zisch: zeitschrift für interdisziplinäre schreibforschung

Wenn das Geschlecht (k)eine Rolle spielt:
Eine Analyse möglicher Verzerrungen in KI-basiertem Schreibfeedback

Julia Kraus (TU Wien), Ognjen Zivkovic (Universität Wien), Dominic Mayer (Universität Wien)

Abstract:
Die zunehmende Integration generativer KI-Systeme in akademische Schreibprozesse wirft grundlegende Fragen nach Fairness, Objektivität und Chancengleichheit auf. Während zahlreiche sozialpsychologische Studien geschlechtsspezifische Bewertungsverzerrungen in menschlichen Beurteilungsprozessen nachweisen, ist bislang unzureichend geklärt, ob und in welchem Ausmaß vergleichbare Effekte auch in KI-basiertem Schreibfeedback auftreten. Die vorliegende Studie untersucht in einem kontrollierten experimentellen Design (n = 1000), ob das Sprachmodell ChatGPT bei der Bewertung identischer studentischer Texte systematische Unterschiede in Abhängigkeit vom binär markierten Geschlecht der vermeintlichen Verfasser*innen aufweist.
Die Ergebnisse zeigen keinen signifikanten Mittelwertunterschied zwischen männlich und weiblich zugeschriebenen Profilen (Welch-t-Test: $p = .083$; Cohen's $d \approx 0.11$). Auch nichtparametrische Tests bestätigen diese Befunde. Allerdings treten seltene negative Extrembewertungen ausschließlich bei männlich kodierten Profilen auf (Fisher-Test: $p = .031$).
Die Befunde sprechen gegen einen starken, globalen Gender Bias im untersuchten Setting, weisen jedoch auf mögliche subtile Verzerrungen in Randbereichen hin. Für die Schreibdidaktik impliziert dies, dass KI-Feedback weder als vollständig neutral noch als per se diskriminierend verstanden werden sollte, sondern als sozio-technisches System, dessen Wirkungen kontextabhängig reflektiert werden müssen.

Keywords: KI-Bias; Gender Bias; Experimentelle Schreibforschung; KI-basiertes Schreibfeedback

Empfohlene Zitierweise:
Kraus, J., Mayer, D., & Zivkovic, O. (2026): Wenn das Geschlecht (k)eine Rolle spielt: Eine Analyse möglicher Verzerrungen in KI-basiertem Schreibfeedback. zisch: zeitschrift für interdisziplinäre schreibforschung, 14, 102-117. DOI: https://doi.org/10.48646/zisch.261416

ISSN: 2709-3778

Wenn das Geschlecht (k)eine Rolle spielt: Eine Analyse möglicher Verzerrungen in KI-basiertem Schreibfeedback

Julia Kraus (TU Wien), Ognjen Zivkovic (Universität Wien), Dominic Mayer (Universität Wien)
Betreuer*in: Katrin Miglar (Universität Wien)

Einleitung

Mit der rasanten Verbreitung generativer KI-Systeme verändert sich die Praxis wissenschaftlichen Schreibens fundamental. Sprachmodelle wie ChatGPT übernehmen zunehmend Funktionen, die traditionell Lehrenden oder Peer-Reviewer*innen vorbehalten waren: Strukturierungshilfe, sprachliche Überarbeitung, argumentative Verdichtung – und nicht zuletzt die Vergabe von Feedback. In dieser Verschiebung liegt eine zentrale Herausforderung für Hochschulen: Wenn Bewertung und Rückmeldung algorithmisch erfolgen, stellt sich die Frage nach deren Fairness in neuer Schärfe.

Bias wird in der Forschung als systematische Verzerrung verstanden, die nicht auf Leistungsunterschiede zurückzuführen ist, sondern auf soziale Kategorien wie z.B. das Geschlecht oder Herkunft (Friedman & Nissenbaum, 1996; Mehrabi et al., 2021). Sozialpsychologische Experimente belegen seit Jahrzehnten, dass selbst identische Leistungen unterschiedlich bewertet werden, wenn sie mit unterschiedlichen sozialen Zuschreibungen assoziert werden. Moss-Racusin et al. (2012) zeigen etwa, dass identische Bewerbungen mit männlichem Namen als kompetenter eingestuft werden als solche mit weiblichem Namen. Bertrand und Mullainathan (2004) kommen zu vergleichbaren Ergebnissen im Arbeitsmarktkontext in Hinblick auf Chancengleichheit bei Bewerbungen.

Neuere Forschung differenziert diese Effekte weiter aus: Bewertungsverzerrungen manifestieren sich nicht ausschließlich in expliziter Benachteiligung, sondern auch in subtilen Feedbackunterschieden, etwa in Form übertrieben positiven Lobs oder unterschiedlicher Kritikintensität (Jampol & Zayas, 2020; Wolfe, 2024). Bewertungspraktiken sind somit tief in implizite soziale Kognitionen eingebettet (Greenwald & Banaji, 1995, 27).

Vor diesem Hintergrund erscheint es plausibel, dass auch KI-Systeme solche Muster reproduzieren könnten. Große Sprachmodelle werden auf umfangreichen Textkorpora trainiert, die gesellschaftliche Stereotype enthalten (Bender et al., 2021). Wenn diese Modelle statistische Regularitäten aus historischen Daten lernen, liegt die Vermutung nahe, dass sie implizite Bewertungslogiken mitübernehmen.

Gleichzeitig unterscheiden sich KI-Systeme grundlegend von menschlichen Akteur*innen: Sie verfügen weder über intentionale Vorurteile noch über bewusste Diskriminierungsabsichten. Ihre „Urteile“ entstehen aus Wahrscheinlichkeitsverteilungen sprachlicher Muster. Bias in KI ist daher nicht psychologisch, sondern statistisch zu verstehen.

Die vorliegende Studie untersucht, ob ChatGPT bei der Bewertung identischer studentischer Texte geschlechtsspezifische Unterschiede aufweist. Das Design orientiert sich an klassischen Audit-Studien (Moss-Racusin et al., 2012), überträgt diese jedoch in einen algorithmischen Kontext. Ziel ist

es, empirisch zu prüfen, ob generatives KI-Feedback geschlechtsbezogene Bewertungsverzerrungen reproduziert – und welche Implikationen sich daraus für eine schreibdidaktische Verantwortungskultur ergeben.

Ergänzend zu diesen normativ-didaktischen Positionierungen hat sich in den vergangenen Jahren eine zunehmend differenzierte Auseinandersetzung mit dem konkreten Einsatz generativer KI im wissenschaftlichen Schreiben etabliert. Praxisorientierte Arbeiten wie Buck (2026) und Lang (2025) systematisieren Einsatzmöglichkeiten von KI entlang des gesamten Schreibprozesses und betonen dabei die Notwendigkeit einer reflektierten, verantwortungsvollen Integration. Schreibdidaktische Beiträge (Schmidt & Seegel, 2024) sowie hochschuldidaktische Analysen (Schulze Heuling et al., 2025) verweisen darüber hinaus auf strukturelle Veränderungen von Feedback-, Bewertungs- und Prüfungskulturen. Erste empirische Studien untersuchten bereits, wie KI den Schreibprozess verändert (Hickisch, 2024) oder ob sich modellinterne Gender-Bias-Signale in studentischen Texten niederschlagen (Wambsganss et al., 2023). Während geschlechtsspezifische Verzerrungen in großen Sprachmodellen in unterschiedlichen Kontexten dokumentiert sind (Döll et al., 2024; Kotek et al., 2023), bleibt bislang weitgehend offen, ob sich solche Effekte auch in der konkreten Bewertung wissenschaftlicher Texte manifestieren. Genau an dieser Schnittstelle zwischen Schreibdidaktik, KI-Anwendung und algorithmischer Fairness setzt die vorliegende Studie an.

Theoretischer Rahmen

Künstliche Intelligenz ist längst nicht mehr nur ein technisches Werkzeug, sondern ein sozio-technisches Phänomen, das tief in gesellschaftliche Strukturen eingreift. Damit KI-Systeme wie ChatGPT oder andere generative Modelle in der Lage sind, Inhalte zu erzeugen und Entscheidungen zu treffen, greifen sie auf große Mengen menschlich erzeugter Daten zurück und damit auch auf die darin enthaltenen sozialen Muster, Bewertungen und Vorurteile. Der folgende Theorieteil bildet den wissenschaftlichen Rahmen, um zu erklären, wie Verzerrungen (Bias) in der KI entstehen. Das Kapitel beleuchtet zuerst die technischen Grundlagen generativer Modelle und deren epistemische Grenzen, anschließend psychologische Mechanismen impliziter Stereotypisierung und letztlich die Verbindung beider Ebenen, bzw. die Frage, wie soziale Vorurteile durch technische Systeme reproduziert werden können.

Technische Grundlagen

Einen guten Überblick, was man unter einer sogenannten generativen KI versteht, bieten Feuerriegel et al. (2023). Diese beschreiben generative KI als einen Teilbereich der künstlichen Intelligenz, welcher auf generativen Modellen basiert und in der Lage ist aus Trainingsdaten neue sinnvolle Inhalte wie Texte, Bilder oder Audios zu erzeugen, die häufig nicht mehr von menschlich erstellten Inhalten unterscheidbar sind (Feuerriegel et al., 2023, 111). Konkret gesagt ist somit eine generative KI, im Gegensatz zur analytischen KI, in der Lage neue Datenobjekte zu produzieren.

Im praktischen Einsatz zeigen sich jedoch bei generativen KI-Systemen trotz ihrer hohen Leistungsfähigkeit erhebliche Herausforderungen. Insbesondere große Sprachmodelle, die auf

umfangreichen Textkorpora aus dem Internet trainiert werden, erzeugen zwar sprachlich kohärente Inhalte, handeln jedoch nicht zwangsläufig im Sinne der Nutzerintention. Häufige Problembereiche sind das Erfinden nicht vorhandener Fakten, das unzuverlässige Befolgen von Anweisungen sowie die Generierung potenziell schädlicher oder verzerrter Inhalte. Ouyang et al. (2022) weisen darauf hin, dass eine bloße Vergrößerung von Sprachmodellen diese Defizite nicht automatisch behebt, da größere Modelle nicht inhärent besser darin sind, die tatsächlichen Absichten der Nutzenden zu verstehen oder korrekt umzusetzen.

Als zentrale Ursache identifizieren die Autor*innen eine strukturelle Fehlanpassung zwischen Trainingsziel und Anwendungszweck. Klassische generative Sprachmodelle werden primär darauf trainiert, das jeweils wahrscheinlichste nächste Token in einem Text vorherzusagen. Dieses statistische Optimierungsziel unterscheidet sich jedoch grundlegend von dem Anspruch, Benutzeranfragen hilfreich, wahrheitsgemäß und sicher zu beantworten. Während das Modell, GPT-3, auf sprachliche Plausibilität optimiert ist, bleibt die Ausrichtung auf menschliche Erwartungen unzureichend, was in der Forschung als Misalignment bezeichnet wird (Ouyang et al., 2022, 2). Um diese Problematik zu adressieren, stellen Ouyang et al. (2022) mit Reinforcement Learning from Human Feedback (RLHF) einen Ansatz vor, welcher generative KI gezielt an menschliche Präferenzen anpasst. Dabei wird das vortrainierte Sprachmodell zunächst durch überwachtes Lernen anhand von Menschen erstellten Beispielantworten angepasst. Anschließend bewerten menschliche Annotatoren mehrere vom Modell erzeugte Antwortvarianten, wodurch Präferenzdaten entstehen. Auf dieser Grundlage wird ein sogenanntes Reward Model trainiert, das menschliche Bewertungen approximiert und im letzten Schritt als Optimierungsziel für ein bestärkendes Lernverfahren dient (2-3). Petroni et al. (2019) berichten ähnliches zu dieser Thematik. Diese sprechen empirische Untersuchungen zu sogenannten Language-Model-Probes an, im Rahmen des LAMA-Benchmarks, welche zeigen, dass Sprachmodelle relationales Wissen in erheblichem Umfang internalisieren, dieses jedoch selektiv, kontextabhängig und ungleichmäßig abrufen. Epistemisch handelt es sich folglich nicht um gesichertes Wissen oder Wahrheitsproduktion, sondern um eine stochastische Rekombination gespeicherter Regularitäten. Diese Einsicht stützt die kritische Charakterisierung großer Sprachmodelle als „stochastic parrots“, welche sprachliche Muster reproduzieren können, ohne deren semantische oder normative Bedeutung zu verstehen.

Das Ergebnis dieses mehrstufigen Trainingsprozesses ist InstructGPT, eine speziell auf Instruktionsbefolgung ausgerichtete Variante des GPT-3-Modells. Die empirischen Ergebnisse zeigen, dass diese Form der Feinabstimmung erhebliche qualitative Verbesserungen bewirkt. In menschlichen Evaluationen wurden die Antworten eines InstructGPT-Modells mit lediglich 1,3 Milliarden Parametern häufiger bevorzugt als jene des ursprünglichen GPT-3-Modells mit 175 Milliarden Parametern (Ouyang et al., 2022, 1). Dies verdeutlicht, dass die Leistungsfähigkeit generativer KI nicht ausschließlich von der Modellgröße abhängt, sondern maßgeblich von der zielgerichteten Ausrichtung auf menschliche Nutzungskontexte beeinflusst wird. Darüber hinaus zeigt InstructGPT eine erhöhte Zuverlässigkeit bei der Befolgung expliziter Anweisungen sowie eine signifikante Reduktion sogenannter Halluzinationen, also der Erzeugung inhaltlich falscher oder nicht durch die Eingabe gedeckter Informationen.

Insbesondere bei Aufgaben mit klar vorgegebenen Antwortmöglichkeiten oder Bewertungskriterien halbiert sich die Häufigkeit erfundener Inhalte im Vergleich zum Basismodell nahezu (ebd.: 3).

Zur konzeptionellen Einordnung definieren die Autor*innen drei zentrale Qualitätskriterien für generative KI-Systeme: Sie sollen hilfreich (helpful), wahrheitsgemäß beziehungsweise ehrlich (honest) und unschädlich (harmless) agieren (Ouyang et al., 2022, 10). Diese Kriterien verdeutlichen, dass technische Leistungsfähigkeit allein nicht ausreicht, sondern durch normative Zielsetzungen ergänzt werden muss, um generative KI verantwortungsvoll einsetzen zu können. Gleichzeitig machen die Ergebnisse deutlich, dass auch ausgerichtete Modelle weiterhin Einschränkungen aufweisen. InstructGPT kann falsche Prämissen ungeprüft übernehmen, übermäßig vorsichtig formulieren oder komplexe Anweisungen nicht vollständig umsetzen. Dennoch zeigt die Studie, dass Reinforcement Learning mit menschlichem Feedback einen wesentlichen Fortschritt für die praktische Nutzbarkeit generativer KI darstellt und einen grundlegenden Baustein moderner Sprachassistenzsysteme bildet (Ouyang et al., 2022, 4). Insgesamt wird ersichtlich, dass generative KI nicht allein durch umfangreiche Trainingsdaten und Modellskalierung verbessert wird, sondern insbesondere durch eine gezielte technische Ausrichtung an menschlichen Erwartungen. Die Kombination aus generativen Modellen und menschlichem Feedback stellt somit eine zentrale Grundlage für heutige und zukünftige Anwendungen generativer KI dar.

Im Zusammenhang mit der technischen Funktionsweise generativer KI ist zudem zu berücksichtigen, dass viele moderne Systeme auf sogenannten Foundation Models basieren. Diese Modelle werden auf sehr breiten, domänenübergreifenden Datensätzen trainiert und können anschließend für eine Vielzahl unterschiedlicher Anwendungen angepasst werden. Durch diese Wiederverwendbarkeit entfalten Foundation Models eine besonders hohe systemische Wirkung, da ihre erlernten Fähigkeiten, aber auch bestehende Risiken und Defizite, in zahlreiche nachgelagerte Anwendungen übertragen werden. Bommasani et al. betonen in diesem Zusammenhang, dass Foundation Models nicht als rein technische Artefakte verstanden werden können, sondern grundsätzlich sozio-technischer Natur sind. Ihr tatsächliches Verhalten entsteht aus dem Zusammenspiel verschiedener Faktoren, darunter die verwendeten Trainingsdaten, die Modellarchitektur, die eingesetzte Recheninfrastruktur, die Gestaltung von Nutzerprompts sowie organisatorische Governance-Strukturen und der jeweilige Einsatzkontext (Bommasani et al., 2021, 1-3).

Geschlechtsbezogene Bewertungsverzerrungen als strukturelle Vorläufer algorithmischer Bias

Seit mehreren Jahrzehnten belegen sozialwissenschaftliche Audit- und Laborexperimente systematische Verzerrungen in Bewertungsprozessen, welche selbst dann auftreten, wenn die zu beurteilenden Inhalte objektiv identisch sind. Ein zentrales Merkmal dieser Befunde ist, dass minimale soziale Zuschreibungen, insbesondere das wahrgenommene Geschlecht, signifikante Auswirkungen auf Kompetenzurteile, Einstellungsentscheidungen und Leistungsbewertungen haben. Ein häufig zitierter Nachweis stammt aus einem doppelblinden Experiment im Hochschulkontext, in dem identische Bewerbungsunterlagen lediglich durch unterschiedlich zugewiesene Vornamen variiert

wurden. Bewerbungen mit männlichem Namen wurden dabei signifikant häufiger als kompetent wahrgenommen, als eher einstellungswürdig beurteilt und mit höheren Gehaltsvorstellungen assoziiert als identische Unterlagen mit einem weiblich zugewiesenen Namen (Moss-Racusin et al., 2012, 16474-16475). Vergleichbare Effekte zeigten sich bereits in früheren Untersuchungen akademischer Rekrutierungsprozesse: Steinpreis et al. (1999) konnten nachweisen, dass identische wissenschaftliche Lebensläufe bei männlicher Zuschreibung konsistent bessere Bewertungen erhielten als bei weiblicher Zuordnung (509–528).

Auch im wissenschaftlichen Publikationswesen lassen sich geschlechtsbezogene Bewertungsasymmetrien empirisch belegen. Wennerås und Wold (1997) zeigten, dass Frauen im Peer-Review-Prozess nachweislich höhere Publikationsleistungen vorweisen mussten, um vergleichbare Förderentscheidungen zu erhalten, was auf strukturelle Verzerrungen in vermeintlich objektiven Begutachtungsverfahren hinweist. Gleichzeitig verdeutlichen Studien zur Wirkung von Anonymisierung, dass solche Bias-Effekte nicht unvermeidlich sind. Besonders eindrücklich zeigen dies die Untersuchungen zu blinden Probespielen in Spitzenorchestern, bei denen die Einführung anonymisierter Auswahlverfahren zu einem signifikanten Anstieg des Frauenanteils führte (Goldin & Rouse, 2000, 715).

Feldexperimente am Arbeitsmarkt bestätigen diese Mechanismen auch außerhalb akademischer Kontexte. Bertrand und Mullainathan (2004) konnten zeigen, dass bereits minimale Unterschiede in Namen, wie ihre ethnische oder geschlechtliche Konnotation, die Rückrufquoten auf Bewerbungen deutlich beeinflussen, obwohl Qualifikation und Berufserfahrung identisch waren. Diese Ergebnisse verdeutlichen die erhebliche Wirkmacht sozialer Kategorien in Bewertungsprozessen, selbst wenn keinerlei leistungsbezogene Differenzen vorliegen.

Diese sogenannten „analogen“ Befunde bilden eine zentrale Vorlage für die Analyse moderner KI-Systeme. Generative Sprachmodelle werden auf großen Mengen historischer Texte, Bewertungen und Entscheidungsdokumentationen trainiert, die genau solche gesellschaftlichen Praktiken widerspiegeln. Entsprechend ist zu erwarten, dass Modelle diese Muster statistisch internalisieren und in ihren Ausgaben reproduzieren – etwa durch unterschiedliche Tonalität, variierende Zuschreibungen von Kompetenz oder divergierende Bewertungsschwerpunkte bei inhaltsgleichen Profilen. Ohne gezielte Gegenmaßnahmen können sich damit bestehende Ungleichheiten in automatisierter Form fortsetzen oder sogar verstärken.

Vor diesem Hintergrund gewinnen Maßnahmen wie bewusste Datenkuratierung, der Einsatz von Blinding-Workflows, systematische Bias-Audits sowie evaluative Kontrollmechanismen besondere Bedeutung. Sie dienen dazu, historisch gewachsene Verzerrungen nicht unreflektiert in technische Systeme zu übertragen, sondern deren Reproduktion im Kontext generativer KI aktiv zu begrenzen.

Implizite soziale Kognition als kognitiver Ursprung systematischer Bewertungsverzerrungen

Einen zentralen theoretischen Erklärungsansatz für persistente Bewertungsverzerrungen liefern Greenwald und Banaji mit ihrer Theorie der Implicit Social Cognition. Die Autor*innen argumentieren, dass ein erheblicher Teil sozialer Wahrnehmungs- und Urteilsprozesse nicht bewusst, sondern implizit

abläuft. Implizite Kognitionen sind demnach dadurch gekennzeichnet, dass frühere Erfahrungen gegenwärtiges Denken und Handeln beeinflussen, ohne dem Subjekt in reflektierbarer Form zugänglich zu sein. Wie Greenwald und Banaji formulieren, besteht das zentrale Merkmal impliziter Kognition darin, „that traces of past experiences affect some performance, even though the influential earlier experience is not remembered in the usual sense – that it is unavailable to self-report or introspection" (Greenwald & Banaji, 1995, 5).

Zur Veranschaulichung dieses Wirkmechanismus verweisen die Autorinnen auf Befunde aus der Gedächtnisforschung, insbesondere auf das Phänomen der sogenannten Restwirkung (priming). In entsprechenden Experimenten zeigt sich, dass Personen bei der Vervollständigung unvollständiger Buchstabenfolgen mit höherer Wahrscheinlichkeit jene Wörter bilden, denen sie zuvor beiläufig ausgesetzt waren. Entscheidendes Merkmal dieses Effekts ist, dass sich die Probandinnen in der Regel weder bewusst an die vorherige Darbietung erinnern noch diese als Ursache ihres Handelns identifizieren können. Gleichwohl beeinflussen die zuvor aktivierten Gedächtnisspuren messbar das spätere Verhalten.

Innerhalb dieses theoretischen Rahmens nimmt die implizite Stereotypisierung eine besondere Stellung ein. Greenwald und Banaji (1995) definieren implizite Stereotype als „the introspectively unidentified (or inaccurately identified) traces of past experience that mediate attributions of qualities to members of a social category" (15). Entscheidend ist dabei, dass diese „past experiences" nicht ausschließlich auf individuell Erlebtem beruhen, sondern auch kulturell geteilte Narrative, gesellschaftliche Rollenbilder und medial vermittelte Überzeugungen umfassen können (Greenwald & Banaji, 1995, 14). Implizite Stereotype entstehen somit nicht nur durch persönliche Erfahrungen, sondern auch durch die fortwährende Exposition gegenüber sozialen Bedeutungszuschreibungen innerhalb einer Gesellschaft.

Diese kognitiven Strukturen führen dazu, dass soziale Kategorien wie Geschlecht oder Ethnie automatisch mit bestimmten Eigenschaften, Kompetenzen oder Bewertungen assoziiert werden. Die entsprechenden Zuschreibungen erfolgen dabei unbewusst und ohne intentionale Steuerung. Individuen projizieren diese erlernten Kategorien auf andere Personen, ohne sich des zugrunde liegenden Einflusses bewusst zu sein. Implizite Stereotypisierung kann somit Urteile, Wahrnehmungen und Entscheidungen beeinflussen, selbst wenn keinerlei explizite Vorurteile vorhanden sind.

Ein zentrales Merkmal impliziter Kognitionen besteht folglich darin, dass sie unabhängig vom bewussten Selbstbild wirken. Greenwald und Banaji (1995) betonen, dass implizite Stereotype auch bei Personen auftreten können, die sich explizit als egalitär oder vorurteilsfrei verstehen. Gerade diese Diskrepanz zwischen expliziten Einstellungen und implizitem Verhalten erklärt, weshalb diskriminierende Bewertungsmuster in Organisationen, Wissenschaft und Arbeitsmärkten selbst dort fortbestehen, wo normative Gleichstellungsziele offiziell anerkannt sind.

Damit liefert die Theorie der impliziten sozialen Kognition einen zentralen psychologischen Erklärungsrahmen für die zuvor dargestellten empirischen Befunde zu geschlechtsbezogenen Bewertungsverzerrungen. Sie macht deutlich, dass solche Effekte nicht primär auf bewusste Diskriminierungsabsicht zurückzuführen sind, sondern auf tief verankerte kognitive

Assoziationsmuster, die sich der bewussten Kontrolle weitgehend entziehen. Diese Einsicht ist insbesondere für die Analyse algorithmischer Systeme relevant, da auch datengetriebene Modelle bestehende implizite Muster reproduzieren können, wenn diese in Trainingsdaten strukturell angelegt sind.

Verbindung der theoretischen Ebenen

Zusammenfassend verdeutlichen die dargestellten theoretischen Ansätze, dass der Bias in generativen KI-Systemen weder als rein technischer Zufall noch als ausschließlich menschliches Fehlurteil zu verstehen ist. Vielmehr entsteht er aus einem komplexen Zusammenspiel zwischen technischen Modellarchitekturen, datengetriebenen Lernprozessen und gesellschaftlich geprägten Kategorisierungsmechanismen. Die Funktionsweise generativer Modelle führt dazu, dass diese statistischen Muster aus ihren Trainingsdaten erlernen und reproduzieren, ohne deren soziale oder normative Bedeutung einordnen zu können. Die technische Struktur solcher Systeme begünstigt somit die Übernahme historisch gewachsener Bewertungslogiken, sofern diese in den zugrunde liegenden Daten enthalten sind.

Aus sozialpsychologischer Perspektive lassen sich diese Muster als Ausdruck impliziter Kognitionen begreifen. Wie die Theorie der Implicit Social Cognition zeigt, wirken kulturell verankerte Stereotype häufig jenseits bewusster Kontrolle und beeinflussen Wahrnehmung, Bewertung und Entscheidungsprozesse selbst dann, wenn Individuen explizit egalitäre Überzeugungen vertreten. Diese impliziten Assoziationen prägen soziale Praktiken, institutionelle Routinen sowie dokumentierte Bewertungsprozesse und schreiben sich dadurch dauerhaft in Textkorpora, Leistungsbeurteilungen und Entscheidungsdaten ein.

Werden generative KI-Systeme auf solchen Daten trainiert, übernehmen sie diese impliziten Verzerrungen in Form statistischer Regularitäten. Die Modelle erlernen nicht Vorurteile im normativen Sinne, sondern Wahrscheinlichkeitsverteilungen sprachlicher Muster, in denen soziale Zuschreibungen bereits enthalten sind. Infolgedessen können sich in automatisierten Textfeedbacks systematische Unterschiede zeigen, etwa in Tonalität, Bewertungsfokus oder Kompetenzzuschreibungen, welche je nach zugeschriebenem Geschlecht oder Namen der Verfasser*innen variieren, obwohl die zugrunde liegende inhaltliche Leistung identisch ist.

Die Verbindung der technischen und sozialpsychologischen Ebenen macht somit deutlich, dass KI gesellschaftliche Machtverhältnisse nicht reproduziert, weil sie bewusst „voreingenommen programmiert“ wäre, sondern weil sie auf Daten operiert, die selbst Ausdruck historischer und kultureller Ungleichheiten sind. Bias entsteht demnach nicht erst im Algorithmus, sondern wird durch dessen Skalierung, Automatisierung und scheinbare Objektivität potenziell verstärkt sichtbar gemacht. Innerhalb dieses theoretischen Rahmens ist die empirische Untersuchung der vorliegenden Arbeit zu verorten. Sie knüpft an die Annahme an, dass geschlechtsbezogene Bewertungsunterschiede in KI-generiertem Feedback nicht als Einzelfälle zu interpretieren sind, sondern als Ergebnis struktureller Zusammenhänge zwischen sozialer Kognition und datenbasierter Modellierung. Die nachfolgende Analyse zielt daher darauf ab, diese Zusammenhänge empirisch zu überprüfen und sichtbar zu machen,

inwiefern generative KI bestehende Bewertungsmuster reproduziert oder transformiert.

Die dargestellten theoretischen Ansätze verdeutlichen, dass Bewertungsverzerrungen weder ausschließlich im Individuum noch ausschließlich im technischen System verortet werden können. Sie entstehen im Zusammenspiel von Daten, kognitiven Mustern und sozialen Kontexten. Vor diesem Hintergrund stellt sich die empirische Frage, ob sich geschlechtsbezogene Bewertungsunterschiede auch in standardisiertem KI-basiertem Schreibfeedback nachweisen lassen.

Methode

In diesem Kapitel werden das Forschungsdesign, die Datenerhebung sowie die Auswertungsverfahren dieser Studie systematisch dargestellt und begründet. Zunächst wird der Untersuchungsgegenstand beschrieben und in Beziehung zur zentralen Forschungsfrage gestellt. Anschließend erfolgt die Begründung der gewählten quantitativen Methodik. Darauf aufbauend werden das experimentelle Design, die Konstruktion der Stichprobe, die eingesetzten Erhebungsinstrumente, der Ablauf der Datenerhebung sowie die spezifischen Schritte der Datenanalyse ausführlich erläutert. Abschließend werden zentrale methodische Reflexionen, Limitationen und relevante Gütekriterien diskutiert, um die Aussagekraft und Belastbarkeit der gewählten Vorgehensweise transparent darzustellen.

Untersuchungsgegenstand

Diese Untersuchung analysiert, ob generative KI-Sprachmodelle bei der automatisierten Bewertung identischer studentischer Texte geschlechterspezifische Verzerrungen aufweisen. Aufbauend auf der Forschungslogik von Moss-Racusin et al. (2012), die zeigten, dass identische Bewerbungsunterlagen je nach zugeschriebenem Geschlecht unterschiedlich beurteilt werden, wird dieses Prinzip auf KI-basierte Textbewertungen übertragen. Die Texte bleiben dabei unverändert, nur der vermeintlichen Autorin bzw. dem vermeintlichen Autor wird jeweils ein anderer „weiblich“ oder „männlich“ gelesener Name und die Anrede „Frau“ bzw. „Herr“ und somit ein fiktives Geschlecht zugeordnet.

Methodenwahl

Zur Beantwortung der Forschungsfrage wird ein quantitatives experimentelles Design gewählt. Das experimentelle Vorgehen ermöglicht es, den kausalen Einfluss der unabhängigen Variable „Geschlecht“ auf die Bewertungen der KI-Modelle präzise zu bestimmen. Indem identische Texte verwendet und ausschließlich Name und Geschlechtsangabe der vermeintlichen Verfasserinnen bzw. Verfasser variiert werden, lässt sich sicherstellen, dass potenzielle Unterschiede in der Bewertung tatsächlich auf diese Manipulation zurückzuführen sind und nicht auf inhaltliche oder formale Merkmale der Texte. Damit ist der gewählte Methodenansatz sowohl inhaltlich als auch methodologisch geeignet, die Forschungsfrage valide zu beantworten.

Forschungsdesign

Das Forschungsdesign folgt einem quantitativen experimentellen Ansatz mit Analyse.

Stichprobe und Stimulusmaterial

Es wurden n = 1.000 fiktive Studierende generiert (500 weiblich, 500 männlich). Für jedes Profil wurde ein Vorname entsprechend der Geschlechtszuweisung sowie eine zufällig generierte Matrikelnummer erstellt und dokumentiert. Als Stimulusmaterial dienten fünf Seminararbeiten. Zwei Arbeiten der Geschichtswissenschaft, zwei der Politikwissenschaft sowie einer der Wirtschaftswissenschaft. Jede Arbeit wurde jeweils 100 weiblichen und 100 männlichen Profilen zugewiesen, sodass pro Arbeit n = 200 Bewertungen entstanden.

Experimentelle Manipulation

Die Texte der Seminararbeiten blieben unverändert. Manipuliert wurden ausschließlich Name, Matrikelnummer und Geschlechtskennzeichnung der vermeintlich verfassenden Person. Diese Variation stellt die unabhängige Variable dar. Ausgestattet mit diesen Informationen, wurden die Texte im jeweiligen Schritt an die KI gesendet.

Experimentelle Durchführung

Die Bewertungen wurden automatisiert über eine Programmierschnittstelle (API) mit dem KI-Sprachmodell GPT-4.5 durchgeführt. Für jeden Fall wurden dem Modell der jeweilige Text sowie Name, Matrikelnummer und die Anrede „Herr“ bzw. „Frau“ übermittelt. Das Modell wurde angewiesen, die Arbeiten auf Grundlage einer Bewertungsvorlage für Bachelorarbeiten der Universität Wien (2016) zu beurteilen, eine Gesamtnote zu vergeben und ein schriftliches Feedback zu erstellen.

Der Temperaturparameter wurde auf null gesetzt, um gleichbleibende Ausgaben zu gewährleisten und zufallsbedingte Varianz auszuschließen. Unterschiede in Bewertung und Feedback können somit ausschließlich auf die Variation der unabhängigen Variable zurückgeführt werden. Alle Anfragen erfolgten ohne Nutzung von Konversationsverläufen, sodass keine Vorinformationen aus vorherigen Interaktionen berücksichtigt wurden.

Datenanalyse

Im ersten Schritt wurden die numerischen Bewertungen statistisch ausgewertet. Grundlage bildet dabei die experimentelle Struktur des Designs, in dem das fiktive Geschlecht als unabhängige Variable und die vergebenen Einzelnoten sowie die Gesamtnote als abhängige Variablen operationalisiert sind. Für jedes Modell wurden die Notenverteilungen zwischen männlich und weiblich zugewiesenen Profilen systematisch miteinander verglichen, um mögliche geschlechtsspezifische Abweichungen zu identifizieren. Ergänzend wurden die Gesamtnoten sowie potenzielle Fehlanwendungen des vorgegebenen Bewertungsschemas analysiert, um Hinweise auf konsistente Benachteiligungen oder strukturelle Muster in der Bewertungspraxis des Sprachmodells zu gewinnen. Dieser quantitative Teil ermöglicht Aussagen darüber, ob statistisch bedeutsame Unterschiede zwischen den beiden Geschlechtsbedingungen bestehen.

Da es sich bei den vergebenen Einzelnoten sowie der Gesamtnote um ordinal skalierte Daten handelt, erfolgte die Auswertung der Zusammenhänge zwischen dem fiktiven Geschlecht der Verfasser*innen

und den vergebenen Bewertungen mithilfe des Spearman-Rangkorrelationskoeffizienten. Dieses nichtparametrische Korrelationsmaß setzt keine Normalverteilung voraus und ist besonders geeignet, monotone Zusammenhänge mit ordinalen Variablen abzubilden.

Für jedes KI-Modell wurde getrennt die Korrelation zwischen dem binär codierten Geschlecht (0 = weiblich, 1 = männlich) und den jeweiligen Einzel- bzw. Gesamtnoten berechnet. Signifikant von null abweichende Korrelationskoeffizienten werden als Hinweis auf systematische geschlechterspezifische Bewertungsunterschiede interpretiert.

Methodenreflexion

Die gewählte methodische Vorgehensweise weist mehrere zentrale Stärken auf, die wesentlich zur Robustheit der Untersuchung beitragen. Durch die konsequente Standardisierung des Untersuchungsszenarios und die Randomisierung der fiktiven Profile wird eine hohe interne Validität erreicht, da die Effekte eindeutig der manipulierten Variable, dem fiktiven Geschlecht der vermeintlichen Verfasserinnen und Verfasser, zugeschrieben werden können (Shadish et al., 2002). Auch die Verwendung identischer Texte für alle experimentellen Bedingungen entspricht der Logik klassischer Vignetten- und Audit-Studien, die darauf abzielen, Bewertungsunterschiede eindeutig auf sozial zugeschriebene Merkmale zurückzuführen (Pager & Shepherd, 2008). Die strikte Standardisierung des Bewertungsprompts trägt zusätzlich zur Objektivität und Replizierbarkeit der Datenerhebung bei. Die Standardisierung wird in der empirischen Sozialforschung als zentrales Kriterium zur Sicherstellung vergleichbarer Messbedingungen hervorgehoben (Diekmann, 2021).

Gütekriterien

Hinsichtlich der Gütekriterien zeigt die Studie eine hohe Objektivität und Reliabilität aufgrund der großen Stichprobe. Die Validität ist im Hinblick auf die interne Logik des Experiments stark ausgeprägt, wenngleich die externe Validität aufgrund der künstlichen Versuchsanordnung und der Dynamik der Modelle begrenzt bleibt. Insgesamt ist die gewählte Vorgehensweise jedoch gut geeignet, um die Forschungsfrage fundiert zu beantworten und Aufschluss über mögliche geschlechterspezifische Verzerrungen in KI-basierten Bewertungsprozessen zu geben.

Methodische Einschränkungen

Es ist jedoch zu beachten, dass die gewählte Methode inhärente Grenzen aufweist, welche für die Interpretation der Ergebnisse berücksichtigt werden muss. Die geschlechtsbezogene Manipulation erfolgt ausschließlich über fiktive Namen und binäre Geschlechtsmarkierungen, wodurch komplexere Dimensionen geschlechtlicher Identität unberücksichtigt bleiben.

Darüber hinaus sind generative KI-Modelle dynamische Systeme, die sich durch kontinuierliche Updates verändern können, was die zeitliche Stabilität der Ergebnisse einschränkt. Auch wenn fünf unterschiedliche Seminararbeiten verwendet werden, bilden sie nicht die volle Vielfalt realer studentischer Schreibleistungen ab, sodass die externe Validität begrenzt bleibt. Ethisch ist das Forschungsdesign grundsätzlich vertretbar, da ausschließlich künstlich generierte Profile verwendet

werden und keine realen Personen unmittelbar betroffen sind.

Ergebnisse

Deskriptive Statistik

Die Analysestichprobe umfasst n = 1000 vollständig auswertbare Fälle (500 weiblich, 500 männlich kodiert). Die durchschnittliche Gesamtnote beträgt M = 1.97 (SD = 0.20). Die Verteilung zeigt eine starke Konzentration auf die Note 2,0 (95,9 % aller Fälle), was auf eine eingeschränkte Varianz der abhängigen Variable hinweist.

Getrennt nach Geschlecht ergeben sich folgende Kennwerte:
Männlich: M = 1.98, SD = 0.20
Weiblich: M = 1.96, SD = 0.20
Die Mittelwertdifferenz beträgt 0.022 Notenpunkte.

Mittelwertvergleich

Ein unabhängiger Welch-t-Test ergibt:
$t(996.33) = 1.74$, $p = .083$.
Das 95-%-Konfidenzintervall [−0.003; 0.047] enthält den Wert 0.
Cohen's $d \approx 0.11$, was nach Cohen (1992) einem sehr kleinen Effekt entspricht.

Nichtparametrische Analyse

Der Mann-Whitney-U-Test bestätigt die Nicht-Signifikanz ($p = .086$).

Regressionsanalyse

Unter Kontrolle des Texttitels bleibt der Geschlechtseffekt klein und statistisch nicht signifikant ($p = .069$). Der Texttitel erklärt jedoch ca. 13 % der Varianz ($R^2 \approx .13$).

Analyse seltener Extremwerte

Die Note 3,0 tritt ausschließlich bei männlich kodierten Profilen auf (6 Fälle vs. 0 Fälle). Ein Fisher-Exact-Test ergibt $p = .031$.

Diskussion

Kein globaler Gender Bias

Die Ergebnisse zeigen keinen signifikanten globalen Geschlechtseffekt in der durchschnittlichen Notenvergabe. Dies steht im Kontrast zu zahlreichen humanen Bewertungsstudien (Moss-Racusin et al., 2012; Steinpreis et al., 1999), die deutliche Unterschiede dokumentieren.
Dieser Befund relativiert pauschale Annahmen, generative KI reproduziere zwangsläufig menschliche Diskriminierungsmuster. Im untersuchten Setting zeigt ChatGPT keine systematische Mittelwertverschiebung zugunsten oder zulasten eines Geschlechts.

Subtile Effekte im Randbereich

Gleichzeitig deutet die Extremwertanalyse auf eine asymmetrische Verteilung seltener negativer Bewertungen hin. Solche Randbereichseffekte erinnern an Forschung zu „gendered white lies“ (Jampol & Zayas, 2020), bei der sich Verzerrungen nicht im Durchschnitt, sondern in qualitativen Nuancen zeigen.

Die geringe Zahl der Fälle erfordert jedoch Zurückhaltung in der Interpretation. Dennoch legt der Befund nahe, dass Bias – sofern vorhanden – nicht als breiter struktureller Effekt, sondern als situativer Ausreißer-Mechanismus auftreten könnte.

Technische Moderation von Bias

Mehrere Faktoren könnten zur Abschwächung potenzieller Verzerrungen beigetragen haben:

- Strikte Standardisierung des Prompts
- Temperaturparameter = 0
- Klare Bewertungskriterien
- Alignment-Mechanismen moderner Sprachmodelle (Ouyang et al., 2022)
- Die Ergebnisse deuten darauf hin, dass strukturierte Bewertungsschemata Varianz reduzieren und potenzielle Verzerrungsspielräume einschränken können.

Theoretische Einordnung

Aus sozialpsychologischer Perspektive lassen sich die Befunde im Lichte impliziter Kognition (Greenwald & Banaji, 1995) interpretieren. Während menschliche Bewertende implizite Stereotype aktivieren, operieren Sprachmodelle auf statistischer Ebene. Sie internalisieren Regularitäten, verfügen jedoch nicht über intentionale Bewertungsabsichten.

Die Studie spricht daher für eine Differenzierung zwischen:

intentionalem Bias (menschlich)

statistischem Bias (algorithmisch)

Nicht jede gesellschaftliche Verzerrung muss zwangsläufig in KI-Ausgaben sichtbar werden – insbesondere nicht in stark standardisierten Settings.

Schreibdidaktische Implikationen

Für die Schreibdidaktik ergeben sich mehrere Konsequenzen:

KI ist weder neutral noch automatisch diskriminierend

Die Ergebnisse zeigen keine starke systematische Benachteiligung. Gleichzeitig machen sie deutlich, dass KI nicht als vollständig kontextfreie Instanz verstanden werden darf.

Bedeutung von AI-Literacy

AI-Literacy umfasst laut Long und Magerko (2020) die Fähigkeit, KI kritisch zu bewerten und ihre

Funktionsweise zu verstehen. Dazu gehört auch die Sensibilität gegenüber möglichen Verzerrungen.
Studierende sollten lernen:

- KI-Feedback nicht unkritisch zu übernehmen
- Bewertungsmaßstäbe transparent zu prüfen
- Feedback mit eigener Urteilskompetenz abzugleichen

Verantwortungskultur

Im Sinne von Brommer et al. (2023) bedarf es einer gemeinsamen Verantwortungskultur. Lehrende sollten KI-Feedback weder verbieten noch unreflektiert integrieren, sondern transparent rahmen.
Die vorliegende Studie liefert hierfür eine empirische Grundlage:
Sie zeigt, dass standardisierte Settings potenzielle Verzerrungen reduzieren können – was didaktisch nutzbar ist.

Limitationen

- Eingeschränkte Notenvarianz
- Binäre Geschlechtskategorisierung
- Untersuchung nur eines Modells
- Zeitgebundene Modellversion
- Keine umfassende qualitative Diskursanalyse!

Conclusio

Die vorliegende Untersuchung findet keinen starken geschlechtsspezifischen Bias in der automatisierten Bewertung identischer studentischer Texte durch ChatGPT-4.5. Gleichzeitig zeigen Randbereichsanalysen, dass subtile Unterschiede nicht ausgeschlossen werden können.
Die Ergebnisse sprechen gegen alarmistische Annahmen einer systematischen algorithmischen Diskriminierung im untersuchten Setting. Zugleich unterstreichen sie die Notwendigkeit kontinuierlicher empirischer Evaluation. Weiterführende Forschung sollte KI-generiertes Textfeedback vertieft im Hinblick auf qualitative Biases untersuchen, um auch subtile, kontextabhängige Verzerrungen differenziert erfassen und bewerten zu können.
Für die Schreibdidaktik bedeutet dies:
KI-Feedback ist ein sozio-technisches Instrument, dessen Einsatz Reflexionskompetenz, Transparenz und institutionelle Rahmung erfordert. Eine verantwortungsvolle Integration generativer KI setzt nicht auf blinde Technikgläubigkeit, sondern auf empirisch informierte Urteilskraft.

Literatur

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? American Economic Review, 94(4), 991–1013. https://doi.org/10.1257/0002828042002561

Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Brommer, S., et al. (2023). Wissenschaftliches Schreiben im Zeitalter von KI gemeinsam verantworten. Hochschulforum Digitalisierung [PDF] https://hochschulforumdigitalisierung.de/wp-content/uploads/2023/11/HFD_DP_27_Schreiben_KI.pdf.

Buck, I. (2026). Wissenschaftliches Schreiben mit KI. utb GmbH. https://doi.org/10.36198/9783838565682

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Döll, M., Döhring, M., & Müller, A. (2024). Evaluating Gender Bias in Large Language Models. arXiv. https://doi.org/10.48550/ARXIV.2411.09826

Diekmann, A. (2021). Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen (10. Aufl.). Rowohlt.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. Business & Information Systems Engineering, 66(1), 111-126. https://doi.org/10.1007/s12599-023-00834-7

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems, 14(3), 330–347. https://doi.org/10.1145/230538.230561

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of “blind” auditions on female musicians. American Economic Review, 90(4), 715–741. https://doi.org/10.1257/aer.90.4.715

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition. Psychological Review, 102(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Hickisch, L. (2024). Maschinenrevolution in der Schreibdidaktik. https://doi.org/10.25365/thesis.76718

Jampol, L., & Zayas, V. (2020). Gendered white lies. Personality and Social Psychology Bulletin, 47(1), 57-69. https://doi.org/10.1177/0146167220916622

Kotek, H., Dockum, R., & Sun, D. Q. (2023). Gender bias and stereotypes in Large Language Models. arXiv. https://doi.org/10.48550/ARXIV.2308.14921

Lang, F. (2025). Künstliche Intelligenz in Seminar- und Abschlussarbeiten: Ein Praxisleitfaden für Studierende mit Handlungsempfehlungen, Prompt-Beispielen und kritischer Einordnung. Springer. https://doi.org/10.1007/978-3-662-71542-0

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3313831.3376727

Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Moss-Racusin, C. A., et al. (2012). Science faculty's subtle gender biases favor male students. PNAS, 109(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. arXiv.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology, 34, 181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740.

Petroni, F., et al. (2019). Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://aclanthology.org/D19-1250.pdf.

Schmidt, R. L., & Seegel, S. (2024). Voneinander lernen – KI-gestütztes wissenschaftliches Schreiben im Team lehren. JoSch - Journal für Schreibwissenschaft, 15, 65–72. https://doi.org/10.3278/JOS2401W006.

Schulze Heuling, D., Jakobi, A. P., Schaal, G. S., & Gerlich, M. (2025). Generative KI und (politik) wissenschaftliches Schreiben: Herausforderung für die Lehre und darüber hinaus. Politische Vierteljahresschrift, 66(4), 913–940. https://doi.org/10.1007/s11615-025-00631-9.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin.

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of CVs. Sex Roles, 41, 509–528. https://doi.org/10.1023/A:1018839203698.

Universität Wien. (2016). Bewertungsvorlage für Seminar- und Bachelorarbeiten [PDF]. https://musikwissenschaft.univie.ac.at/fileadmin/user_upload/i_musikwissenschaft/Studium/Bewertungsvorlage_Seminar-_und_Bachelorarbeiten_ausfuellbares_PDF-Formular.pdf.

Wambsganss, T., Su, X., Swamy, V., Neshaei, S. P., Rietsche, R., & Käser, T. (2023). Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance. arXiv. https://doi.org/10.48550/ARXIV.2311.03311.

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. Nature, 387(6631), 341–343. https://doi.org/10.1038/387341a0.

Wolfe, J. (2024). What educational psychology can teach us about providing feedback to Black students. College Composition and Communication, 75(4), 759–783. https://doi.org/10.58680/ccc2024754759.