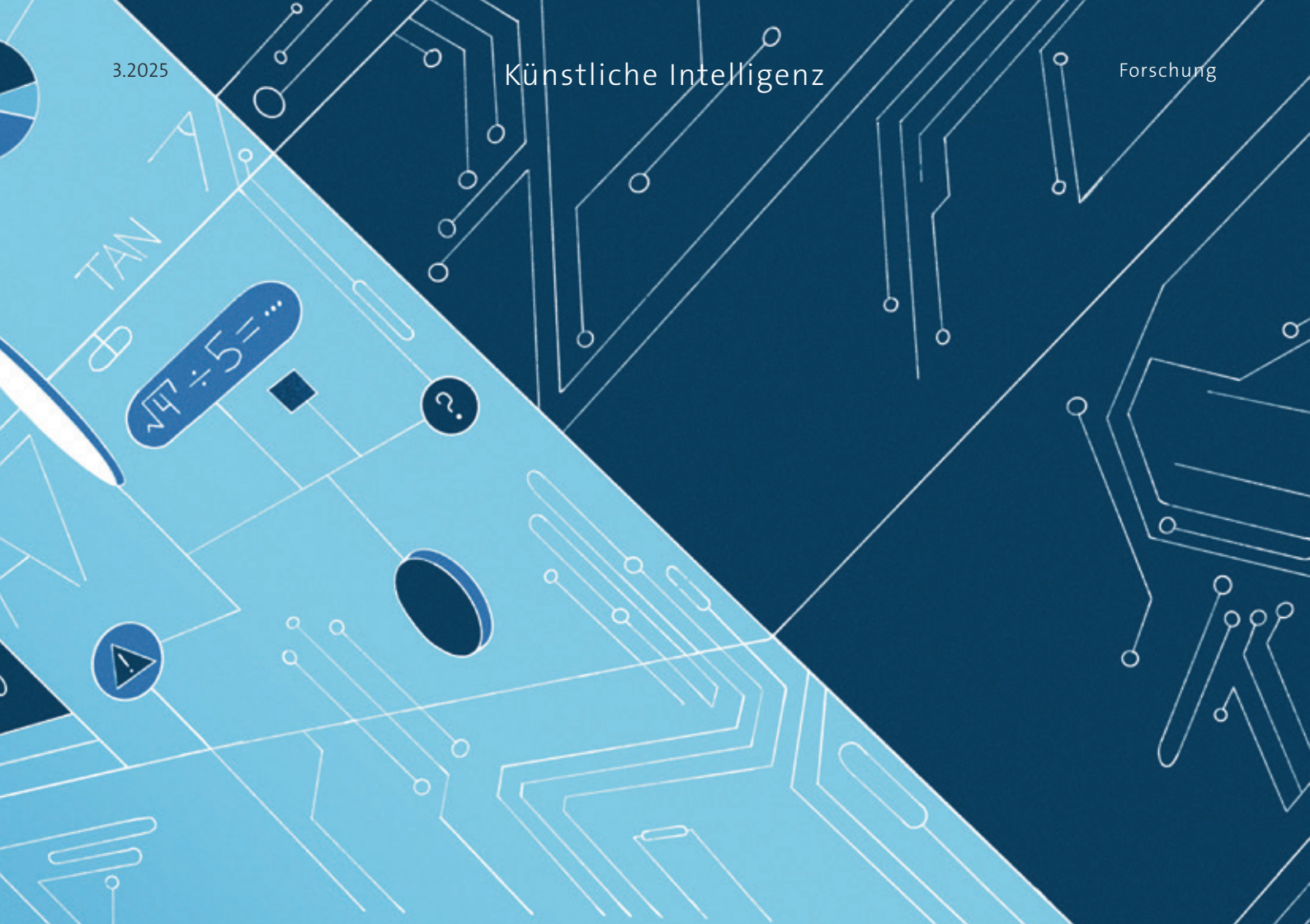
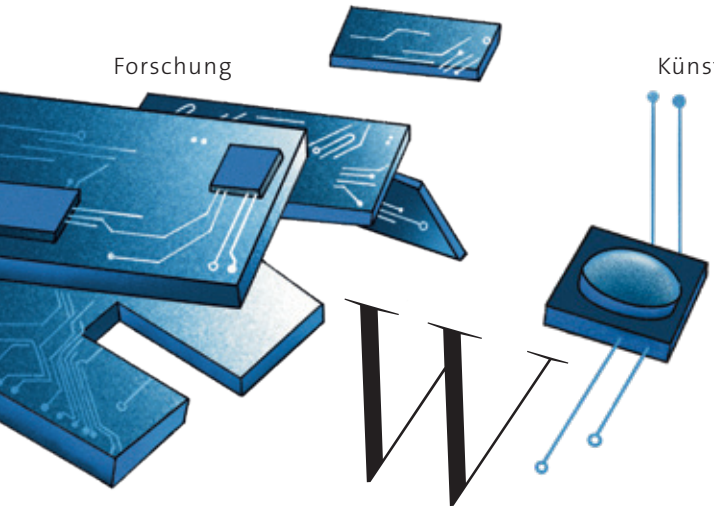
Alle Illustrationen: Natalia Drahun für Akademie Aktuell



Richtig prompten: Tipps für den Umgang mit Generativer KI

Von **Sonja Niemann**

ChatGPT, DeepSeek und Co. haben in vielen Berufsfeldern Einzug gehalten, auch in der Wissenschaft. Was bei der Verwendung zu beachten ist, und wie man **die besten Ergebnisse** erhält.



Wie viele Trends und Neuheiten erweitert auch Generative Künstliche Intelligenz (GenKI) unseren Wortschatz. Social Media hat „Feed", „Tweet", „Follow" und viele weitere Begriffe in den Sprachgebrauch gebracht, und Generative KI ergänzt ihn nun um „Prompt", „Zero- und Few-Shot", „Transformer" und andere Ausdrücke. „Prompt" lässt sich wörtlich mit Eingabeaufforderung übersetzen, das Wort beschreibt also die Interaktion mit sprachbasierter Generativer KI. Es gibt viel Forschung im Bereich Prompt Engineering, zugleich ist es fast unmöglich, jede Entwicklung genau mitzuverfolgen. Welches Modell ist für welche Aufgabe am besten geeignet? Was ist Zero- und Few-Shot Prompting? Und sind meine geteilten Informationen nun sicher oder nicht? Ziel dieses Artikels ist es, einen Überblick über die richtige Anwendung Generativer KI zu geben.

Die Grundlagen des Promptings

Egal, in welchem Anwendungsfeld – um gute Ergebnisse zu erzielen, sollten folgende grundlegende Fragen in einem Prompt beantwortet werden:

1. Welche Aufgabe soll übernommen werden?
2. Welche Rolle soll das gewählte Sprachmodell einnehmen?
3. Wie soll der Output, also das Ergebnis, aussehen?

Diese Fragen können als Leitlinien dienen, um den richtigen Prompt zu formulieren; das kann durchaus mehrere Wiederholungen in Anspruch nehmen. Frage 1 hilft, um zu ordnen, was man genau erledigt haben möchte. Generative KI-Modelle basieren auf Wahrscheinlichkeiten: Stark vereinfacht bedeutet das, dass die wahrscheinlichste Antwort auf die Frage berechnet wird. Was die sogenannten Transformer-Modelle wie ChatGPT und Co. besonders leistungsfähig gemacht hat, ist, dass sie die Reihenfolge der Wörter als relevanten Faktor für die Berechnung der Antwort miteinbeziehen.

Die Konsequenz für Prompt-Anfragen: Es ist empfehlenswert, so exakt wie möglich zu beschreiben, was erledigt werden soll, und genau über den Prompt nachzudenken. Ein einfaches Beispiel: Frage 1 könnte in einem Prompt so beantwortet werden: „Schreibe eine Einleitung für einen Artikel zum Thema Resozialisierung von Strafgefangenen." Etwas mehr Kontext kann hinzugefügt werden, indem man genauer beschreibt, wo der Artikel veröffentlicht werden soll, z. B. in einem Publikumsmagazin oder einer wissenschaftlichen Fachzeitschrift.

Frage 2 soll klären, aus welcher Perspektive die Antwort gegeben wird. Dem KI-Modell kann man z. B. etwas vorgeben wie „Du bist eine Wissenschaftlerin im Bereich Sozialpsychologie, die Erkenntnisse aus der Forschung allgemeinverständlich darstellen kann."

Schließlich neigen einige Sprachmodelle zu sehr langen Antworten. Um dies zu vermeiden oder spezielle Wünsche für die Form des Outputs zu berücksichtigen, lohnt es sich, Frage 3 zu beantworten. Auch hier kann man unterschiedlich konkret werden – von „Halte dich kurz" bis hin zu „Gib zwei Absätze aus. Im ersten Absatz geht es um die aktuelle Relevanz des Themas, im zweiten um die weitere Struktur des Artikels."

Von Prompt Engineering bis Chain of Thought

Die vorgestellten Fragen können erweitert und verbessert werden – je nach Aufgabenfeld haben sich verschiedene Vorgehensweisen entwickelt. Mit Prompt Engineering hat sich ein Begriff etabliert, der komplex klingt. Die Begriffe Zero-, One- oder Few-Shot Prompting beschreiben nichts anderes als die Anzahl der Beispiele, die dem Sprachmodell mit dem Prompt übermittelt werden. Für einen Artikel oder eine Einleitung können das eine (One-Shot) oder mehrere (Few-Shot) Referenzen sein. Benötigt man wiederum Code, der Daten sortiert, können Input- und Output-Beispiele gegeben werden. Dieses Vorgehen lohnt sich besonders für komplexere Aufgaben.

Ein weiterer Begriff, der häufiger verwendet wird, ist Chain of Thought (CoT) Prompting. Die Idee dabei ist, den Prompt an menschliche Gedankenketten anzupassen, also kleinere Teilprobleme zu formulieren, die der Reihe nach abgearbeitet werden können. Die einfachste Form ist es, am Ende des Prompts das Modell aufzufordern, in Schritten zu denken. Weitere Variationen des CoT-Promptings schreiben die einzelnen Schritte im

Die Nutzung Generativer KI kann nicht losgelöst von der Problematik des Datenschutzes betrachtet werden.

Prompt aus. Manche gehen so weit, eine bestimmte Struktur für diese Abfolge vorzugeben, was als Structured Chain of Thought (SCoT) bezeichnet wird. Für einige Bereiche kann sich dieser Prozess bis zu einem guten Prompt auszahlen, etwa, wenn man statistisches Verständnis hat, aber zum ersten Mal das Programm R anstatt SPSS ausprobieren möchte. Um CoT- oder SCoT-Prompts zu schreiben, braucht es nicht nur eine konkrete Vorstellung, welche Aufgabe genau übernommen werden soll, sondern auch ein inhaltliches Verständnis. Am Ende ist es wichtig, selbst beurteilen zu können, ob der Output richtig ist und qualitativ für den Anwendungszweck ausreicht, denn: Generative KI macht nach wie vor Fehler!

Generative KI in der Wissenschaft: Darauf sollten Forschende achten

Setzt man Generative KI ein, ist nicht nur das Ergebnis entscheidend, es spielen auch ethische Aspekte und Datenschutzfragen eine Rolle. Allein in der Forschung gibt es eine Vielzahl von möglichen Anwendungsfällen. Bei fast allen besteht das Problem der Replizierbarkeit, die Frage nach der eigenen Leistung und die Sorge, dass innovative Ideen weitergegeben werden. Einige Modelle eignen sich besser für bestimmte Aufgaben; so verfügen unter anderem OpenAI, Mistral oder DeepSeek über spezielle Varianten, die durch weiteres Training auf Programmcode-Generierung spezialisiert sind. Möchte man also eine spezifische Aufgabe lösen, die über Textgenerierung hinausgeht, lohnt es sich, einige Modelle zu vergleichen. Die meisten Modelle bieten auch die Möglichkeit, ein eigenes Fine Tuning vorzunehmen, beispielsweise, um Textannotationen erstellen zu lassen. Fine Tuning beschreibt generell einen Prozess, bei dem ein vorhandenes Modell gezielt für eine spezifische Aufgabe weitertrainiert wird.

Die Nutzung Generativer KI kann nicht losgelöst von der Problematik des Datenschutzes betrachtet werden, denn: Werden ChatGPT, DeepSeek und Co. online genutzt, werden Daten außerhalb der EU weiterverarbeitet und gegebenenfalls auch zum Training neuer Modelle verwendet. Das stellt vor allem dann ein Problem dar, wenn sensible Daten verarbeitet werden, seien es innovative Ideen aus der eigenen Forschung, Daten von Studienteilnehmenden oder Forschungsanträge. Die einfachste Lösung können API-Zugänge sein. Die Server der Anbieter werden dann über eine Code-Schnittstelle angesprochen anstatt über das Online-Interface. Dies macht es auch einfacher, Daten in größeren Mengen automatisiert an das Modell zu senden. OpenAI verspricht z. B., dass Daten, die über eine API-Schnittstelle zu den Servern gelangen, nicht gespeichert oder für Trainingszwecke verwendet werden. Weitere Datenschutzprobleme können Nutzende umgehen, wenn sie die Modelle lokal ausführen. Einige Modelle, etwa von Llama (Meta) und DeepSeek, sind frei verfügbar, man kann sie theoretisch lokal laufen lassen, also auf dem privaten Rechner oder Server ausführen und somit die Kommunikation zu den Servern der Anbieter vermeiden.

Zwei Faktoren stellen jedoch oft einen Engpass dar: die Rechenkapazität und die Modellgröße. Generative KI-Modelle haben für jede Ausgabe einen enorm hohen Rechenaufwand. Damit eine Antwort nicht zu lange dauert, werden einige Prozesse gleichzeitig ausgeführt. Nicht jede Art von Hardware kann diese parallelen Berechnungen ausführen, entscheidend hierfür ist ein Grafikprozessor (engl. graphic processing unit – GPU). Nicht jeder Rechner hat jedoch einen Grafikprozessor, und auch bei den Rechnern mit solcher Hardware gibt es unterschiedlich leistungsstarke Optionen. Der zweite Faktor ist die Modellgröße. Hinter den genauen Modellbezeichnungen ist meist eine Parameter-Angabe zu finden, zum Beispiel Llama 3.3 - 70B. 70B steht für 70 Billionen Parameter. Je größer die Zahl der Parameter, desto größer sind das Modell und der Rechenaufwand, aber desto besser ist auch das Ergebnis. Wer Modelle lokal ausführen möchte, muss also in Erfahrung bringen, welche Rechenkapazität vorhanden ist und wie groß das Modell ist, das benutzt werden soll.

Es gibt weitere Anwendungsmöglichkeiten, bei denen keine sensiblen Daten weitergegeben werden müssen. Immer häufiger wird Generative KI z. B. beim Erstellen von Personas für Studien genutzt, um Beschreibungen zu vereinheitlichen oder Bilder für die Personas zu erstellen. Dabei kann man aus vielen verschiedenen Stilen auswählen und hat den Vorteil, einzelne Merkmale verändern zu können, statt die ganze Persona auszutauschen, wie es häufig bei Stockfotos der Fall ist. Aber: Bildgenerierung ist aktuell nur mit einem sehr kleinen täglichen Limit kostenlos.

Fazit: Klare Erwartungen – besseres Ergebnis

Wer Generative KI einsetzen will, muss sich vorab ein paar Gedanken machen, um ein gutes Ergebnis zu erzielen. Das fängt bei der Auswahl des Modells an, geht über Restriktionen wie etwa Datenschutzaspekte und endet bei der Formulierung des Prompts. Ob Nutzende sich tiefer mit dem Prompt Engineering vertraut machen und den perfekten SCoT-Prompt ausarbeiten oder sich nur gut überlegen, welche Aufgabe sie erledigen möchten, und ausreichend Kontextinformationen an das Sprachmodell übergeben, bleibt jeder und jedem selbst überlassen.

Sonja Niemann
ist wissenschaftliche Referentin in der Abteilung Forschung am Bayerischen Forschungsinstitut für Digitale Transformation (bidt) der Bayerischen Akademie der Wissenschaften. Sie arbeitet im Forschungsschwerpunkt „Mensch und generative KI: Trust in Co-Creation“ insbesondere zum Thema Code-Generierung. Mehr dazu unter: bidt.digital/generative-ki