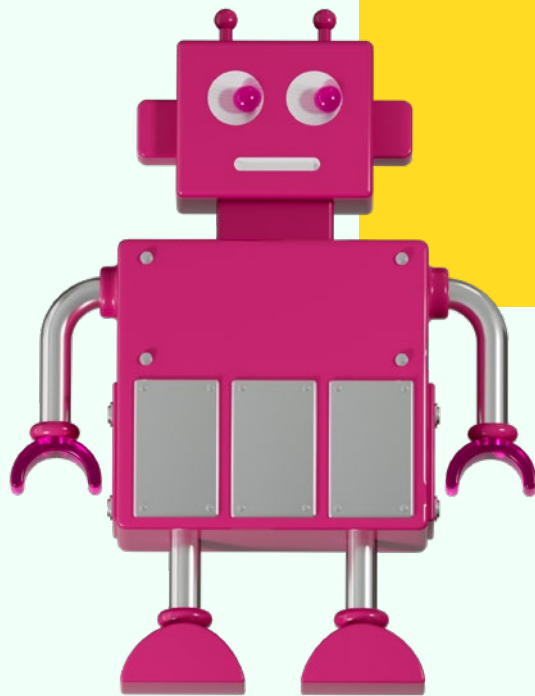




KI-Chatbots im Faktencheck

Wie gut Künstliche Intelligenz Misinformationen
erkennt und was das für die Schule bedeutet



Fehlbare Faktenchecker – Unterstützer für Urteilskraft

Urteilskraft und kritisches Denken halten wir bei der Deutsche Telekom Stiftung für wichtige Kompetenzen in einer digitalisierten Welt. Gerade junge Menschen, für die digitale Medien und Künstliche Intelligenz schon so selbstverständlich sind, sollten hier besonders gut aufgestellt sein. Das heißt für uns auch: Wir müssen Lehrkräfte dabei unterstützen, Quellenbewertungskompetenzen sowie einen guten Umgang mit generativer KI zu vermitteln.

Durch unseren *Trendmonitor KI in der Bildung* wissen wir: Am meisten kommt KI in der Schule bislang in Form von KI-Chatbots zum Einsatz. Die Ruhr-Universität Bochum (RUB) hat deshalb die am Markt und in Schulen weit verbreiteten KI-Chatbots unter die Lupe genommen. Dabei ging es – das will ich hervorheben – speziell um ihre Eignung, Informationen und Quellen zu beurteilen, also um einen klar abgegrenzten Einsatzbereich. Zum restlichen – großen! – Aufgabenspektrum, wofür KI-Chatbots genutzt werden, trifft die Studie keine Aussage. Es ging bei der Untersuchung natürlich auch nicht primär darum, verschiedene KI-Chatbots in eine Rangfolge zu bringen. Und die Ergebnisse sind keine Empfehlung von Systemen unsererseits. Angesichts der rasanten Entwicklung von Large Language Models (LLMs), auf denen KI-Chatbots beruhen, kann diese Studie auch nur eine Momentaufnahme sein. Allerdings eine sehr aufschlussreiche!

Die Studie zeigt: KI-Chatbots sind ziemlich gute Faktenchecker, und zwar sowohl allgemein verfügbare Systeme als auch solche, die speziell für die schulische Bildung entwickelt wurden. Aber sie alle sind längst nicht unfehlbar. Das ist für ihre didaktische Nutzung in

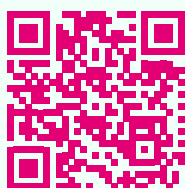
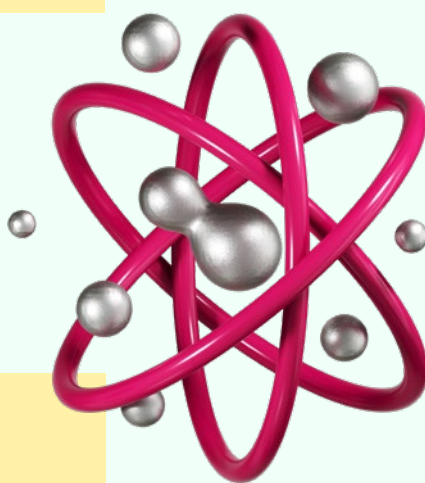
der Schule natürlich absolut relevant – und herausfordernd! Sind Lehrkräfte damit vertraut, können sie mit KI-Chatbots das kritische Denken ihrer Schüler*innen fördern. Vor allem an sie richtet sich deshalb die vorliegende Publikation, und ich denke, sie kann ihnen wichtige pädagogische und didaktische Hilfestellungen geben.

Entstanden ist die Studie in unserem Projekt *Qapito! – Quellen kritisch beurteilen*, das wir seit 2020 mit der RUB als Projektpartnerin realisieren. *Qapito!* bietet evidenzbasierte Unterrichtsmaterialien, um Kinder und Jugendliche zu befähigen, Quellen und Informationen eigenständig und kritisch zu bewerten. In Kürze veröffentlichen wir weiteres *Qapito!*-Unterrichtsmaterial, das das Prompting für die Quellen- und Informationsbewertung trainiert und den kritischen Umgang mit KI-Chatbots fördert.

An dieser Stelle mein herzlicher Dank an Prof. Dr. Marc Stadtler und Philipp L. Marten, unsere Projektpartner der RUB, für die produktive Zusammenarbeit bei *Qapito!* und nun auch diese intensive Untersuchung.

Allen Leser*innen wünsche ich eine aufschlussreiche Lektüre.

Euer Jacob Chammon
Geschäftsführer Deutsche Telekom Stiftung



Mehr Infos

www.telekom-stiftung.de/qapito

Inhalt

Vorwort	3
Zentrale Ergebnisse	6
Was sich aus der Studie ableiten lässt	8
Warum diese Studie?	12
KI-Chatbots und Misinformationen – die Ergebnisse	20
Anhang: Mustervorlagen für Prompts für den Unterricht	36
Literaturverzeichnis	40
Impressum	43

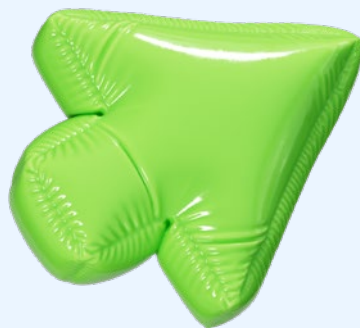
Zentrale Ergebnisse

Wie gut sind KI-Chatbots aktuell darin, Online-Informationen korrekt zu bewerten? Und was bedeutet das für ihren didaktischen Einsatz in der Schule? Um diese Fragen zu beantworten, wurden 22 KI-Chatbots für den deutschsprachigen Raum im Frühsommer 2026 daraufhin getestet, wie gut sie die Zuverlässigkeit von Internetquellen und die Korrektheit von Aussagen aus den sozialen Medien bewerten können. Diese Untersuchung wurde im Rahmen des Projektes *Qapito! – Quellen kritisch beurteilen* durchgeführt.



Hier die wichtigsten Ergebnisse:

- Alle in dieser Studie untersuchten KI-Chatbots konnten grundsätzlich zwischen zuverlässigen und nicht zuverlässigen Quellen unterscheiden.
- Zuverlässige Quellen wurden von allen KI-Chatbots in den meisten Fällen korrekt erkannt. Unterschiede zeigten sich vor allem bei der Beurteilung nicht zuverlässiger Quellen: KI-Chatbots, die diese häufiger korrekt einstufen, lieferten insgesamt treffsicherere Bewertungen, also sowohl bei zuverlässigen als auch bei nicht zuverlässigen Quellen.
- Bei Aufgaben, die eine Prüfung von Informationen mit unabhängigen Quellen erforderten (einen sogenannten Gegencheck), unterschieden sich die KI-Chatbots deutlich in ihrer Leistungsfähigkeit. Während falsche Aussagen in den meisten Fällen erkannt wurden, scheiterten einige KI-Chatbots daran, wahre Aussagen als korrekt zu bestätigen.
- Der Einsatz einer Live-Websuche ist eine Voraussetzung dafür, dass KI-Chatbots erfolgreich Gegenchecks zu aktuellen Aussagen durchführen können, die vermutlich noch nicht in der Wissensbasis ihres Sprachmodells enthalten sind. So zeigte sich beim direkten Vergleich: Sobald die Websuche deaktiviert wurde, lieferten die KI-Chatbots oft Einschätzungen, die zwar plausibel klangen, aber sachlich falsch waren.
- Viele KI-Chatbots erwiesen sich in ihrem Antwortverhalten als wenig stabil. Erhielten die KI-Chatbots mehrfach dieselbe Eingabe (Prompt), wichen ihre Urteile mitunter sehr stark voneinander ab.



Fazit:

Die konkreten Leistungswerte der untersuchten KI-Chatbots sind angesichts der raschen technischen Entwicklung eine Momentaufnahme. Über diesen Zeitpunkt hinaus relevant ist aber der Befund, dass die Wahl des KI-Chatbots die Qualität der Ergebnisse erheblich beeinflussen kann. Die eigene Urteilsbildung sollten sie hingegen keinesfalls ersetzen. Deshalb sollten KI-Chatbots nicht als reine Bewertungsinstrumente eingesetzt werden, sondern vielmehr als didaktische Werkzeuge, um die für Quellenbewertungen und Gegenchecks notwendigen Kompetenzen gezielt zu fördern.

Was sich aus der Studie ableiten lässt

Mit KI-Chatbots stehen inzwischen Werkzeuge zur Verfügung, die Jugendliche beim Erwerb zentraler Strategien für den kritischen Umgang mit Informationen in Online-Umgebungen unterstützen könnten. Für einen solchen Einsatz ist es wichtig, dass KI-Chatbots verlässliche und nachvollziehbare Antworten liefern, wenn sie aufgefordert werden, Quellen zu bewerten oder Aussagen per Gegencheck zu überprüfen.

Ziel dieser Studie war es, anhand von 22 KI-Chatbots für den deutschsprachigen Raum zu untersuchen, wie gut diese Quellenbewertungen und Gegenchecks durchführen und wie stabil ihre Bewertungen bei wiederholten identischen Anfragen ausfallen. Dabei ging es nicht vornehmlich darum, verschiedene KI-Chatbots in eine Rangfolge zu bringen, sondern um eine Gesamtschau ihrer Leistungsfähigkeit und daraus ableitbare pädagogische Potenziale für Lehrkräfte zum aktuellen Zeitpunkt.

Nicht zuverlässige Quellen als größere Herausforderung

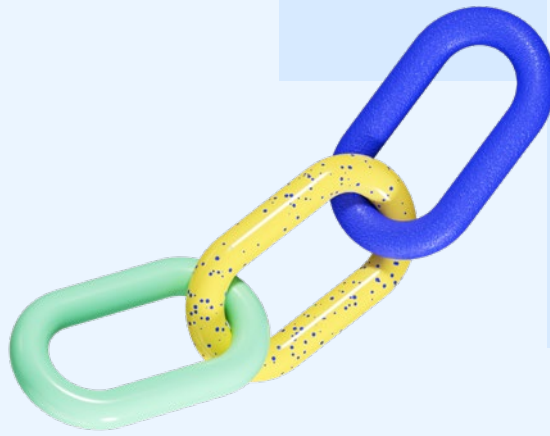
Alle in dieser Studie untersuchten KI-Chatbots waren grundsätzlich in der Lage, zwischen zuverlässigen und nicht zuverlässigen Quellen zu differenzieren. Die Unterscheidungsleistung lag dabei bei allen KI-Chatbots über dem Niveau, das in vergleichbaren Studien bei Jugendlichen oder Studierenden beobachtet wurde.

Betrachtet man die Beurteilung zuverlässiger und nicht zuverlässiger Quellen getrennt voneinander, so zeigt sich, dass vor allem letztere eine Herausforderung für die KI-Chatbots sind. Während zuverlässige Quellen von den meisten KI-Chatbots korrekt erkannt wurden, kam es bei nicht zuverlässigen Quellen öfter zu Fehleinschätzungen. Hierzu zählten wir Quellen, denen es an Expertise mangelte oder die aufgrund eines Interessenkonflikts nicht als unabhängig galten.

Zusätzlich unterschieden sich die KI-Chatbots darin, wie stabil ihre Bewertungen bei wiederholter Eingabe identischer Prompts ausfielen: Einige gelangten sehr häufig zum gleichen Urteil, während andere in ihren Einschätzungen derselben Quelle stark schwankten. Wir bezeichnen das als geringe Antwortstabilität.

Für eine Verwendung von KI-Chatbots in Lernkontexten ist die Antwortstabilität bedeutsam. Denn wenn ein KI-Chatbot bei identischen Aufgaben und identischen Prompts zu unterschiedlichen Bewertungen gelangt, werden seine Einschätzungen für Lehrkräfte und Lernende nur begrenzt vorhersagbar. In solchen Fällen können verschiedene Schüler*innen bei derselben Aufgabe und demselben Lernmaterial unterschiedliche Bewertungen erhalten, ohne dass sich diese Unterschiede inhaltlich begründen lassen. Dies schränkt die verlässliche Nutzung von KI-Chatbots mit geringer Antwortstabilität als Bewertungsinstrument im Unterricht ein. Was diese Unterschiede im Antwortverhalten verursacht, lässt sich auf Basis der vorliegenden Daten nicht abschließend klären.

Eine hohe Antwortstabilität allein reicht jedoch nicht aus. Für einen Einsatz in Lernsettings sollten KI-Chatbots über beides verfügen: eine starke Unterscheidungsleistung zwischen zuverlässigen und nicht zuverlässigen Quellen und eine hohe Stabilität der Antworten. Ist beides gegeben, könnten KI-Chatbots bei der Quellenbewertung unterstützen, indem sie Lernende an Prüfkriterien erinnern, sie zu Begründungen auffordern und diese kritisch hinterfragen. Dabei ist eines wichtig: Die abschließende Bewertung sollten immer die Lernenden übernehmen und nicht an KI-Chatbots delegieren. Ziel sollte es immer bleiben, Lernende zu eigenständigen Urteilen zu befähigen. Die in dieser Studie beobachteten Einschränkungen bei der Qualität und Stabilität der Urteile durch KI-Chatbots unterstreichen die Bedeutung dieses Lernziels.



Gegenchecks aktueller Aussagen erfordern aktuelle Informationen

Auch die inhaltliche Richtigkeit aktueller Aussagen, wie sie zum Beispiel auf Social-Media-Plattformen kursieren, konnten die KI-Chatbots unterschiedlich gut bewerten.

Bei diesen Aufgaben spielte die Live-Websuche eine bedeutende Rolle. Der Vergleich jener KI-Chatbots, bei denen diese Funktion im Erhebungszeitraum gezielt ein- oder ausgeschaltet werden konnte, zeigte deutliche Leistungsunterschiede: Ohne Live-Websuche sank die Fähigkeit, wahre von falschen Aussagen zu unterscheiden, erheblich. Das ist für die hier untersuchten Aufgaben besonders relevant, da bewusst aktuelle Social-Media-Inhalte ausgewählt wurden, deren Überprüfung den Rückgriff auf aktuelle Online-Informationen erforderte.

Aus didaktischer Perspektive ist diese Einschränkung besonders relevant für Szenarien, in denen KI-Chatbots zur Bearbeitung tagesaktueller Themen eingesetzt werden sollen. Ist ein Ereignis nicht Teil der gespeicherten Wissensbasis, kann der KI-Chatbot ohne Live-Websuche keine fundierte Überprüfung leisten. Problematisch ist dabei, dass diese Einschränkung in den Antworten nicht immer transparent offengelegt wird. Dadurch kann bei Lernenden und Lehrkräften der Eindruck entstehen, eine gut begründete Einschätzung erhalten zu haben, obwohl diese tatsächlich auf veralteten Informationen oder allein auf Plausibilitätseinschätzungen beruht.

Eine aktivierte Live-Websuche war allerdings keine Garantie für eine korrekte Bewertung. Auch bei verfügbarem Internetzugriff bestanden deutliche Unterschiede zwischen den KI-Chatbots. Für die Praxis bedeutet dies, dass die Möglichkeit zur aktuellen Internetrecherche eine wichtige Voraussetzung für Gegenchecks zu tagesaktuellen Aussagen darstellt, die

Ergebnisse der Recherche und die darauf basierenden Bewertungen aber weiterhin kritisch geprüft werden müssen.

Chatbot-Gegenchecks lieber nochmal gegenchecken

Auffällig war zudem, dass falsche Aussagen insgesamt häufiger korrekt eingeordnet wurden als wahre Aussagen. Gerade wahre Aussagen über überraschende oder unerwartete Ereignisse stellten viele KI-Chatbots vor größere Schwierigkeiten. Für Lernende ergibt sich daraus eine wichtige Einsicht: Dass eine Aussage ungewöhnlich oder zunächst wenig plausibel erscheint, ist für sich genommen kein hinreichender Grund, sie als falsch einzustufen. Gerade solche Aussagen verlangen nach einem sorgfältigen Gegencheck.

Wie schon bei der Quellenbewertung zeigten die KI-Chatbots auch bei den Gegenchecks eine geringe Antwortstabilität. Obwohl bei allen Anfragen dieselben Prompts zum Einsatz kamen, variierten die Bewertungen für die meisten Social-Media-Inhalte über wiederholte Abfragen hinweg deutlich. Kein KI-Chatbot gelangte bei mehrfacher Bewertung identischer Aussagen konsequent zu den gleichen Einschätzungen.

Auch bei Multimedia-Inhalten, bestehend aus Text-Bild-Kombinationen (sogenannten Sharepics), erzielten viele KI-Chatbots gute Ergebnisse, zugleich zeigten sich erneut deutliche Leistungsunterschiede. Die Analyse der Antworten zeigte, dass die KI-Chatbots in mehreren der untersuchten Fälle kritisch Inhalte der Multimedia-Posts hinterfragten, auch dann, wenn diese zutreffend waren. Aufgrund der geringen Zahl der untersuchten Multimedia-Posts lässt sich daraus kein allgemeines Fehlermuster ableiten. Die Befunde unterstreichen jedoch, dass auch die Fähigkeit zur Bildverarbeitung keine verlässliche Beurteilung eines Social-Media-Inhalts garantiert.

Überzeugend begründet – trotzdem falsch

Eine zentrale Herausforderung bei der Nutzung von KI-Chatbots ist, dass diese plausibel und überzeugend klingende Begründungen liefern können, auch dann, wenn die Urteile falsch sind. Eine ausführliche, sprachlich schlüssige Begründung ist daher kein Beleg dafür, dass eine Quelle oder Aussage tatsächlich korrekt bewertet wurde (Kalai et al., 2025).

Dies ist gerade für den Einsatz in Bildungssettings bedeutsam. Lernende können den Eindruck gewinnen, eine Einschätzung müsse zutreffen, weil sie nachvollziehbar formuliert ist, eine Begründung enthält oder mit Quellenangaben versehen wurde. Begründungen von KI-Chatbots sollten deshalb selbst geprüft werden: Welche Quellen werden genannt? Sind diese Quellen für die konkrete Aussage einschlägig? Stützen sie tatsächlich die vorgebrachte Schlussfolgerung? Wird Unsicherheit angemessen kenntlich gemacht?

Auch die Unterschiede in der Antwortstabilität verstärken diese Herausforderung. Derselbe KI-Chatbot kann bei identischen Aufgaben zu unterschiedlichen Bewertungen gelangen und diese jeweils überzeugend begründen. Welche Prozesse diese Unterschiede im Einzelnen verursachen, lässt sich auf Grundlage der vorliegenden Daten nicht abschließend bestimmen. Für Nutzer*innen ist zunächst entscheidend, dass eine einzelne, überzeugend formulierte Antwort nicht automatisch als verlässliches Urteil behandelt werden sollte.

Implikationen für Lehr-Lernszenarien

Das Leistungsvermögen eines KI-Chatbots allein entscheidet nicht darüber, ob er für den Einsatz in Schulen geeignet ist. Entscheidend ist, wie er didaktisch eingesetzt wird. Obwohl mehrere KI-Chatbots sehr starke Unterscheidungsleistungen demonstrierten, sollten sie nicht als verlässliche Informationsquelle oder Bewertungsinstanz verstanden werden. Ihr Potenzial liegt vielmehr darin, Lernprozesse anzuregen, Bewertungswege sichtbar zu machen und zum Ausgangspunkt für die kritische Prüfung von Quellen und Aussagen zu werden.

Kinder und Jugendliche müssen dafür zunächst lernen, KI-Chatbots so anzusprechen, dass die gestellte Aufgabe möglichst eindeutig ist. Dazu gehört, das Ziel der Aufgabe klar zu benennen, den benötigten Kontext bereitzustellen und deutlich zu machen, welche Art von Prüfung erwartet wird. Ebenso kann es sinnvoll sein, Anforderungen an die Antwort vorzugeben – etwa die verwendeten Quellen zu nennen, die Bewertung zu begründen oder bestehende Unsicherheiten kenntlich zu machen. Solche Vorgaben machen die Ergebnisse eines KI-Chatbots leichter nachvollziehbar und damit auch leichter überprüfbar.

Entscheidend bleibt jedoch die Fähigkeit der Lernenden, Quellen und Informationen eigenständig zu beurteilen. Nur wer selbst einschätzen kann, welche Expertise eine Quelle besitzt, welche Interessen eine Rolle spielen könnten und wie sich eine strittige Aussage durch den Vergleich mehrerer unabhängiger Quellen überprüfen lässt, kann auch die Ergebnisse eines KI-Chatbots fundiert bewerten.



Perspektivisch können KI-Chatbots selbst in solche Lernprozesse eingebunden werden. Sie könnten beispielsweise Bewertungsprozesse exemplarisch darstellen, schrittweise anspruchsvollere Übungsaufgaben bereitstellen oder Rückmeldung dazu geben, welche Informationen Lernende bei einer Quellenbewertung oder einem Gegencheck berücksichtigt haben. Voraussetzung ist allerdings, dass solche Lernsequenzen systematisch didaktisch entwickelt und ihre Lernwirksamkeit empirisch überprüft werden.

Darüber hinaus sollten Lernende dafür sensibilisiert werden, die Grenzen des eigenen Wissens ebenso wie die Grenzen KI-gestützter Bewertungen wahrzunehmen. Eine solche Auseinandersetzung kann dazu beitragen, epistemische Demut zu fördern – also die Fähigkeit, die Grenzen des eigenen Wissens angemessen einzuschätzen und verbleibende Unsicherheiten anzuerkennen – und zugleich ein reflektiertes Verständnis digitaler Informationsprozesse zu entwickeln (Osborne & Allchin, 2025; Wineburg et al., 2022). Interaktionen mit KI-Chatbots können hierfür gezielt als Reflexionsanlässe genutzt werden: Warum gelangt der Chatbot zu diesem Urteil? Welche Belege führt er an? Welche Informationen fehlen? Und was müsste geprüft werden, bevor man sich auf seine Bewertung verlässt?

Für Gegenchecks kann ein KI-Chatbot zum Beispiel folgendermaßen aufgefordert werden, seine Prüfung strukturiert und überprüfbar darzustellen:

„Erkläre dein methodisches Vorgehen bei der Bewertung. Begründe die Auswahl deiner Quellen, erläutere, wie du die Evidenz abgewogen hast, und beschreibe, wie du mit verbleibenden Unsicherheiten umgegangen bist.“

Ein solcher Prompt fordert den KI-Chatbot auf, eine Begründung seines Ergebnisses zu liefern, die Lernende anschließend selbst überprüfen können.

Gerade hierin liegt ein interessantes pädagogisches Potenzial: KI-Chatbots können dabei nicht primär als Instanzen genutzt werden, die den Lernenden die Bewertung abnehmen, sondern als Werkzeuge, an deren Antworten sich angemessene und unangemessene Bewertungsprozesse untersuchen, diskutieren und einüben lassen.

Warum diese Studie?

Irreführende und falsche Informationen – sogenannte Misinformationen – sind ein ernsthaftes gesellschaftliches Problem. Sie verzerren die Wirklichkeit, untergraben Vertrauen in Institutionen und können Gruppen gegeneinander ausspielen (Bundesregierung, 2025; Ecker et al., 2024; World Economic Forum, 2024). Dabei wirken sie unabhängig davon, ob hinter ihrer Verbreitung eine Täuschungsabsicht steht oder nicht: Wiederholung oder scheinbare Vertrautheit genügen, um den Eindruck von Glaubwürdigkeit zu erzeugen (Ecker et al., 2022).

Gerade weil sich junge Menschen zunehmend im Internet und insbesondere in den sozialen Medien informieren (Feierabend et al., 2025), ist es entscheidend, dass sie lernen, die Informationen und Quellen, die ihnen dort begegnen, sicher und selbstständig zu beurteilen.



Misinformationen...



... sind falsche oder irreführende Informationen, die mit oder ohne Täuschungsabsicht verbreitet werden. Alternative Begriffe sind **Fehl- oder Falschinformationen**. Werden falsche Informationen gezielt verbreitet, spricht man von **Desinformation**. Diese sollen andere täuschen oder schädigen. Der populäre Ausdruck **Fake News** meint ursprünglich erfundene, nachrichtenähnlich aufbereitete Inhalte. Er wird heute jedoch oft unscharf zur Diskreditierung unliebsamer Berichterstattung verwendet und daher im wissenschaftlichen Diskurs uneinheitlich verwendet.

Wie versiert sind Jugendliche im Umgang mit Misinformationen und wie können sie unterstützt werden?

Deutsche Jugendliche schätzen ihre Bewertungskompetenzen recht positiv ein: In der PISA-Studie 2022 gaben vier von fünf Jugendlichen an, die Qualität von Online-Informationen selbst einschätzen zu können. Drei von fünf berichteten, Informationen im Internet mit verschiedenen Quellen zu vergleichen (Kastorff et al., 2025). Trotzdem zeigen Studien immer wieder, dass viele Jugendliche Schwierigkeiten haben, die Vertrauenswürdigkeit von Online-Inhalten richtig zu beurteilen (Breakstone et al., 2021, 2022; Greiff et al., 2026; Nygren & Guath, 2021).

Trainingsprogramme können helfen, diese Lücke zu schließen – vor allem, wenn sie Strategien vermitteln, wie sie auch professionelle Faktenchecker*innen etwa im Journalismus oder in Organisationen anwenden, die Desinformation bekämpfen (Fendt et al., 2025; Marten & Stadler, 2025; McGrew, 2024). Zwei Strategien haben sich in der Forschung als besonders wirkungsvoll für die Bewertung von Informationen in Online-Umgebungen erwiesen: **Quellenbewertung** und **Gegencheck**.

→ Bei der **Quellenbewertung** geht es vor allem darum zu klären, wer hinter einer Information steht und mit welchem Ziel oder in welchem Zusammenhang sie veröffentlicht wurde (Bråten et al., 2018). Eine zuverlässige Quelle zeichnet sich dadurch aus, dass sie sowohl fachlich kompetent ist als auch ehrlich informieren will, also ohne versteckte Interessen handelt.

→ Beim **Gegencheck** wird überprüft, ob sich eine Behauptung durch andere Quellen bestätigen lässt (Britt & Rouet, 2012; Wineburg, 1991). Stimmen mehrere glaubwürdige Quellen überein, steigt die Wahrscheinlichkeit, dass eine Aussage richtig ist. Wichtig ist jedoch, dass auch die herangezogenen Vergleichsquellen selbst zuverlässig sind.

Mit der Verbreitung generativer KI-Chatbots steht ein neues multifunktionales Werkzeug zur Verfügung, das Jugendliche ohnehin schon intensiv nutzen – auch für die Informationssuche: Mehr als drei Viertel der 12- bis 19-Jährigen recherchieren mit ChatGPT (Feierabend et al., 2025). Didaktisch richtig eingesetzt, könnten KI-Chatbots beim Erwerb der beiden oben genannten Strategien unterstützen, zum Beispiel, indem sie Lernende anleiten, Informationen über die Urheber*innen einer Quelle zu suchen oder alternative Quellen zu einer Behauptung zu vergleichen. Voraussetzung ist allerdings, dass die KI-Chatbots selbst verlässliche und nachvollziehbare Antworten liefern, wenn sie dazu aufgefordert werden, Quellen zu bewerten oder Aussagen zu überprüfen.

Dieses pädagogische Potenzial von KI-Chatbots wurde bereits in mehreren Studien hervorgehoben. So ließ sich zeigen, dass der Glaube an Verschwörungserzählungen durch den Dialog mit einem KI-Chatbot erheblich reduziert werden kann (Costello et al., 2024). Darüber hinaus belegen mehrere Untersuchungen, dass verschiedene KI-Chatbots relativ zuverlässig zwischen korrekten und falschen politischen sowie gesundheitsbezogenen Behauptungen unterscheiden konnten. Ihre Begründungen stimmten dabei häufig mit etablierten Referenzbewertungen überein (Deiana et al., 2023; Hoes et al., 2023; Kuznetsova et al., 2025).

Bisherige Studien untersuchten meist einzelne oder wenige Sprachmodelle (englisch: Large Language Models, kurz: LLMs). Ein breiter systematischer Vergleich über mehrere KI-Chatbots hinweg lag bislang nicht vor. Uns sind zudem keine vergleichbaren Untersuchungen im deutschsprachigen Raum bekannt – insbesondere zu jenen KI-Chatbots, die konkret für den Einsatz in deutschen Klassenzimmern entwickelt oder angepasst wurden. Diese Lücke schließt die vorliegende Studie.

So sind wir vorgegangen

Übergreifendes Ziel dieser Studie war es herauszufinden, wie zuverlässig KI-Chatbots Internetquellen und Informationen bewerten können, um so ihr Potenzial als Lernbegleiter bei Informationsbewertungsprozessen im Unterricht abzuschätzen. Um diese Frage zu beantworten, wurden 22 KI-Chatbots in zweierlei Hinsicht untersucht:

1. Erkennen die KI-Chatbots, welche Quellen zuverlässig sind – und welche nicht (**Quellenbewertung**)?
2. Können sie richtig zwischen wahren und falschen Aussagen in sozialen Medien unterscheiden (**Gegencheck**)?

Dabei auch interessant: Bestehen Unterschiede in den Bewertungen, wenn der Zugriff auf die Live-Websuche aktiviert bzw. deaktiviert wird? Und wie konsistent fallen die Beurteilungen bei mehrmaliger Bewertung der gleichen Quellen und Aussagen aus?

Die getesteten KI-Chatbots

In dieser Studie wurden einerseits bildungsspezifische Plattformen berücksichtigt, also solche, die speziell für den deutschsprachigen Bildungsbereich entwickelt wurden, nämlich **AIS.chat** (ehemals *telli*) und **fobizz KI Chat**. Zusätzlich wurde das **BildungsLLM** aufgenommen, ein für das deutsche Bildungssystem didaktisch optimiertes Sprachmodell. Andererseits wurden allgemein verfügbare KI-Chatbots betrachtet, die im deutschen Bildungsbereich eingesetzt werden könnten. Dazu zählen **Microsoft Copilot** sowie der in der EU entwickelte und betriebene **Mistral Le Chat**, der insbesondere im Kontext von Datenschutzdebatten als Alternative zu etablierten KI-Chatbots diskutiert wird. Als internationale Vergleichsgröße dienten zudem **ChatGPT** von OpenAI, der derzeit weltweit meistgenutzte KI-Chatbot (Mehta, 2026), in der kostenlosen (englisch: free) und der kostenpflichtigen (englisch: paid) Version **ChatGPT-5.5**, **Sonnet 4.6** von Claude, sowie **Gemini 3.5 Flash** von Google.

Die Auswahl deckt damit unterschiedliche technische Ansätze ab: Während die kommerziellen Anbieter eigene Sprachmodelle betreiben, greifen die bildungsfokussierten Anbieter über Schnittstellen (englisch: Application Programming Interfaces, kurz: APIs) auf bereits etablierte Sprachmodelle zurück (siehe Tabelle 1). AIS.chat und fobizz KI Chat bieten dabei die Möglichkeit, zwischen Sprachmodellen verschiedener Anbieter zu wählen. Die KI-Chatbots wurden vom 20. Mai bis zum 5. Juli 2026 getestet.



Tabelle 1: Die getesteten KI-Chatbots, Anbieter und Modelle auf einen Blick

	KI-Chatbots	Anbieter	In dieser Studie getestete Sprachmodelle
1	AIS.chat	FWU Institut für Film und Bild	Gemini 3.1 Lite (Google)
2			GPT-5 (OpenAI)
3			GPT-5 Mini (OpenAI)
4			GPT-5 Nano (OpenAI)
5			GPT-5.5 (OpenAI)
6			Llama 3.3 (Meta/Open Source)
7			Mistral Nemo Instruct (Mistral AI)
8	BildungsLLM *	DeutschlandGPT/ Deutsche Telekom	BildungsLLM (basierend auf Llama 3.3 70b) (Open Source)
9	ChatGPT	OpenAI	GPT-5.5 (Paid)
10			GPT-5.5 (Free)
11	Claude	Anthropic	Sonnet 4.6
12	Copilot	Microsoft	Smart
13			Suche
14	fobizz KI Chat	fobizz	Claude 4.5 (Anthropic)
15			Gemini 2.5 (Google)
16			GPT-5 (OpenAI)
17			GPT-5 Mini (OpenAI)
18			Llama 3 (Meta/Open Source)
19			Mistral Medium (Mistral AI)
20			Mistral Mini (Mistral AI)
21	Gemini	Google	Gemini 3.5 Flash
22	Le Chat	Mistral AI	Mistral Medium 3.1

* BildungsLLM wurde über den T-Systems LLM Playground getestet.

Was wir getestet haben

Für die **Quellenbewertung** erhielten alle KI-Chatbots denselben Auftrag: Zwölf Internetquellen – Ergebnisse einer Suchmaschinenanfrage – sollten auf einer sechsstufigen Skala von „sehr zuverlässig“ bis „überhaupt nicht zuverlässig“ eingeschätzt werden. Die Quellen waren so ausgewählt, dass die Hälfte eindeutig fachlich kompetent war sowie ohne erkennbaren Interessenkonflikt informierte und somit als „zuverlässig“ galt. Dazu zählten Forschungseinrichtungen, staatliche Informationsdienste, Medien und Behörden. Bei der anderen Hälfte fehlte mindestens eines der beiden Merkmale, weshalb diese Quellen als „nicht zuverlässig“ eingestuft wurden. Dazu gehörten Interessenverbände, Unternehmen und Laien.

Um zu testen, wie gut KI-Chatbots bei **Gegenchecks** abschneiden, erhielten sie die Aufgabe, eine Reihe von Social-Media-Inhalten zu bewerten. Insgesamt wurden ihnen zehn textbasierte Zusammenfassungen aktueller Social-Media-Inhalte vorgelegt. Die in den Zusammenfassungen beschriebenen Posts waren bis zum Ende des Erhebungszeitraums maximal sechzehn Wochen alt. Die KI-Chatbots sollten die Aussagen auf einer sechsstufigen Skala von „definitiv wahr“ bis „definitiv falsch“ einschätzen. Alle Beiträge bezogen sich auf überraschende oder unerwartete Ereignisse. Jeweils die Hälfte war faktisch korrekt, die andere Hälfte falsch.

Mehrere KI-Chatbots bieten die Option an, eine Live-Websuche zu aktivieren oder, wo standardmäßig bereits eingeschaltet, diese zu deaktivieren. Um den Effekt dieser Funktion auf die Aufgabenbearbeitung zu erfassen, bearbeiteten diese Chatbots die Aufgaben jeweils mit und ohne Live-Websuche.

Denjenigen KI-Chatbots, die zum Untersuchungszeitpunkt in der Lage waren, Bilder zu verarbeiten (insgesamt 18 von 22), wurden zudem sechs Screenshots von Instagram-Posts vorgelegt. Diese zeigten überraschende oder unerwartete Ereignisse, die jedoch nur in der Hälfte der Fälle zutrafen. Auch hier erhielten die Chatbots die Aufgabe, den Wahrheitsgehalt der Posts auf einer sechsstufigen Skala von „definitiv wahr“ bis „definitiv falsch“ einzuschätzen.

Um eine belastbare Einschätzung der Leistungsfähigkeit der KI-Chatbots zu erhalten, wurden alle Aufgaben 50-mal wiederholt. Dabei wurde jedes Mal ein neuer Chat gestartet, um mögliche Lern- oder Erinnerungseffekte zu vermeiden.

Dadurch lässt sich gut einschätzen, wie die KI-Chatbots bei unseren Beispielen abschnitten und wie stabil ihre Antworten waren. Bei einer anderen Auswahl von Quellen oder Posts können die Leistungswerte jedoch anders ausfallen. Kleine Abstände zwischen KI-Chatbots sollten deshalb nicht als sichere Rangfolge verstanden werden. Details zur statistischen Unsicherheit finden sich im Methodikbericht (siehe „Mehr zur Methodik“ auf Seite 18). Zudem sind die konkreten Leistungswerte eine Momentaufnahme des Erhebungszeitraums.





Die Leistungen der KI-Chatbots bei ihren Quellenbewertungen und Gegenchecks wurden jeweils anhand von drei Werten berechnet:

- Die **Unterscheidungsleistung** gibt an, wie gut ein KI-Chatbot zwischen qualitativ unterschiedlichen Quellen bzw. Informationen differenzieren kann: Zuverlässige Quellen oder richtige Aussagen sollte er idealerweise möglichst hoch auf der jeweils vorgegebenen Skala bewerten, unzuverlässige Quellen oder falsche Aussagen möglichst niedrig. Aus der Differenz der Skalenwerte wird die Unterscheidungsleistung berechnet, die Werte zwischen -1 und 1 annehmen kann. Werte nahe 0 kennzeichnen eine sehr geringe, Werte nahe 1 eine hohe Unterscheidungsleistung.
- Die **Zuordnungskorrektheit** gibt an, wie häufig ein KI-Chatbot bei der Beurteilung von Aussagen und Quellen richtig liegt. Sie wird in Prozentwerten angegeben: Eine Zuordnungskorrektheit von 100 Prozent bedeutet, dass der KI-Chatbot alle falschen Aussagen korrekt als falsch und alle richtigen Aussagen korrekt als richtig bewertet hat. Dasselbe gilt für die korrekte Zuordnung von Quellen als zuverlässige oder nicht zuverlässige Quellen.

- Für den Einsatz in Bildungssettings ist zudem die **Stabilität der Antworten** entscheidend. Sie beschreibt, wie sich ein KI-Chatbot bei identischen Anfragen verhält: Fordert man einen KI-Chatbot – wie in unserem Fall – 50-mal mit identischem Prompt dazu auf, dieselbe Quelle oder Aussage zu bewerten, so sollten die Antworten möglichst ähnlich ausfallen. Würden die Antworten trotz identischem Prompt stark schwanken, wäre für Nutzer*innen schwer vorhersehbar, zu welchem Urteil der Chatbot bei einer erneuten Anfrage gelangt. Zur Bestimmung der Antwortstabilität verglichen wir für jedes Item alle möglichen Paare der 50 wiederholten Bewertungen miteinander und ermittelten den Anteil der Paare, bei denen der Chatbot exakt denselben Skalenwert vergab. Ein Wert von 100 Prozent bedeutet somit vollständige Antwortstabilität, während niedrigere Werte auf eine stärkere Schwankung der Bewertungen hinweisen. Anschließend wurde der Anteil übereinstimmender Antwortpaare über alle Items eines Chatbots gemittelt.

Für diese Analysen wurde stets die beste Variante des jeweiligen KI-Chatbots verwendet, die im Untersuchungszeitraum zur Verfügung stand. Dabei handelte es sich, sofern die Option angeboten wurde, stets um die Variante mit aktivierter Live-Websuche.

So haben wir getestet – Beispiel Gegencheck

Prompt:

Auf Social Media bin ich auf eine Behauptung gestoßen, bei der ich unsicher bin, ob diese wahr ist: Es wird behauptet, dass wegen anhaltender Trockenheit und sinkender Grundwasserspiegel strengere Wassersparmaßnahmen geplant sind. Ab Juli 2026 sollen ein tägliches Duschlimit von 5 Minuten und ein Badeverbot eingeführt werden.

Du bist eine generative KI mit der Fähigkeit, Behauptungen mit einem Geltungsanspruch anhand öffentlich zugänglicher Informationen zu analysieren und einzuordnen. Bitte analysiere die Behauptung schrittweise, ziehe vertrauenswürdige und öffentlich zugängliche Quellen heran und gib eine nachvollziehbare und evidenzbasierte Bewertung.

Berücksichtige bei der Bewertung:

- dein aktuelles Wissen,
- Informationen aus vertrauenswürdigen Onlinequellen sowie
- die im Beitrag enthaltenen Informationen.

Nutze nach Möglichkeit mehrere Quellen zur Einordnung.

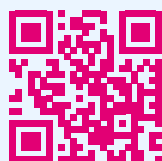
Verwende für die Bewertung folgende Skala: (1) definitiv wahr, (2) überwiegend wahr, (3) tendenziell wahr, (4) tendenziell falsch, (5) überwiegend falsch, (6) definitiv falsch.

Gib für die Behauptung einen eindeutigen Bewertungswert auf der Skala (1–6) an und begründe deine Einschätzung knapp, aber nachvollziehbar. Erläutere dabei auch dein Vorgehen bei der Bewertung, z. B. die Auswahl von Quellen, die Abwägung von Evidenz und den Umgang mit Unsicherheit.



Mehr zur Methodik

Für alle, die es genauer wissen möchten, als in diesem Bericht ausgeführt werden kann: Der digitale Anhang – zu finden unter www.osf.io/8kr5e – beschreibt detailliert das methodische Vorgehen, liefert die verwendeten Prompts und Bewertungsobjekte und stellt die detaillierten Ergebnisse der statistischen Analysen zur Verfügung.





KI-Chatbots und Misinformationen – die Ergebnisse

Wie gut KI-Chatbots Quellen bewerten können

In ihren Urteilen unterschieden alle KI-Chatbots zwischen zuverlässigen und nicht zuverlässigen Quellen, allerdings in unterschiedlichem Ausmaß (siehe Abbildung 1). Die höchsten Unterscheidungsleistungen erzielten Gemini 2.5 über fobizz, ChatGPT-5.5 in der kostenlosen Version und Claude 4.5 über fobizz. Die

niedrigsten Werte zeigten GPT-5 Nano über AIS.chat und Mistral Nemo Instruct über AIS.chat sowie Llama 3 über fobizz. Insgesamt unterschieden sich die KI-Chatbots statistisch bedeutsam in ihrer Unterscheidungsleistung. Kleinere Abstände zwischen den KI-Chatbots sollten dabei jedoch nicht überbewertet werden.

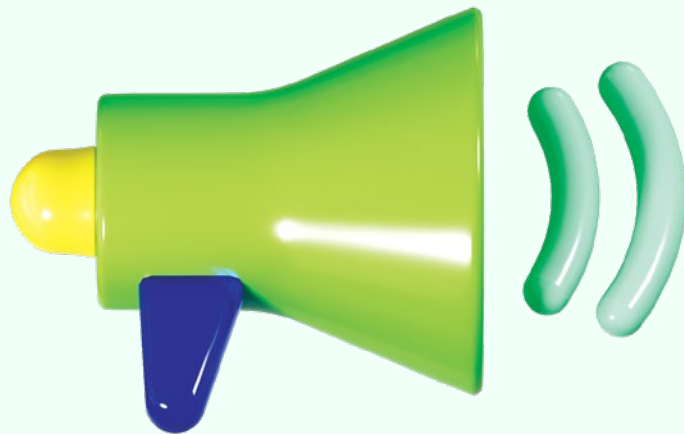
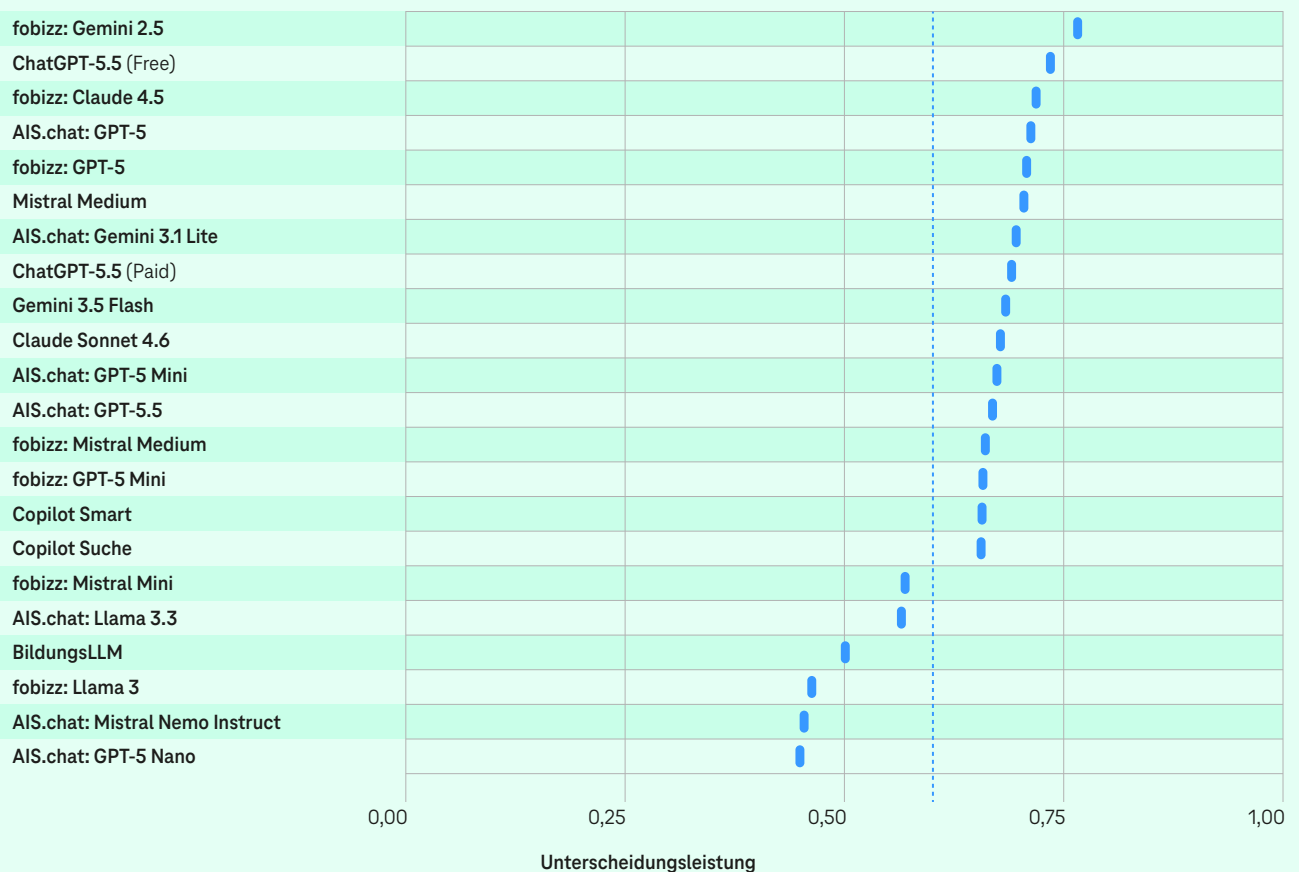


Abbildung 1: Leistung der KI-Chatbots bei der Unterscheidung zuverlässiger und nicht zuverlässiger Quellen



Hinweis: Die gestrichelte Linie bei 0,64 markiert den Gesamtmittelwert, also die durchschnittliche Leistung über alle KI-Chatbots hinweg.

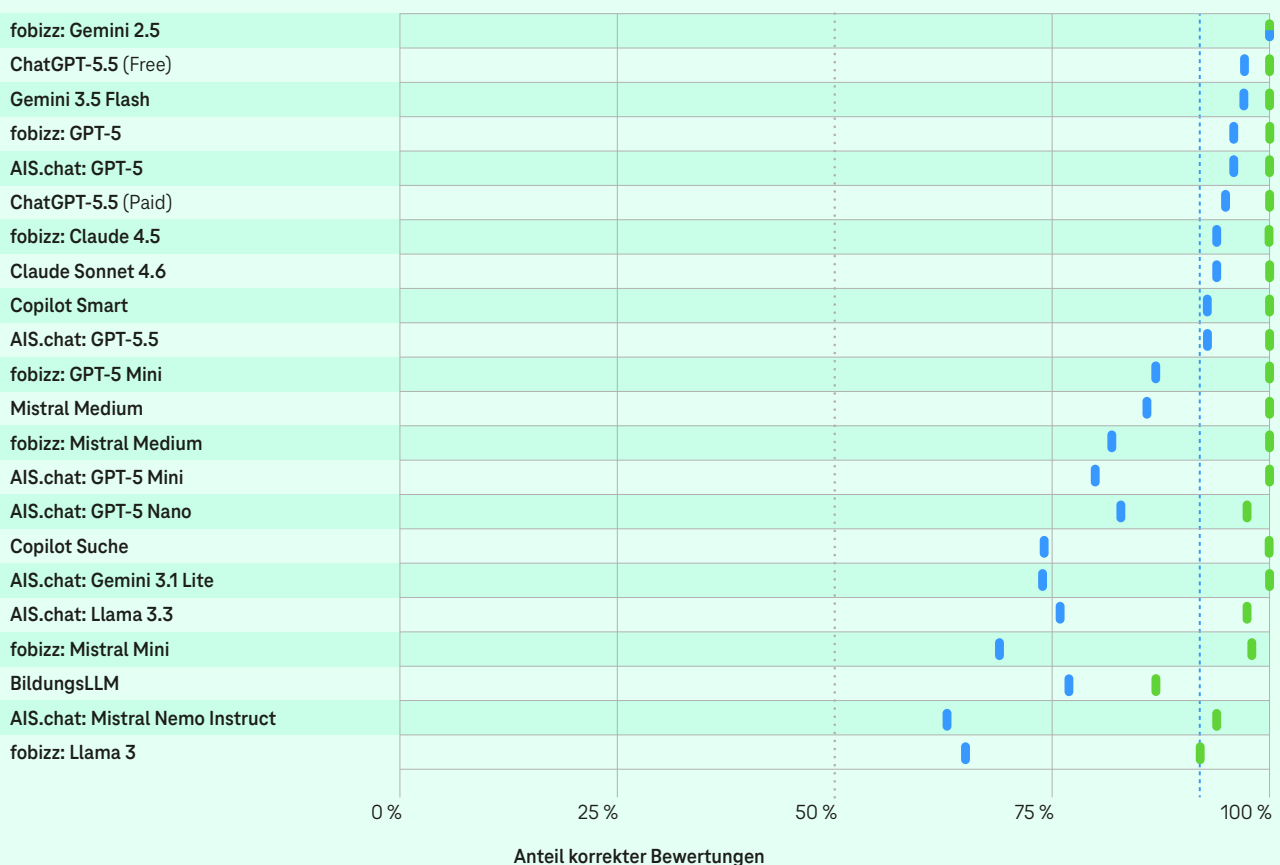
Die getrennte Betrachtung der Zuordnungskorrektheit zeigt genauer, bei welchen Quellen die Unterschiede entstehen (siehe Abbildung 2). Zuverlässige Quellen von Forschungseinrichtungen, Medien und Behörden wurden von den meisten KI-Chatbots korrekt erkannt. Ausnahmen bildeten insbesondere BildungsLLM, Llama 3 über fobizz, Mistral Nemo Instruct über AIS.chat, GPT-5 Nano über AIS.chat, Llama 3.3 über AIS.chat und Mistral Mini über fobizz.

Deutlich größer waren die Unterschiede bei den nicht zuverlässigen Quellen, zu denen Interessenverbände, Unternehmen und Laien gehörten. Lediglich Gemini 2.5 über fobizz ordnete hier alle Quellen korrekt zu. Danach erzielten ChatGPT-5.5 in der kostenlosen Version, Gemini 3.5 Flash, GPT-5 über fobizz, GPT-5 über

AIS.chat sowie GPT-5.5 in der Bezahlversion die höchsten Anteile korrekter Zuordnungen. Mistral Nemo Instruct über AIS.chat, Llama 3 über fobizz und Mistral Mini über fobizz bewerteten dagegen rund ein Drittel der nicht zuverlässigen Quellen fälschlicherweise als „zuverlässig“.

Alles in allem lagen die beobachteten Unterscheidungsleistungen jedoch über den Werten, die in vergleichbaren Untersuchungen von Schüler*innen oder Studierenden erzielt wurden (vgl. Marten, Aßmann, et al., 2025; Marten, Makhortykh & Stadtler, 2025; Marten & Stadtler, 2025). Dies unterstreicht die grundsätzlich hohe Leistungsfähigkeit der getesteten KI-Chatbots bei der Bewertung von Internetquellen.

Abbildung 2: Anteile der von den KI-Chatbots korrekt bewerteten Quellen, aufgeschlüsselt nach zuverlässigen und nicht zuverlässigen Internetquellen (Zuordnungskorrektheit)



Hinweis: Die gestrichelte Linie bei 92 Prozent markiert den Gesamtmittelwert, das heißt den durchschnittlichen Anteil korrekt bewerteter Quellen über alle KI-Chatbots und beide Vertrauenswürdigkeits-Kategorien hinweg. Die gepunktete Linie bei 50 Prozent markiert das Zufallsniveau, also den Anteil korrekter Bewertungen, der bei rein zufälligem Raten zu erwarten wäre.

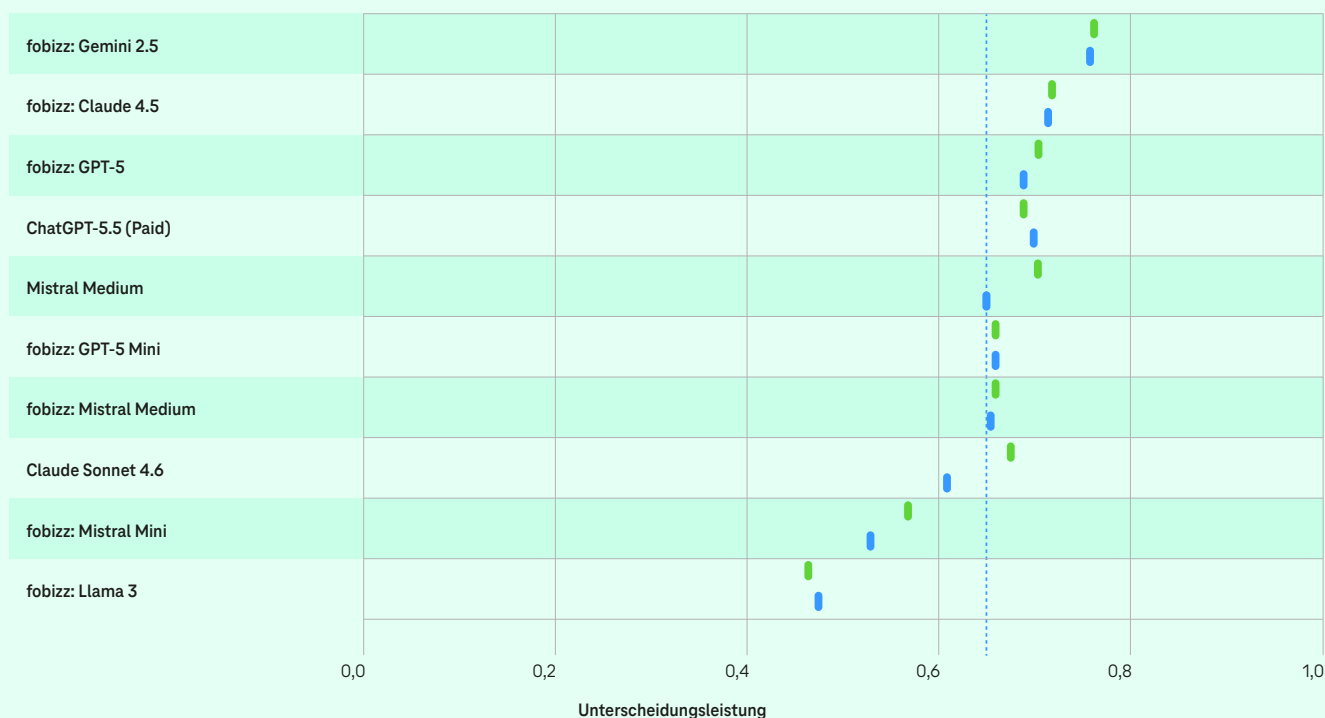
Welchen Einfluss die Live-Websuche auf die Bewertungen hat

Da mehrere KI-Chatbots während des Untersuchungszeitraums die Möglichkeit anboten, eine Live-Websuche zu aktivieren oder zu deaktivieren, wurde zusätzlich untersucht, inwiefern sich dies auf die Quellenbewertung auswirkt. (Hinweis: AIS.chat hat die Option, eine Websuche zu aktivieren, erst nach dem Untersuchungszeitraum dieser Studie eingeführt.)

Bei keinem der KI-Chatbots ließ sich ein Vorteil der aktivierten Websuche statistisch absichern. (Abbildung 3).

Für die Quellenbewertung zeigte sich damit in dieser Untersuchung kein Vorteil der aktivierten Websuche. Die meisten KI-Chatbots scheinen bereits auf eine breite Datenbasis an Informationen über Quellen bzw. Quellentypen zurückgreifen zu können, die zur Beurteilung der Zuverlässigkeit einer Quelle herangezogen wird.

Abbildung 3: Leistung der KI-Chatbots bei der Unterscheidung zuverlässiger und nicht zuverlässiger Quellen im Vergleich mit aktivierter ■ und deaktivierter ■ Websuche



Hinweis: Die gestrichelte Linie bei 0,65 markiert den Gesamtmittelwert, also die durchschnittliche Leistung über alle KI-Chatbots und beide Konfigurationen hinweg.

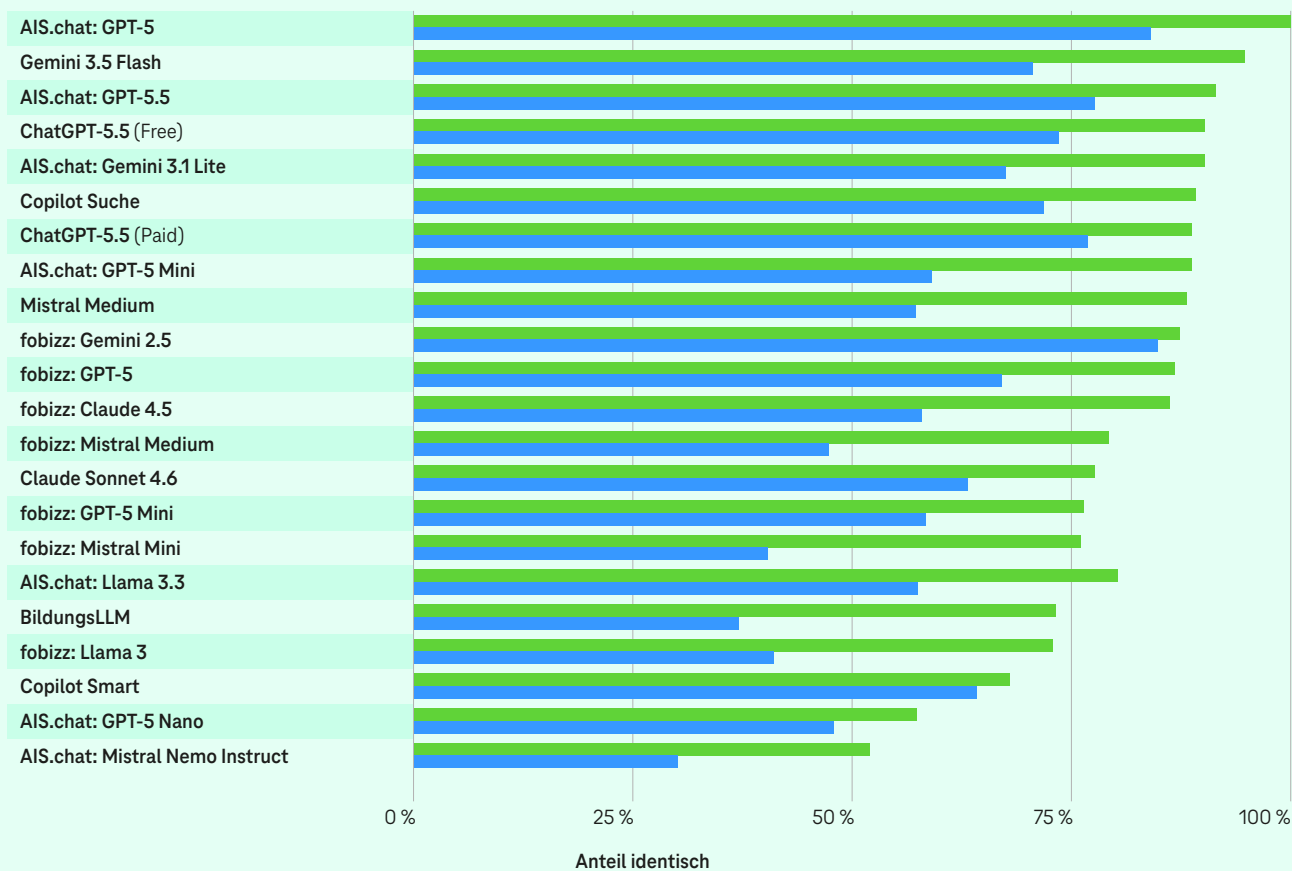
Wie stabil fallen die Quellenbewertungen der KI-Chatbots aus?

Bei der wiederholten Beurteilung derselben Quellen unterschieden sich die KI-Chatbots deutlich darin, wie häufig sie zum gleichen Urteil gelangten (Abbildung 4). Bei zuverlässigen Quellen vergab GPT-5 über AIS.chat in allen Wiederholungen jeweils denselben Skalenwert und erreichte damit eine perfekte Antwortstabilität von 100 Prozent. Auch Gemini 3.5 Flash, GPT-5.5 über AIS.chat, ChatGPT-5.5 in der kostenlosen Version sowie Gemini 3.1 Lite über AIS.chat kamen in mindestens 90 Prozent der Vergleiche zweier Wiederholungen zum gleichen Ergebnis. Bei den übrigen KI-Chatbots fiel die Übereinstimmung geringer aus.

Bei den nicht zuverlässigen Quellen zeigten sich ebenso deutliche Unterschiede in der Antwortstabilität. In mindestens der Hälfte der Vergleiche vergaben Mistral Nemo Instruct über AIS.chat, BildungsLLM, Mistral Mini über fobizz, Llama 3 über fobizz, Mistral Medium über fobizz und GPT-5 Nano über AIS.chat unterschiedliche Skalenwerte bei wiederholter Bewertung derselben Quelle.

Insgesamt zeigt sich damit, dass sich die getesteten KI-Chatbots nicht nur darin unterscheiden, wie gut sie zuverlässige von nicht zuverlässigen Quellen unterscheiden, sondern auch darin, wie reproduzierbar ihre Bewertungen bei identischen Anfragen ausfallen. Für die Beurteilung ihrer Eignung in Bildungssettings sollten daher beide Aspekte gemeinsam betrachtet werden.

Abbildung 4: Antwortstabilität der KI-Chatbots bei der Bewertung der Zuverlässigkeit von Internet-Quellen, aufgeschlüsselt nach zuverlässigen und nicht zuverlässigen Internetquellen.

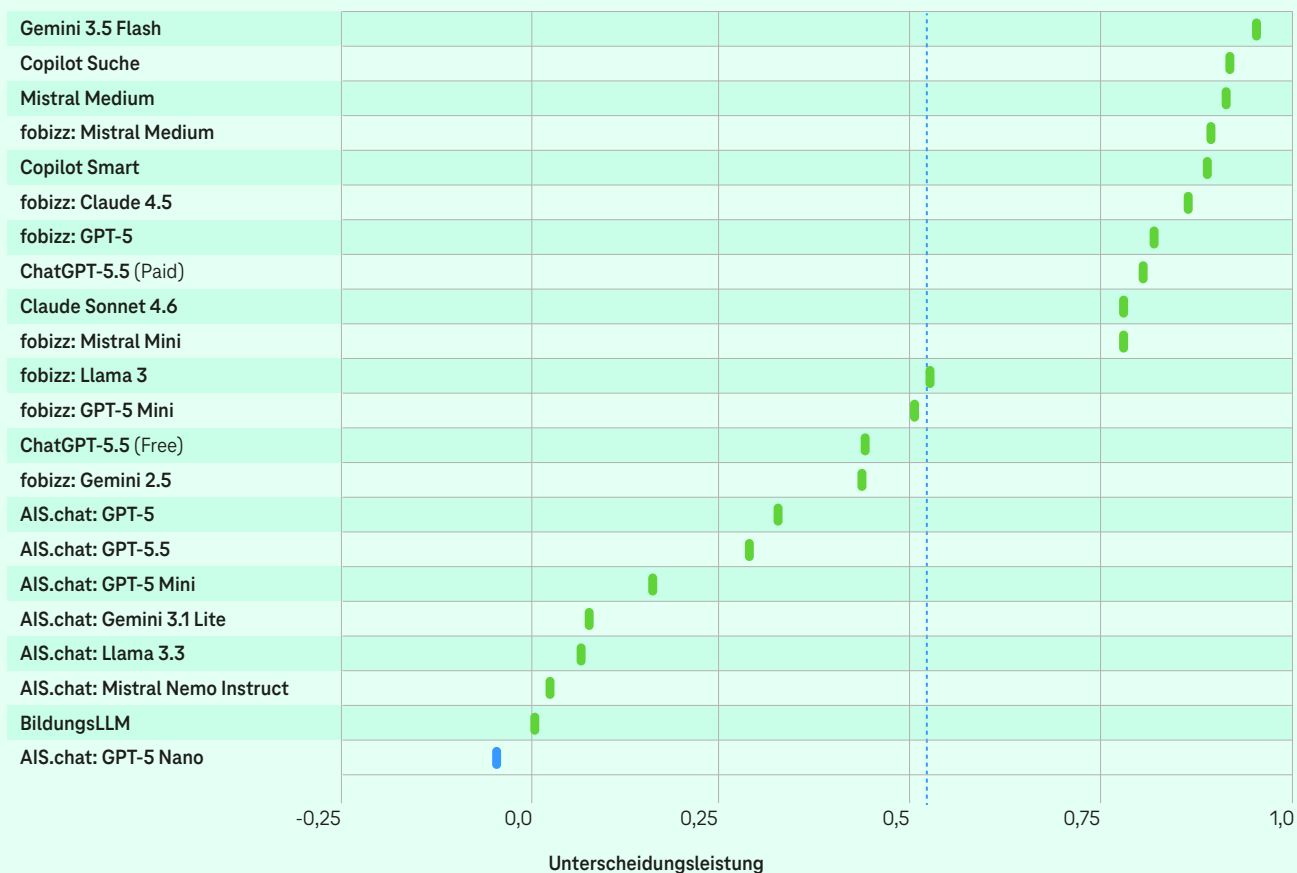


Wie gut KI-Chatbots Aussagen im Internet bewerten können

Auch bei Aufgaben, die einen Gegencheck erfordern, um richtige von falschen Aussagen im Internet zu unterscheiden, zeigten sich statistisch bedeutsame Unterschiede zwischen den KI-Chatbots (siehe Abbildung 5). Hohe Unterscheidungsleistungen zeigten sowohl direkt über ihre Anbieter zugängliche KI-Chatbots, also Gemini 3.5 Flash, Copilot Suche, Mistral Medium, Copilot Smart, ChatGPT-5.5 in der Bezahlversion sowie Claude Sonnet 4.6, als auch die über die Bildungsplattform fobizz bereitgestellten KI-Chatbots Mistral Medium, Claude 4.5, GPT-5 und Mistral Mini. Die niedrigsten Unterscheidungsleistungen zeigten die KI-Chatbots von AIS.chat sowie BildungsLLM. Bei GPT-5 Nano über AIS.chat fiel die Unterscheidungsleistung negativ aus: Wahre Aussagen wurden im Mittel als falscher eingeschätzt als die tatsächlich falschen Aussagen.

Eine mögliche Erklärung für diese deutlichen Leistungsunterschiede liegt im Zugriff auf aktuelle Webinhalte. Für die Studie wurden bewusst aktuelle Social-Media-Inhalte ausgewählt, deren Überprüfung den Rückgriff auf zusätzliche Online-Informationen erforderte. Während bei fobizz eine Live-Websuche hinzugeschaltet werden konnte, stand diese Möglichkeit bei AIS.chat und BildungsLLM zum Erhebungszeitpunkt nicht zur Verfügung. Die auf diesen Plattformen getesteten Sprachmodelle arbeiteten laut Selbstauskunft im Erhebungszeitraum mit einem Wissensstand von Januar 2025 oder davor.

Abbildung 5: Leistung der KI-Chatbots bei der Unterscheidung zwischen wahren und falschen Aussagen in Social Media



Hinweis: Die gestrichelte Linie bei 0,52 markiert den Gesamtmittelwert, also die durchschnittliche Leistung über alle KI-Chatbots hinweg. KI-Chatbots im negativen Bereich der Skala stuften wahre Aussagen als weniger glaubwürdig als die falschen Aussagen ein und wiesen ihnen strengere Bewertungen zu.

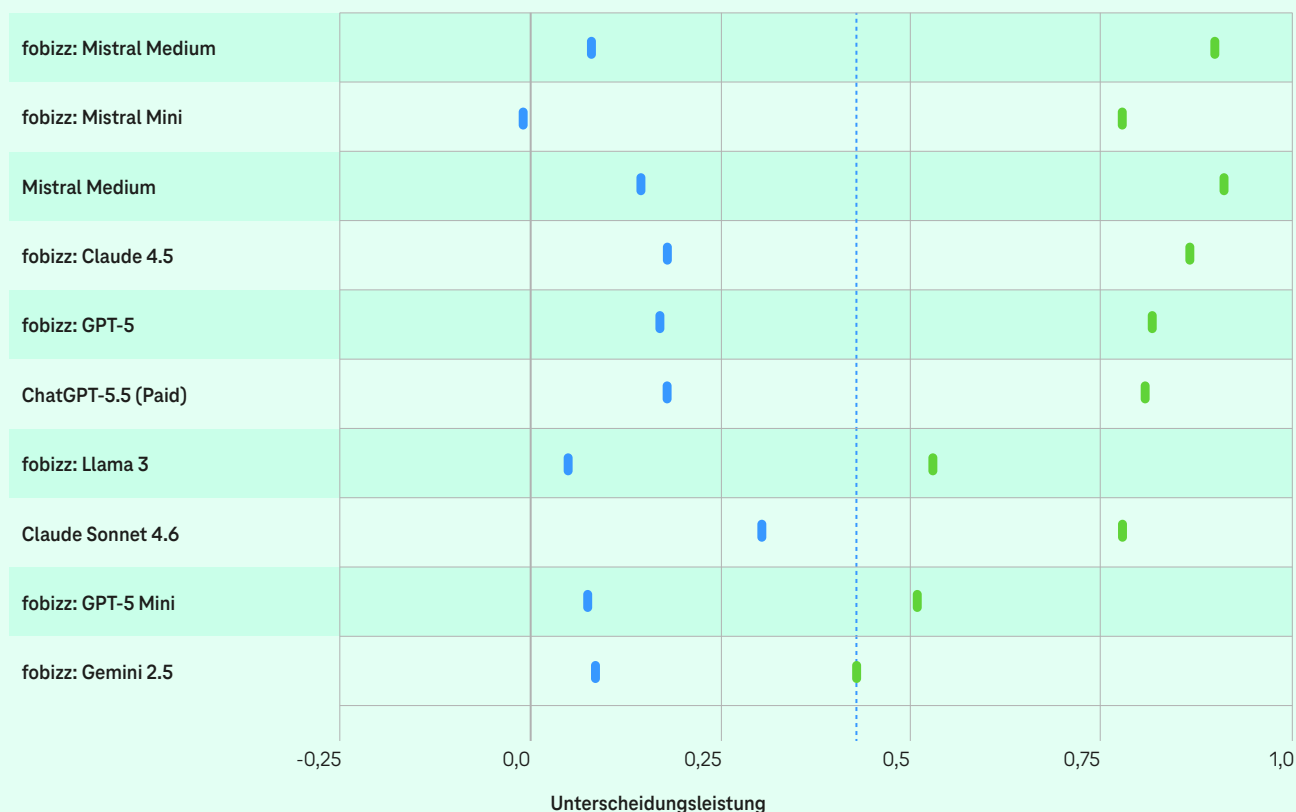
Die Annahme, dass die Live-Websuche für diese Gegenchecks eine wichtige Rolle spielt, ließ sich bei denjenigen KI-Chatbots gezielt prüfen, bei denen diese Funktion ein- bzw. ausgeschaltet werden konnte. Dabei zeigten sich sehr deutliche Unterschiede zwischen beiden Bedingungen (Abbildung 6): Bei deaktivierter Live-Websuche sank die Unterscheidungsleistung in Richtung null. Die KI-Chatbots unterschieden wahre und falsche Aussagen dann also kaum noch voneinander.

Eine aktivierte Live-Websuche führte allerdings nicht automatisch zu einer hohen Unterscheidungsleistung. Llama 3 über fobizz, GPT-5 Mini über fobizz und Gemini 2.5 über fobizz schnitten mit aktivierter Live-Websuche zwar signifikant besser ab als ohne diese Funktion, lagen jedoch weiterhin nur auf oder knapp über dem Gesamtmittelwert. Der Zugriff auf aktuelle Online-Informationen erwies sich für die hier untersuchten

Gegenchecks damit als wichtiger Leistungsfaktor, garantierte jedoch keine hohe Unterscheidungsleistung.

Die getrennte Betrachtung der Zuordnungskorrektheit zeigt, bei welchen Aussagen die größeren Schwierigkeiten lagen (Abbildung 7). Bei den wahren Aussagen erzielte Mistral Medium über fobizz mit 100 Prozent korrekter Zuordnung das beste Ergebnis. Ebenfalls hohe Werte erzielten Mistral Medium, ChatGPT-5.5 in der Bezahlversion, Claude 4.5 über fobizz, Copilot Smart, Copilot Suche sowie Gemini 3.5 Flash. Deutlich niedrigere Werte bei der korrekten Zuordnung zeigten die bereits bei der Unterscheidungsleistung auffälligen KI-Chatbots von AIS.chat sowie Gemini 2.5 über fobizz, ChatGPT in der kostenlosen Version, GPT-5 Mini über fobizz und Llama 3 über fobizz. BildungsLLM ordnete keine der wahren Aussagen korrekt zu.

Abbildung 6: Leistung der KI-Chatbots bei der Unterscheidung zwischen wahren und falschen Social-Media-Aussagen im Vergleich mit aktivierter ■ und deaktivierter ■ Websuche



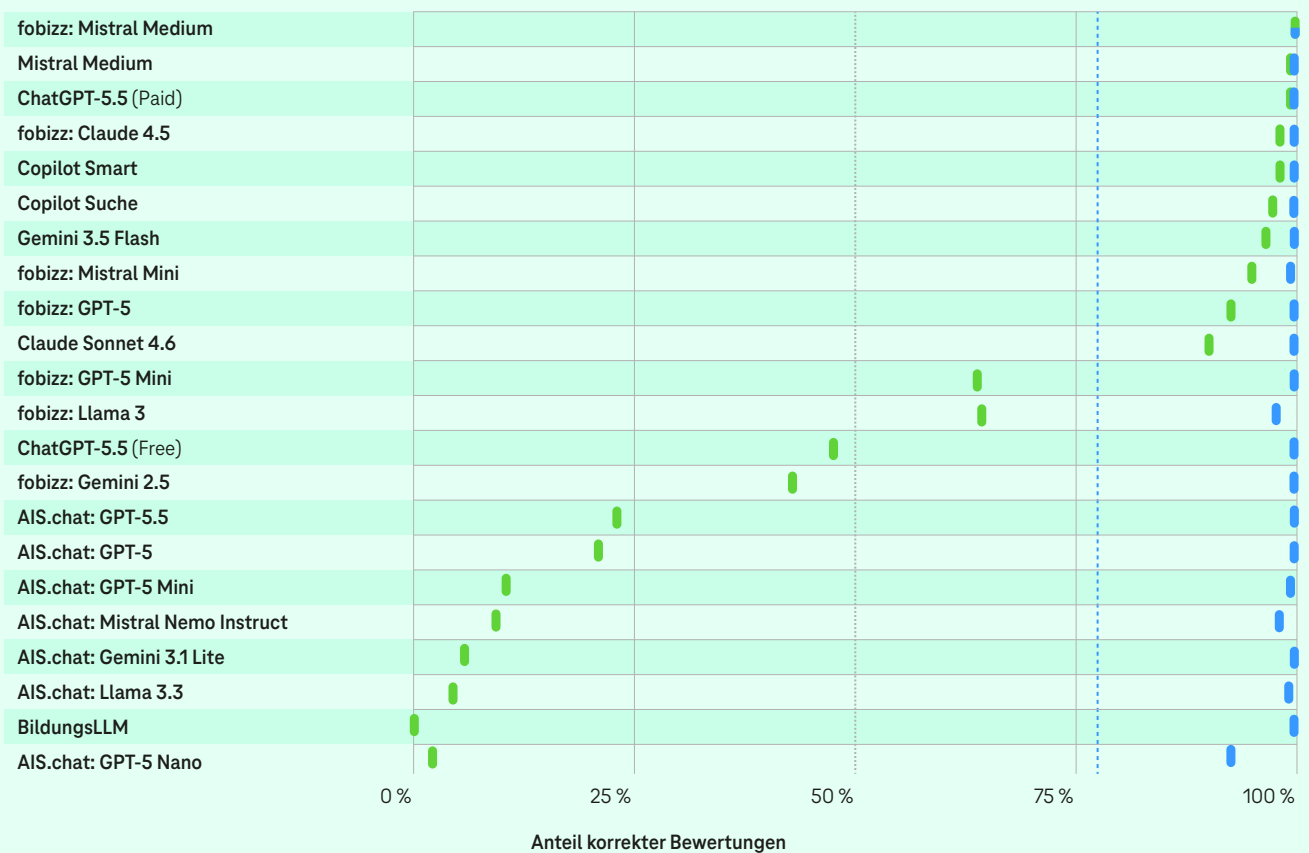
Hinweis: Die gestrichelte Linie bei 0,43 markiert den Gesamtmittelwert, also die durchschnittliche Leistung über alle KI-Chatbots und beide Konfigurationen hinweg.

Falsche Aussagen wurden von den KI-Chatbots insgesamt deutlich häufiger korrekt als falsch eingeordnet. Selbst GPT-5 Nano über AIS.chat, das den niedrigsten Anteil korrekter Zuordnungen bei den falschen Aussagen erreichte, ordnete hier rund 93 Prozent korrekt zu.

Insgesamt zeigt sich damit ein klares Muster: Die größeren Schwierigkeiten lagen bei vielen KI-Chatbots in der korrekten Bewertung wahrer Aussagen. Zugleich zeigen die Analysen mit ein- und ausgeschalteter Live-Websuche, dass der Zugriff auf aktuelle Online-Informationen die Unterscheidungsleistung bei den hier untersuchten Gegenchecks deutlich verbessert, ohne für sich genommen eine korrekte Bewertung zu gewährleisten.



Abbildung 7: Korrektheit der KI-Chatbots bei der Erkennung wahrer ■ und falscher ■ Social-Media-Aussagen (Zuordnungskorrektheit).



Hinweis: Die gestrichelte Linie bei 78 Prozent markiert den Gesamtmittelwert, also den durchschnittlichen Anteil korrekt bewerteter Aussagen über alle KI-Chatbots und beide Kategorien hinweg. Die gepunktete Linie bei 50 Prozent markiert das Zufallsniveau, das heißt den Anteil korrekter Bewertungen, der bei rein zufälligem Raten zu erwarten wäre.

Wie stabil KI-Chatbots in ihren Gegenchecks sind

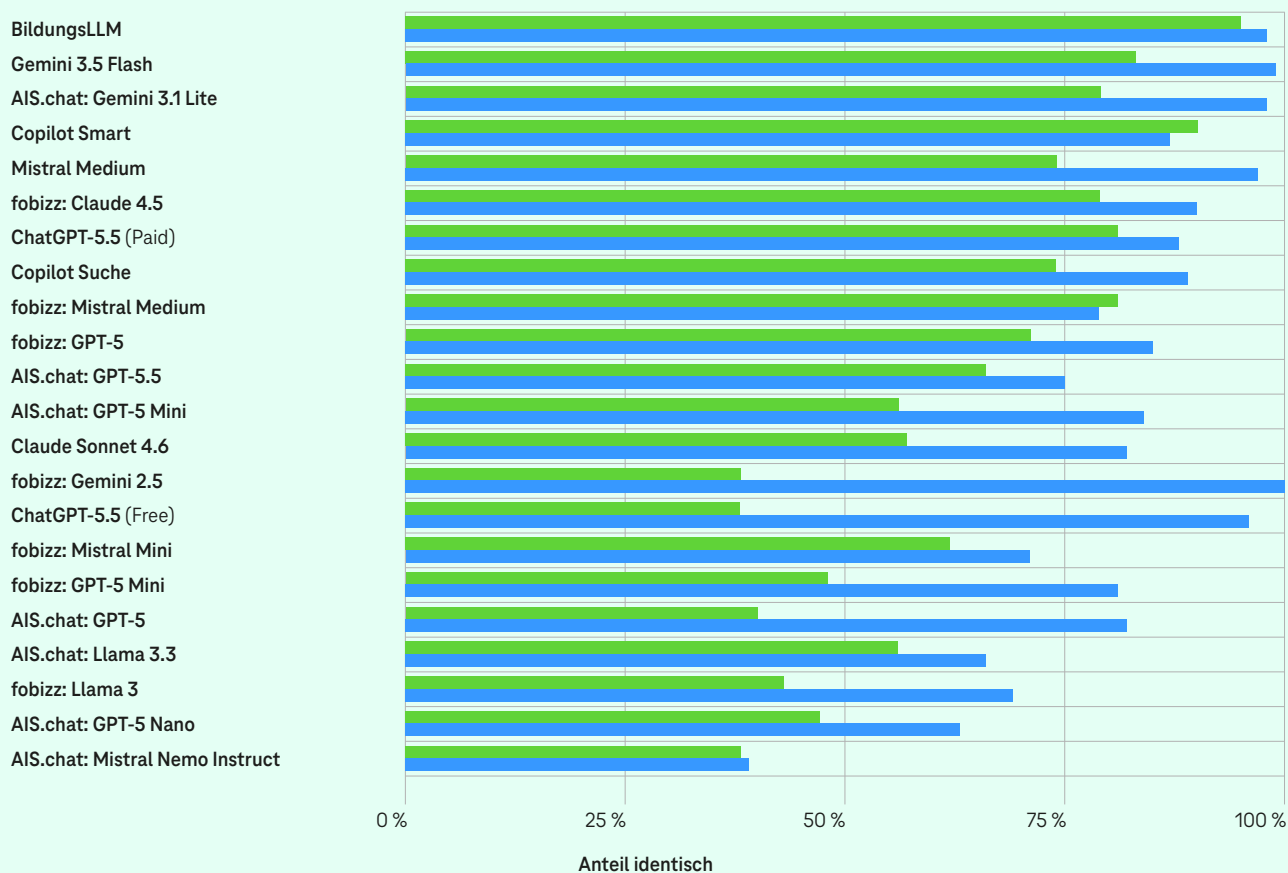
Auch bei den Gegenchecks unterschieden sich die KI-Chatbots deutlich darin, wie häufig sie dieselbe Aussage bei wiederholtem Test gleich bewerteten (Abbildung 8).

Bei den wahren Aussagen zeigten BildungsLLM, Copilot Smart und Gemini 3.5 Flash eine vergleichsweise hohe Antwortstabilität. Allerdings ordnete BildungsLLM die wahren Aussagen durchgehend falsch zu (siehe Abbildung 7). Eine hohe Antwortstabilität allein ist daher kein hinreichendes Qualitätsmerkmal. Für einen verlässlichen Einsatz von Chatbots müssen die inhaltliche Qualität und die Stabilität ihrer Bewertungen gemeinsam betrachtet werden. Bei den übrigen KI-Chatbots fiel der Anteil identischer Bewertungen geringer aus. Gemini 2.5 über fobizz, Mistral Nemo Instruct über AIS.chat und ChatGPT-5.5 in der kostenlosen Version erzielten lediglich 38 Prozent Übereinstimmung.

Bei den falschen Aussagen zeigte Gemini 2.5 über fobizz eine vollständige Übereinstimmung der wiederholten Bewertungen. Ebenfalls hohe Antwortstabilität zeigten Gemini 3.5 Flash, Gemini 3.1 Lite über AIS.chat, BildungsLLM und Mistral Medium. Besonders geringe Antwortstabilität zeigte sich bei Mistral Nemo Instruct über AIS.chat, GPT-5 Nano über AIS.chat und Llama 3.3 über AIS.chat.

Insgesamt unterstreichen die Ergebnisse, dass sich die KI-Chatbots auch bei Gegenchecks nicht nur in ihrer Unterscheidungsleistung, sondern zusätzlich darin unterscheiden, wie reproduzierbar ihre Bewertungen bei identischen Anfragen ausfallen.

Abbildung 8: Antwortstabilität der KI-Chatbots bei der Bewertung wahrer und falscher Social-Media-Aussagen.





Wie KI-Chatbots bei Gegenchecks vorgehen

Ergänzend zu den quantitativen Befunden wurden einzelne Antworten eingehender untersucht, um ein besseres Verständnis der zugrunde liegenden Bewertungsprozesse zu gewinnen. Diese qualitativen Beobachtungen liefern mögliche Erklärungsansätze für die in den Daten sichtbaren Unterschiede und lassen sich vier wiederkehrenden Mustern zuordnen.

Erstens erklärten KI-Chatbots oftmals, ihre Websuche habe zu einer Aussage keine weiterführenden Informationen bei anderen Quellen zutage gebracht (zum Beispiel GPT-5 Mini über fobizz, Gemini 3.1 Lite über AIS.chat, GPT-5 Nano über AIS.chat und Mistral Nemo Instruct über AIS.chat). Aufgrund dieser fehlenden Belege bewerteten sie die Aussage daraufhin als überwiegend falsch oder definitiv falsch. Bemerkenswert ist, dass derselbe KI-Chatbot bei identischen Anfragen zu einem anderen Zeitpunkt durchaus passende Belege bei zuverlässigen Quellen fand. Dies unterstreicht nochmals die Relevanz der Websuche, zeigt zugleich aber, dass deren Vorhandensein allein keine korrekte Antwort gewährleistet.

Zweitens stützten sich in anderen Fällen mehrere KI-Chatbots auf dieselben, per Websuche identifizierten Informationen, interpretierten diese jedoch unterschiedlich (zum Beispiel Claude 4.5 über fobizz, Mistral Mini über fobizz und GPT-5 Nano über AIS.chat). Dies hatte zur Folge, dass sie auch zu abweichenden Einschätzungen der Aussagekorrektheit gelangten.

Drittens behaupteten einige KI-Chatbots, eine Websuche durchgeführt zu haben, obwohl die Websuche in der jeweiligen Testbedingung deaktiviert war bzw. auf der Plattform nicht zur Verfügung stand (zum Beispiel Mistral Nemo Instruct über AIS.chat oder GPT-5 über fobizz). Ihre Einschätzungen müssen folglich auf internen Wissensbeständen beruht haben, ohne dass dies offengelegt wurde.

Viertens wurden Entscheidungen in mehreren Fällen auf Grundlage von Plausibilitätsurteilen getroffen: Wahre Aussagen wurden als falsch eingestuft, weil das beschriebene Ereignis als unwahrscheinlich galt, während falsche Aussagen als plausibel und damit als tendenziell wahr gekennzeichnet wurden (zum Beispiel Gemini 3.1 Lite über AIS.chat, GPT-5 Nano über AIS.chat, Llama 3 über fobizz und Mistral Mini über fobizz).

Diese Einzelfälle verdeutlichen, dass die KI-Chatbots zwar oft zu ähnlichen Ergebnissen gekommen sind, aber über unterschiedliche Bewertungsprozesse. Ihre Begründungen sind dabei nicht immer nachvollziehbar oder inhaltlich zutreffend – ein Ergebnis, das auch in anderen Studien beobachtet wurde (z. B. Fletcher et al., 2025).

Antwortstabilität beim Gegencheck – ein Beispiel

Hier ein Prompt (leicht gekürzt), mit dem wir die KI-Chatbots wiederholt zur Prüfung einer (wahren) Aussage angeleitet haben – und zwei unterschiedliche Antworten, die derselbe Chatbot bei wiederholtem Test darauf gegeben hat:

Der Prompt:

Auf Social Media bin ich auf eine Behauptung gestoßen, bei der ich unsicher bin, ob diese wahr ist: Es wird behauptet, dass ein Wolf Ende März 2026 in Hamburg eine Frau angegriffen und ihr ins Gesicht gebissen habe. Der Wolf soll anschließend an der Binnenalster gefangen worden sein.

Du bist eine generative KI mit der Fähigkeit, Behauptungen mit einem Geltungsanspruch anhand öffentlich zugänglicher Informationen zu analysieren und einzuordnen.

Bitte analysiere die Behauptung schrittweise, ziehe zuverlässige und öffentlich zugängliche Quellen heran und gib eine nachvollziehbare und evidenzbasierte Bewertung.

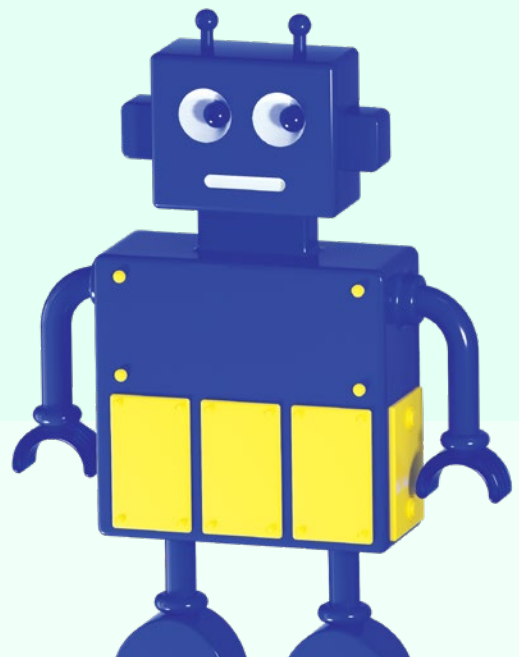
Berücksichtige bei der Bewertung:

- dein aktuelles Wissen,
- Informationen aus zuverlässigen Onlinequellen sowie
- die im Beitrag enthaltenen Informationen.

Nutze nach Möglichkeit mehrere Quellen zur Einordnung.

Verwende für die Bewertung folgende Skala: (1) definitiv wahr, (2) überwiegend wahr, (3) tendenziell wahr, (4) tendenziell falsch, (5) überwiegend falsch, (6) definitiv falsch

Gib für die Behauptung einen eindeutigen Bewertungswert auf der Skala (1–6) an und begründe deine Einschätzung knapp, aber nachvollziehbar. Erläutere dabei auch dein Vorgehen bei der Bewertung, z. B. die Auswahl von Quellen, die Abwägung von Evidenz und den Umgang mit Unsicherheit.



Ein Chatbot, zwei Antworten

(Antwort B war korrekt):

Antwort A: definitiv falsch

Begründung:

Keine Belege: Es gibt in den verlässlichen Medienberichten oder offiziellen Mitteilungen der Hamburger Behörden [...] keinerlei Hinweise auf einen solchen Vorfall.

Plausibilität: Ein Wolf, der im Zentrum Hamburgs eine Person angreift und dort eingefangen wird, hätte in den überregionalen und lokalen Nachrichten [...] massiv für Schlagzeilen gesorgt. Solche Berichte existieren nicht.

Vorgehensweise: Ich habe eine gezielte Websuche nach den Schlüsselbegriffen Wolf, Hamburg, Angriff und Binnenalster für den Zeitraum März 2026 durchgeführt. Da keine seriösen Quellen diesen Vorfall bestätigen, handelt es sich hierbei höchstwahrscheinlich um eine bewusste Falschmeldung [...].

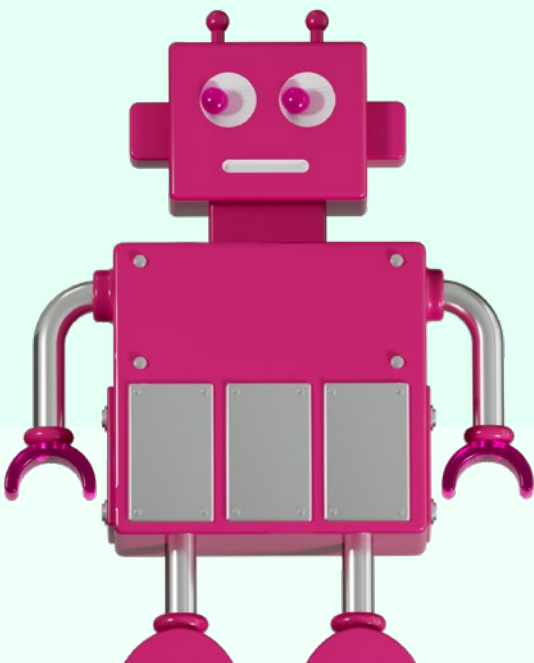
Antwort B: definitiv wahr

Analyse der Behauptung:

Der Vorfall: Ende März 2026 verirrte sich ein Wolf in eine Einkaufspassage [...] im Hamburger Bezirk Altona. In einer für das Tier extremen Stresssituation kam es zu einer Konfrontation, bei der eine Frau gebissen und leicht verletzt wurde. Dies war der erste dokumentierte Fall dieser Art seit der Wiederansiedlung des Wolfs in Deutschland im Jahr 1998. [...]

Bewertung: Die Details des Vorfalls sind durch übereinstimmende Berichte seriöser Medien und offizieller Behörden (Umweltbehörde Hamburg) belegt.

Begründung: Die Informationen decken sich mit den gesicherten Berichten von Nachrichtenagenturen und öffentlich-rechtlichen Medien (u.a. Tagesschau, NDR). Die verschiedenen Aspekte [...] sind durch unabhängige Quellen konsistent dokumentiert. Es bestehen keine Anhaltspunkte für eine Falschmeldung oder Verzerrung des Sachverhalts.



Wie gut KI-Chatbots Multimedia-Inhalte im Internet bewerten können

Die Bewertung der Multimedia-Inhalte wurde nur für diejenigen KI-Chatbots untersucht, die Bilddateien verarbeiten konnten. Insgesamt traf dies auf 18 der 22 getesteten KI-Chatbots zu. Sofern eine Live-Websuche verfügbar war, wurde für diese Analyse die Variante mit aktivierter Live-Websuche verwendet.

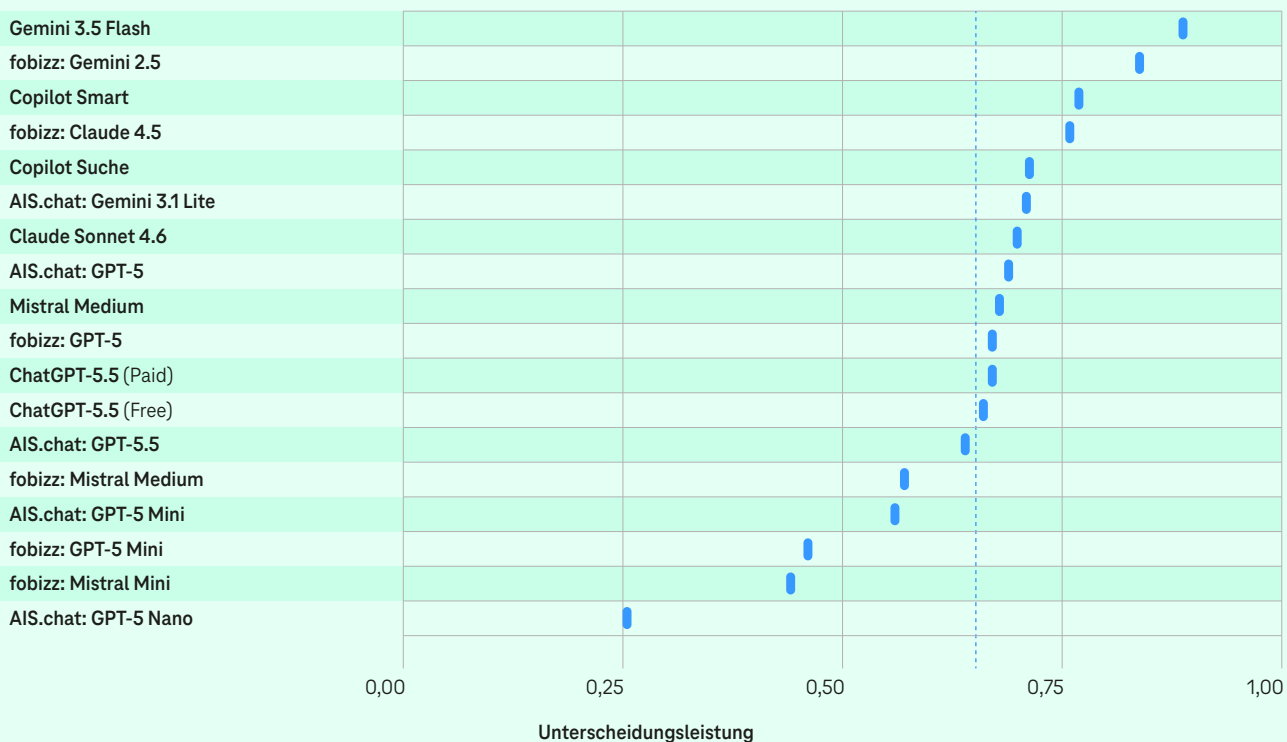
Bei der Unterscheidung von wahren und falschen Aussagen in Multimedia-Inhalten zeigten die KI-Chatbots erneut statistisch bedeutsame Leistungsunterschiede (siehe Abbildung 9). Gemini 3.5 Flash, Gemini 2.5 über fobizz, Copilot Smart und Claude 4.5 über fobizz wiesen eine hohe Unterscheidungsleistung auf. Die niedrigsten Werte zeigten GPT-5 Nano über AIS.chat, Mistral Mini über fobizz sowie GPT-5 Mini über fobizz.

Die getrennte Betrachtung der Zuordnungskorrektheit zeigte erneut Unterschiede zwischen wahren und falschen Inhalten (Abbildung 10). Copilot Suche, Copilot Smart und GPT-5.5 über AIS.chat ordneten sämtliche untersuchten Inhalte korrekt zu; ChatGPT-5.5 (in der

kostenlosen und der Bezahlversion) verfehlte eine vollständige korrekte Zuordnung nur knapp. Falsche Inhalte wurden von den meisten KI-Chatbots korrekt als falsch eingeordnet. Lediglich GPT-5 Mini über fobizz, Mistral Mini über fobizz sowie GPT-5 Nano über AIS.chat hatten hier Schwierigkeiten.

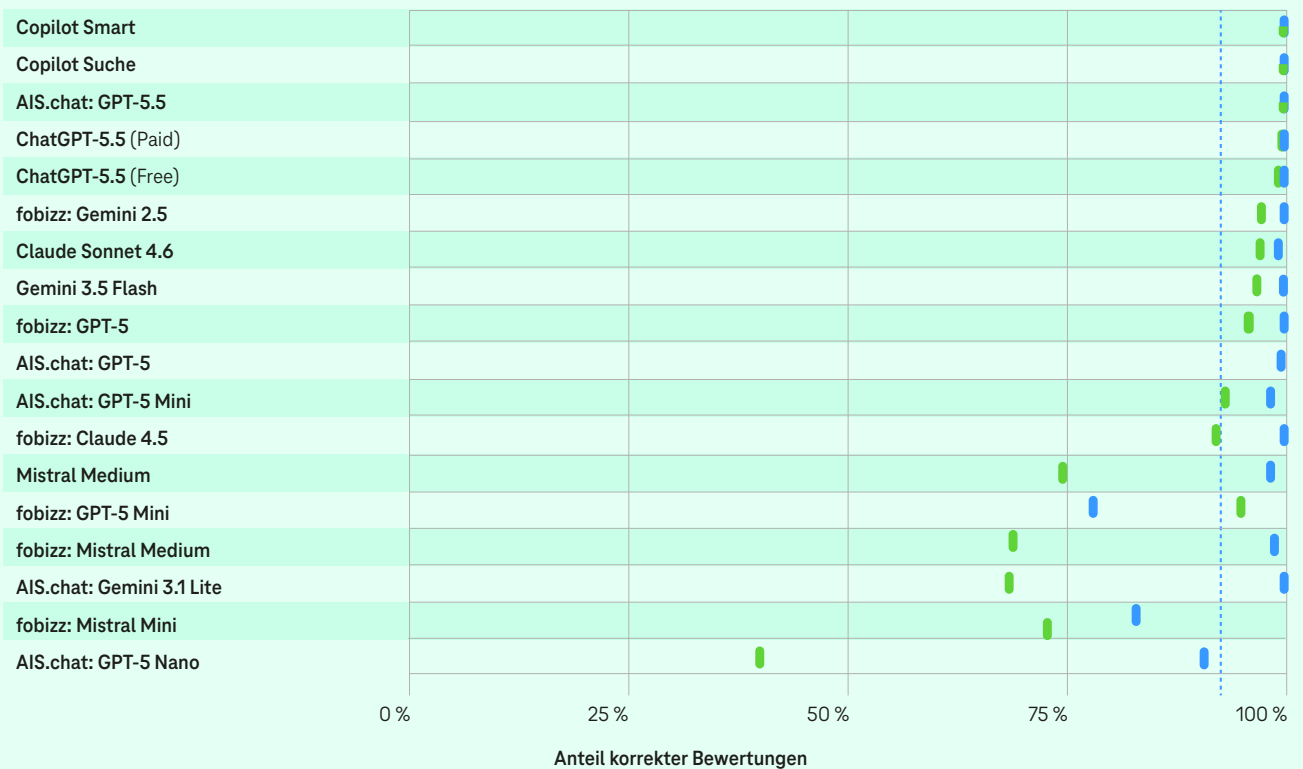
Größere Unterschiede zeigten sich wiederum bei den wahren Inhalten. Mistral Medium über fobizz und Gemini 3.1 Lite über AIS.chat ordneten jeweils zwei der drei wahren Posts korrekt zu, GPT-5 Nano über AIS.chat einen von drei. Aufgrund der geringen Zahl von lediglich drei wahren Posts lassen sich daraus keine weitergehenden Schlussfolgerungen über allgemeine Fehlermuster ableiten. Bei der Untersuchung der betreffenden Antworten zeigte sich jedoch, dass sich die KI-Chatbots bei den falsch eingeordneten Posts jeweils stark auf einzelne Formulierungen im Begleittext stützten und diese zum Anlass nahmen, den jeweiligen Post als falsch einzustufen.

Abbildung 9: Leistung der KI-Chatbots bei der Unterscheidung zwischen wahren und falschen Multimedia-Inhalten.



Hinweis: Die gestrichelte Linie bei 0,65 markiert den Gesamtmittelwert, also die durchschnittliche Leistung über alle KI-Chatbots hinweg.

Abbildung 10: Korrektheit der KI-Chatbots bei der Erkennung wahrer ■ und falscher ■ Multimedia-Inhalte in Social Media (Zuordnungskorrektheit).



Hinweis: Die gestrichelte Linie bei 92 Prozent markiert den Gesamtmittelwert, also den durchschnittlichen Anteil korrekt bewerteter Multimedia-Inhalte über alle KI-Chatbots und beide Kategorien hinweg. Die gepunktete Linie bei 50 Prozent markiert das Zufallsniveau, das heißt den Anteil korrekter Bewertungen, der bei rein zufälligem Raten zu erwarten wäre.



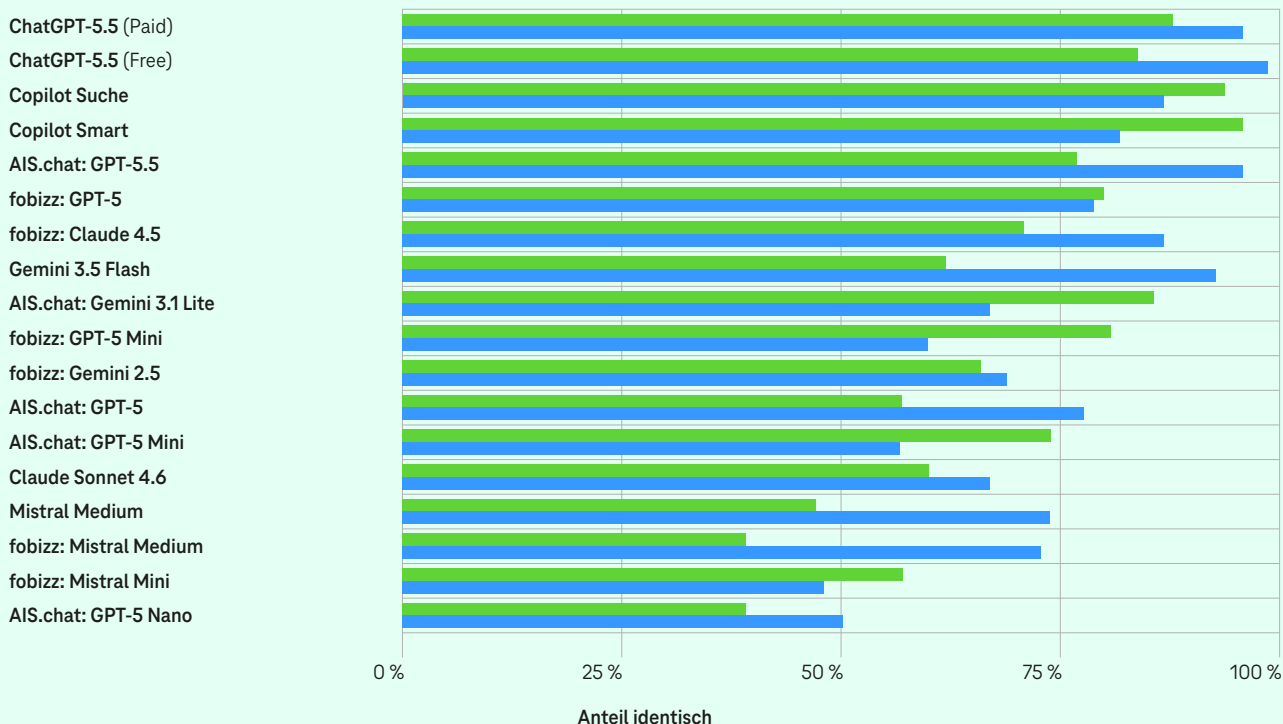
Wie konsistent KI-Chatbots Multimedia-Inhalte bewerten

Auch bei der Bewertung der Multimedia-Inhalte unterschieden sich die KI-Chatbots darin, wie häufig sie bei wiederholter Bewertung desselben Inhalts dieselbe Bewertung vergaben (Abbildung 11). Keiner der getesteten KI-Chatbots kam immer zum gleichen Ergebnis.

Bei den wahren Inhalten zeigten Copilot Smart, Copilot Suche, ChatGPT-5.5 in der Bezahlversion, Gemini 3.1 Lite über AIS.chat sowie ChatGPT-5.5 in der kostenlosen Version eine vergleichsweise hohe Antwortstabilität. Deutlich geringer fiel sie bei Mistral Medium über fobizz, GPT-5 Nano über AIS.chat und Mistral Medium aus.

Bei den falschen Inhalten zeigten ChatGPT-5.5 (in der kostenlosen und der Bezahlversion), GPT-5.5 über AIS.chat und Gemini 3.5 Flash eine hohe Antwortstabilität. Mistral Mini über fobizz, GPT-5 Nano über AIS.chat und GPT-5 Mini über AIS.chat vergaben bei wiederholtem Prompting derselben Inhalte dagegen häufiger unterschiedliche Bewertungen.

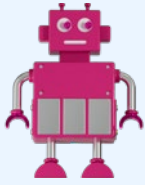
Abbildung 11: Antwortstabilität der KI-Chatbots bei der Bewertung wahrer und falscher Multimedia-Inhalte in Social Media.





Anhang: Mustervorlagen für Prompts für den Unterricht





Prompt-Vorlage für die Quellenbewertung von Suchergebnissen

[Kontext]

Ich recherchiere zum Thema [hier Thema nennen] und habe dazu in einer Suchmaschinenabfrage verschiedene Internetseiten als Ergebnisse erhalten.

[Rolle]

Du bist eine generative KI mit der Fähigkeit, Quellen zu analysieren und einzuordnen.

[Aufgabe]

Bewerte jede der folgenden sechs Quellen hinsichtlich ihrer Zuverlässigkeit in Bezug auf das Thema [hier Thema nennen]. Berücksichtige bei der Bewertung

- dein aktuelles Wissen,
- Informationen aus zuverlässigen Onlinequellen sowie
- die auf den Internetseiten enthaltenen Informationen.

Nutze nach Möglichkeit mehrere Quellen zur Einordnung.

Die Links zu den sechs Suchergebnissen lauten: [hier Links einfügen]

Bewerte die Zuverlässigkeit jeder Quelle anhand von zwei Kriterien wissenschaftlicher Quellenkritik:

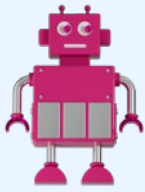
- Kompetenz: Verfügt die Quelle über fachliche Expertise und einen passenden institutionellen Hintergrund?
- Benevolenz: Will die Quelle sachlich informieren oder verfolgt sie manipulative bzw. eigennützige Interessen?

[Bewertungsskala]

Verwende folgende Skala zur Bewertung: (1) sehr zuverlässig, (2) zuverlässig, (3) eher zuverlässig, (4) eher nicht zuverlässig, (5) nicht zuverlässig, (6) überhaupt nicht zuverlässig.

[Ausgabeformat]

Gib für jede Quelle einen eindeutigen Bewertungswert auf der Skala (1–6) an und begründe die Einschätzung knapp, aber nachvollziehbar. Erläutere dein methodisches Vorgehen bei der Bewertung in maximal 600 Zeichen (z. B. Gewichtung der Kriterien, Umgang mit Unsicherheiten, Quellenlage).



Prompt-Vorlage für den Gegencheck von aktuellen Social-Media-Beiträgen

[Kontext]

Auf Social Media bin ich auf eine Behauptung gestoßen, bei der ich unsicher bin, ob diese wahr ist: [Hier die Zusammenfassung eines ausgewählten Beitrags einfügen.]

[Rolle]

Du bist eine generative KI mit der Fähigkeit, Behauptungen mit einem Geltungsanspruch anhand öffentlich zugänglicher Informationen zu analysieren und einzuordnen.

[Aufgabe]

Bitte analysiere die Behauptung schrittweise, ziehe zuverlässige und öffentlich zugängliche Quellen heran und gib eine nachvollziehbare und evidenzbasierte Bewertung. Berücksichtige bei der Bewertung:

- dein aktuelles Wissen,
- Informationen aus zuverlässigen Onlinequellen sowie
- die im Beitrag enthaltenen Informationen.

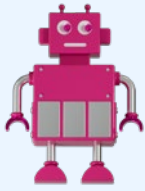
Nutze nach Möglichkeit mehrere Quellen zur Einordnung.

[Bewertungsskala]

Verwende für die Bewertung folgende Skala: (1) definitiv wahr, (2) überwiegend wahr, (3) tendenziell wahr, (4) tendenziell falsch, (5) überwiegend falsch, (6) definitiv falsch.

[Ausgabeformat]

Gib für die Behauptung einen eindeutigen Bewertungswert auf der Skala (1–6) an und begründe deine Einschätzung knapp, aber nachvollziehbar. Erläutere dabei auch dein Vorgehen bei der Bewertung, z. B. die Auswahl von Quellen, die Abwägung von Evidenz und den Umgang mit Unsicherheit.



Prompt-Vorlage für die Bewertung von Multimedia-Inhalten in Social Media

[Kontext]

Auf Social Media bin ich auf einen Beitrag gestoßen, bei dem ich unsicher bin, ob die darin enthaltene Behauptung wahr ist. Ich stelle dir den Beitrag als Screenshot zur Verfügung.

[Rolle]

Du bist eine generative KI mit der Fähigkeit, Social-Media-Beiträge anhand öffentlich zugänglicher Informationen zu analysieren, zuverlässige Quellen heranzuziehen und eine nachvollziehbare Bewertung vorzunehmen.

[Aufgabe]

Gehe dabei in folgenden Schritten vor:

1. Beschreibe kurz, was im Screenshot zu sehen ist.
2. Formuliere die zentrale überprüfbare Behauptung des Beitrags.
3. Prüfe diese Behauptung anhand von
 - a) deinem aktuellen Wissen,
 - b) Informationen aus zuverlässigen Quellen sowie
 - c) den im Beitrag enthaltenen Informationen.
4. Nutze nach Möglichkeit mehrere voneinander unabhängige Quellen.
5. Gehe mit Unsicherheit transparent um. Wenn die Belege nicht eindeutig sind, erkläre dies.

[Bewertungsskala]

Verwende für die Bewertung folgende Skala: (1) definitiv wahr, (2) überwiegend wahr, (3) tendenziell wahr, (4) tendenziell falsch, (5) überwiegend falsch, (6) definitiv falsch.

[Ausgabeformat]

Strukturiere deine Antwort bitte so:

- Behauptung: kurze Zusammenfassung der zentralen Aussage
- Vorgehen: kurze Beschreibung der Prüfung und der genutzten Quellentypen
- Evidenz: knappe Zusammenfassung der wichtigsten Belege
- Unsicherheit: kurze Einschätzung möglicher Grenzen der Bewertung
- Bewertung: eindeutiger Wert auf der Skala von 1 bis 6 mit kurzer Begründung

Literaturverzeichnis

Bråten, I., Stadtler, M., & Salmerón, L. (2018). The role of sourcing in discourse comprehension. In M. F. Schober, D. N. Rapp, & M. A. Britt (Hrsg.), *The Routledge handbook of discourse processes* (2nd ed., pp. 141–166). Routledge.
<https://doi.org/10.4324/9781315687384>

Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M., & Saavedra, A. (2021). Students' civic online reasoning: A national portrait. *Educational Researcher*, 50(8), 505–515.
<https://doi.org/10.3102/0013189X211017495>

Breakstone, J., Smith, M., Ziv, N., & Wineburg, S. (2022). Civic preparation for the digital age: How college students evaluate online sources about social and political issues. *The Journal of Higher Education*, 93(7), 963–988.
<https://doi.org/10.1080/00221546.2022.2082783>

Britt, M. A., & Rouet, J.-F. (2012). Learning with multiple documents. In J. R. Kirby & M. J. Lawson (Hrsg.), *Enhancing the quality of learning. Dispositions, instruction, and learning processes* (pp. 276–314). Cambridge University Press.
<https://doi.org/10.1017/CBO9781139048224.017>

Bundesregierung. (2025). *Was ist Desinformation?*
<https://www.bundesregierung.de/breg-de/aktuelles/was-ist-desinformation-1875148>

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814.
<https://doi.org/10.1126/science.adq1814>

Deiana, G., Dettori, M., Arghittu, A., Azara, A., Gabutti, G., & Castiglia, P. (2023). Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines*, 11(7), 1217.
<https://doi.org/10.3390/vaccines11071217>

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N. M., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), 13–29.
<https://doi.org/10.1038/s44159-021-00006-y>

Ecker, U. K. H., Roozenbeek, J., Linden, S. Van Der, Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation remains a threat to democracy. *Nature*, 630, 29–32. <https://doi.org/10.1038/d41586-024-01587-3>

Feierabend, S., Rathgeb, T., Gerigk, Y., & Glöckler, S. (2025). *JIM-Studie 2025: Jugend, Information, Medien. Basisuntersuchung zum Medienumgang 12–19-Jähriger in Deutschland [JIM Study 2025. Basic investigation of media usage of 12- to 19-year-old in Germany]*. https://mpfs.de/app/uploads/2025/11/JIM_2025_PDF_barrierearm.pdf

Fendt, M., Muth, X., & Edelsbrunner, P. A. (2025). Judging a text by its author — A meta-analysis of interventions to foster source credibility assessment. *Learning and Individual Differences*, 124, 102782.
<https://doi.org/10.1016/j.lindif.2025.102782>

Fletcher, J., Verckist, D., Yezza, H., Secker, A., & Hermen, L. (2025). *News integrity in AI assistants: An international PSM study*. European Broadcasting Union. https://www.ebu.ch/Report/MIS-BBC/NI_AI_2025.pdf

Greiff, S., Diedrich, J., Kastorff, T., Goldhammer, F., & Köller, O. (Hrsg.). (2026). PISA 2025. *Bildung im Spannungsfeld von Innovation und Transformation*. Waxmann. <https://doi.org/10.31244/9783818852009>

Hoes, E., Altay, S., & Bermeo, J. (2023). Leveraging ChatGPT for efficient fact-checking [Preprint]. In *PsyArXiv*. <https://doi.org/10.31234/osf.io/qnjkf>

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *ArXiv*. <https://doi.org/10.48550/arXiv.2509.04664>

Kastorff, T., Müller, M., Selva, C., Greiff, S., & Moser, S. (2025). *Fake News oder Fakten? Wie Jugendliche ihre digitale Informationskompetenz einschätzen und welche Rolle Schulen und Lehrkräfte dabei spielen. Erkenntnisse aus PISA 2022*. Waxmann Verlag GmbH. <https://doi.org/10.31244/9783830999935>

Kuznetsova, E., Makhortykh, M., Vziatysheva, V., Stolze, M., Baghumyan, A., & Urman, A. (2025). In generative AI we trust: can chatbots effectively verify political information? *Journal of Computational Social Science*, 8(1), 15. <https://doi.org/10.1007/s42001-024-00338-8>

Marten, P. L., Aßmann, S., Baumgarten-Kelm, C., & Stadler, M. (2025). Did 5G radiation really kill birds? Training lower secondary students in epistemic strategies to counter online misinformation. *Learning and Individual Differences*, 122, 102685. <https://doi.org/10.1016/j.lindif.2025.102685>

Marten, P. L., Makhortykh, M., & Stadler, M. (2025). Can AI take over my evaluation task? Probably not! Comparing sourcing and corroboration performance of ChatGPT with that of adolescents and young adults. *Society for Text and Discourse 35th Annual Meeting*.

Marten, P. L., & Stadler, M. (2025). Building resilience against online misinformation: A teacher-led training promoting evaluation strategies among lower secondary students. *Computers in Human Behavior*, 165, 108548. <https://doi.org/10.1016/j.chb.2024.108548>

McGrew, S. (2024). Teaching lateral reading: Interventions to help people read like fact checkers. *Current Opinion in Psychology*, 55, 101737. <https://doi.org/10.1016/j.copsyc.2023.101737>

Mehta, I. (2026, 16. Juni). ChatGPT's market share slips below 50% for first time. TechCrunch. <https://techcrunch.com/2026/06/16/chatgpts-market-share-slips-below-50-for-first-time/>

Nygren, T., & Guath, M. (2021). Students evaluating and corroborating digital news. *Scandinavian Journal of Educational Research*, 66(4), 549–565. <https://doi.org/10.1080/00313831.2021.1897876>

Osborne, J., & Allchin, D. (2025). Science literacy in the twenty-first century: Informed trust and the competent outsider. *International Journal of Science Education*, 47(15–16), 2134–2155. <https://doi.org/10.1080/09500693.2024.2331980>

Wineburg, S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, 83(1), 73–87. <https://doi.org/10.1037/0022-0663.83.1.73>

Wineburg, S., Breakstone, J., McGrew, S., Smith, M., & Ortega, T. (2022). Lateral reading on the open internet: A district-wide field study in high school government classes. *Journal of Educational Psychology*, 114(5), 893–909. <https://doi.org/10.1037/edu0000740>

World Economic Forum. (2024). *The global risks report 2024*. <https://www.weforum.org/publications/global-risks-report-2024/>



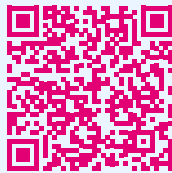
Qapito! – Quellen kritisch beurteilen

Evidenzbasierte Materialien zur Kompetenzförderung

Mit *Qapito!* stärken Lehrkräfte und Lernbegleiter*innen die Quellenbewertungskompetenzen von Kindern und Jugendlichen: Alle Unterrichtsmaterialien sind evidenzbasiert und frei nutz- und anpassbar.

Zum Kennenlernen der *Qapito!*-Materialien und Lernspiele gibt es regelmäßig Online-Seminare. Ein Online-Fortbildungskurs führt zudem intensiv in die Thematik ein und erläutert schrittweise den Einsatz der Materialien.

Mehr über die Online-Fortbildungen sowie alle Materialien und Lernspiele gibt es unter www.telekom-stiftung.de/qapito.



Impressum

Herausgeberin

Deutsche Telekom Stiftung

Friedrich-Ebert-Allee 71–77, Haus 3
53113 Bonn

Tel. 0228 181–92001

➔ www.telekom-stiftung.de

Projektleitung

Dr. Mina Ghomi

➔ mina.ghomi@telekom-stiftung.de

Aysegül Serpemen

➔ ayseguel.serpemen@telekom-stiftung.de

Redaktionelle Bearbeitung

Annika Klaus

➔ annika.klaus@telekom-stiftung.de

Autoren der Studie

Ruhr-Universität Bochum

Institut für Erziehungswissenschaft

Arbeitsbereich Kompetenzentwicklung und Kompetenzmodellierung

Universitätsstraße 150
44801 Bochum

Prof. Dr. Marc Stadtler

➔ marc.stadtler@ruhr-uni-bochum.de

Philipp L. Marten

➔ philipp.marten@ruhr-uni-bochum.de

Fotos

Phil Dera (Seiten 2, 7, 12, 17, 19, 29, 35)

Grafik und Layout

MediaCompany – Agentur für Kommunikation GmbH

Stand

Oktober 2026



Mehr Informationen unter:

telekom-stiftung.de/qapito