

Leitlinie zur Nutzung von Künstlicher Intelligenz in der Begutachtung

Präambel

Die wissenschaftliche Begutachtung von Förderanträgen ist ein integraler Bestandteil zur Ermittlung wissenschaftlicher Qualität im Rahmen der Forschungsförderung durch die Deutsche Forschungsgemeinschaft (DFG). (Generative) Künstliche Intelligenz (im Weiteren „KI-Systeme“) ist inzwischen weithin verfügbar und kann ohne besondere technische Vorkenntnisse eingesetzt werden. Mit Blick auf schützenswerte rechtliche Interessen der antragstellenden Personen, die Vertraulichkeit des Begutachtungsverfahrens, allgemeine Transparenzanforderungen sowie die wissenschaftliche Integrität der DFG-Förderverfahren insgesamt ist der Einsatz von KI-Systemen ausschließlich entlang der im Weiteren näher ausgeführten Prinzipien zulässig.

Diese Nutzungs- und Transparenzleitlinie zielt darauf ab, für Gutachter*innen beim Einsatz von KI-Systemen in der Begutachtung Rahmenbedingungen zu setzen und den verantwortungsvollen, transparenten und – u. a. mit Blick auf die KI-Verordnung (EU) 2024/1689, die Datenschutz-Grundverordnung (DSGVO) und das Urheberrechtsgesetz (UrhG) – rechtskonformen Einsatz von KI-Systemen zu gewährleisten. Die inhaltliche Verantwortung für das Gutachten verbleibt bei den Gutachter*innen.

Einsatzzweck

Die menschliche Analyse und wissenschaftliche Einordnung des Antrags ist Kernbestandteil der Begutachtung. Die Nutzung von KI-Systemen bei der Erstellung von Gutachten ist daher ausschließlich im unterstützenden Sinn gestattet. Alle inhaltlichen Positionierungen mit Blick auf die Förderempfehlung obliegen allein den Gutachter*innen.

*So darf das KI-System zum Beispiel relevante Literatur mit Blick auf den zu begutachtenden Antrag heraussuchen, eigene Stichworte oder Notizen der Gutachter*innen, aus denen die intentionale Richtung des Gutachtens bereits eindeutig hervorgeht, zu einem ausformulierten Fließtext zusammenführen oder Anregungen zur strukturellen oder sprachlichen Überarbeitung eigener Inhalte liefern.*

Unzulässig wäre es hingegen, maßgebliche Teile des Gutachtens automatisch und unreflektiert, d. h. ohne eigene fachliche Prüfung, generieren zu lassen.

Leitende Prinzipien bei der Verwendung von KI

1. Vertraulichkeit

KI-Systeme dürfen nur zum Einsatz kommen, wenn sichergestellt ist, dass die vertraulichen Inhalte von Förderanträgen nicht unkontrolliert verarbeitet, gespeichert oder für andere Zwecke als die Durchführung der wissenschaftlichen Begutachtung verwendet werden. Die unterstützende Verwendung von KI im Rahmen der Gutachtenerstellung darf nicht zur Preisgabe vertraulicher Antragsinhalte führen. Diese Informationen werden der DFG ausschließlich zur Begutachtung, Bewertung und Entscheidung überlassen.

Insbesondere bei cloudbasierten KI-Systemen kann es bei der Eingabe zu einer nachgelagerten Verarbeitung zu Trainingszwecken kommen. Diese ist nicht vom ursprünglichen Zweck der Überlassung durch die antragstellenden Personen gedeckt und daher unzulässig.

*Daher dürfen die betroffenen Antragsinhalte von Gutachter*innen nur in KI-Systeme eingespeist werden, wenn sichergestellt ist, dass die Antragsinhalte von den Anbietern des KI-Systems nicht dauerhaft und über den konkreten Einsatzzweck hinaus gespeichert werden. Hierfür muss eine der folgenden Anforderungen erfüllt sein:*

- *Das KI-System ist auf dem eigenen Endgerät installiert und die Prozessierung erfolgt ausschließlich auf dem Endgerät ohne Cloud-Computing oder der Übertragung von Verarbeitungsergebnissen in die Cloud.*
- *Das KI-System wird von einer vertrauenswürdigen Einrichtung (beispielsweise einer Hochschule oder öffentlichen Forschungseinrichtung) betrieben („gehostet“).*

- *Cloudbasierte nicht kommerzielle oder kommerzielle KI-Systeme (beispielsweise ChatGPT, Claude, Gemini) dürfen nur genutzt werden, wenn eine vertragliche Lizenzvereinbarung die Datensicherheit garantiert, eine Nutzung der eingegebenen Daten durch den Anbieter (etwa zu Trainingszwecken) ausgeschlossen ist und diese Voraussetzungen durch entsprechende Sicherheitseinstellungen (zum Beispiel „kein logging“, „keine Speicherung“) aktiv gesetzt wurden.*
- *Es wird empfohlen, solche KI-Systeme zu verwenden, die auf transparent dokumentierten und öffentlich verfügbaren KI-Modellen mit öffentlich einsehbaren Trainingsdaten beruhen.*

2. Transparenz

Wenn bei der Erstellung des Gutachtens ein KI-System verwendet wird, muss dies unter Nennung der spezifischen Einsatzzwecke offengelegt werden.

Der Einsatz des KI-Systems kann mit wenigen erläuternden Sätzen beschreiben werden. Eine konkrete Kennzeichnung der betroffenen Textpassagen im Gutachten ist nicht erforderlich. Unter „Offenlegung“ sind dabei Angaben zu verstehen, zu welchem der oben genannten zulässigen Einsatzzwecke und in welchem Umfang das KI-System eingesetzt wurde.

Eingesetzte KI, die sich nicht auf den (wissenschaftlichen) Inhalt des Gutachtens auswirkt (beispielsweise Grammatik-, Stil-, Rechtschreibprüfung, Übersetzungsprogramme), muss nicht angegeben werden.

3. Qualitätssicherung

Eine unreflektierte Übernahme KI-generierter Textteile ist unzulässig. KI-generierte Texte müssen stets kritisch gewürdigt, angepasst und auf inhaltliche Richtigkeit, Aktualität sowie mögliche Verzerrungen (Bias-Strukturen) hin überprüft werden.

Da generative KI-Modelle auf probabilistischen Verfahren basieren, können sie sachlich falsche, unausgewogene oder voreingenommene Inhalte produzieren. Die Modelle basieren auf öffentlich zugänglichen Trainingsdaten, die auf Grundlage von (stochastischen) Wahrscheinlichkeitsberechnungen gesellschaftliche, kulturelle oder auch wissenschaftliche Diskurse abbilden, ohne hierzu eine wissenschaftlich-inhaltlich begründete Position zu beziehen. Dadurch perpetuieren sie die in den Trainingsdaten enthaltenen Bias, etwa Geschlechter- oder Diversitätsbias. Mit Blick auf wissenschaftliche Erkenntnisprozesse können sie auf dem Etablierten aufsetzende Konzepte gegebenenfalls positiver bewerten als neuartige Forschungsansätze, die zu Paradigmenwechseln führen könnten (Status-quo-Bias). Insgesamt kann generative KI unzutreffende oder verzerrte Inhalte erzeugen. Ohne kritische Überprüfung durch einen Menschen („human control“) ist die wissenschaftliche Qualität des Gutachtens sowie die Integrität des DFG-Begutachtungssystems gefährdet. Bei schlüssig scheinenden KI-generierten Texten kann zudem „automation bias“ auftreten.

4. Verantwortung

Die inhaltliche Verantwortung für das Gutachten liegt uneingeschränkt bei der Person, die das Gutachten erstellt hat. Dies gilt auch dann, wenn Textteile unter Zuhilfenahme von KI-Systemen generiert worden sind. Eine verantwortliche wissenschaftliche Begutachtung erfordert fachliche Kompetenz und wissenschaftsethische Urteilsfähigkeit, die KI-Systeme systembedingt nicht mitbringen. Die Verantwortung darf daher nicht auf ein KI-System übertragen werden, sie verbleibt bei den Gutachter*innen.

*Diese Prinzipien werden im DFG-Kodex „[Leitlinien zur Sicherung guter wissenschaftlicher Praxis](#)“ konkret benannt. Leitlinie 16 stellt die Einhaltung der vertraulichen Behandlung fremder Inhalte, die zur Begutachtung oder Bewertung vorgelegt werden, als zentrales Element des wissenschaftlichen Urteilsbildungsprozesses heraus. Wissenschaftler*innen übermitteln unveröffentlichte wissenschaftliche Ideen, Forschungsdaten, Abbildungen und methodische Details, die oft das Ergebnis jahrelanger Arbeit sind und daher einem besonderen Schutzbedürfnis unterliegen. Der Schutz dieser Inhalte ist entscheidend, um die Rechte der Autor*innen zu wahren, das Vertrauen in einen wissenschaftsgeleiteten Begutachtungsprozess zu sichern und wissenschaftliches Fehlverhalten zu verhindern.*

*Leitlinie 16 benennt überdies die Neutralität von Begutachtungen als wichtige Grundlage für Glaubwürdigkeit, Fairness und Qualität in wissenschaftlichen Entscheidungsprozessen. Sie bildet die unverzichtbare Grundlage für einen fairen und transparenten Urteilsprozess in der Wissenschaft. Nur wenn Gutachter*innen ihre Positionierungen unabhängig, sachlich und unbeeinflusst abgeben, können Wissenschaft und Gesellschaft auf die Ergebnisse solcher Prozesse vertrauen.*

*Nach § 9 Nr. 2 der [DFG-Verfahrensordnung zum Umgang mit wissenschaftlichem Fehlverhalten](#) (VerfOWF) können Gutachter*innen die Vertraulichkeit verletzen und damit wissenschaftliches Fehlverhalten verwirklichen, wenn sie wissenschaftliche Inhalte, die ihnen zur Begutachtung übermittelt wurden, unbefugt an Dritte weitergeben. Dies kann gegebenenfalls auch durch Einspeisen in Large Language Models (LLMs) erfolgen, sofern die Inhalte als Trainingsdaten genutzt werden.*