Blogbeitrag vom 15. Dezember 2023

KI-TEXT IN PRÜFUNGSARBEITEN ERKENNEN

Probleme und Lösungen

Matthis Kepser

Vermehrt tauchen an Schulen und Hochschulen schriftliche Prüfungsarbeiten auf, bei denen die Lehrenden den Verdacht haben, dass sie in großen Teilen KI-generiert worden sind. Aber wie kann ein überzeugender argumentativer Nachweis gelingen? Der folgende Beitrag zeigt dazu zwei Wege auf: der Einsatz von sogenannten Detection Tools am Beispiel von *GPT Radar* und die händische Analyse von KI-typischen Textmerkmalen. Beide Verfahren sind für sich fehleranfällig, aber zusammen ergeben sie doch eine recht plausible und verlässliche Beurteilungsgrundlage. Dabei ist jedoch zu bedenken, dass sich KI-gestützte Textgeneratoren auf der einen Seite und ebenso Detection Tools auf der anderen kontinuierlich weiterentwickeln. Auch können Schüler/-innen und Studierende versuchen, die Herkunft von KI-generiertem Text bewusst zu verschleiern.

VK:KIWA

Problemstellung

Die Entwicklung von Chatbots auf der Basis von vortrainierten Large Language Models (LLMs) birgt ohne Zweifel große Chancen für die Bewältigung von Schreibaufgaben aller Art, seien es fiktionale, pragmatische oder auch wissenschaftliche Texte. Schüler/-innen und Studierende können, ja sollen sie nutzen, um sich in der Planungsphase bei der Ideengewinnung („inventio") und argumentativen Anordnung ihrer Überlegungen („dispositio") inspirieren zu lassen. Auch ihr Einsatz im Sinne eines formalen und stilistischen Lektorats ist in der Überarbeitungsphase wohl hinzunehmen, selbst wenn fraglich ist, ob die Autoren bzw. Autorinnen dadurch ihre diesbezügliche Schreibkompetenzen optimieren. Dies würde einen äußerst reflektierten Umgang mit den KI-generierten Verbesserungsvorschlägen voraussetzen. Hinzunehmen ist eine solche Schreibpraxis aus prüfungsrechtlicher Sicht vor allem deshalb, weil ihr Nachweis ex post facto so gut wie unmöglich ist. Auch Eltern, Geschwister, Mitschüler/-innen, Kommilitonen oder gar bezahlte Lektorate könnten zu einer solchen Schreibberatung beigetragen haben. In dieser Hinsicht gleichen AI-Writing Tools sogar Bildungs- und Chancenbenachteiligungen aus: Schüler/-innen und Studierende mit einem entsprechenden Umfeld konnten schon immer ihre Arbeiten gegenlesen lassen und mit formal sowie stilistisch ansprechenden Ergebnissen glänzen, ohne dass erkennbar war, was davon ihren eigenen Fähigkeiten und Fertigkeiten zuzuschreiben ist (vgl. auch Schindler 2023, 18).

Problematisch wird der Einsatz von *ChatGPT* und seinen Verwandten dann, wenn die Ausformulierung („elocutio") von ihnen übernommen wird. Schreibdidaktisch betrachtet nehmen sich die Autoren/Autorinnen damit die Chance, ihre eigenen Writing-Skills zu verbessern. Denn was für das Lesen gilt, gilt auch für das Schreiben: Schreiben lernt man durch Schreiben (Deskilling; vgl. Schindler 2023, 8). Zudem kommt dem Schreiben eine epistemische Funktion zu, wie hinlänglich bekannt ist: Erst während des Schreibens entwickeln sich häufig Ideen und Argumentationsstränge, an die man in den vorangegangenen Phasen gar nicht gedacht hatte (vgl. ebd., 9). Darüber hinaus ist erheblich in Zweifel zu ziehen, ob ein überwiegend von KI verfasster Text noch als „eigenständig verfasste" Arbeit zu werten ist. Zwar handelt es sich bei solchen Konvoluten nicht wirklich um Plagiate, denn KI-generierte Texte sind Unikate (vgl. Salden et al. 2023, 9) und haben nach allgemeiner Rechtsauffassung keinen Autor, dessen Urheberrechte durch eine ungekennzeichnete Übernahme verletzt werden (vgl. Hoeren 2023, 23). Auf der anderen Seite ist eine solche Arbeit aber wohl nicht im Sinne einer Erklärung verfasst, die zumindest von Studierenden regelhaft an den meisten Universitäten und Hochschulen eingefordert wird: „Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe." [1]. Zwar könnte man argumentieren, dass es auch bereits vor den Chatbots elektronische Unterstützungsmöglichkeiten wie die in Textverarbeitungssysteme integrierte Rechtschreib- und Grammatikprüfung gab, deren Einsatz nicht angegeben werden musste. AI-Writing Tools sind aber weitaus potentere Hilfsmittel, denn sie können Schreibende nicht nur unterstützen, sondern vollständig ersetzen. KI-generierter Text müsste also stets als solcher gekennzeichnet werden (vgl. Hoeren 2023, 23). [2]

Der Nachweis für einen (ungekennzeichneten) Einsatz eines AI-Writing Tools als „Hilfsmittel" ist indessen schwierig und stellt Lehrende und Prüfungsausschüsse vor erhebliche Herausforderungen. Bisweilen wird auf diversen Nachrichtenkanälen kolportiert und auch von Studierenden angenommen, dass ein solcher überhaupt nicht gelingen kann. Schließlich handelt es

sich bei KI-generiertem Text um eine einmalige, nicht-reproduzierbare maschinelle Antwort auf eine mit Prompts übergebene Schreibanweisung. Dem ist aber nicht so.

KI Detection Tools am Beispiel von GPT Radar

Schon früh kam die Idee auf, KIs mit ihren eigenen Waffen zu schlagen, also KI-generierten Text mit Hilfe von KI-Chatbots zu entlarven. Dafür einfach die bereits vorhandenen Tools zu verwenden, erwies sich schnell als der falsche Weg. So konnte etwa die bekannteste App *ChatGPT* von ihr selbst erzeugten Text nicht als solchen einigermaßen verlässlich erkennen. Im Januar 2023 stellte *OpenAI*, der Entwickler von *ChatGPT*, daher eine eigenständig trainierte Software vor, von der man sich erhoffte, das Problem zu lösen (*AI classifier*). Der Platzhirsch unter den KI-Chatbots blieb damit nicht alleine; heute tummeln sich mehr als ein Dutzend solcher Detection Tools auf dem Markt. Die Ergebnisse waren allerdings ernüchternd: In einer groß angelegten internationalen Studie (Weber-Wulff et al. 2023) kamen die Autoren und die Autorinnen zu dem Ergebnis:

> *Our findings do not confirm the claims presented by the systems. They too often present false positives and false negatives. […] Therefore, our conclusion is that the systems we tested should not be used in academic settings. (ebd., 30).*

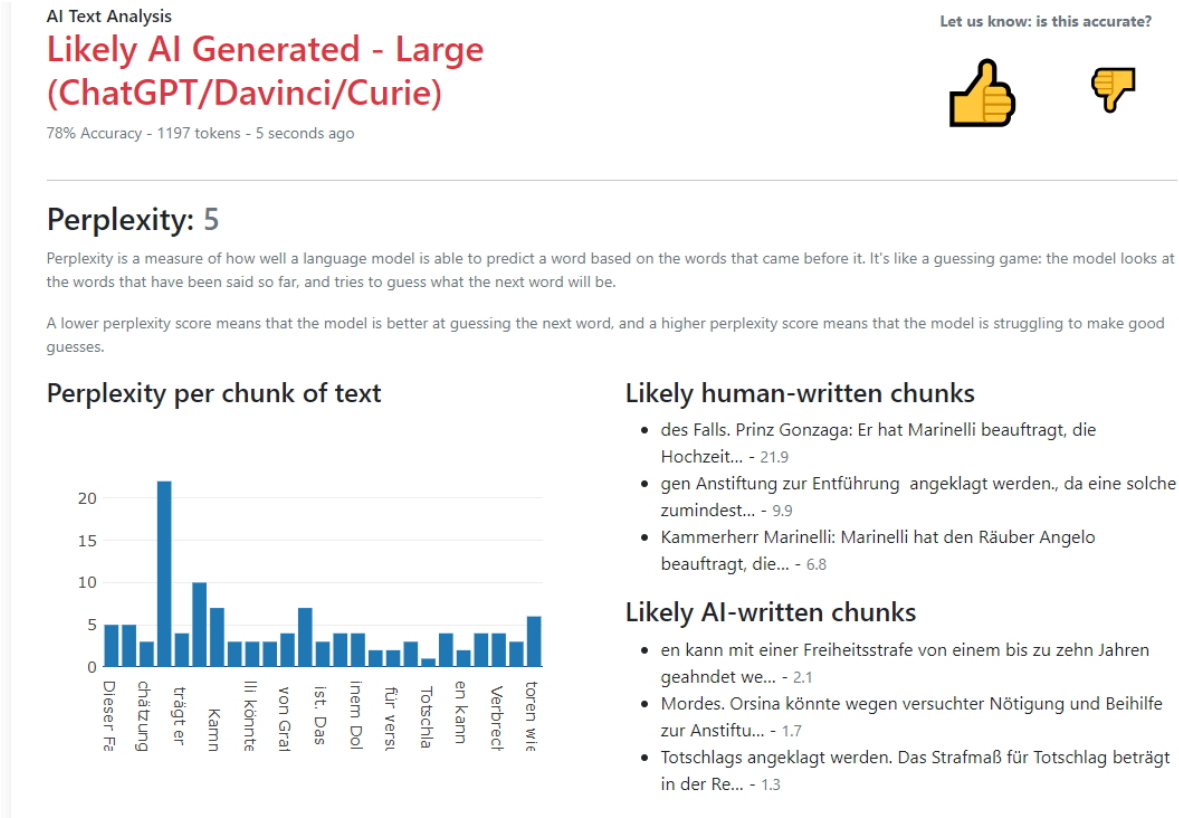
OpenAI zog die Konsequenzen und nahm ihren *AI classifier* Ende Juli 2023 vom Netz. Daraus den Schluss zu ziehen, dass AI Detection prinzipiell nicht funktioniert, wäre allerdings vorschnell. Zum einen zeigte die erwähnte Untersuchung durchaus große Unterschiede in der Genauigkeit, mit der die 14 getesteten Systeme menschlich verfasste von KI generierten Texten unterscheiden konnten. Beachtlich gute Ergebnisse lieferte etwa *Turnitin*. Zum anderen sind Detection Tools wie auch die KI-Chatbots selbstlernende Systeme, von denen man annehmen kann, dass sie ihre Fähigkeiten in kürzester Zeit ausbauen können [3]. Entsprechend sind empirische Studien wie die von Weber-Wulff et al. (2023) wahrscheinlich sehr schnell überholt. Vorgestellt wird im Folgenden eine App, die zwar in verschiedenen Tech-Blogs seit Anfang 2023 immer wieder positiv erwähnt worden ist, aber nicht zu den von Weber-Wulff et al. getesteten Tools gehört: *GPT Radar* des italienischen Entwicklers *Neuraltext*.

Die Arbeitsweise von *GPT Radar* und anderen Detection Tools ist im Prinzip einfach zu verstehen: Bekanntlich erzeugen Chat-Bots Ausgaben in natürlicher Sprache auf der Grundlage von stochastischen Modellen. Ihre „Intelligenz“ verdanken sie grob gesagt üppigen Berechnungen, mit welcher Wahrscheinlichkeit ein Wort einem anderen Wort folgt. *GPT Radar* ermittelt ebenfalls solche Wahrscheinlichkeiten in einem gegebenen Text (Token Probability) und vergleicht sie mit denen, die bei KI-generierten Texten üblich sind (Token Probability Distribution) [4]. Beides wird nachvollziehbar rückgemeldet. Das Ergebnis ist ein Score, der bei GPT-Radar „Perplexity“ genannt wird. Je höher dieser Wert ist, desto wahrscheinlicher ist es, dass der Text von einem menschlichen Autor stammt (vgl. Abb. 1). Zusätzlich wird noch ein Genauigkeitswert (Accuracy) ermittelt. Dabei ist zu koinzidieren: *GPT Radar* arbeitet auf der Grundlage von *Chat GPT 3.5*.

Der Autor hat das Tool einigen experimentellen Tests unterzogen, die dieses Vorgehen veranschaulichen [5]. Ausgangspunkt war eine bei *Fanfiktion.de* veröffentliche Kurzgeschichte

[6], die in praktisch jeder Hinsicht fehlerhaft ist (Rechtschreibung, Interpunktion, Grammatik, Kohärenz, Spannungserzeugung etc.). Die Autorin gibt in ihrem Profil an, weiblich und 22 Jahre alt zu sein.

Durch den Radar geschickt, wird eine Perplexity von 30 rückgemeldet: ein mit höchster Wahrscheinlichkeit von einem Menschen verfasster Text. Lässt man den Text durch *ChatGPT 3.5* in formaler Hinsicht verbessern (Rechtschreibung, Grammatik, Interpunktion), sinkt der Wert auf 16. Bei einer erneuten automatischen Revision im Hinblick auf lokale und globale Kohärenz erfolgt eine Bewertung mit Perplexity 10; die Aufforderung, den Text spannender zu machen, führt zu einem Ergebnis, das die Perplexity 9 bekommt. Je größer also der Eingriff der KI in den ursprünglich menschlichen Text ist, desto mehr sinkt der Perplexity-Wert.



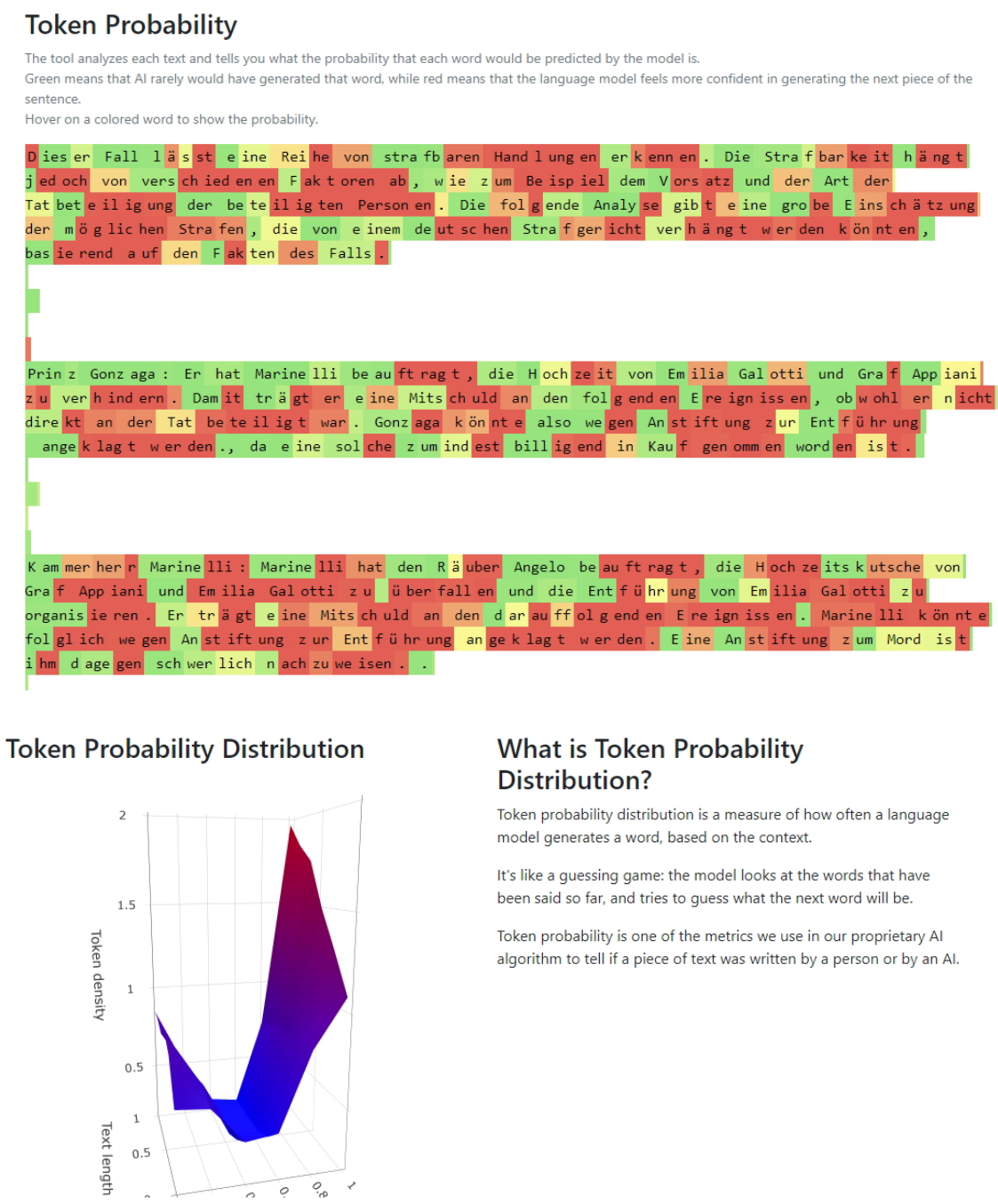


Abb. 1: Beispiel für eine Response von GPT Radar. Die Darstellung der Token Probability ist eingekürzt worden [7].

Ein Perplexity-Wert von 5 und geringer führt regelhaft zu einer Einstufung als „Likely AI Generated - Large (ChatGPT/Davinci/Curie)“. Welches Wahrscheinlichkeitsmodell dahinter steckt, wird auf den Seiten des Betreibers leider nicht angegeben, auch nicht die angenommene Irrtumswahrscheinlichkeit. Der Wert scheint aber sehr plausibel: Eine Fanfiktion-Kurzgeschichte, die mit inhaltlichen Vorgaben aus dem Testtext komplett von *ChatGPT 3.5* erzeugt worden ist, bekommt genau diese Einstufung. Mit denselben Vorgaben als Prompt wurde auch eine Geschichte bei *Bing* angefordert; Microsofts neuer Browser *Edge*hat bekanntlich *ChatGPT 4.0* unter *Bing* implementiert. Das Ergebnis ist verblüffend gut. Der Text ist nicht nur deutlich länger, sondern auch narrativ, dramaturgisch und stilistisch erheblich

besser, als der mit dem Vorgängermodell erzeugte. Mit solchen Eigenschaften bewirbt *OpenAI* auch die Weiterentwicklung:

> *GPT-4 is more creative and collaborative than ever before. It can generate, edit, and iterate with users on creative and technical writing tasks, such as composing songs, writing screenplays, or learning a user's writing style. (https://openai.com/gpt-4).*

Diesen Text bewertet *GPT Radar* sogar nur mit einer Perplexity von 3 und stuft ihn korrekt als mit sehr hoher Wahrscheinlichkeit KI-generiert ein. Das Experiment kann leicht repliziert werden. Dabei darf man nicht erwarten, stets exakt dieselben Perplexity-Werte zu erreichen, aber sie dürften nicht weit davon abweichen.

GPT Radar wurde vom Autor mit über 50 weiteren Texten bzw. Textauszügen konfrontiert. Zwei Drittel davon stammten aus mutmaßlich menschlichen Quellen, denn diesbezügliche Fehler hätten bei einer möglichen prüfungsrechtlichen Begutachtung schwerwiegendere Konsequenzen. Darunter befanden sich Zeitungsartikel aus der *Süddeutschen Zeitung* und der *Bild*, Artikel aus fachdidaktischen Zeitschriften bzw. Buchpublikationen, juristische Kommentare, studentische Hausarbeiten aus den Gebieten Fachdidaktik, Jura und Betriebswissenschaft (eingereicht vor Ende 2022), Wikipedia-Artikel, Gesetzestexte, Verordnungen, Satzungen sowie (fiktionale) Geschichten, die über die Plattform *fanfiktion.de*publiziert worden sind. KI-generierte Texte waren Zeitungsartikel zu Ereignissen aus der Vergangenheit, eine juristische Fallbeurteilung, Vorschläge für Dissertationsthemen, Unterrichtskonzepte, didaktische Begründungen, automatisch erstellte Abstracts zu wissenschaftlichen Artikeln, Wikipedia-Artikel zu einem bei Wikipedia nicht verzeichneten Lemma und (weitere) Fanfiktion-Erzählungen. Dafür sind sowohl *ChatGPT 3.5* also auch *4.0* eingesetzt worden.

Bei den KI-verfassten Texten kam *GPT Radar* in einem einzigen Fall zu einem falschen Ergebnis: Eine KI-erzeugte Fanfiction-Kurzgeschichte wurde mit einem Perplexity-Wert von 6 und „Likely Human Generated" eingestuft. Freilich lag das Tool mit seiner Analyse nur knapp daneben. Unterschiede in der Beurteilung von ChatGPT 3.5 vs. ChatGPT 4.0 konnten nicht systematisch ausgemacht werden; tendenziell bekamen Texte des neueren Modells 4.0 sogar niedrigere Perplexity Scores.

Die mutmaßlich von Menschen verfassten Texte wurde alle korrekt eingestuft – mit einer einzigen, auch systematisch zu replizierenden Ausnahme: Rechtsnormen wie Verordnungen, Satzungen und Gesetzestexte. Dass ausgerechnet beispielsweise die ersten 17 Artikel des Grundgesetzes mit einer Perplexity von 3 bewertet wurden und damit hochgradig unter den Verdacht gestellt worden sind, von einer KI formuliert zu sein, ist schon kurios. [8] Der Fehler betrifft nur die Rechtsnormen selbst nicht ihre Kommentierungen: Diesbezügliche Auszüge aus juristischen Fachtexten der Bundesrepublik und der Schweiz wurden mit Perplexity Scores zwischen 6 und 11 sowie dem Urteil „Likely Human Generated" belegt.

Getestet wurde auch, ob ein von *Quillbot* paraphrasierter Wikipedia-Artikel (Lemma „Fan-Fiktion") als maschinengeneriert erkannt wird. Hier versagte das Tool: Original und automatisch paraphrasierter Text wurden beide mit einer Perplexity von 7 als „Likely Human

Generated“ bewertet. Allerdings war mit der Paraphrase auch ein Plagiatscanner nicht zu überlisten: *PlagAware* meldete über 43% Übereinstimmung mit dem korrekt erkannten Wikipedia-Artikel zurück.

Als Zwischenfazit kann festgestellt werden: Eine Überprüfung von Texten auf den Verdacht KI-generierter Passagen mit Hilfe von *GPT Radar* führt durchaus zu mehrheitlich korrekten Ergebnissen. Sich alleine auf das Tool zu verlassen, ist indes nicht angezeigt. So gilt es zu bedenken, dass KI-erzeugter Text soweit nachbearbeitet worden sein kann, dass eine automatische Detektion zu falschen Ergebnissen führt. Davor haben auch bereits Weber-Wuff et al. (2023, 30) gewarnt. Der Versuch, *GPT Radar* durch zahlreiche nachträglich eingeflochtene Grammatik-, Rechtschreib- und Interpunktionsfehler zu irritieren, gelang interessanterweise nicht: Zwar wurde der entsprechende KI-Text dann mit einer Perplexity von 7 statt 3 bewertet, aber trotzdem als „Likely AI Generated“ eingestuft. Der Perplexity-Wert ist also offensichtlich nicht alleine für die Gesamtbeurteilung entscheidend; korrigierend greift hier wohl die Genauigkeit (Accuracy) ein. Auch formal korrekte stilistische Eingriffe in KI-erzeugte Texte (Umgang ca. 20 %) oder eine Paraphrasierung durch *Quillbot* konnten zwar den Perplexity-Wert um eine Stufe heben, aber nicht über einen Perplexity-Wert von 5. Nur in einem einzigen Fall führt die menschliche Revision zu einem Perplexity Score von 6 und dem Urteil: „Likely Human Generated“. Man muss also offensichtlich KI-erzeugten Text schon erheblich nachbearbeiten, um seinen Ursprung zu verschleiern. Dann freilich hat aus prüfungsrechtlicher Perspektive immerhin eine eigenständige Auseinandersetzung mit einem KI-generierten Textvorschlag stattgefunden (vgl. Hoeren 2023, 26).

Jenseits dessen ist darauf hinzuweisen, dass Urteile durch ein AI Detection Tool prinzipiell fehleranfällig sind, denn sie beruhen auf Wahrscheinlichkeitsberechnungen, die immer auch Irrtümern unterliegen können. Bisher fehlen uns verlässliche Daten zu anzunehmenden Irrtumswahrscheinlichkeiten, die mutmaßlich auch textsortenspezifisch erhoben werden müssten (vgl. oben, Rechtsnormen). Aber selbst dann, wenn die in den Sozialwissenschaften als akzeptabel angenommene Irrtumswahrscheinlichkeit von $\alpha = 5\%$ erreicht werden würde, gilt: Bei gewichtigen Entscheidungen, etwa einer auszusprechenden Exmatrikulation wegen eines Verstoßes gegen die Eigenständigkeitserklärung, bleibt eine ausschließliche Argumentation mit einem AI Detection-Ergebnis problematisch.

Händische Analyse von mutmaßlich KI-verfassten Texten

Wenn Prüfer/-innen den Verdacht äußern, dass eine Prüfungsarbeit mehrheitlich mit KI-Unterstützung verfasst wurde, werden sie dafür mit Sicherheit keine statistische Analyse im Kopf vorgenommen haben. Eine interessante Frage ist also, was eine solche intuitive Einschätzung eigentlich motivieren kann. Schon vor Aufkommen der KIs waren falsche Anführungszeichen ein mögliches Indiz dafür, dass hier mit Copy-and-Paste gearbeitet worden ist. Der Layout-Fehler entsteht bei direkten Übernahmen aus Bildschirmtexten (“”, statt „“) und kann Prüfer/-innen dafür sensibilisieren, dem korrekten Umgang mit fremden Gedanken nachzugehen. Die im Folgenden aufgelisteten Verdachtsmomente sind hingegen KI-systemimmanent. Ein erstes, sehr sicheres Anzeichen dafür, dass eine KI zumindest das Lektorat übernommen hat, sind formvollendete Texte: Selbst professionellen Schreiber/-innen

unterlaufen bisweilen kleine Rechtschreib-, Interpunktions- oder Grammatikfehler – nicht aber den KIs.

Eine Gruppe von Auffälligkeiten ist mutmaßlich dem Schreibtraining der neuronalen Netze geschuldet:

- Stilistisch neigen ChatGPT und Verwandte zu persistenten Übertreibungen: Buzzwords und Hochwertbegriffe sowie hyperbolisch wertende Adjektive (z. B. „formvollendet", „tiefgründig") werden immer wieder und meist völlig gleichlautend verwendet. Zu schulischen Textsorten, aber auch wissenschaftlichen Texten passen sie meist nicht, vielleicht mit Ausnahme von „Antragsprosa". Ihre Verwendung ist eher bei marketing-tauglichen Texten zu erwarten, deren Textmodelle wohl eine wesentliche Grundlage der AI-Writing Tools gewesen sind.
- Bei der Textproduktion setzen KI-Textgeneratoren auffällig häufig „Closer" ein, also kurze Zusammenfassungen. Sie tun das selbst bei Kapiteln, deren Länge eine Zusammenfassung nicht nötig machen. Manchmal sind solche Closer fast ebenso lang, wie der Text, auf den sie sich beziehen. Dies ist mutmaßlich dem Umstand geschuldet, dass längere Ausführungen von den Autoren/Autorinnen in mehreren Schritten generiert worden sind: ChatGPT und Verwandte erlauben nur eine begrenzte Textausgabe von 2000 bis 4000 Tokens. Stilistisch finden sich auch hier oft hyperbolische Attribute und Hochwertbegriffe. [9]
- KI-Textgeneratoren gendern meist nur auf explizite Aufforderung. Schlecht oder gar nicht gegenderte Texte deuten darauf hin, dass sich Autoren und Autorinnen nicht einmal in dieser Hinsicht darum bemüht haben, KI-Texte zu überprüfen und zu überarbeiten.
- Fachtermini werden bisweilen nicht korrekt verwendet. Bei fachdidaktischen Anfragen verwendet ChatGPT 3.5 z. B. regelhaft „pädagogisch", wo es „didaktisch" oder „fachdidaktisch" heißen müsste. Solche Fehler unterlaufen selbst Studierenden im fortgeführten Bachelorstudium sehr selten. Dies mag damit zusammenhängen, dass AI-Writing Tools englischsprachig trainiert worden sind und eine Übersetzung ins Deutsche nicht adäquat gelingt.
- Manchmal werden Texte erzeugt, die im Hinblick auf Satzbau und Wortschatz einen ausgesprochen szientistischen Duktus aufweisen. Sie scheinen zu den anzunehmenden Fähigkeiten des vorgeblichen Autors (Schüler/-innen, Studierende) nicht so recht zu passen.

Eine weitere Gruppe betrifft KI-typische Fehler, die den mangelhaften intellektuellen Fähigkeiten der Tools zuzuschreiben sind. Künstlich ‚intelligent' sind die „stochastischen Papageien" (Bender et al. 2021) eben nur beschränkt:

- Am offensichtlichsten sind diesbezüglich sogenannte KI-„Halluzinationen": Sie betreffen die Argumentation mit Hilfe erfundener Quellen (Werke, Autoren/ Autorinnen, angeblichen historischen Fakten etc.). Solches nachzuweisen, gelingt oft schon mit einer einfachen Abfrage über die bekannten Suchmaschinen.

- Erstaunlich schwach ist auch ihre Fähigkeit, Inhaltsangaben bzw. Zusammenfassungen zu schreiben. Bei Inhaltsangaben zu Texten, die der KI nicht direkt vorgegeben werden („Schreibe eine Zusammenfassung von Lessings Abhandlung zum Nutzen der Fabel in der Schule"), kommt es sehr schnell zu den oben erwähnten Halluzinationen. Bei vorgegebenen Texten sieht die Zusammenfassung auf den ersten Blick beeindruckend aus; auf den zweiten stellt man häufig fest: Wesentliches fehlt. Das zeigt, was in den Fachdidaktiken der Sprachen schon lange bekannt ist: Eine gute Inhaltsangabe zu verfassen, ist eine Aufgabe, deren Lösung oft sehr komplexer Entscheidungen bedarf. Dies gilt insbesondere, aber nicht nur für Inhaltsangaben zu literarischen Texten in allen medialen Spielarten.

- Auffällig sind weiterhin erhebliche Schwächen in der Argumentation. Um sie zu entdecken, muss man sich als Leser/-in allerdings von der häufig beeindruckenden Textoberfläche KI-generierter Texte lösen. Erst eine genaue Analyse fördert dann von künstlicher Intelligenz künstlich produzierten Unsinn zu Tage. Dass hier auch eine gesellschaftspolitisch hochbrisante Problematik vorliegt, sei nur am Rande erwähnt: KIs werden nicht selten dazu benutzt, um schnell Texte für Plattformen der sozialen Netzwerke zu produzieren, die dann geteilt werden.

Eine letzte Gruppe von Auffälligkeiten liegt ursächlich in Vorschriften, die den Tools von deren Entwicklern/ Entwicklerinnen mitgegeben worden sind, um justiziable Konflikte zu vermeiden. Sie sind sozusagen dem Stammhirn der KIs eingeschrieben und gegen Eingriffe durch Trainingsmaßnahmen ausgesprochen gut geschützt, um millionenschweren Klagen aus dem Weg zu gehen. Das betrifft u. a. den Schutz von Persönlichkeitsrechten, den Jugendschutz, v.a. aber Urheberrechte. Folgende KI-typische Auffälligkeiten können bei Prüfungsaufgaben darauf zurückgeführt werden:

- Argumentiert wird mit globalen Behauptungen ohne Belege bzw. konkrete Beispiele, z. B. „Von der Literatur bis hin zu Theaterstücken, Filmen und nun auch Graphic Novels scheint die Figur immer wieder die Vorstellungskraft von Künstlern und Publikum gleichermaßen anzuregen." Welche Theaterstücke, Filme etc. das sein sollen, wird nicht genannt.

- Es gibt keine direkten Zitate in Anführungszeichen und mit genauem Stellennachweis, natürlich auch keine Blockzitate.

- Bei indirekten Zitaten und Zusammenfassungen fremder Gedanken gibt es zwar bisweilen grobe Quellenangaben, sie sind aber nicht stellengenau, d. h. ohne Seitennachweis. Zu diesem Verhalten neigt insbesondere das Model *ChatGPT* 4.0.

- Ironischerweise attestieren Textkontrollen durch Plagiatscanner wie *PlagAware* eine nahezu blütenweiße Weste.

Es bleibt anzumerken, dass natürlich nicht eine einzelne dieser Auffälligkeiten ausreicht, um eine Arbeit unter den Verdacht zu stellen, dass hier mit einer KI gearbeitet worden ist. Viele davon sind auch in studentischen Texten wiederzufinden, die mit Sicherheit nicht KI-generiert sind, z. B., weil ihre Entstehung auf vor Ende 2022 datiert werden kann. Ihr gehäuftes Auftreten

erlaubt selbst ohne Täuschungsvorwurf die Vergabe einer schlechten Zensur oder sogar ihre Zurückweisung mit dem Prädikat „mangelhaft“. Finden sie sich aber in Kombinationen, die KI-typisch sind (formal fehlerfreier Text, KI-Halluzinationen, zahlreiche Closer, fehlende direkte Zitate etc.), ist der Verdacht durchaus gerechtfertigt. Kann man ihn dann noch mit einer Analyse durch ein gutes AI-Detection Tool erhärten, ist er auch unbedingt auszusprechen: Eine eigenständig verfasste Arbeit liegt nicht vor; durchaus kann von einem wissenschaftlichen Fehlverhalten gesprochen werden (vgl. Hoeren 2023, 23 sowie ausführlich 30-32).

Fazit

Der Beitrag konnte zeigen, dass Prüfer/-innen gegenüber KI-generierten Prüfungslösungen keineswegs machtlos sind. Eine händische Analyse in Kombination mit einem guten AI Detection Tool erlaubt eine sehr plausible Einschätzung, dass eine Arbeit unter Verwendung eines nicht angegebenen Hilfsmittels entstanden ist. Dabei ist jedoch in Rechnung zu stellen, dass sich sowohl KI-gestützte Textgeneratoren als auch AI Detection Tools stetig fortentwickeln werden.

Wünschenswert wäre freilich, dass solche Arbeiten erst gar nicht auf den Schreibtischen der Prüfenden landen. Insofern ist den vielen Stimmen unbedingt zuzustimmen, die eine vermehrte Aufklärungsarbeit an Schulen und Hochschulen zu einem verantwortungsvollen Umgang mit den neuen Writing Tools einfordern. Was hier zu vermitteln wäre, haben etwa Salden et al. (2023, 13) wie folgt zusammengefasst:

- Studierende können die Funktionsweise KI-basierter Schreibtools erklären sowie beurteilen, welche Möglichkeiten und Grenzen die Unterstützung durch solche Tools hat.
- Studierende können KI-basierte Schreibtools im Sinne von Schreibassistenzsystemen für ihre akademische Textproduktion einsetzen, so dass diese auf Wunsch bei der Formulierung, bei der Schärfung von Ideen und Argumentation sowie bei der stilistischen Überarbeitung unterstützen. Dies beinhaltet die Kompetenz, den Textoutput von KI-Schreibwerkzeugen zu reflektieren, zu redigieren und in eigene Textstrukturen zu integrieren.
- Studierende können die rechtlichen Rahmenbedingungen zum Einsatz KI-basierter Schreibtools (z. B. Kennzeichnungspflichten) erklären und im Rahmen der eigenen Textproduktion anwenden.

Zu ergänzen wäre im Sinne der Transparenz, dass Schüler/-innen und Studierende expressis verbis auf die prinzipielle Möglichkeit zur Detektion von Verstößen hinzuweisen sind. AI-Writing Tools können für das schulische und wissenschaftliche Schreiben von großem Nutzen sein; sie können aber auch zum eigenen Schaden eingesetzt werden.

Anmerkungen

1. Urheberrechtliche Erklärung für schriftliche Arbeiten an der Universität Bremen, Fachbereich10; https://www.uni-bremen.de/fileadmin/user_upload/fachbereiche/fb10/fb10/Formulare/Selbststaendigkeitserklaerung_Hausarbeit.pdf

2. Das Rechtsgutachten von Hoeren (2023, 23 u. 38) empfiehlt, einen entsprechenden Passus in urheberrechtliche Erklärungen expressis verbis aufzunehmen, auch wenn das nicht unbedingt nötig wäre. Dem schließt sich der Autor an.

3. Die Annahme einer stetigen Verbesserung von AI Detection Tools steht unter dem Vorbehalt, dass umgekehrt KI-gestützte Textgeneratoren deren Arbeitsweise nicht systematisch zu torpedieren versuchen (vgl. Sadasivan et al. 2023, 21). Derzeit scheint keines der auf dem Markt befindlichen AI Writing Tools darauf trainiert worden zu sein, die Herkunft von KI-generiertem Text bewusst verschleiern zu wollen. Denkbar ist das aber und man müsste dann mit einem digitalen Wettrüsten rechnen, wie es zwischen Viren-Entwicklern/Entwicklerinnen und Programmieren/Programmierinnen von Antivirus-Software stattfindet.

4. Nach Angaben des Herstellers werden dafür etwa 400 Wörter benötigt. Zum Testen können Anfragen mit 2000 Tokens (nicht übereinstimmend mit Wörtern) pro Server getätigt werden. Wechselt man den Server z. B. via VPN sind entsprechend mehr Anfragen möglich. Im kostenpflichtigen Angebot sind Anfragen im Umfang von 4000 Tokens erlaubt. Ganze Hausarbeiten können also nicht übergeben werden; man wird sich auf Stichproben beschränken. Die Preisgestaltung ist moderat: Es gibt kein Abo, sondern frei wählbare Creditvolumen; pro Wort werden ca. 0.08 Euro berechnet. Leider ist der Server mittlerweile stark ausgelastet, sodass man des Öfteren recht lange Response-Zeiten und Time Outs in Kauf nehmen muss. Bezüglich des Datenschutzes findet man keine Angaben auf den Seiten des Betreibers. Ob und wie eingegebene Texte gegen eine Weitergabe geschützt sind, blieb daher unklar. Auch weiß man nicht, wo die verwendeten Server stehen. Als europäisches Unternehmen ist der Betreiber eigentlich verpflichtet, die Datenschutzrichtlinien der Europäischen Union einzuhalten.

5. Getestet wurden hier wie im Folgenden fast ausschließlich deutschsprachige Texte. Bislang wurden in größeren Studien m.W. nur englische Texte mit Detection Tools untersucht.

6. Vgl. https://www.fanfiktion.de/s/5b99312d0009d450d6ca693/1/Das-Komische-parket-Graf-Bleck

7. Das hier verwendete Material stammt aus einer menschlich nachbearbeiteten ChatGPT 3.5-Antwort auf eine Anfrage zur strafrechtlichen Beurteilung des Falls „Emilia Galotti." Auch dieses Experiment war im Hinblick auf die Fähigkeiten des Tools interessant. Übergeben wurden alle strafrechtlich möglicherweise relevanten Fakten aus Gotthold Ephraim Lessings gleichlautendem bürgerlichen Trauerspiel. Der diesbezügliche Prompt lautete: „Angenommen, der Fall käme vor ein deutsches Strafgericht. Wer der Beteiligten würde für welches Vergehen gemäß welchen Paragraphen bestraft und wie hoch würde

wohl in etwa das Strafmaß sein?“. Das System lieferte eine für Laien durchaus plausible strafrechtliche Einschätzung. Eine Beurteilung der Ausführungen durch zwei Juristinnen und einen Juristen deckte indes erhebliche Mängel auf. Eine Replikation der im März 2023 durchgeführten Abfrage war im November 2023 nur sehr beschränkt möglich: ChatGPT 3.5 erläuterte vorweg: „Ich bin kein Rechtsanwalt, aber ich kann versuchen, einige rechtliche Aspekte des Falls zu skizzieren. Beachten Sie bitte, dass meine Antwort nicht als professionelle rechtliche Beratung betrachtet werden sollte. [...] Die genauen Paragraphen des deutschen Strafrechts hängen von den Umständen ab und könnten von einem Anwalt genauer bestimmt werden.“. Offensichtlich sind hier in der Zwischenzeit durch die Entwickler weitere Vorsichtmaßnahmen implementiert worden.

8. Das Phänomen lässt sich auch für andere Sprachen und Kulturen replizieren. Articel I, Section 1-3 der Verfassung der Vereinigten Staaten (The Constitution of the United States) bekam einen Perplexity Score von 2, der (besonders lange) Artikel 7 der französischen Verfassung (Constitution) wurde mit Perplexity 3 ebenfalls unter den Verdacht gestellt wurde, KI-generiert zu sein. Eine Ausnahme bildeten die ersten acht Artikel der Irischen Verfassung in englischer Sprache (Constitution of Ireland) mit einem Perplexity Wert von 9. Dagegen wurde die offizielle deutsche Übersetzung auf den Seiten der EU (https://www.verfassungen.eu/irl/verf37-i.htm) als „Likely AI Generated“ mit 5 bewertet. Weil hier noch alte Rechtschreibung verwendet worden ist, handelt es sich mit Sicherheit um eine menschliche Übersetzung. Eine Erklärung für das Phänomen zu finden, ist schwierig: Es liegt nahe, Trainingseffekte zu vermuten, denn zu den Texten, mit denen ChatGPT trainiert worden ist, gehörten mit Sicherheit auch Rechtsnormen. Allerdings werden sie nur einen Bruchteil des millionenschweren Korpus ausgemacht haben und bei keiner anderen Textsorte konnte der Autor einen derart systematischen Fehler ausmachen. Hier gibt es also Forschungsbedarf!

9. Bei fiktionalen Texten wird der Closer durch die KI nur am Ende der Geschichte gesetzt. Er enthält dann fast immer moralische Botschaften, die man der Erzählung entnehmen soll. Der Struwwelpeter ist digital wieder auferstanden!

LITERATUR

Bender, Emily et al. (2021): „On the Dangers of Stochastic Parrots: Can Language Models be too big?“ In: *FaccT 21 – Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623, https://dl.acm.org/doi/10.1145/3442188.3445922.

Hoeren, Thomas (2023): Rechtsgutachten zum Umgang mit KI-Software im Hochschulkontext. In: Salden/ Leschke (Hg.), 22-40.

Sadasivan, Vinu Sankar et al.: (2023): Can AI-Generated Text be Reliably Detected? In: *Department of Computer Science University of Maryland*, arXiv:2303.11156, https://doi.org/10.48550/arXiv.2303.11156.

Salden, Peter/ Leschke, Jonas (Hg.) (2023): *Didaktische und rechtliche Perspektiven auf KI-gestütztes Schreiben in der Hochschulbildung. Zentrum für Wissenschaftsdidaktik der Ruhr-Universität Bochum.* https://doi.org/10.13154/294-9734.

Salden, Peter/ Lordick, Nadine/ Wiethoff, Maike (2023): KI-basierte Schreibwerkzeuge in der Hochschule: Eine Einführung. In: Salden/ Leschke (Hg.), 4-21.

Schindler, Kirsten (2023): Chatgpt oder Überlegungen zu den Veränderungen des Schreibens in der Schule. In: *MiDu* 2, DOI:10.18716/OJS/MIDU/2023.2.5.

Weber-Wulff, Deborah et al. (2023): Testing of Detection Tools for AI-Generated Texts. In: *International Journal of Educational Technology. Higher Education,* DOI:10.48550/arXiv.2306.15666.

Impressum

Herausgeber: Virtuelles Kompetenzzentrum – Schreiben lehren und lernen mit Künstlicher Intelligenz (VK:KIWA)

Redaktion: Kirsten Schindler & Nicolaus Wilder

Satz: Nicolaus Wilder

DOI: 10.5281/zenodo.10396374