Bundesministerium für
wirtschaftliche Zusammenarbeit
und Entwicklung

Ergebnisbericht Testphase KI

Finanziert von der Europäischen Union
NextGenerationEU

DATALAB

Zum Produkt „GPT-4U“ und diesem Bericht haben beigetragen:

Dr. David Wisniewski (Projektmanagement, Schulungen, Support, Umfrage, Steuerung Software Entwicklung), Dr. Peter Kumberger (Planung Cloud Infrastruktur, Steuerung Software Entwicklung, Projektmanagement, Support), Tanja Hagemann (Projektmanagement, Support), Lea Smidt (Support), Parham Kouloubandi (Support), sowie Dr. Iliya Nickelt (Chief Data Scientist des BMZ)

Projektgruppe Datenlabor

Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (BMZ)

Stresemannstraße 94, 10963 Berlin

Internet: https://www.bmz.de

Inhaltsverzeichnis

Executive Summary

Zur Erfüllung der Aufgaben des BMZ muss täglich eine wachsende Zahl an Dokumenten analysiert und bearbeitet werden. Die Anwendung Künstlicher Intelligenz (KI) bietet für diese Situation Unterstützung, sogenannte „große Sprachmodelle“ (oder auch Chat-KI) können bei der Arbeit mit Dokumenten gezielt entlasten. Besonders effizient sind Chat-KIs bei Aufgaben wie der Zusammenfassung von langen Texten und der Extraktion einzelner Fakten (siehe Beispiel im Kasten). Auch Textentwürfe sind ein mögliches Einsatzgebiet. KI kann dabei gebräuchliche Sprachen gut verarbeiten.

Um Richtlinien für eine eigenständige Nutzung von Chat-KI durch Mitarbeitende des BMZ zu erproben, wurde im August 2023 eine personell begrenzte Testphase beschlossen und in einem Zeitraum von knapp drei Monaten durchgeführt. Dafür wurde den Teilnehmenden aus allen Abteilungen und dem Leitungsstab die Möglichkeit der eigenständigen Nutzung einer Chat-KI ermöglicht. Eine Risikobetrachtung der Projektgruppe Datenlabor ergab, dass mithilfe einer BMZ-eigenen Anwendung Risiken in der Nutzung generativer KI im Vergleich zu kommerziell verfügbaren Angeboten deutlich reduziert werden können (siehe Kapitel 3). Deshalb hat die Projektgruppe Datenlabor vorab einen besonders sicheren Zugang für das BMZ entwickelt: GPT-4U. Mit in der Testphase weiterentwickelten Nutzungsrichtlinien und begleitenden Veranstaltungen wurden den Teilnehmenden die sichere und verantwortungsvolle Nutzung vermittelt.

Beispiel: Extraktion relevanter Informationen

Problem: Mitarbeitende brauchen viel Zeit, aus langen Dokumenten nur einzelne, für aktuelle Prozesse relevante Passagen zu extrahieren.

Wie können große Sprachmodelle helfen? Die KI kann große Mengen an Text verarbeiten und Informationen zu vorgegebenen Themen herausfiltern. So können sich Mitarbeitende leichter auf die relevanten Textinhalte fokussieren und diese gezielt weiterverarbeiten.

Das Ergebnis der Testphase zeigt, dass sich GPT-4U als nutzbringendes Werkzeug in vielen Bereichen der täglichen Arbeit des BMZ erwiesen hat. Die Einsparung an Arbeitszeit im Umgang mit großen Textmengen wird durch die Teilnehmenden der Testphase auf rund 25% geschätzt. Von den Teilnehmenden wurde ein klarer Bedarf nach dieser Form der KI-Unterstützung formuliert. Die Qualität der Antworten, die GPT-4U generiert hat, wurde allgemein als hoch eingeschätzt. Die Hälfte aller Teilnehmenden beantwortete die Frage, ob sie GPT-4U weiter einsetzen würden, mit der höchstmöglichen Punktzahl: „sehr wahrscheinlich“ (Abbildung 2). Diese Ergebnisse sind für die Einführung einer neuen Software äußerst positiv.

Die Nutzungsrichtlinien zur sicheren Nutzung wurden gut verstanden. Teilnehmende konnten Risiken erkennen und eigenständig minimieren. Dank der begleitenden Schulungen der Projektgruppe Datenlabor, der erworbenen Erfahrung im Umgang mit der Chat-KI und der Kenntnis von Arbeitsprozessen im BMZ konnten die Teilnehmenden die Möglichkeiten und Grenzen großer Sprachmodelle im BMZ fundiert beurteilen.

Ausgangssituation: Wozu Künstliche Intelligenz?

Die Analyse von Dokumenten und die Extraktion von Informationen nehmen bei den täglichen Aufgaben der Mitarbeitenden des BMZ einen wesentlichen Raum ein. So müssen z.B. verschiedene Berichte gelesen und geprüft werden. Ebenso wird ein großer Anteil der täglichen Arbeitszeit in die Erstellung und Bearbeitung von Dokumenten investiert. Dabei ist über die Zeit ein erhebliches und kontinuierliches Wachstum der Textmengen zu verzeichnen. Texte müssen von den Mitarbeitenden des BMZ zusammengefasst werden, Informationen müssen aus den Inhalten extrahiert und in geeigneter Weise für die jeweilige Aufgabe neu zusammengesetzt und dokumentiert werden. Durch Umformulierungen werden zielgruppenspezifische Anpassungen an den Texten vorgenommen. Zielgruppen bzw. Adressaten der Texte variieren je nach Anlass und Thema. Dabei sind bestimmte Aspekte der Texte hervorzuheben, das Format der Texte ist an Vorlagen anzupassen und die Wortwahl muss an der Zielgruppe ausgerichtet werden.

Im November 2022 wurden große Sprachmodelle erstmals einer breiten Öffentlichkeit zugänglich gemacht. Dabei stellte sich heraus, dass große Sprachmodelle eine gute Unterstützung bei verschiedenen textuellen Bearbeitungen sein können. Diese Modelle sind ein Typ generativer KI, also solcher KI, die aus Eingaben in natürlicher Sprache („Prompts“) neue Inhalte generieren können. Als Quellen nutzen diese Modelle zuvor antrainiertes Wissen aus Informationen aus dem Internet oder aus Dokumenten. Große Sprachmodelle basieren auf Wahrscheinlichkeiten, wobei bei der Generierung von Texten das nächste, wahrscheinlichste Wort hinzugefügt wird (Eingabe: „Das BMZ ist ein Bundesministerium in“, Ausgabe: „Deutschland“). Es besteht die Möglichkeit, dass in einigen Fällen fiktive, d.h. nicht Fakten entsprechende Ergebnisse produziert werden, sogenanntes „Halluzinieren“.

Das in der Testphase eingesetzte Sprachmodell GPT-4 war zum damaligen Zeitpunkt eines der weltweit besten verfügbaren Modell mit hoher Belastbarkeit der Antworten. Halluzination, also das Erfinden von plausibel klingenden aber inkorrekten Ausgaben des Sprachmodells, sind im Vergleich zu Vorgängermodellen sehr selten geworden. Des Weiteren ermöglicht die BMZ-Version es, durch Quellenangaben die Korrektheit von Aussagen der Chat-KI zu verifizieren.

Große Sprachmodelle haben u.a. folgende Einsatzbereiche in Bezug auf die Aufgaben des BMZ:

- Zusammenfassung langer fremdsprachiger Texte
- Extraktion spezifischer Inhalte aus Texten
- Verfassen von Texten
- Vergleichen von Texten und Erstellen von Gegenüberstellungen
- Informations- und Dokumentenrecherche

Da Apps mit integrierten großen Sprachmodellen insgesamt intuitiv zu bedienen sind, schätzen wir den Aufwand für die Einarbeitung zur Nutzung der App als überschaubar ein. Nach den Erfahrungen aus der Testphase, sind wir zuversichtlich, dass BMZ-Mitarbeitende nach wenigen Stunden Einweisung und ohne technisches Vorwissen in der Lage sein sollten, die BMZ-Lösung GPT-4U in der täglichen Arbeit zu verwenden (vergleiche Abbildung 1).

Beschreibung des Vorhabens „Testphase KI“

Die Projektgruppe Datenlabor hat von Oktober bis Dezember 2023 eine Testphase mit ausgewählten Organisationseinheiten durchgeführt. Dabei stand die für das BMZ relevanteste Anwendung generativer KI, d.h. Unterstützung bei textzentrierten Arbeiten durch Chat-KI, im Fokus. Die Nutzung von KI war ausschließlich den Mitarbeitenden der ausgewählten Pilot-Organisationseinheiten erlaubt. Anderen Mitarbeitenden war die KI-Nutzung nicht gestattet.

1. KI-Anwendung „ChatGPT“: Risiken der Nutzung

Mit dem Erfolg kommerziell verfügbarer großer Sprachmodelle ab November 2022 wurde der Einsatz in vielen Bundesressorts diskutiert. Dabei wurden verschiedene Risiken in der Nutzung von ChatGPT identifiziert (s. hierzu auch **Error! Reference source not found.**), insbesondere jedoch:

- **Daten**: Bei der Nutzung kommerzieller Produkte können Eingaben der Nutzenden („Prompts“) und hochgeladene Dokumente von den anbietenden Unternehmen zur Verbesserung der Chat-KI herangezogen werden.
- **Halluzinationen**: Aufgrund der Algorithmen und Verwendung von Wahrscheinlichkeiten bei der Erzeugung von Antworten des Modells auf die Eingaben der Nutzenden besteht die Möglichkeit, dass Antworten nicht den Fakten entsprechen.
- **Infrastruktur**: Bei der Verwendung kommerzieller Produkte ist man gänzlich auf die IT-Infrastruktur der anbietenden Unternehmen angewiesen. Diese beinhaltet oft Server außerhalb der EU.
- **Zugangskontrolle**: Bei der Verwendung kommerzieller Produkte muss sich jede*r Nutzende in der Regel mit Namen und einer E-Mail-Adresse registrieren.

Nach Betrachtung der Risiken hat die Projektgruppe Datenlabor beschlossen, eine eigene Chat-KI zu entwickeln: „GPT-4U“. Somit konnte den Risiken kommerzieller Produkte Rechnung getragen werden. In Kapitel 3.2 stellen wir GPT-4U vor. In Kapitel 3.3 stellen wir dar, wie GPT-4U die identifizierten Risiken minimiert.

2. Die BMZ-eigene KI-Anwendung „GPT-4U“

Für die Testgruppe hat die Projektgruppe Datenlabor ein eigenes, dezidiertes und abgesichertes Chat-Interface als Webseite für den Zugang zum großen Sprachmodell GPT-4 zur Verfügung gestellt, „GPT-4U“. Dieser Zugang erlaubte nur die Testung großer Sprachmodelle von OpenAI, also generativer KI, um Texte zu erzeugen. Andere generative KI, z. B. für die Erzeugung und Bearbeitung von Bildern steht nicht im Fokus der täglichen Arbeit im BMZ und wurde in der Testphase nicht betrachtet.

Die durch die Projektgruppe Datenlabor entwickelte prototypische Anwendung „GPT-4U“ gab Teilnehmenden der Testphase einen niedrigschwelligen Zugang zu einem der damals besten großen Sprachmodelle: GPT-4. Dieses Sprachmodell erzeugt besonders hochwertige Antworten. GPT-4U trägt gleichzeitig den Risiken Rechnung, die sich bei der Nutzung der kommerziellen Version von „ChatGPT“ ergeben (siehe Kapitel 1). Mitarbeitende können auf einfache Weise eine neue Konversation (Chat) mit dem großen Sprachmodell starten und Fragen stellen oder Anweisungen geben („Prompts“). GPT-4U generiert daraus eine Antwort und gibt diese in einem

Chatfenster aus. Zur Erleichterung der Arbeit mit Dokumenten, stellt die Anwendung unter anderem den Upload von PDF-Dokumenten zur Verfügung. Die gesicherte Webseite, die den Zugang zu GPT-4 erlaubt, zeigt **Error! Reference source not found.**.

Nach Auftakt-Veranstaltungen und Einführungen in GPT-4U durch die Projektgruppe Datenlabor haben die Teilnehmenden das Sprachmodell meist eigenständig erprobt und Erfahrungen mit der Nutzung gesammelt. In der Anwendung ist es Nutzenden möglich, den Grad an Konsistenz oder Kreativität (auch Temperatur genannt) einzustellen: Während präzise Analysen konsistentere Antworten erfordern, benötigt Ideengenerierung für z.B. die Tagesordnung eines Workshops mehr Kreativität. Nutzende können über die Anwendung auf Wunsch auch Quellen aus dem Internet verarbeiten. So können beispielsweise die wichtigsten entwicklungspolitischen Meldungen von einer News-Website extrahiert werden. Weiterhin können Nutzende eigene Dokumente hochladen und befragen („Chat with your documents“). So können z.B. aus langen Gesprächsprotokollen schnell die Positionen der einzelnen Akteure extrahiert werden.

Die Anwendung GPT-4U gibt mit jeder Antwort zur Verifizierbarkeit des Ergebnisses die Quelle der Information an, auf denen die Antworten beruhen (entweder antrainiertes „Weltwissen“, eine von den Anwendenden angeforderte Website mit URL, oder ein hochgeladenes Dokument mit Dateinamen).

Die Projektgruppe Datenlabor hat GPT-4U nach Rückmeldungen der Nutzenden während der Testphase kontinuierlich und behutsam verbessert. Die Ergebnisse der Weiterentwicklung wurden den Teilnehmenden regelmäßig bereitgestellt, so dass Feedback zu neuen Funktionen direkt eingeholt werden konnte. Anforderungen der Nutzenden wurden schnell in die Anwendung integriert, um GPT-4U sukzessive zu verbessern.

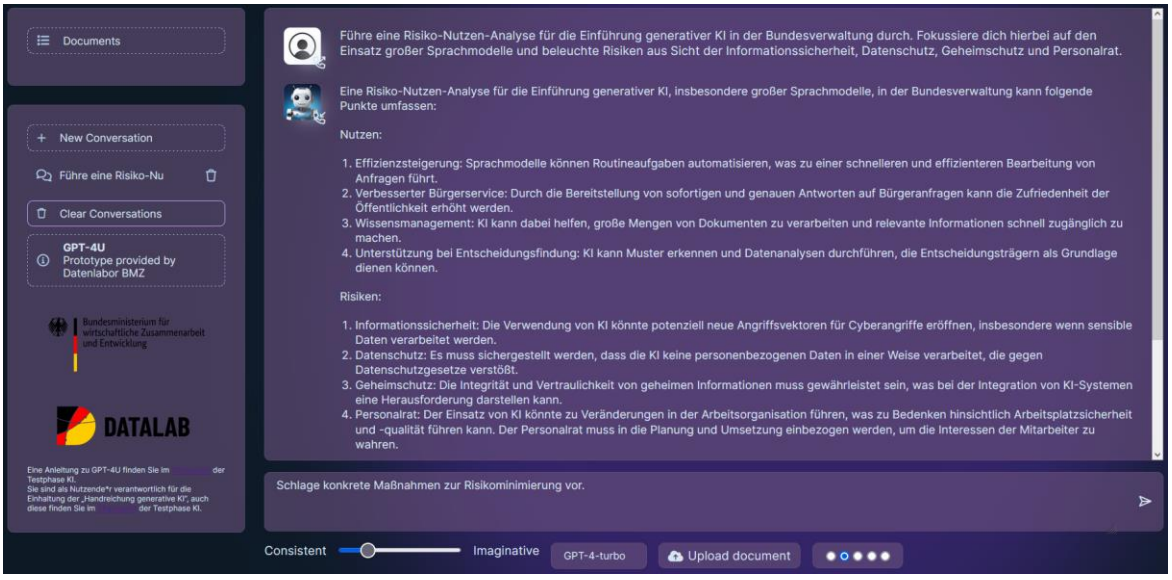

Abbildung 1: Oberfläche der Anwendung "GPT-4U"

3. Risiken begrenzen

Die Projektgruppe Datenlabor hat für die Testphase ganz bewusst eine eigene Anwendung entwickelt, anstatt die öffentlich verfügbare Anwendung „ChatGPT“ zu verwenden. Die Entwicklung einer eigenen Anwendung unter Verwendung von GPT-4 bietet insbesondere folgende Vorteile für die Sicherheit der Nutzung (siehe **Error! Reference source not found.** für eine Übersicht):

- **Daten:** Die Architektur von GPT-4U stellte sicher, dass die Eingaben der Nutzenden und die Ausgaben des Sprachmodells (Chatverlauf) zu keinem Zeitpunkt außerhalb des Browsers des Nutzenden gespeichert wurden. In der Cloud bzw. auf Seiten von OpenAI findet keine Speicherung statt. Dokumente müssen aus technischen Gründen auf einen von der Projektgruppe Datenlabor konfigurierten Server hochgeladen werden. Diese Dokumente wurden niemals zu OpenAI hochgeladen, und der Zugriff war auf die aktuell nutzende Person beschränkt. GPT-4U löschte Dokumente automatisch nach Gebrauch, spätestens nach einem Tag. Wichtig ist auch, dass keine Inhalte von Anfragen, Texten oder Dokumenten dem Hersteller OpenAI für das weitere Training des zugrundeliegenden Sprachmodells zur Verfügung gestellt wurden.

- **Halluzinationen**: Diesem Risiko wurde durch die Angabe von Quellen in GPT-4U begegnet. Ausgenommen vom verwendetem Weltwissen als Quelle können Nutzende anhand von Quellangaben erkennen, welches Dokument oder welche Website vom Modell für die Erzeugung jeder einzelnen Antwort herangezogen wurde. Dies ermöglicht Mitarbeitenden die Qualitätskontrolle generierter Antworten.

- **Infrastruktur**: Anders als ChatGPT verwendet GPT-4U IT-Infrastruktur, die vollständig innerhalb der EU betrieben wird (Zentral-West-Deutschland, Schweden). Die Azure IT-Infrastruktur in der EU ist CSA STAR Level 2 attestiert, in Deutschland zudem C5-testiert.

- **Zugangskontrolle**: Mit GPT-4U unterliegt die Zugangskontrolle dem ITZBund bzw. der Projektgruppe Datenlabor. Teilnehmende Organisationseinheiten haben pseudonymisierte Accounts erhalten. Dienstliche Email-Adressen mussten somit nicht an Drittanbietende übertragen werden. Einzelne Nutzende haben dabei zu keinem Zeitpunkt Zugriff zu den Daten (z.B. Chatverläufe, Dokumente) Anderer.

- **Sprachmodell**: GPT-4U nutzte das damals beste Sprachmodell, GPT-4-turbo, wodurch die bestmöglichen Antworten generiert wurden.

- Die Anwendung ist so konfiguriert, dass **Nutzungsverhalten** einzelner Nutzender nicht überwacht werden konnte.

Die Projektgruppe Datenlabor behielt so die Kontrolle über die Abläufe und konnte auf die spezifischen Anforderungen des BMZ eingehen, sowohl aus Sicherheits- als auch aus Nutzenden-Perspektive.

Neben den technischen Schritten zur Risikominimierung wurden auch einige organisatorische Maßnahmen getroffen werden. Die Projektgruppe Datenlabor hat Schulungen zur Verwendung von GPT-4U konzipiert, sowie klare Nutzungsrichtlinien formuliert. Dort wird darauf hingewiesen, dass KI-generierte Ergebnisse immer durch einen Menschen geprüft werden müssen. Dadurch wird das Risiko der fehlenden Faktizität oder der Nutzung eines Programmcodes mit Sicherheitslücken adressiert. Auch das Verwenden vertraulicher und sensibler Daten ist in den Nutzungsrichtlinien geregelt.

VERGLEICH GPT-4U UND CHATGPT		
	GPT-4U	**ChatGPT**
Systeme	Front-End auf einer durch ITZBund administrierten Azure Umgebung mit Schnittstelle Azure OpenAI zum Zugriff auf das Sprachmodell. Alle IT-Infrastruktur innerhalb der EU (aktuell Deutschland und Schweden).	Komplette Anwendung auf Infrastruktur der Firma OpenAI. Weltweite Verteilung der IT-Infrastruktur.
Zugang	Accounts vergeben durch ITZBund und administriert durch die Projektgruppe Datenlabor.	Accounts bei OpenAI, Übermittlung von dienstlichen E-Mails und Namen notwendig.
Speicherung und Nutzung von Eingaben / Ausgaben	Eingaben / Ausgaben werden nur im lokalen Browser gespeichert. Eingaben / Ausgaben werden nie zum weiteren Training des Sprachmodells verwendet.	Ohne aktives Eingreifen der Nutzenden werden Eingaben und Ausgaben gespeichert und zum weiteren Training des Sprachmodells verwendet (Opt-Out).
Quelle der Antwort	Rückmeldung über die Quelle zur Erzeugung der Antworten: antrainiertes Wissen, Website, hochgeladenes Dokument.	Keine transparente Angabe über genutzte Quellen.
Dokumente	Automatisches und kurzfristiges Löschen hochgeladener Dokumente.	Mögliche Weiternutzung hochgeladener Inhalte durch OpenAI.
Kontrolle	Projektgruppe Datenlabor und ITZBund, mit Business-Vertragskonditionen zu Microsoft und OpenAI: für BMZ keine Speicherung, keine Filterung.	Nutzungsbedingungen von OpenAI.
Oberfläche	Variabel, aktuell ohne personenbezogene Anmeldung.	E-Mail-Adresse, ggf. Telefonnummer und Kreditkarte.
Qualität	Jeweils bestes verfügbares Modell ca. sechs Wochen nach Release, aktuell GPT-4 Turbo.	kostenfreie Version nur GPT 3.5.

Tabelle 1: Vergleich unterschiedlicher Aspekte von GPT-4U und ChatGPT

Aktuelle große Sprachmodelle werden nicht kontinuierlich trainiert. Dies führt dazu, dass die Informationen ohne Zugriff auf das Internet nicht tagesaktuell sind, sondern nur bis zu einem bestimmten Datum reichen. Zum Teil weist GPT-4U auf diese Tatsache hin. Für tagesaktuelle Informationen kann GPT-4U auf Internetquellen zugreifen. Die Nutzung dieses Features wurde den Teilnehmenden in einer Veranstaltung vermittelt. Dabei muss darauf hingewiesen werden, dass ein geringes Risiko bei der Verwendung von Quellen aus dem Internet besteht. Hier können Nutzende dazu verleitet werden, über Links auf schadhafte Webseiten geleitet zu werden, oder persönliche Daten auf Webseiten einzugeben. Um diesem Risiko zu begegnen, gelten für die

Nutzenden die Regeln für die normale Nutzung des Internets. Unbekannten Links darf nicht gefolgt, persönliche Daten nicht auf unbekannten Webseiten eingegeben werden.

Auch die verantwortungsvolle und transparente Nutzung generativer KI wird durch verschiedene organisatorische Maßnahmen gefördert. Alle KI-generierten Inhalte müssen klar markiert werden. Weiterhin müssen alle Ausgaben der Chat-KI durch Mitarbeitende qualitätsgesichert werden. Für diesen Schritt ist die Verwendung einer KI nicht erlaubt. Auch in den Schulungen der Projektgruppe Datenlabor werden Best Practices für den ethischen und verantwortungsvollen Umgang mit KI vermittelt. Und letztlich stellen auch die Entscheidungsprozesse im BMZ eine weitere Stufe der Qualitätssicherung dar. Wichtige Entscheidungen werden stets durch Mitarbeitende verschiedener Einheiten und Hierarchiestufen zusammen erarbeitet.

4. Begleitung der Testphase durch die Projektgruppe Datenlabor

Die eigenständige Nutzung von GPT-4U durch die Teilnehmenden der Testphase wurde eng von der Projektgruppe Datenlabor begleitet (s. Abbildung 2). Die Testphase wurde am 04.10.2023 durch einen Auftakt-Workshop gestartet. Dabei wurde in die Verwendung von GPT-4U eingeführt und Hinweise zur Verwendung gegeben. Mit den Teilnehmenden wurden praktische Anwendungsfälle für ihre Arbeit mit GPT-4U identifiziert. Zwei Erfahrungsaustausche nach jedem Drittel der Testphase gaben den Teilnehmenden die Möglichkeit, Feedback zu geben und Fragen zu stellen. Neue Funktionalitäten, die im Laufe der Testphase GPT-4U hinzugefügt wurden, wurden erläutert und demonstriert. Begleitende Schulungen vermittelten den Teilnehmenden sowohl erforderliches theoretisches als auch praktisches Wissen zum Umgang mit großen Sprachmodellen. Jede Woche konnten Teilnehmende in einer „KI-Sprechstunde“ Fragen zu GPT-4U und zu KI im Allgemeinen stellen. Zusätzlich erfasste die Projektgruppe Datenlabor in individuellen Beratungsgesprächen die Bedarfe einzelner Referate. Die Testphase wurde am 13.12.2023 mit einem Abschlussworkshop beendet.

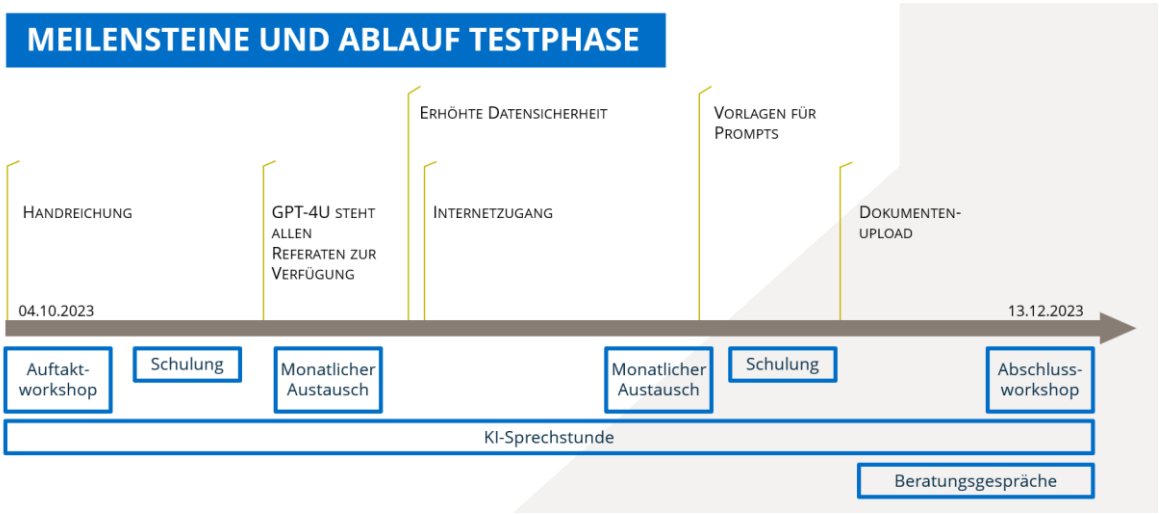

Abbildung 2: Ablauf Testphase

Um die Testphase zu begleiten hat die Projektgruppe Datenlabor eine Reihe von Dokumenten erstellt. Eine Klickanleitung erklärte Teilnehmenden die Funktionalitäten von GPT-4U. Nutzungsrichtlinien waren über das Interface von GPT-4U sofort zugänglich und erläuterten zentral Aspekte von Datenschutz, Vertraulichkeit und IT-Sicherheit. Alle Dokumente wurden in der Testphase basierend auf Teilnehmenden-Feedback kontinuierlich verbessert.

5. Ziele der Testphase

Ziel er Testphase war es, die eigenständige Nutzung generativer KI durch BMZ-Mitarbeitende zu erproben. Es wurden Antworten für die folgenden vier Themenbereiche gesucht (siehe **Error! Reference source not found.**): Bedarfsanalyse, inhaltliche Bewertung, Verbesserung der Nutzungsrichtlinien und Weiterentwicklung der KI-Nutzung im BMZ.

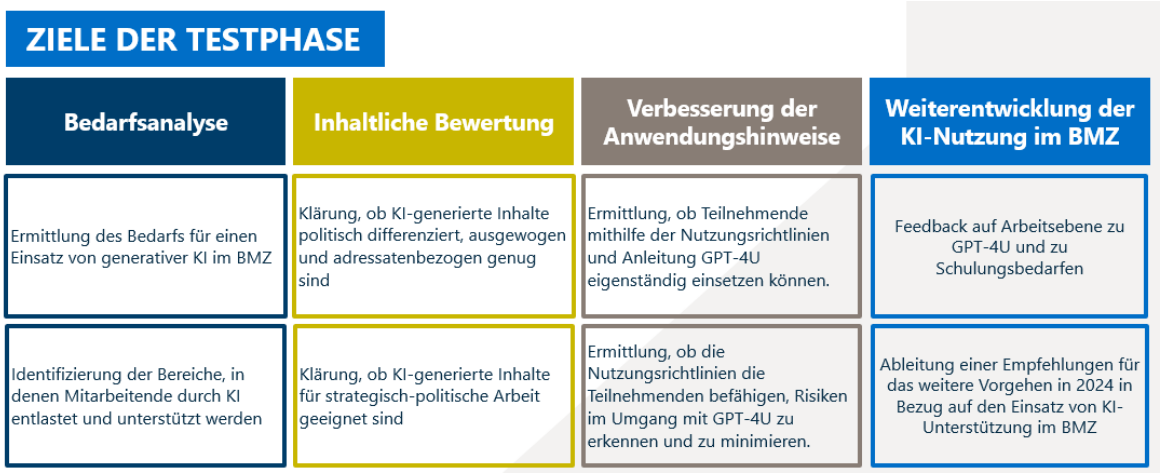

Abbildung 3: Zielsetzungen der Testphase

Ergebnisse der Testphase

Insgesamt haben rund 60 Mitarbeitende aktiv an der Testphase teilgenommen. Die hier präsentierten Ergebnisse beruhen auf zwei Umfragen, sowie auf vielen direkten Gesprächen mit den Teilnehmenden. An den anonymen Umfragen zu Beginn und zum Ende der Testphase nahmen 35 bzw. 33 Personen teil. Das ist in etwa die Hälfte der aktiv teilnehmenden Personen. Umfragen und Gespräche zielten auf die vier anfangs identifizierten Themenbereiche ab: Bedarfsanalyse, Inhaltliche Bewertung, Verbesserung der Anwendungshinweise, Weiterentwicklung der KI-Nutzung im BMZ. Zusätzlich wurden auch die Themen Nutzung von GPT-4U und Lernerfahrungen erfasst. Im Folgenden stellen wir die Ergebnisse jedes Bereichs vor.

6. Nutzung und Lernen

Die Teilnehmenden der Testphase sind Expert*innen in ihren jeweiligen Fachgebieten, sie kennen die Prozesse im BMZ sehr gut. Sie sollten den Nutzen von GPT-4U im Kontext dieser Prozesse erproben und bewerten. Eine anonyme Umfrage wurde durchgeführt, um zu verstehen, wie GPT-4U in der Testphase eingesetzt wurde und ob Teilnehmende einen Lernfortschritt erzielt haben. Denn viele der zentralen Fragen im Bereich inhaltliche Bewertung oder Verbesserung der

Anwendungshinweise lassen sich nur sinnvoll beantworten, wenn GPT-4U auch aktiv genutzt wurde.
Insgesamt wurden während der Testphase mehr als 2500 Fragen an GPT-4U gestellt. Das bedeutet mehr als 54 Fragen pro Werktag. GPT-4U hat in der Testphase mehr als 5000 DIN-A4 Seiten verarbeitet. Laut Umfrage haben rund zwei Drittel der Teilnehmenden GPT-4U einmal pro Woche oder öfter eingesetzt, rund ein Drittel 2-3 Mal pro Monat oder seltener. Einem noch häufigeren Einsatz stand vor allem Zeitmangel durch hohe Arbeitsbelastung im Weg. Dabei haben rund 70% der Teilnehmenden GPT-4U aktiv in einen Arbeitsprozess integriert (siehe Abbildung 1). Dies geht über das bloße Experimentieren mit einer neuen Anwendung hinaus, und hat die Projektgruppe Datenlabor in dieser Deutlichkeit überrascht. Die Teilnehmenden haben GPT-4U viel schneller als erwartet in Arbeitsprozessen genutzt. Dies deutet auf schnelle Lernerfolge der Teilnehmenden hin. Vor allem ging es hier um das Erstellen, Analysieren und Zusammenfassen von – zum Teil fremdsprachlichen – Texten, sowie Rechercheaufgaben. Die Nutzung von GPT-4U in weiteren Arbeitsprozessen wurde u.a. durch die fehlende Integration in bestehende Anwendungen verhindert, wie beispielsweise Microsoft Outlook.

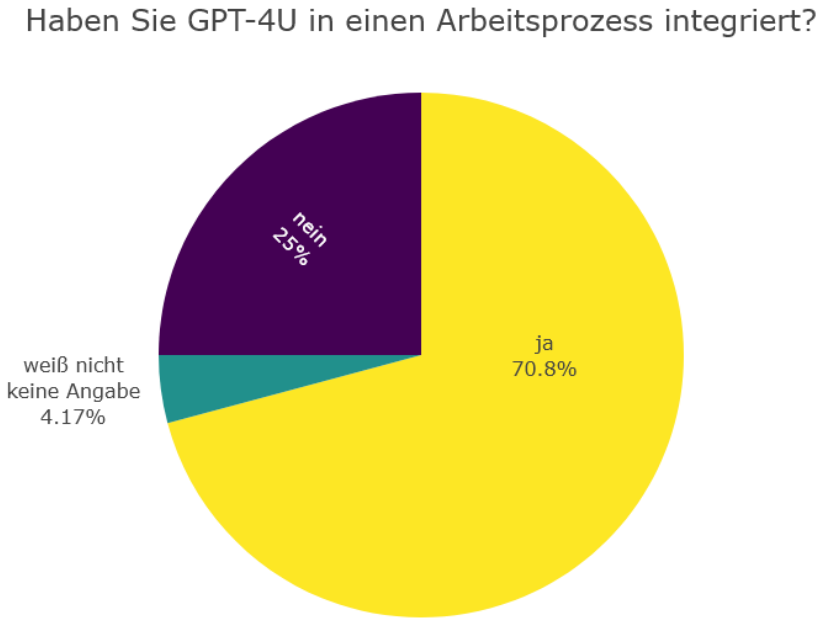

Abbildung 1: Wurde GPT-4U bereits aktiv in einen Arbeitsprozess integriert? Die in diesem Ergebnisbericht dargestellten Antworten beruhen auf den Rückmeldungen von 33 Personen.

Mehr als 80% der Teilnehmenden gaben an, in der Testphase etwas oder viel neues theoretisches Wissen erworben zu haben. Rund 75% gaben an, einige oder viele neue praktische Fähigkeiten im Umgang mit großen Sprachmodellen erworben zu haben. Die wichtigste neu erlernte Fähigkeit war das Schreiben von Eingaben (Prompts). Dieses Wissen hat die Projektgruppe Datenlabor in eigens konzipierten Schulungen vermittelt, zu denen die Teilnehmenden positives Feedback gaben. Analog der Verwendung anderer technischer Anwendungen existiert auch bei der Verwendung von GPT-4U eine Lernkurve, um sich mit der Anwendung vertraut zu machen. Möglichkeiten und Grenzen von KI haben die Teilnehmende im Laufe des Testphase einzuschätzen gelernt.

Bewertung
Teilnehmende haben GPT-4U mehrheitlich regelmäßig eingesetzt, können große Sprachmodelle also basierend auf diesen konkreten Erfahrungen bewerten. Dabei wurde die Anwendung oft aktiv

in Arbeitsprozesse integriert und Teilnehmende haben im Laufe der Testphase neues theoretisches und praktisches Wissen zur Verwendung generativer KI erworben. In Kombination mit ihrer Expertise zu Arbeitsprozessen im BMZ sind die Teilnehmenden also sehr gut qualifiziert, fundierte Aussagen zum Einsatz generativer KI im BMZ zu treffen.

7. Bedarfsanalyse

Nachdem die Teilnehmenden mehrere Monate Erfahrung im Umgang mit großen Sprachmodellen und mit GPT-4U gesammelt haben, haben wir gefragt: Haben Sie Bedarf nach einer solchen Anwendung? Wenn ja, für welche Aufgaben würden Sie diese einsetzen? Wir sehen, dass Teilnehmende GPT-4U wahrscheinlich bis sehr wahrscheinlich einsetzen würden. In diesem Fall würde GPT-4U insbesondere für die Zusammenfassung von Texten sowie für die Extraktion von Informationen aus Texten genutzt werden. Obwohl GPT-4U prinzipiell dazu fähig wäre, stehen z.B. die Umformatierung von Texten und das Finden von Argumenten / Gegenargumenten aktuell nicht im Fokus (siehe Abbildung 2).

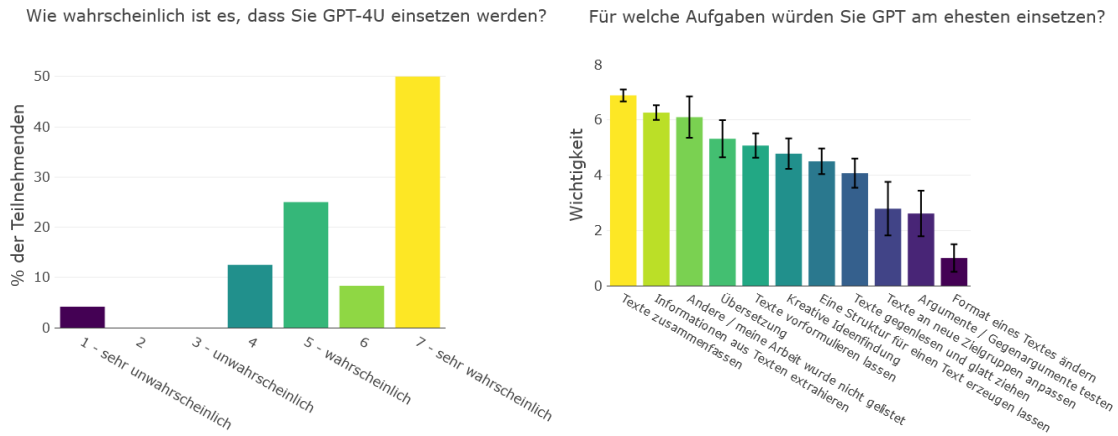

Abbildung 2: Wie wahrscheinlich ist ein Einsatz großer Sprachmodelle? Und für welche Aufgaben würde es eingesetzt? Für die Skala „Wichtigkeit“ wurden Teilnehmende gebeten, die Aufgaben in eine Rangfolge zu bringen. Wichtigkeit ist hier der inverse durchschnittliche Rangplatz.

Bewertung
Teilnehmende formulieren einen klaren Bedarf nach großen Sprachmodellen. Eine deutliche Mehrheit würde diese regelmäßig einsetzen. Viele Teilnehmende haben auch rückgemeldet, dass Sie in der Testphase nicht genug Zeit gefunden haben, um das volle Potential großer Sprachmodelle auszuschöpfen. Da sich Aufgaben wie Textzusammenfassung und Informationsextraktion relativ leicht umsetzen lassen, ist es deshalb zu erwarten, dass Bedarfe zunächst in diesen Feldern gesehen werden. Mit zunehmender Erfahrung werden sich die Bedarfe mutmaßlich auf weitere Aufgaben ausweiten.

Interessant ist, dass viele Teilnehmende „Andere / meine Arbeit wurde nicht gelistet“ als wichtigen Anwendungsfall benannt haben. Währen der Testphase hat die Projektgruppe Datenlabor beobachtet, dass viele Referate sehr spezifische Bedarfe an große Sprachmodelle haben. Diese Bedarfe haben sich jedoch erst im Laufe der Testphase, mit steigender Erfahrung der Teilnehmenden, ergeben. Für diese spezifischen Fälle kann die Projektgruppe Datenlabor spezialisierte Anwendungen auf Grundlage von GPT-4U entwickeln, sowie in Zukunft gezielte Schulungen anbieten.

8. Inhaltliche Bewertung

Für den Einsatz großer Sprachmodelle muss die Qualität der Ausgaben den Ansprüchen der BMZ-Mitarbeitenden genügen. Ausgaben müssen ausgewogen, spezifisch und adressatenbezogen genug sein.

Im Allgemeinen wird die Qualität der Antworten als eher hoch eingeschätzt, Ausgaben sind oft ausgewogen. Aussagen mit einem zu starken Bias sind also nicht aufgefallen. Für kurze Texte entspricht die Qualität eher den Erwartungen der Nutzenden als bei längeren Texten. Bei Spezifität und Adressatenbezogenheit zeigt sich ein differenzierteres Bild. Einige Teilnehmende sehen Ausgaben als oft adressatenbezogen genug, andere jedoch nur als teilweise adressatenbezogen. Ein ähnlich zweigeteiltes Bild zeigt sich beim Thema Spezifität, einige sehen Ausgaben als oft spezifisch, andere als nur teilweise spezifisch genug an (siehe Abbildung 3).

Bewertung
Grundsätzlich wird die Qualität der Ausgaben als ausreichend hoch bewertet, Antworten sind oft ausgewogen. Feedback zur Spezifität und Adressatenbezogenheit ist jedoch polarisierter. Dies lässt sich damit erklären, dass GPT-4U für unterschiedliche Arbeiten benutzt wurde. Für Unterstützung bei der Erstellung eines Curriculusm für einen Workshop ist z.B. ein hohes Maß an Spezifität erforderlich. Das Curriculum muss z.B. klar an die Zielgruppe angepasst werden. Im Gegensatz hierzu sind die Anforderungen an Spezifität und Adressatenbezogenheit bei z.B. der Zusammenfassung eines Dokuments weniger hoch. Je nachdem auf welche Aufgaben Teilnehmende sich in der Testphase fokussiert haben, sind hier also unterschiedliche Bewertungen zu erwarten.

In diesem Kontext ist zu beachten, dass die Qualität der Ausgaben großer Sprachmodelle von vielen Faktoren abhängt. Einer hiervon ist die Konfiguration der Anwendung durch Entwickler*innen. Große Sprachmodelle können so konfiguriert werden, dass sie z.B. besonders ausgewogen oder besonders vorsichtig antworten. Sie können mit BMZ-spezifischen Texten vorbereitet und fokussiert werden, um die Konventionen des BMZ genauer zu folgen. Mit dieser Konfiguration begann die Projektgruppe Datenlabor während der Testphase, diese kann jedoch durch weitere Iterationen noch verbessert werden.

Auch die Eingaben der Nutzenden haben einen großen Einfluss auf die Qualität der Ausgaben. Für hochwertige Antworten müssen Prompts spezifisch sein und ausreichend Kontextinformationen enthalten. Eine kurze, allgemeine Frage wie „Welche Ziele verfolgt das BMZ?“ wird zu einer sehr allgemeinen Antwort führen. Die Projektgruppe Datenlabor hat während der Testphase mehrere Schulungen zum „Prompt Engineering“, also der Formulierung von Eingaben in große Sprachmodelle, angeboten. Diese Fähigkeit bedarf jedoch einiger Übung. Nicht alle Teilnehmenden konnten im Rahmen der Testphase genügend Zeit in die Erstellung hochwertiger Prompts investieren. Als zusätzliche Hilfe hat die Projektgruppe Datenlabor deswegen Prompt-Vorlagen für Standard-Fälle entwickelt, getestet, und den Teilnehmenden zur Verfügung gestellt. Dies hat die Nutzung hochwertiger Prompts erleichtert. Es wird erwartet, dass die Qualität der Ausgaben mit zunehmender Erfahrung der Nutzenden steigen wird.

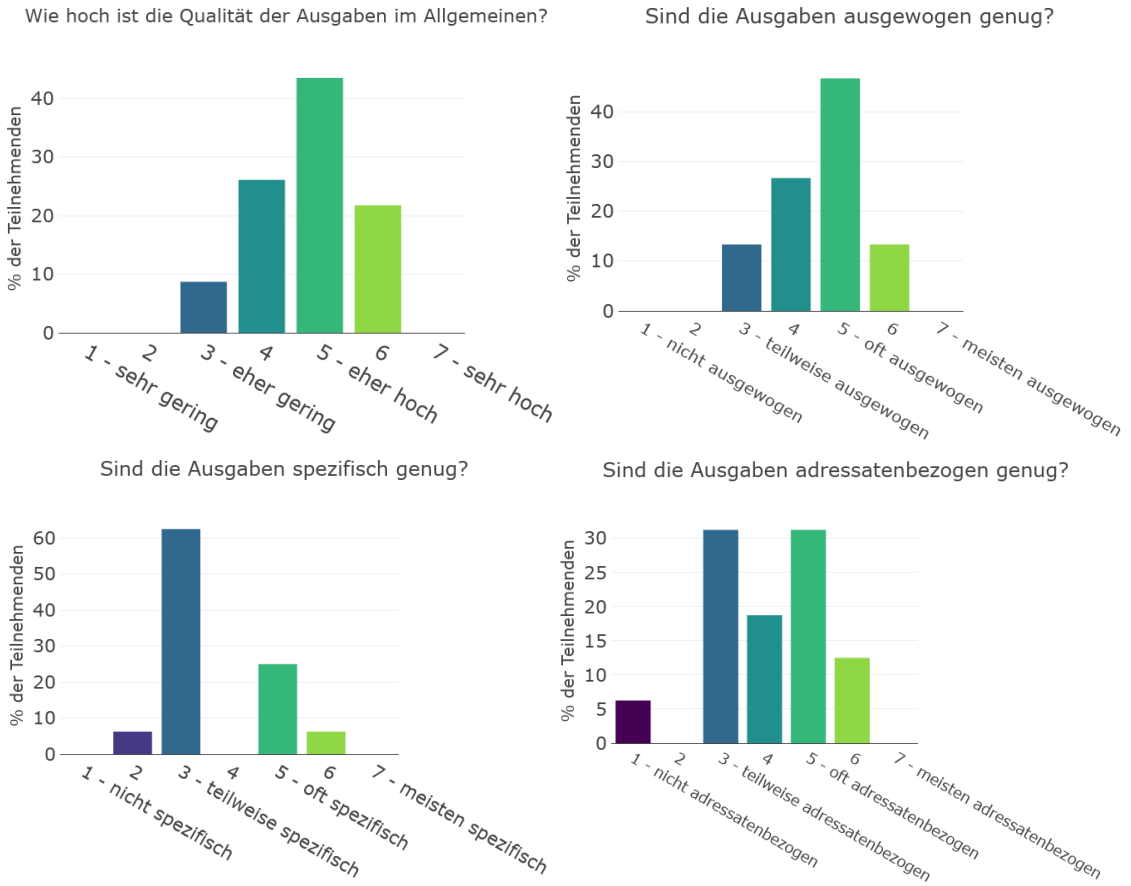

Abbildung 3: Wie hoch ist die Qualität der Ausgaben von GPT-4U? Sind Ausgaben ausgewogen, spezifisch und adressatenbezogen genug?

9. Verbesserung der Nutzungsrichtlinien

Für den sicheren Einsatz großer Sprachmodelle bedarf es klarer Regeln. Welche Informationen dürfen verarbeitet werden? Welche nicht? Nur wenn diese Richtlinien von Nutzenden verstanden und umgesetzt werden, ist ein eigenständiger Einsatz großer Sprachmodelle durch Mitarbeitende des BMZ zu empfehlen. Deshalb ist es wichtig, dass die Nutzungsrichtlinien zum Einsatz generativer KI klar formuliert ist.

Während der Testphase wurden die Nutzungsrichtlinien basierend auf dem Teilnehmenden-Feedback überarbeitet. Zum Ende der Testphase schätzen Teilnehmende diese als verständlich bis sehr verständlich ein. Risiken, die sich beim Einsatz Künstlicher Intelligenz ergeben, wurden ausreichend erläutert, so dass sich alle Anwendenden in der Lage sahen, gut bis sehr gut die Risiken zu erkennen und Maßnahmen zu ergreifen, um die Risiken zu minimieren. Richtlinien zu Datenschutz, Datensicherheit und Ethik konnten am Ende der Testphase nach Einschätzung der Teilnehmenden gut umgesetzt werden, deutlich besser als am Anfang (siehe Abbildung 4). Wissen hierzu erhielten Teilnehmende aus den Nutzungsrichtlinien, sowie aus den Veranstaltungen während der Testphase.

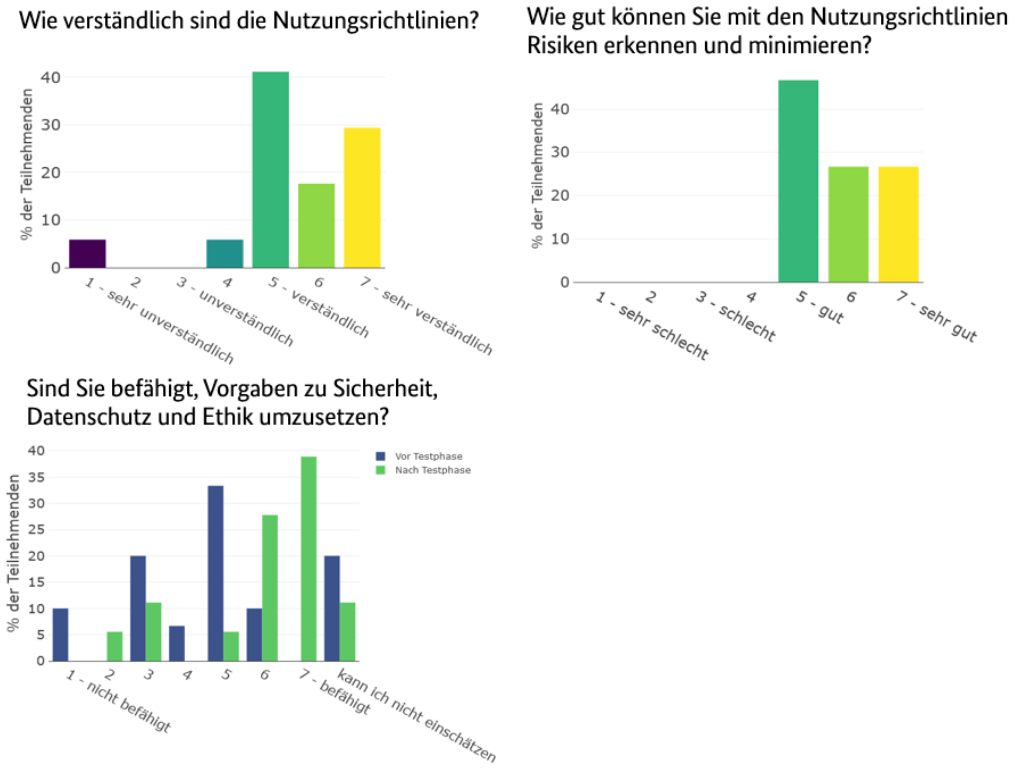

Abbildung 4: Sind die Nutzungsrichtlinien verständlich? Können mit ihrer Hilfe Risiken eigenständig erkannt und minimiert werden? Können Vorgaben zu Sicherheit, Datenschutz und Ethik umgesetzt werden?

Bewertung
Die Einschätzung der Folgen für Datenschutz und Sicherheit beim Einsatz einer Chat-KI wie GPT-4U ist wichtig, und muss vor dem eigenständigen Einsatz genau verstanden und erlernt werden. Workshops und die Nutzungsrichtlinien haben Teilnehmenden geholfen, diese Einschätzungen vorzunehmen und GPT-4U eigenständig und sicher einzusetzen.

10. Weitere Nutzung von Chat-KI im BMZ

Die Teilnehmenden der Testphase sind aktuell die Mitarbeitenden mit der höchsten Expertise im Einsatz generativer KI für BMZ-typische Aufgaben. Daher können sie sehr gut einschätzen, ob generative KI und große Sprachmodelle die Arbeit im BMZ im Allgemeinen verbessern könnten. Eine deutliche Mehrheit (83.3%) glaubt, KI könne die Arbeit im BMZ verbessern. Keine teilnehmende Person glaubt, KI könne die Arbeit nicht verbessern. Auch auf die Frage „Wird GPT Arbeit leichter oder schwerer machen?“ antwortet eine klare Mehrheit, dass die Arbeit aus Ihrer Sicht durch große Sprachmodelle einfacher wird (siehe Abbildung 5). Die **durch den Einsatz großer Sprachmodelle eingesparte Arbeitszeit wird** von den Teilnehmenden der Testphase **auf rund 25% geschätzt** (Unsicherheitsbereich + - 3%).

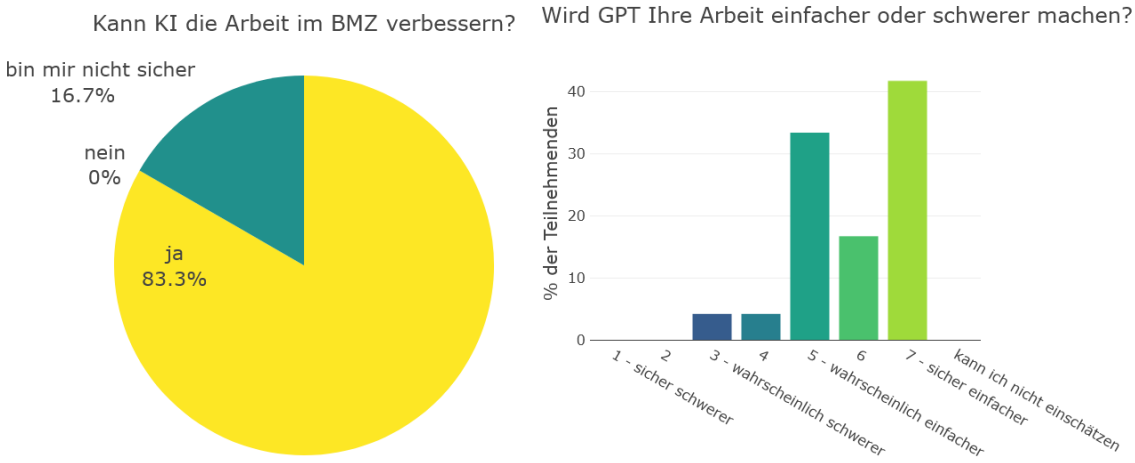

Abbildung 5: Kann generative KI die Arbeit im BMZ verbessern? Oder macht sie diese eher schwerer?

Bewertung

Die Teilnehmenden der Testphase sehen ein hohes Potential für den Einsatz generativer KI im BMZ. Wichtig ist, dass diese Einschätzung auf Kenntnis sowohl der Arbeit im BMZ als auch großer Sprachmodelle beruht. Die Schätzung einer Zeitersparnis von 25% ist als realistisch einzustufen. Aktuell gehen Expert*innen davon aus, dass große Sprachmodelle ca. 40% an Zeit einsparen können (Shakked & Zhang, *Science*, 2023, https://doi.org/10.1126/science.adh2586)Dies hängt jedoch stark von der Erfahrung der Nutzenden, sowie von den konkret genutzten Anwendungen ab.

Beispiel: Textentwürfe

Problem: Im BMZ müssen oft standardisierte Texte geschrieben werden. Dabei werden unterschiedliche Inhalte in immer gleichen Formaten kommuniziert.

Wie können große Sprachmodelle helfen? Die KI kann das Format standardisierter Texte anhand einiger guter Beispiele verstehen. Danach können beliebige fachliche Inhalte in diesem Format ausgegeben werden. Die automatische Erstellung eines Textentwurfs spart Zeit beim Verfassen von Texten.

Es ist wichtig, BMZ-Mitarbeitende in Zukunft im Einsatz großer Sprachmodelle zu schulen. Da sich diese Technologie aktuell noch sehr schnell verändert, kann nur durch kontinuierliche Schulung und Weiterentwicklung der KI-Werkzeuge sichergestellt werden, dass wir im BMZ das volle Potential ausschöpfen. Die Projektgruppe Datenlabor hat ein entsprechendes Schulungsangebot für den Bereich KI aufgrund der Erfahrungen aus der Testphase und der allgemein hohen Nachfrage zum Thema im Jahr 2023 konzeptioniert.

11. Zusammenfassung der Ergebnisse

GPT-4U hat sich als nutzbringendes Werkzeug in vielen Bereichen der täglichen Aufgaben für die Teilnehmenden der Testphase erwiesen. Vor allem bei Textzusammenfassung und Informationsextraktion war GPT-4U ein hilfreicher Assistent. Teilnehmende haben jedoch nicht immer ausreichend Zeit gefunden, um das Schreiben hochwertiger Prompts zu üben. Hier hat die Projektgruppe Datenlabor deshalb spezifische Unterstützungsangebote entwickelt, z.B. Prompt-

Vorlagen für Standard-Fälle. Um dem Bedarf nach der Arbeit mit Dokumenten gerecht zu werden, wurde das Hochladen von Dokumenten in die Anwendung GPT-4U integriert. Die Qualität der Ausgaben wird allgemein als gut und ausgewogen eingeschätzt. In den Bereichen Spezifität und Adressatenbezogenheit ist jedoch noch Verbesserungspotential vorhanden. Man kann aber davon ausgehen, dass sich mit zunehmender Erfahrung der Nutzenden und Optimierung der Anwendung diese Bereiche in Zukunft verbessern werden. Die Nutzungsrichtlinien haben Teilnehmenden dazu verholfen, eigenständig Risiken bei der Nutzung großer Sprachmodelle zu erkennen und zu minimieren. **Allgemein sehen die Teilnehmenden ein großes Potential für den Einsatz großer Sprachmodelle im BMZ.** Und als Expert*innen für BMZ-Prozesse und nun auch erfahren im Einsatz großer Sprachmodelle kann dieser Einschätzung ein hohes Gewicht gegeben werden.

Die Testphase zur Erprobung des Einsatzes generativer KI im BMZ hat gezeigt, dass bereits mithilfe des Prototypen wesentlicher Nutzen für die Mitarbeitenden geschaffen werden konnte. Bei textzentrierten Aufgaben wie Generierung, Analyse, Bearbeitung und Informationsextraktion gehen erste Schätzungen der Nutzenden von einer **Arbeitszeitersparnis von 25%** aus. Die Einschätzung der Teilnehmenden ist, dass BMZ von einer Integration generativer KI in die tägliche Arbeit profitieren wird.

Ausblick

GPT-4U wurde für die Testphase innerhalb weniger Monate prototypisch erstellt und weiterentwickelt. Anforderungen aus den Feedback-Runden und Austauschen mit den Teilnehmenden der Testphasen wurden in Wochenzyklen in die Software integriert. Die Verbesserungen bestanden im Wesentlichen in der Möglichkeit einer Internet-Abfrage, einem Dokumentenupload und kontinuierlich durch Konfigurationen verbesserte Datensicherheit. Dadurch ist eine Anwendung entstanden, welche fachlich die grundlegenden Anforderungen der Teilnehmenden umfasst.

Ein wichtiges Feedback aus der Testphase ist, dass verschiedene Facheinheiten oft sehr spezifische, meist nur innerhalb der Facheinheit relevante Anforderungen an große Sprachmodelle stellen. Für diese Fälle ist die Entwicklung spezialisierter Anwendungen, auch basierend auf großen Sprachmodellen, notwendig. Diese enthalten beispielsweise vordefinierte und zuvor erprobte Prompts. Auch angepasste User Interfaces können dabei helfen, diese Spezial-Anforderungen besser zu unterstützen, und so die Qualität der Ergebnisse zu verbessern.

Darüber hinaus erfordert die Nutzung generativer KI das Erlernen neuer Fähigkeiten, analog etwa zur Nutzung von Suchmaschinen. Die Teilnehmenden haben diese Fähigkeiten während der Testphase verbessert; durch die Schulungs- und Gesprächsangebote der Projektgruppe Datenlabor, sowie durch eigene Nutzung. Dieses Schulungsprogramm sollte weitergeführt und ausgebaut werden. Mit GPT-4U kann Mitarbeitenden eine hilfreiche Anwendung zur Verfügung gestellt werden, und durch gezielte Schulungen gäbe es genug Raum für das Erlernen eines sicheren Umgangs mit dieser KI-Anwendung.

“First Movers will be rewarded and the global race is already on without any question. Our future competitiveness depends on AI adoption in our daily businesses, and Europe must up its game and show the way to responsible use of AI.”

Ursula von der Leyen, Präsidentin der Europäischen Kommission
WEF, Davos, 16.01.2024

Die Arbeit der Datenlabore des Bundes wird

finanziert von der Europäischen Union

über den Deutschen Aufbau- und Resilienzplan für

NextGenerationEU