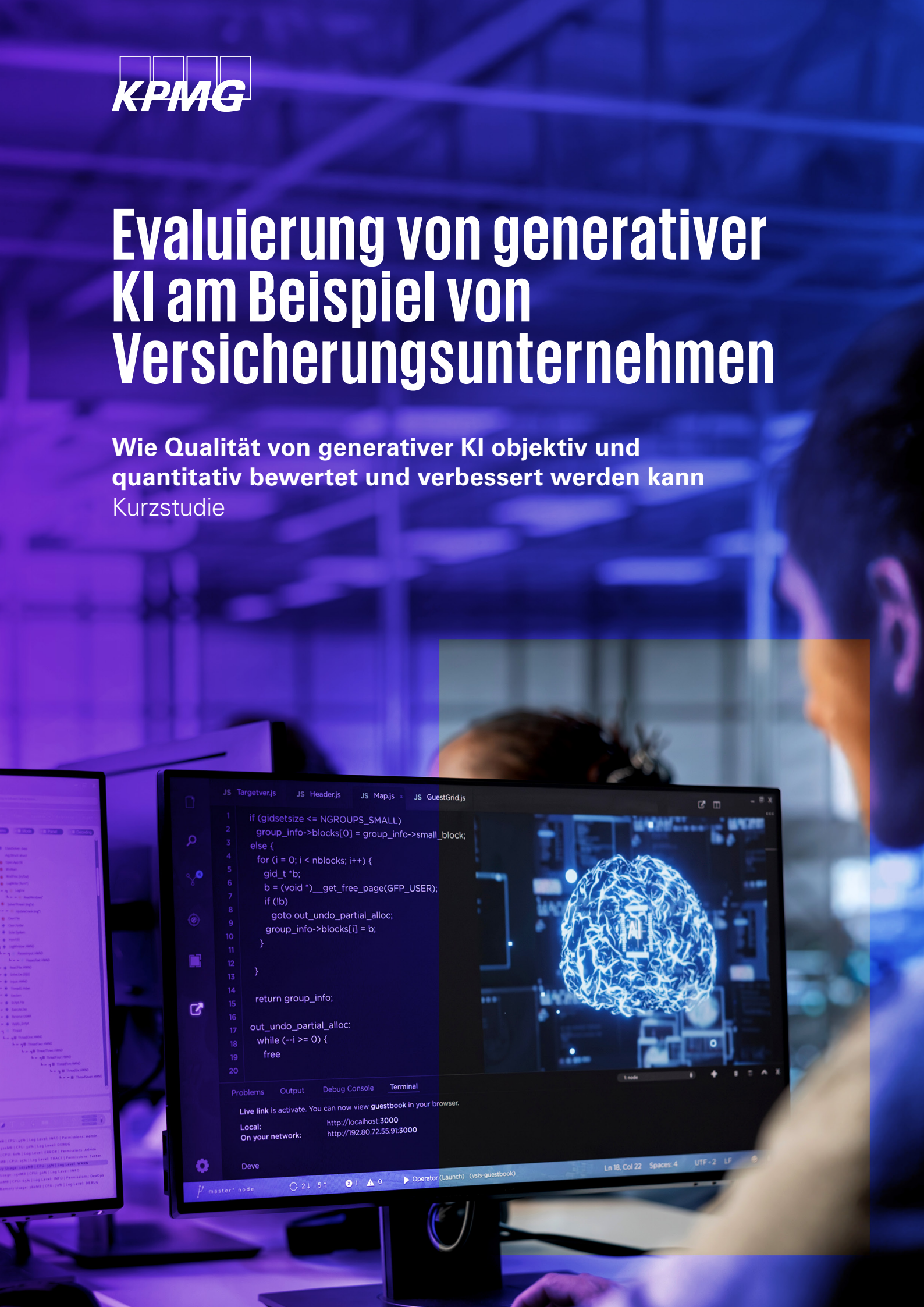
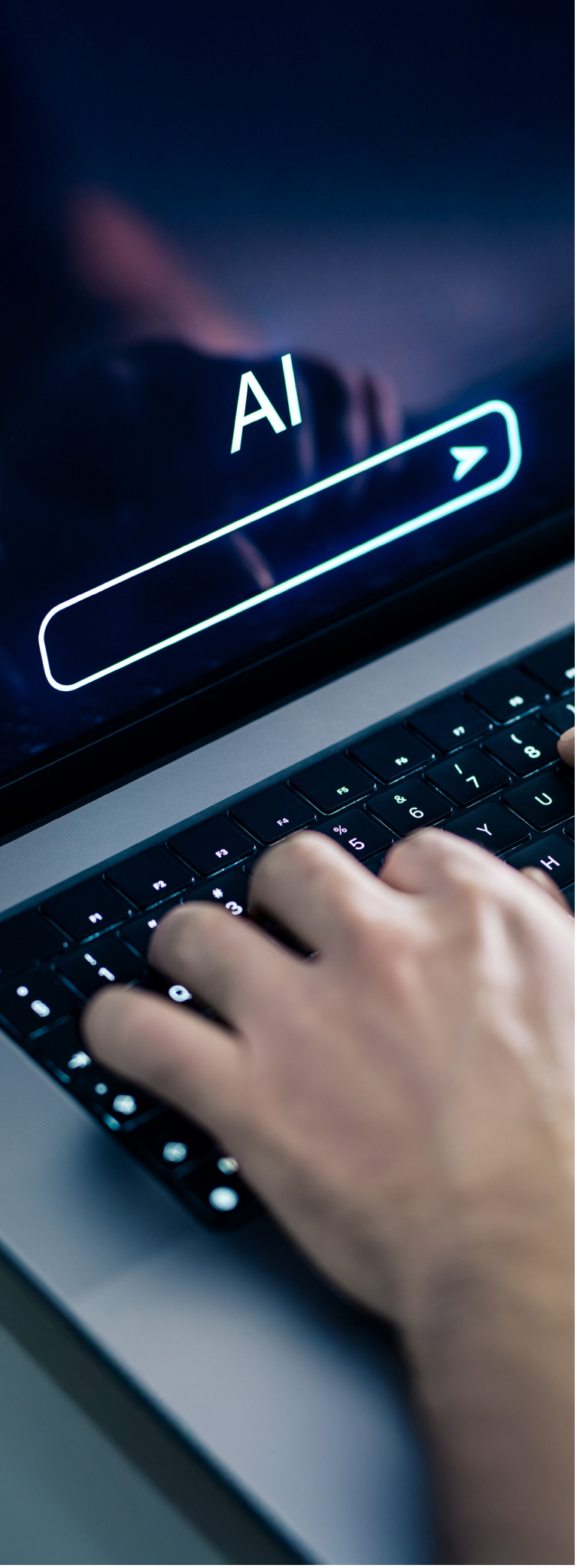
KPMG

Evaluierung von generativer KI am Beispiel von Versicherungsunternehmen

Wie Qualität von generativer KI objektiv und quantitativ bewertet und verbessert werden kann

Kurzstudie





Einleitung

Derzeit revolutioniert generative künstliche Intelligenz (generative KI) zahlreiche Branchen, darunter auch die Versicherungsbranche, deren Aktuariate, deren Rechnungswesen und deren Risikomanagement. Generative KI verspricht, komplexe Prozesse, Datenanalysen und Prognosen effizienter und präziser zu gestalten. Doch trotz ihrer beeindruckenden Fähigkeiten bleibt die Frage bestehen: Wie zuverlässig sind die Ergebnisse, die von generativer KI erzeugt werden? Halluzination sowie fehlerhafte Ergebnisse können Anwender entmutigen und den Einsatz von generativer KI bremsen und sogar blockieren.

Um diese Frage zu beantworten, haben wir eine umfassende Studie durchgeführt, in der die von generativer KI erzeugten Texte mit denen von menschlichen Fachexpertinnen und -experten verglichen werden. Unsere Untersuchung konzentriert sich auf die Versicherungswirtschaft, wo präzise Datenanalysen und fundierte Entscheidungen von maßgeblicher Bedeutung sind. Die Studie zielt zum einen darauf ab, die Qualität und Genauigkeit der von KI generierten Texte zu bewerten und zu verstehen, wie diese mit den von erfahrenen Expertinnen und Experten erstellten Inhalten übereinstimmen. Zum andern stellt sie aber auch eine Methodik für die Qualitätssicherung bei der Implementierung von Anwendungen der generativen KI bereit.

Die Ergebnisse unserer Studie wurden am Beispiel der Versicherungsbranche erarbeitet, sind aber auf Banken und Asset-Manager sowie andere Industrien übertragbar.

Methodik

Unsere Studie untersucht die Qualität der durch generative KI verfassten Texte mithilfe eines Abweichungsmaßes, das diese Texte mit den Vergleichstexten aus öffentlichen Berichten von Fachexpertinnen und -experten der Versicherungsindustrie vergleicht. Weil die Texte auf der gleichen Datengrundlage basieren, ist ein direkter Vergleich möglich. Dadurch lassen sich quantitative Abweichungen systematisch untersuchen. Unsere Methodik folgt dabei Kryscinski et al. (2020), welche wir als Werkzeug zur Evaluierung der inhaltlichen und faktischen Korrektheit von generativer KI erzeugten Texten im Vergleich zu menschlich erstellten Texten nutzen (im Folgenden als Genauigkeitsmaß bezeichnet).[1] Wir haben zudem andere alternative Maße getestet, welche jedoch unseren Qualitäts- und Konsistenzanforderungen nicht genügten. Im Gegensatz zu herkömmlichen Studien zur Evaluierung solcher Genauigkeiten fokussiert sich unsere Studie auf die Bereiche Aktuariat, Rechnungswesen und Risikomanagement. Diese Bereiche sind von besonderem Interesse, da sie zum einen ein hohes Potenzial für die Technologie besitzen, zugleich in diesen Bereichen aber auch eine hohe Genauigkeit unter der Verwendung von nicht alltäglichen Fachbegriffen vorausgesetzt wird.

Neben der objektiven Bewertung durch unser Genauigkeitsmaß setzen wir eine Kontrolle durch Fachexpertinnen und -experten als Qualitätsschleife ein. Aktuarinnen und Aktuare, Risikomanager:innen und Wirtschaftsprüfer:innen von KPMG haben die mit generativer KI erzeugten Texte auf inhaltliche Richtigkeit geprüft. Durch diese Absicherung gewährleisten wir, dass die Ergebnisse unserer Studie sowohl statistisch als auch fachlich fundiert sind.

Durch das mehrstufige Verfahren kann diese Studie qualitätsgesicherte Aussagen zu den Ergebnissen generativer KI treffen und gleichzeitig quantitativ und objektiv aufzeigen, welche Maßnahmen zu signifikanten Verbesserungen führen. Die folgenden Maßnahmen sollen hier evaluiert werden:

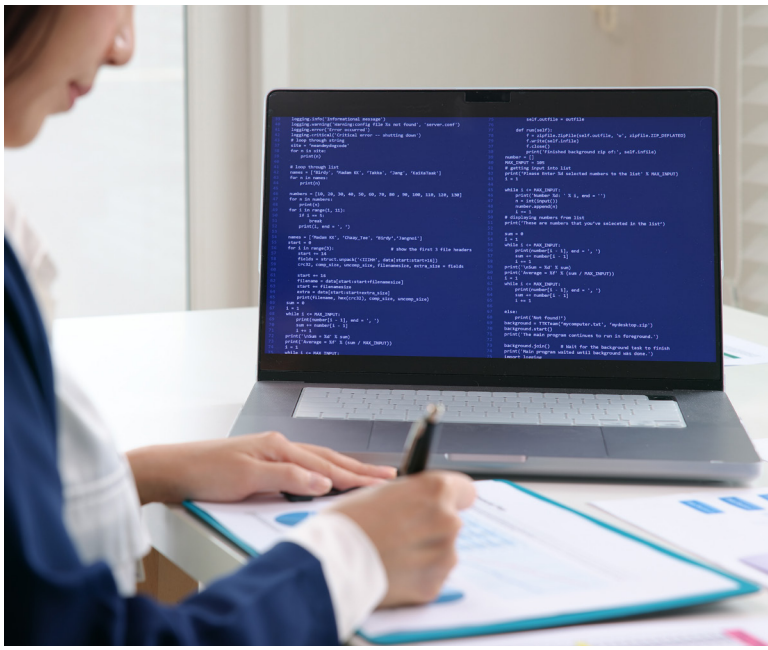
- Prompt Engineering
- Meta-Prompting.

[1] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Con-sistency of Abstractive Text Summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

Prompt Engineering bezieht sich auf die Technik, bei der spezifische Eingaben oder Anweisungen an die generative KI gegeben werden, um die gewünschten Ergebnisse zu erzielen. Durch gezieltes Prompting kann die KI dazu gebracht werden, präzisere und kontextuell relevante Texte zu generieren. In dieser Studie wird eine Prompting-Datenbank von KPMG verwendet, um die Eingaben zu optimieren und die Qualität der generierten Inhalte zu verbessern.

Meta-Prompting erweitert Prompt Engineering um komplexere und mehrschichtige Eingaben, mit der die KI die Struktur und den Kontext der gewünschten Ergebnisse besser verarbeiten kann. Die Meta-Prompting-Datenbank von KPMG versorgt die KI mit umfassenderen und strategischeren Anweisungen, was zu einer höheren Genauigkeit und Kohärenz der generierten Texte führt.

Für diese Studie wurden über 200 Texte aus der Versicherungsindustrie ausgewertet. Alle Texte hatten dabei einen Fokus auf die folgenden Bereiche: Aktuariat, Rechnungswesen, ESG- oder Kreditrisikomanagement sowie allgemeine Bereiche des Risikomanagements. Die Studie umfasst bewusst Fachbegriffe, bei denen die korrekte Anwendung durch eine generative KI grundsätzlich zu hinterfragen ist. Das lässt sich mit der seltenen Verwendung im täglichen Sprachgebrauch und entsprechend auch in den Trainingsdaten des generativen KI-Modells erklären.



Ergebnisse

Die Auswertung des Genauigkeitsmaßes in der folgenden Tabelle zeigt, dass die generative KI nicht ohne Weiteres eine hohe Genauigkeit und Faktentreue liefert. Die Qualität kann aber signifikant gesteigert werden. Eine angemessene Genauigkeit von mehr als 95 Prozent wird lediglich durch das vollständige Zusammenspiel aus Prompt Engineering und Meta Prompting erreicht. Somit zeigt diese Studie, dass der sinnvolle Einsatz generativer KI in komplexen Sachverhalten professionell geplant und implementiert werden muss, um echten Mehrwert zu schaffen.

Diese Ergebnisse zeigen, dass durch Einsatz generativer KI Genauigkeiten erreichen können, die den hohen Ansprüchen eines Unternehmens genügen. Um diese Qualität zu erreichen, ist jedoch eine Vorleistung notwendig, die auf den jeweiligen Anwendungsfall angepasst wird.

Maßnahme	Genauigkeitsmaß	Abweichung zur vorherigen Zeile (p-Wert)	Prüfung Korrektheit durch Fachexperten
Einfaches Prompting	57%	n/a	60%
Prompt Engineering	76%	19%-Punkte (0,1)	78%
Prompting Engineering inkl. Meta-Prompts	98%	**22%-Punkte*** (8,54e-05)	99%

*Werte gerundet: für p<0,05, für p<0,01 und für p<0,001

Fazit

Unsere Studie zeigt, dass generative KI auch in anspruchsvollen Situationen vielversprechende Ergebnisse liefern kann. Um qualitativ hochwertige Ergebnisse zu erzeugen, sind aber zusätzliche Schritte neben der bloßen technischen Bereitstellung der KI-Lösung nötig. Dies gilt vor allem für Branchen und Bereiche, die in ihrer Arbeit besonders häufig auf Fachbegriffe zurückgreifen.

Um Mehrwerte zu schaffen, sollten Unternehmen bei der Implementierung und Nutzung von generativer KI Folgendes berücksichtigen: Eine konsequente Fokussierung auf Qualität, Genauigkeit und Zuverlässigkeit ist entscheidend, um die gewünschten Ergebnisse zu erzielen und Vertrauen in KI-Lösungen aufzubauen. Fachwissen und Expertise in den Geschäftsprozessen sollten bei der effektiven Implementierung von generativer KI unterstützen, um die Technologie optimal zu nutzen. Gezielt eingesetzt können Prozesse optimiert und automatisiert werden, was zu Effizienzsteigerungen und Kosteneinsparungen führen kann. Es ist wichtig, dass alle Strategien und Methoden regelmäßig überprüft und angepasst werden, um den maximalen Nutzen aus generativer KI zu ziehen.

Eine Maßnahme kann dabei der Aufbau eigener sogenannter Bibliotheken bzw. Libraries sein, die konkrete (Meta-)Prompts Verfügung stellen.
So nutzte diese Studie beispielsweise folgende Libraries:

- KPMG Prompt Library für das Prompt Engineering
- KPMG Meta-Prompt Library für das Meta-Prompting.

Ausblick

Für Anwendungsfälle, die über öffentliche Daten hinausgehen, bietet sich ein RAG-Ansatz (Retrieval Augmented Generation) an, um die Qualität im Hinblick auf das Verwenden von internen Informationen zu steigern. RAG ist eine Technik, die die generative KI mit zusätzlichen Informationen aus zusätzlichen Datenquellen versorgt. Diese Methode kombiniert die Fähigkeit der KI, Texte zu generieren, mit der Fähigkeit, relevante Informationen aus einer Datenbank abzurufen. Dadurch wird die Faktentreue und Relevanz der generierten Inhalte erhöht. RAG kann beispielsweise eingesetzt werden, um sicherzustellen, dass die generierten Texte nicht nur kreativ, sondern auch faktisch korrekt sind.

Um erfolgreich zu sein, sollten Unternehmen bei der Implementierung von KI-Strategien auf verschiedene Aspekte achten: Dazu gehört die Erarbeitung einer klaren strategischen Ausrichtung für den Einsatz von KI und die Ausarbeitung und Priorisierung passender Use Cases, die in den Unternehmensprozessen umgesetzt werden können. Es ist wichtig, Anwendungen von KI gezielt in die Unternehmensprozesse zu integrieren, um Effizienz und Innovation zu fördern. Zudem sollte eine KI-Governance etabliert werden, die auch gesetzliche Vorgaben wie den EU AI Act berücksichtigt. Schließlich ist die Qualitätssicherung, einschließlich der Methoden zur objektiven Bewertung der von generativer KI erzeugten Inhalte, von großer Bedeutung. Unternehmen sollten sicherstellen, dass diese Aspekte sorgfältig geplant und umgesetzt werden, um den maximalen Nutzen aus KI-Anwendungen zu ziehen.

Kontakt

KPMG AG
Wirtschaftsprüfungsgesellschaft

Dr. Fabian Bohnert
Director, Financial Services
T +49 170 7016615
fbohnert@kpmg.com

Hauke Lehnhoff
Manager, Financial Services
T +49 69 9587 4404
hlehnhoff@kpmg.com

Einige oder alle der hier beschriebenen Leistungen sind möglicherweise für KPMG-Prüfungsmandanten und deren verbundene Unternehmen unzulässig.

www.kpmg.de

www.kpmg.de/socialmedia