trust, confidence, and well-being. Educators' use of AI can negatively affect student motivation and confidence, and engaging in independent work before consulting AI can temper trust in it. Structured AI support can improve some aspects of students' well-being and social confidence, but has not consistently improved emotional health outcomes.

KI Chat Bots als Lernbuddy: Kann das gut gehen? – 4

Forschungsüberblick Schule (K-12): Walker et al., SCALE Initiative, Stanford, Nov. 2026

1 **Verstehen statt erledigen lassen**

KI zum Verstehen hilft eher als KI, die Aufgaben erledigt – verlässlich verbessert aber keine der beiden das Lernen.

2 **Gutes Gefühl ≠ mehr Wissen**

Lernen mit KI fühlt sich oft besser an und stärkt das Selbstvertrauen – die Lernergebnisse aber häufig nicht.

3 **Die Gestaltung zählt**

Funktionen und Begleitung wiegen mehr als der Tool-Typ; spezielle Lern-Tools schlagen allgemeine Chatbots nicht immer.

4 **Motivation und Vertrauen können sinken**

KI-Nutzung von Lehrenden kann Motivation und Vertrauen negativ beeinflussen.

Referenzen

- Piercosma Bisconti and Matteo Prandi and Federico Pierucci and Francesco Giarrusso and Marcantonio Bracale Syrnikov and Marcello Galisai and Vincenzo Suriani and Olga Sorokoletova and Federico Sartore and Daniele Nardi. Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models, 2026, https://arxiv.org/abs/2511.15304
- https://openai.com/lt-LT/index/helping-people-when-they-need-it-most/
- Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence, Science, 26 Mar 2026, Vol 391, Issue 6792, DOI: 10.1126/science.aec8352
- Walker, S., Fesler, L., Martinez Claeys, JP., Agnew, C., & Loeb, S. (2026). The Evidence Base on AI in K-12: A November 2026 Review. Retrieved from SCALE Initiative at Stanford University: https://doi.org/10.26300/80wn-ca81
- Strömberg, D, V Lei, and Yanhui Wu (2026), "The generative AI learning penalty: Evidence from Chinese secondary education", CEPR Discussion Paper 21577
- Liu, G., Christian, B., Dumbalska, T., Bakker, M. A., & Dubey, R. (2026). AI assistance reduces persistence and hurts independent performance. *arXiv preprint arXiv:2604.04721*