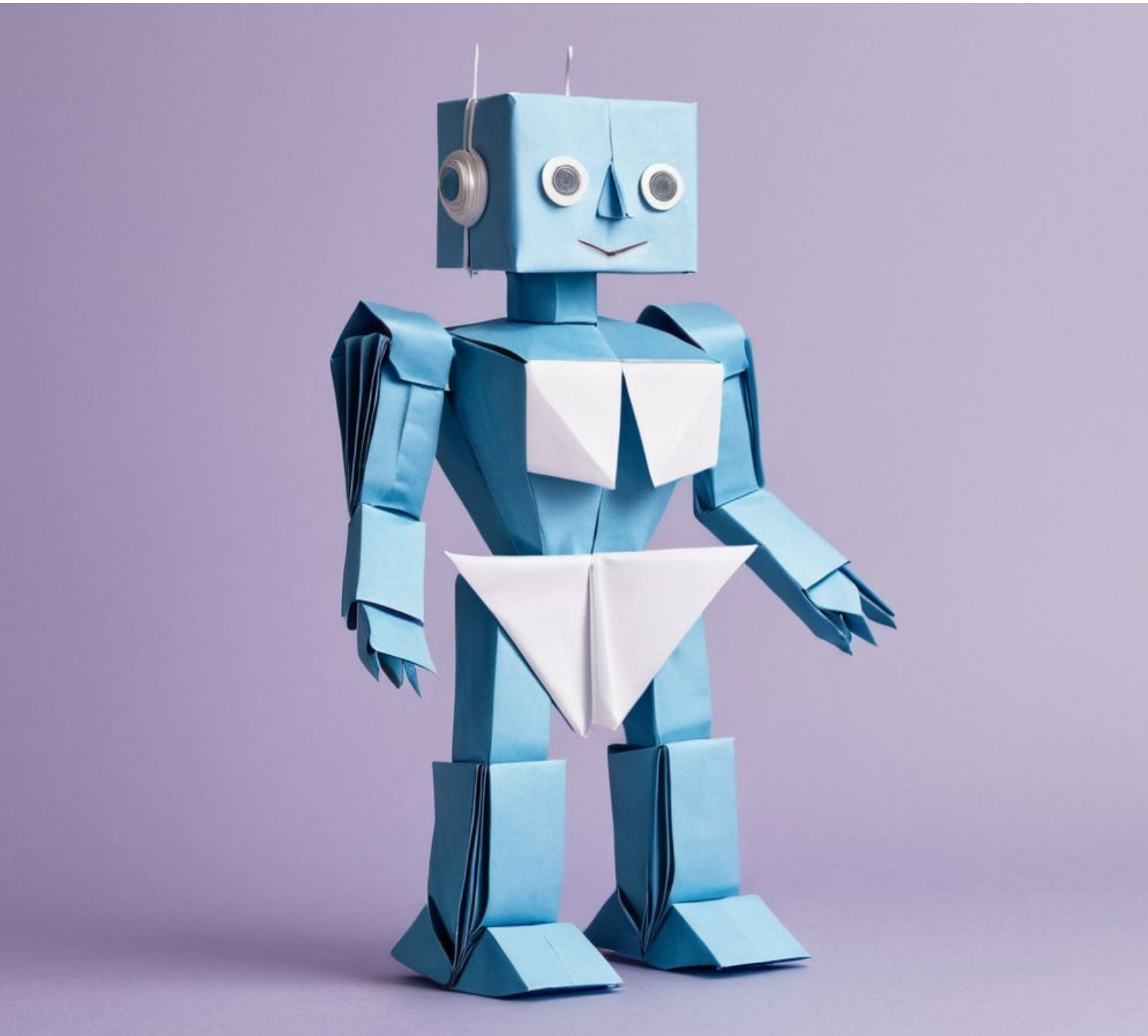
FZI

— Generative KI und ihr Einsatz im beruflichen Umfeld



– Inhaltsverzeichnis

1 Einleitung 3
2 Rechtliche Aspekte 5
2.1 Urheberrecht 5
2.2 AI Act 6
2.3 Datenschutz und Compliance 7
2.4 Lizenzbestimmungen und Nutzungsbedingungen 7
2.5 Wasserzeichen in GKI-Ergebnissen 8
3 Wichtige Einschränkungen 9
3.1 Begrenztes Kontextverständnis 9
3.2 Mangelnde Transparenz 9
3.3 Bias oder „Vorurteile“ 9
3.4 Halluzinationen 9
4 Klärung funktionaler Anforderungen 10
4.1 Skalierbarkeit und Latenz 10
4.2 Interpretierbarkeit 10
4.3 Evaluierung und Überwachung 10
4.4 Datenschutz und Compliance 10
5 Technische Umsetzung und Effizienz 11
5.1 Einsatz einer lokalen GPU-Instanz 11
5.2 Einsatz einer fremdbetriebenen GPU-Instanz 11
5.3 Nutzung von APIs 11
5.4 Ökologische und ökonomische Betrachtung 11
5.4.1 Energiebedarf 12
5.4.2 Hardwarebedarf 12
5.4.3 Datenbedarf 12
6 Zusammenfassung und Checkliste 13
7 Changelog 14
8 Kontakt 15

1 Einleitung

Die rasante Entwicklung von Künstlicher Intelligenz (KI) hat in den letzten Jahren eine innovative Klasse von Systemen hervorgebracht, die als generative KI (GKI) bekannt ist. Diese neuartige Technologie zeigt das Potenzial, nicht nur bestehende Arbeitsweisen zu optimieren, sondern auch die Grundlagen unserer Interaktion mit Computern und Maschinen grundlegend zu verändern. Generative KI geht über die traditionelle Anwendung von Künstlicher Intelligenz hinaus, indem sie in der Lage ist, eigenständig Inhalte und Informationen zu erstellen, ohne auf vordefinierte Muster oder Datensätze beschränkt zu sein. Diese revolutionäre Eigenschaft eröffnet eine Fülle von Möglichkeiten für den beruflichen Kontext, die von der Automatisierung repetitiver Aufgaben bis hin zur kreativen Generierung von Inhalten reichen.

Warum ist generative KI so revolutionär?

Generative KI markiert einen Wendepunkt in der Geschichte der Künstlichen Intelligenz, da sie die Fähigkeit besitzt, kreativ zu denken und neuartige Lösungen zu generieren. Im Gegensatz zu herkömmlichen KI-Systemen, die auf Trainingsdaten basieren und Muster erkennen, ist GKI in der Lage, eigenständig Inhalte zu erzeugen, die über das Gelernte hinausgehen. Diese Autonomie ermöglicht nicht nur eine verbesserte Anpassung an spezifische Anforderungen im beruflichen Umfeld, sondern fördert auch die Entstehung von Innovationen, die durch traditionelle Methoden nur schwer realisierbar wären.

Ein weiterer entscheidender Aspekt der Revolution liegt in der Flexibilität von generativer KI. Diese Systeme sind in der Lage, verschiedene Aufgaben und Szenarien abzudecken, was eine breite Palette von Anwendungen im beruflichen Kontext ermöglicht. Von der Erstellung von Texten über die Gestaltung von Grafiken bis hin zur Entwicklung von Softwarecode – generative KI revolutioniert die Art und Weise, wie wir Arbeit verstehen und durchführen.

In diesem White-Paper werden wir die fundamentale Bedeutung von generativer KI im beruflichen Kontext erkunden und dabei die verschiedenen Anwendungsbereiche, Herausforderungen und potenziellen Vorteile dieser bahnbrechenden Technologie beleuchten.

Wer etwas Erfahrung mit Texterzeugnissen einer generativen KI (GKI) besitzt, ahnt es bereits: „Autor“ dieser Einleitung ist kein Mensch, sondern eine generative KI. Im vorliegenden Fall hat ChatGPT eine Erklärung seiner selbst geliefert und die besonderen Eigenschaften einer GKI aufgezeigt. Hierzu war lediglich eine knappe Anweisung erforderlich:

>_ Ich möchte ein White Paper zum Thema generative KI und Einsatz von GKI im beruflichen Kontext schreiben. Bitte schreibe eine knappe Einleitung zum Thema generative KI und erkläre, warum das Thema so revolutionär ist.

Auch wenn GKI mittlerweile beeindruckende Leistungen hervorbringt, sollte der Einsatz dieser Werkzeuge – insbesondere im beruflichen Umfeld – mit Bedacht erfolgen. Einfache Bedienung und auf den ersten Blick oft überraschend gute Ergebnisse täuschen leicht darüber hinweg, dass die Nutzung von GKI auch mit Problemen und Fallstricken verbunden ist.

In diesem White Paper sollen daher neben dem Potenzial der GKI auch die damit verbundenen Herausforderungen betrachtet werden.

- Rechtliche Aspekte
- Wichtige Einschränkungen
- Klärung funktionaler Anforderungen
- Technische Umsetzung und Effizienz

Dieses White Paper soll Ihnen einen kurzen Überblick über das komplexe Themengebiet der GKI vermitteln und die Herausforderungen aufzeigen. Sprechen Sie uns an, wenn Sie GKI für Produkte oder Prozesse in Ihrem Unternehmen einsetzen möchten. Wir beraten und unterstützen Sie dabei, GKI gewinnbringend in Ihr Unternehmen zu integrieren.

2 Rechtliche Aspekte

Die rapide Entwicklung von GKI lässt auch das Recht nicht unberührt. Im Folgenden werden die Herausforderungen und Chancen erörtert, die sich im Spannungsfeld zwischen Künstlicher Intelligenz und Recht ergeben. Der Abschnitt wird kontinuierlich an die sich wandelnden rechtlichen Rahmenbedingungen angepasst.

2.1 Urheberrecht

Das Urheberrecht stellt nicht das Werk selbst, sondern den Schöpfer eines Werkes in den Mittelpunkt (Schöpferprinzip). Laut § 7 UrhG ist der Schöpfer eines Werkes Urheber. Das Urheberrecht schützt den Urheber, seine geistige und persönliche Beziehung zum Werk und seine Verwertungsinteressen. Eine grundlegende Voraussetzung, dass ein Werk dem Urheberrecht unterliegt, ist dessen Einordnung als persönliche geistige Schöpfung (§ 2 Abs. 2 UrhG). Nur dann genießt das Werk urheberrechtlichen Schutz.

2.1.1 Training einer KI

Die Problematik im Zusammenhang mit dem Training einer GKI besteht nicht darin, dass die GKI urheberrechtlich geschützte Werke „sieht“, sondern in der Speicherung der Trainingsdaten sowie der damit einhergehenden Vervielfältigung. Dies stellt einen Verstoß gegen das ausschließlich dem Urheber zustehende Vervielfältigungsrecht dar (§§ 15 Abs. 1 Nr. 1, § 16 UrhG).

Eine Möglichkeit, urheberrechtlich geschütztes Material dennoch für das Training zu verwenden, besteht in der Beschaffung einer Lizenz. Dies dürfte sich allerdings bei einer großen Datenmenge schwierig gestalten. Eine weitere Option bietet die Schranke des § 44b UrhG, die Data und Text Mining regelt. Text und Data Mining meint die „automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen“ (§ 44b Abs. 1 UrhG).

Bedingung für das Data Mining ist, dass der Urheber dieser Nutzung nicht explizit in maschinenlesbarerer Form widerspricht und die Löschpflichten eingehalten werden. Die Kennzeichnung von online zugänglichen Werken mit einem Nutzungsvorbehalt durch den Rechteinhaber ist allerdings mit Schwierigkeiten verbunden. Es empfiehlt sich daher, potenzielle Trainingsdaten mit besonderer Sorgfalt zu erheben.

2.1.2 Prompts

Wie allgemeine Texte können auch diese Eingaben und Anweisungen an eine GKI die Hürde des Urheberrechts überschreiten. Dies kann beispielsweise der Fall sein, wenn die formulierten Prompts über eine reine Arbeitsanweisung hinausgehen und durch ein besonderes Maß an Individualität und Kreativität gekennzeichnet sind. Ein abgeleitetes Recht an dem generierten Erzeugnis der GKI ergibt sich hieraus jedoch nicht.

Des Weiteren ist zu berücksichtigen, dass Arbeitsanweisungen an eine GKI nicht nur mittels selbstverfasster Texte möglich sind, sondern auch mithilfe von fremden Bildern oder Texten. In diesem Fall ist zu kontrollieren, ob die erforderlichen Nutzungsrechte vorliegen.

2.1.3 Output

KI-generierte Erzeugnisse genießen nach dem gegenwärtigen Stand keinen urheberrechtlichen Schutz. Das grundrechtlich fundierte Urheberrecht ist ein Schutzrecht für menschlich-geistiges Schaffen (Wandtke/Bullinger, § 7 Rn. 18). Grundsätzlich sind die Erzeugnisse einer GKI daher gemeinfrei.

Die Imitation eines besonderen Stils durch eine GKI ist rechtlich nicht relevant (Wandtke/Bullinger, Urheberrecht, § 2 Rn. 40). Problematisch ist es jedoch, wenn der Output einer GKI konkrete urheberrechtlich geschützte Werke enthält. In diesem Fall könnten eine urheberrechtlich relevante Vervielfältigung oder Bearbeitung vorliegen. Eine Dimension, die der EuGH hinsichtlich der Verwendung einer urheberrechtlich geschützten Tonspur zur Bewertung herangezogen hat, war die Wiedererkennbarkeit des konkreten Werkes. Untersucht wurde, inwieweit das ursprüngliche Werk im neuen Erzeugnis erkennbar ist, und ob es durch die neue Verwendung „verblasst“ (EuGH, Urteil vom 29.07.2019 - C-476/17).

2.2 AI Act

Ziel des AI Act ist es, einen einheitlichen Rechtsrahmen für die Entwicklung, das Inverkehrbringen und die Nutzung von Systemen der Künstlichen Intelligenz in der Europäischen Union zu schaffen (Erwägungsgrund 1 AI Act). Der AI Act basiert auf einem abgestuften Ansatz, der sich durch folgende Devise auszeichnet: „Je risikoreicher ein KI-System ist, desto umfangreicher sind die Anforderungen, die es erfüllen muss“.

2.2.1 Risikobasierter Ansatz

Die Verordnung differenziert zwischen „verbotenen KI-Systemen“, „KI-Systemen mit hohem Risiko“, „Allzweck-KI“ und „KI-Systemen mit minimalem Risiko“. Im Mittelpunkt der Verordnung steht die Regulierung von Hochrisikosystemen. Darunter sind KI-Systeme zu verstehen, die aufgrund ihrer Zweckbestimmung ein hohes Risiko bergen, die Gesundheit und Sicherheit oder die Grundrechte von Personen zu schädigen, indem sie unter anderem das Ergebnis der Entscheidungsfindung wesentlich beeinflussen (Erwägungsgrund 52, Art. 6 Abs. 3 AI Act).

Eine erste Übersicht über Pflichten und Anforderungen an Hochrisikosysteme findet sich in Art. 16 AI Act. Zu den Anforderungen zählen unter anderem die Erstellung einer technischen Dokumentation, eine möglichst fehlerfreie und vollständige Aufbereitung der Trainingsdaten, die Protokollierung während des Lebenszyklus des Systems sowie ein angemessenes Maß an Cybersicherheit und Robustheit.

2.2.2 AI Act und GKI

„*Große generative KI-Modelle sind ein typisches Beispiel für ein KI-Modell mit allgemeinem Verwendungszweck, da sie eine flexible Erzeugung von Inhalten ermöglichen, etwa in Form von Text-, Audio-, Bild- oder Videoinhalten, die leicht ein breites Spektrum unterschiedlicher Aufgaben umfassen können (GPAI)*“.

Mit dieser Definition versucht die Europäische Union, GKI wie ChatGPT oder DALL-E zu beschreiben. Im weiteren Verlauf differenziert der AI Act diese „GPAI-Modelle“ weiter aus, indem er zwischen Modellen mit und ohne „systemische Risiken“ unterscheidet (Art 51 Abs. 1 AI Act). Die Entscheidung darüber, ob ein „systemisches Risiko“ besteht, obliegt der Kommission. Zudem gilt eine Vermutungsregel, nach der ein GPAI-Modell ein „systemisches Risiko“ birgt, sofern „die kumulierte Menge der für sein Training verwendeten Berechnungen, gemessen in Gleitkommaoperationen, mehr als 1025“ beträgt (Art. 51 Abs. 2 AI Act). Des Weiteren sind Anbieter von GPAI-Modellen dazu verpflichtet, eine technische Dokumentation des Modells, einschließlich Trainings- und Testverfahren, zu erstellen.

Von besonderer Relevanz ist, dass Anbieter von GPAI-Modellen anderen KI-Anbietern, die das GPAI-Modell integrieren, eine detaillierte Zusammenfassung der Fähigkeiten und Grenzen anfertigen sowie Informationen über verwendete Trainingsinhalte veröffentlichen müssen (Art. 53 Abs. 1 lit. b), d) AI Act). So soll die Verantwortung für die Anforderungen des AI Act entlang der „KI-Wertschöpfungskette“ aufrechterhalten werden (Art. 25 AI Act). Diese Dokumentation dient auch der Transparenz gegenüber Behörden, die auf Anfrage diese Unterlagen über GPAI-Modelle verlangen können (Art. 53 Abs. 1 lit. a AI Act).

Die Dokumentations- und Transparenzpflichten finden keine Anwendung auf GPAI-Modelle, welche unter Verwendung einer freien und quelloffenen Lizenz zur Verfügung gestellt werden (Art. 53 Abs. 2 AI Act), sofern keine systemischen Risiken vorliegen. KI-Systeme und -Modelle, die ausschließlich zu Forschungs- und Entwicklungszwecken entwickelt werden, sind von der Verordnung gänzlich ausgenommen (Art. 2 Abs. 6, 8 AI Act), sofern sie nicht in Verkehr gebracht oder nur zu Forschungszwecken in Betrieb genommen werden.

2.2.3 Zeitplan des AI Act

Der AI Act sieht einen gestaffelten Zeitplan vor. Während verbotene KI-Systeme spätestens nach sechs Monaten außer Betrieb genommen werden müssen, greifen die Pflichten für Anbieter von GPAI-Modellen 12 Monate nach dem Inkrafttreten des AI Act. Anforderungen und Pflichten für Hochrisikosysteme gelten frühestens nach 24 Monaten.

Unternehmen, die KI entwickeln, anbieten und betreiben, sollten einzelne KI-Systeme laufend daraufhin überprüfen, in welche Risikokategorie sie fallen. Dabei ist Vorsicht geboten, wenn GPAI-Modelle in Hochrisikobereichen eingesetzt werden (vgl. Anhang III AI Act). Die zuständige Behörde (AI-Büro), die von der Kommission speziell für die Aufsicht und Durchsetzung eingerichtet wurde, wird künftig Informationen und Leitfäden zum Umgang mit dem AI Act herausgeben.

2.3 Datenschutz und Compliance

Mittlerweile steht eine Vielzahl von Produkten und Diensten für verschiedene Anwendungen zur Verfügung – oftmals sogar mit der Möglichkeit, GKI-Ergebnisse im begrenzten Umfang kostenlos zu generieren. Dank dieser Dienste kann GKI genutzt werden, ohne dass zuvor eine eigene Infrastruktur aufgebaut werden muss. Sie sind daher in vielen Fällen sehr komfortabel und niedrigschwellig einsetzbar. Im Gegensatz zu lokal aufgesetzten Sprachmodellen bringen die fremd gehosteten Sprachmodelle jedoch wichtige Aspekte im Hinblick auf Datenschutz und Compliance ins Spiel: Da keine Datenverarbeitungsvereinbarung vorliegt und die genaue Nutzung der Daten unklar ist, sollten keine personenbezogenen Daten von Mitarbeitenden, Geschäftspartnern oder sonstigen Personen sowie keine vertraulichen/geheimen Informationen eingegeben werden. Dies gilt insbesondere für kostenfreie Dienste.

2.4 Lizenzbestimmungen und Nutzungsbedingungen

Unkomplizierte Nutzung und niedrigschwellige Angebote verleiten dazu, die Lizenzbedingungen auf die leichte Schulter zu nehmen. Wie bei jeder anderen Art von Software sind die vorgegebenen Lizenzbestimmungen und Nutzungsbedingungen des eingesetzten Modells zu beachten. Hierbei wird oftmals zwischen privatem Einsatz, Forschung und kommerzieller Nutzung unterschieden. Vor dem Einsatz der Software oder API muss sichergestellt werden, dass diese für den vorgesehenen Zweck verwendet werden darf.

2.5 Wasserzeichen in GKI-Ergebnissen

Es ist eine fahrlässige Hoffnung, die Nutzung einer GKI möge unbemerkt bleiben. Insbesondere bei Sprachmodellen besteht die Möglichkeit, vom Nutzer unbemerkt ein Wasserzeichen zu integrieren. Hierzu wird der generierte Text von der GKI subtil manipuliert, sodass sich die Anpassungen weder auf die Textqualität noch auf den Inhalt auswirken, aber dennoch algorithmisch erkennbar bleiben. Sprachmodelle generieren ihre Inhalte mithilfe von sogenannten Tokens, die einem Prozess der Wahrscheinlichkeitsverteilungs folgen. Die Nutzung von Wasserzeichen-Technologien kann diese Zufälligkeit auf eine bestimmte Art und Weise verändern, sodass das nächste Token pseudorandomisiert gewählt wird. Ob Diensteanbieter Wasserzeichen einsetzen und von einer möglichen Erkennung Gebrauch machen werden, ist nicht in jedem Fall klar ersichtlich.

3 Wichtige Einschränkungen

Auch wenn die Ergebnisse der GKI beeindruckend sind und die technische Weiterentwicklung schnell voranschreitet, bestehen aktuell noch Einschränkungen, die in der Praxis zu berücksichtigen sind.

3.1 Begrenztes Kontextverständnis

Trotz ihrer beachtlichen Fähigkeiten beim „Verstehen“ und Generieren von Sprache zeigen Sprachmodelle nur eine begrenzte Fähigkeit zum Verständnis der realen Welt, was zu potenziellen Ungenauigkeiten oder unsinnigen Antworten führen kann. So kann es vorkommen, dass die generierten Texte auf einen anderen Kontext abzielen und daher für den konkreten Kontext unpassend sind.

3.2 Mangelnde Transparenz

Aufgrund ihrer Komplexität und Größe agieren große Sprachmodelle wie eine Blackbox und machen es damit schwierig bis unmöglich, die Gründe für bestimmte Ergebnisse oder Entscheidungen nachzuvollziehen. Auch wenn GKI sich teilweise selbst erklären kann, stammen auch diese Erklärungen aus einer Blackbox, deren Vorgehensweise für Menschen in der Regel undurchschaubar bleibt. Die grundsätzliche Funktionsweise mag klar sein – trotzdem können einzelne Antworten aufgrund der enormen Größe der Modelle nicht oder nur mit erheblichem Aufwand nachvollzogen werden.

3.3 Bias oder „Vorurteile“

Große Sprachmodelle, die mit riesigen Datenmengen trainiert wurden, können (unbeabsichtigt) die in den Quelldaten enthaltenen Vorurteile, sogenannte „Biases“ übernehmen. Folglich können die Modelle einen Output generieren, der potenziell voreingenommen oder verzerrt ist. Daher ist insbesondere beim Einsatz von GKI in Bereichen, in denen Personen betroffen sind (beispielsweise Personalentscheidungen), kritisch auf solche Biases zu achten.

3.4 Halluzinationen

GKI-Modelle erzeugen Neues auf der Basis von Gesehenem oder Gelerntem. Dies ist bei künstlich generierten Bildern oftmals deutlich erkennbar, nicht jedoch bei generierten Texten. Wichtig ist daher das Bewusstsein dafür, dass Sprachmodelle Antworten frei erfinden können. Auch wenn ein Text sehr plausibel und überzeugend klingt, kann es sich um eine reine „Halluzination“ der GKI handeln. Aus diesem Grund ist beim Einsatz von GKI in kritischen Anwendungen große Vorsicht geboten. Bei kritischen Entscheidungen und Informationen sollte nicht ausschließlich auf ein Sprachmodell vertraut werden. Dies gilt insbesondere im Falle von möglicherweise größeren Auswirkungen auf Mensch, Leben oder Gesundheit. Aktuelle GKI-Modelle eignen sich sehr gut für „kreative“ Anwendungen, aber beispielsweise noch nicht für Fälle, in denen die faktische Korrektheit der Ergebnisse unverzichtbar ist, wie etwa in juristischen Texten.

4 Klärung funktionaler Anforderungen

Insbesondere im beruflichen Kontext darf GKI nicht unbedarft eingesetzt werden. Eine sichere Nutzung setzt die Prüfung der funktionalen Anforderungen an das System voraus. Im Folgenden ein Überblick über wichtige Punkte, die vorab zu klären sind.

4.1 Skalierbarkeit und Latenz

Die Leistungsfähigkeit eines Sprachmodells muss gemäß den Anforderungen des Anwendungsfalls bewertet werden. Insbesondere entsteht durch große Sprachmodelle ein enormer Ressourcenbedarf (speziell an GPU-Speicher), sodass beim lokalen Einsatz in der Regel nur niedrige Batch Sizes möglich sind. Dies führt dazu, dass die Antworten des Sprachmodells mehr Zeit benötigen. Daher sollten die Anforderungen an die Antwortzeit und Skalierbarkeit des Modells frühzeitig berücksichtigt werden, um sicherzustellen, dass es den erwarteten Durchsatz erzielen kann.

4.2 Interpretierbarkeit

Die Ergebnisse von Sprachmodellen sind auch für Experten oft weder interpretierbar noch nachvollziehbar, da die verwendeten neuronalen Netze zu groß und komplex sind. In bestimmten Anwendungsfällen kann es jedoch von großer Bedeutung sein, die Entscheidungsfindung des Modells nachvollziehen zu können. Dies gilt insbesondere in kritischen Bereichen, beispielsweise im Gesundheitswesen oder Rechtswesen. Hier muss der Einsatz von Sprachmodellen daher kritisch hinterfragt werden.

4.3 Evaluierung und Überwachung

Die Leistung eines Sprachmodells muss kontinuierlich überwacht und bewertet werden, um sicherzustellen, dass es die funktionalen Anforderungen weiterhin erfüllt. In der Praxis kann dies beispielsweise über entsprechende Benchmark-Datensätze erfolgen, sodass die Performance im Zeitverlauf gemessen und bewertet wird. Dies ist vor allem auch beim Einsatz von APIs wichtig, da sich hier die dahinterliegenden Modelle ändern können. Hiervon sind auch große Anbieter betroffen.

4.4 Datenschutz und Compliance

Sprachmodelle können bereits im Trainingsprozess Zugriff auf sensible Informationen erhalten oder diese im anschließenden Betrieb verarbeiten. Angemessene Sicherheitsvorkehrungen sind daher unverzichtbar, um den Zugriff oder Transfer von sensiblen Daten zu beschränken und die Privatsphäre zu schützen. Dies betrifft insbesondere den Einsatz von APIs und Dienstleistern.

5 Technische Umsetzung und Effizienz

GKI-Modelle – auch vermeintlich „kleine“ – haben einen enormen Ressourcenbedarf. Je nach geplanter Anwendung und funktionalen Anforderungen sind lokale oder fremdbetriebene Instanzen besser geeignet. Eine Abwägung muss stets im Hinblick auf die Voraussetzungen und Anforderungen erfolgen.

5.1 Einsatz einer lokalen GPU-Instanz

Für lokale Instanzen empfehlen sich GPUs mit großem Speicher, sodass möglichst große GKI-Modelle geladen werden können. Während es inzwischen auch kleinere Modelle gibt, die auf dem System-RAM und auch ohne dedizierte GPU oder ausschließlich mit CPU laufen, ist deren Performance (im Sinne von generierter Qualität und Antwortzeit) noch stark eingeschränkt. Idealerweise werden für Sprachmodelle Grafikkarten mit 80 GB VRAM verwendet. Je nach Größe des Sprachmodells und Anwendungsfall sollte das Sprachmodell auf mehrere GPUs verteilt werden, sodass mit größeren Batch Sizes gearbeitet werden kann. Entsprechende Systeme kosten in der Regel mehrere Zehntausend bis Hunderttausend Euro.

5.2 Einsatz einer fremdbetriebenen GPU-Instanz

Sprachmodelle können auch „klassisch“ auf Remote-Servern betrieben werden. Hier können nach Bedarf über externe, weltweit agierende Dienstleister leistungsstarke Hardwareressourcen gebucht werden, um Hosting und Skalierung von großen KI-Modellen sicherzustellen. Während die Einstiegsinvestition bei lokalen Instanzen sehr hoch ist, fallen für den Einsatz fremdbetriebener Instanzen vergleichsweise höhere Kosten pro Betriebsstunde an. Diese bewegen sich in der Regel im hohen einstelligen Euro-Bereich pro GPU-Stunde. Preis und Leistung sollten daher im Voraus sorgfältig verglichen werden.

5.3 Nutzung von APIs

Neben der Möglichkeit, die GKI-Modelle auf lokaler oder fremdbetriebener Hardware laufen zu lassen, werden auch – in der Regel kostenpflichtige – APIs angeboten. Drei aktuelle, bekannte APIs sind (Reihenfolge ohne Wertung):

- OpenAI API (https://platform.openai.com/docs/api-reference/introduction)
- Huggingface API (https://huggingface.co/inference-api)
- Aleph Alpha Luminous (https://docs.aleph-alpha.com/docs/introduction/luminous/)

Darüber hinaus gibt es weitere Angebote für den manuellen Gebrauch, wie den OpenAI Playground (https://chat.openai.com/) und weitere Webseiten, die Sprachmodelle nutzen und sich potenziell für kleinere manuelle Anfragen eignen, wie beispielsweise https://www.perplexity.ai/.

5.4 Ökologische und ökonomische Betrachtung

Um Inhalte aus GKI-Modellen zu generieren, sind zahlreiche Rechenschritte erforderlich. Dies macht den Einsatz von GKI in aller Regel äußerst ressourcenintensiv.

5.4.1 Energiebedarf

GKI-Modelle erfordern enorme Rechenleistung und sind äußerst ressourcenintensiv. Das Training großer Modelle erfolgt in großen Rechenzentren mit einem erheblichen Stromverbrauch. Auch nach dem Training ist der Energiebedarf groß, da jede einzelne Anfrage eine Vielzahl an Rechenoperationen beinhaltet.

5.4.2 Hardwarebedarf

Die Erstellung und Nutzung von GKI-Modellen führen zu einer erhöhten Nachfrage nach Hardware, insbesondere nach leistungsfähigen GPUs und Servern. Dies schlägt sich im Preis für diese Hardware nieder.

5.4.3 Datenbedarf

Erstellung und Training von GKI-Modellen erfordern enorme Datenmengen, um geeignete Ergebnisse zu erzielen. Diese Daten müssen erfasst, gespeichert und nutzbar gemacht werden. Auch für diesen Schritt werden technische und personelle Ressourcen benötigt.

6 Zusammenfassung und Checkliste

In diesem White Paper fassen wir die zentralen Herausforderungen und Aspekte zusammen, die beim Einsatz von GKI zu berücksichtigen sind. GKI kann und wird die Arbeit revolutionieren. Ihre Nutzung fordert jedoch sorgfältige Überlegung und Planung.

Wenn der Einsatz von GKI für Sie infrage kommt, enthält die folgende Checkliste die wichtigsten Punkte, die im Vorfeld zu klären sind:

- Konzeptionelle Eignung für die Anwendung
  - ☑ Passt GKI überhaupt sinnvoll zur Anwendung?
  - ☑ Stehen auch weniger komplexe Modelle und Algorithmen zur Verfügung, die besser geeignet wären?
  - ☑ Sind Erklärbarkeit und Nachvollziehbarkeit wichtige Kriterien?
  - ☑ Dürfen die vorhandenen Daten überhaupt für GKI verwendet werden?
- Kritikalität der Anwendungen
  - ☑ Wie kritisch ist die geplante Anwendung?
  - ☑ Welche Entscheidungen sind davon betroffen?
  - ☑ Soll eine Haftung für die Entscheidungen übernommen werden?
  - ☑ Wie ist mit möglichen Vorbehalten umzugehen?
  - ☑ Welche Absicherungen sind erforderlich?
- Technische Ressourcen
  - ☑ Gibt es lokale Ressourcen, um die angedachten GKI-Modelle zu betreiben?
  - ☑ Falls nicht – ist eine entsprechende Investition denkbar?
  - ☑ Ist alternativ der Einsatz von fremdbetriebener Hardware oder Diensten sinnvoll abbildbar?
- Lizenzen und Rechtliches
  - ☑ Sind die angedachten Modelle für den geplanten Einsatz vorgesehen?
  - ☑ Welche lizenzrechtlichen Vorgaben oder Nutzungsbedingungen müssen beachtet werden?

Diese Aufzählung ist sicherlich nicht abschließend und deckt nicht alle relevanten Aspekte vollständig ab. Sie soll jedoch als Einstieg dienen, um einen gewinnbringenden Einsatz von GKI zu planen und umzusetzen.

Wenn Sie über die Möglichkeit nachdenken, GKI in Ihrem Unternehmen einzusetzen, nehmen Sie gerne Kontakt mit uns auf.

Disclaimer: Inhalte dieses White Papers wurden teilweise mit GKI erstellt. Dies gilt insbesondere für die Einleitung und die Titelgrafik.

7 Changelog

- Version 1.0: Initiale Veröffentlichung
- Version 1.1: Ergänzung urheberrechtlicher Aspekte und Anforderungen aus dem AI Act

8 Kontakt

Dr.-Ing. Fabian Rigoll

+49 721 9654-552

rigoll@fzi.de

Hauptsitz Karlsruhe

Dr.-Ing. Steffen Thoma

+49 721 9654-840

thoma@fzi.de

Hauptsitz Karlsruhe

Maria Rill, M. Sc.

+49 721 9654-646

m.rill@fzi.de

Hauptsitz Karlsruhe

FZI Forschungszentrum Informatik

Haid-und-Neu-Str. 10–14

76131 Karlsruhe

+49 721 9654-0

fzi@fzi.de

www.fzi.de

EDIH AICS.
EUROPEAN DIGITAL INNOVATION HUB
Artificial Intelligence & CyberSecurity

Co-funded by the European Union