ZUKUNFTS ZENTRUM SÜD

f-bb Forschungsinstitut Betriebliche Bildung

Zukunfts-Zentren

f-bb-online

Dominique Dauser & Markus Utomo

KI-Chatbots Marke Eigenbau?!

Whitepaper mit Hintergrundinformationen, Empfehlungen und Praxistipps

Gefördert durch:

Bundesministerium für Arbeit und Soziales

Kofinanziert von der Europäischen Union

Baden-Württemberg Ministerium für Wirtschaft, Arbeit und Tourismus

Bayerisches Staatsministerium für Wirtschaft, Landesentwicklung und Energie

Impressum

f-bb-online
Schriftenreihe des Forschungsinstituts Betriebliche Bildung (f-bb)
ISSN 2197-8026

Herausgegeben von
Dr. Iris Pfeiffer
Forschungsinstitut Betriebliche Bildung (f-bb) gGmbH
Rollnerstraße 14
90408 Nürnberg
www.f-bb.de

Das Forschungsinstitut Betriebliche Bildung (f-bb) arbeitet seit 2003 an der Weiterentwicklung des Systems der beruflichen Bildung durch Forschung in Deutschland und international. Das Leistungsspektrum umfasst die Durchführung von Modellversuchen, Gestaltungs- und Transferprojekten, die wissenschaftliche Begleitung von Förderprogrammen, die Evaluation von Verordnungen und Maßnahmen sowie die Umsetzung von Fallstudien, empirischen Erhebungen und Analysen.

Förderung
Das Projekt „Zukunftszentrum Süd“ wird im Rahmen des Programms „Zukunftszentren“ durch das Bundesministerium für Arbeit und Soziales und die Europäische Union über den Europäischen Sozialfonds Plus (ESF Plus) sowie anteilig durch die jeweiligen Landesministerien für Wirtschaft in Bayern und Baden-Württemberg gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin und dem Autor.

Autorinnen und Autoren
Dominique Dauser & Markus Utomo

Erscheinungsjahr
2025

Titelbild
Generiert mit Image Gen + auf HuggingChat

Diese Publikation ist frei verfügbar zum Download
unter www.f-bb.de/

Zitiervorschlag
Dauser, D., & Utomo, M. (2025): KI-Chatbots Marke Eigenbau?! Whitepaper mit Hintergrundinformationen, Empfehlungen und Praxistipps. f-bb-online 01/25

Inhalt

1. Einleitung

Das Zukunftszentrum Süd unterstützt kleine und mittelständische Unternehmen (KMU) mit Sitz in Bayern oder Baden-Württemberg bei der digitalen Transformation. Dazu gehört auch, KMU die Vorteile des Einsatzes von KI-Chatbots und deren praktischen Nutzen über einen **KI-Experimentierraum** erlebbar zu machen (vgl. https://zukunftszentrum-sued.de/ki-experimentierraum/). Dort können Interessierte KI-Chatbots ausprobieren, verschiedene Möglichkeiten der Umsetzung kennenlernen und sich gleichzeitig über das vielfältige Angebot des Zukunftszentrums informieren.

Die Chatbots des Zukunftszentrums Süd fungieren als **Beratungslotsen**. Vorgestellt werden ein klassischer regelbasierter Klickbot (Klara) sowie zwei KI-Chatbots, die beide auf hochentwickelten *Open-Source LLM-Systemen (vgl. Glossar)* basieren (Max und Sophie). Während „Klara – die Logikexpertin" nur Antworten auf vordefinierte Fragen findet, können Fach- und Führungskräfte aus KMU „Max – dem geselligen Berater" und „Sophie – der geduldigen Coachin" über eine freie Spracheingabe schildern, mit welchen Herausforderungen sie gerade konfrontiert sind. Im Gesprächsverlauf erhalten sie dann Vorschläge für mögliche Lösungen. Die KI-Chatbots fragen aber auch gezielt Daten der Unternehmen wie Standort, Branchenzugehörigkeit und Beratungsanliegen ab, um auf konkrete Hilfsangebote verweisen zu können; wobei Max nur über das Zukunftszentrum Auskunft geben kann, während Sophie auch auf Informationen über weitere geförderte Initiativen zurückgreifen kann. Zudem unterscheiden sich die „Charaktere" der KI-Chatbots und damit ihr Rollenverhalten und ihre Sprechweise.

Das Zukunftszentrum Süd hat diese KI-Chatbots in Zusammenarbeit mit dem Designstudio Markus Utomo Design als externen Dienstleister entwickelt. Das vorliegende Whitepaper gibt einen Einblick in den agilen Entwicklungsprozess und zeigt an diesem Beispiel praxisnah, wie KMU vorgehen können, wenn sie selbst einen KI-Chatbot implementieren möchten. Als Einstieg in das Thema zeigt HuggingChat (https://huggingface.co/chat/) Neulingen anhand von praktischen Anwendungsfällen den Mehrwert von KI-Chatbots auf. Darüber hinaus gibt der Open-Source-KI-Chatbot Empfehlungen, wie KMU den für sie passenden Chatbot finden und was bei der Implementierung zu beachten ist. Zudem erfahren sie einige technische Details zum Aufbau und zur Funktionsweise eines KI-Chatbots.

2. Mehrwert und Anwendungsfälle: HuggingChat informiert

Wer könnte besser Auskunft über KI-Chatbots geben als ein KI-Chatbot? HuggingChat (https://huggingface.co/chat/) ist zugänglich über die Open-Source-KI-Plattform, mit deren Hilfe man sich gut über KI-Chatbots informieren kann. Man muss nur die richtigen Fragen stellen, um aussagekräftige Antworten zu erhalten, wie die folgenden Ausführungen zeigen:

HuggingChat wie auch ChatGPT und Co. sind beeindruckende KI-Chatbot-Plattformen, die eine Vielzahl von Anwendungsfällen abdecken können. Über die Plattformen können sogar persönliche Assistenten konfiguriert werden, die für bestimmte Aufgaben wie Bildgenerierung, Coding oder Webdesign optimiert sind. Dennoch gibt es gute Gründe für KMU, ihre eigenen benutzerdefinierten Chatbots zu entwickeln (siehe Kasten unten).

Vorteile benutzerdefinierter KI-Chatbots gegenüber HuggingChat, ChatGPT und Co.

1. Markenindividualität und -anpassung: Unternehmen möchten ihre Marke, Persönlichkeit und Werte widerspiegeln, was durch benutzerdefinierte Chatbots ermöglicht wird.
2. Benutzerdefinierte Funktionen: Jedes Unternehmen hat einzigartige Anforderungen. Eigenständige Chatbots können an spezifische Bedürfnisse angepasst werden und branchenspezifisches Fachwissen bieten.
3. Datenschutz und Datensicherheit: Unternehmen können durch eigene Chatbots die Datenschutz- und Sicherheitsmaßnahmen direkt kontrollieren, was für sensible Kundendaten entscheidend ist.
4. Integration in vorhandene Systeme: Benutzerdefinierte Chatbots können in vorhandene CRM-Software, Helpdesks, Zahlungsplattformen und interne Datenbanken integriert werden, um eine nahtlose Benutzererfahrung zu bieten.
5. Fortlaufende Verbesserung: Unternehmen können ihre eigenen Chatbots kontinuierlich optimieren und an die sich verändernden Anforderungen anpassen, indem sie Feedback und Analysen nutzen.
6. Wettbewerbsvorteil: Eigenständige Chatbots können Unternehmen, insbesondere KMU, dabei helfen, sich von ihren Konkurrenten abzuheben und einen einzigartigen und innovativen Ansatz für den Kundenservice zu bieten.
7. Kosteneffizienz: KMU haben oft begrenzte Ressourcen und benutzerdefinierte Chatbots können helfen, Kosten im Vergleich zu umfangreichen Kundendienstteams zu senken, während gleichzeitig ein hohes Maß an Service aufrechterhalten wird.
8. Spezialisiertes Fachwissen: Benutzerdefinierte Chatbots können mit branchenspezifischem Fachwissen entwickelt werden, wodurch sie für bestimmte Anwendungsfälle oder Branchen besser geeignet sind.
9. Personalisierung: Unternehmen können ihre eigenen Chatbots personalisieren, um eine persönliche und ansprechende Benutzererfahrung zu bieten, die auf ihre Zielgruppe zugeschnitten ist.
10. Lernende Systeme: Eigenständige Chatbots können so konzipiert werden, dass sie mit der Zeit lernen und sich weiterentwickeln, wodurch sie intelligenter und reaktionsfähiger werden und so die Kundenzufriedenheit steigern.

2.1. Wozu braucht ein KMU einen KI-Chatbot?

Ein KI-Chatbot ist für KMU eine vielseitige und lohnende Investition: Im Kundenservice können KI-Chatbots rund um die Uhr Soforthilfe leisten, häufig gestellte Fragen beantworten und so die Kundenzufriedenheit erhöhen. Gleichzeitig können sie die Arbeitsbelastung von Beschäftigten reduzieren, indem sie sich wiederholende Aufgaben übernehmen. KI-Chatbots können auch intern eingesetzt werden, um die Kommunikation und Zusammenarbeit in

Unternehmen zu verbessern, Beschäftigten bei allgemeinen Anfragen zu helfen und sogar bei der Personalentwicklung und -schulung zu unterstützen. Durch die Automatisierung von Routineaufgaben, die Bereitstellung personalisierter Interaktionen und die Analyse von Kundendaten ermöglichen es KI-Chatbots KMU, fundiertere Entscheidungen zu treffen, ihre Produkte und Dienstleistungen besser an Kundenwünsche anzupassen und ihre Ressourcen effizienter einzusetzen. Insgesamt können KI-Chatbots KMU dabei helfen, wettbewerbsfähiger zu werden, ihre Betriebsabläufe zu optimieren und eine bessere Kundenzufriedenheit zu erreichen (vgl. Kasten).

Einsatzmöglichkeiten von KI-Chatbots in KMU

- Kundenservice und Support
- Verkaufs- und Marketing-Unterstützung
- Interne Kommunikation und Zusammenarbeit
- Personalbeschaffung und -schulung
- Datenerfassung und -analyse
- Automatisierung von Routineaufgaben
- Sprachübersetzung
- Bestandsverwaltung und Echtzeit-Updates
- Finanzverwaltung und Budgetplanung
- Identifikation und Gewinnung von Kunden
- Personalisierung des Kundenerlebnisses
- …

2.2. Welche KI-Chatbots passen zu KMU?

KMU können von verschiedenen Arten von KI-Chatbots profitieren, die auf ihre spezifischen Bedürfnisse zugeschnitten sind. Die Wahl des „richtigen" KI-Chatbots hängt von den spezifischen Anforderungen des KMU, dem verfügbaren Budget, der Komplexität der Aufgaben und der erforderlichen Anpassung ab (vgl. Kasten unten).

Arten von Chatbots und mögliche Einsatzgebiete

- **Regelbasierte Chatbots:** z.B. Kundensupport, Auftragsverfolgung und -aktualisierung, Terminplanung und -verwaltung, Dateneingabe und -abruf
- **KI-Chatbots mit maschinellem Lernen:** z.B. Personalisierte Produktempfehlungen, Verkaufs- und Marketingunterstützung, Kundenanalyse und -verständnis, Sprach- und Texterkennung für komplexere Anfragen
- **KI-Chatbots mit natürlicher Sprachverarbeitung (Natural Language Processing, NLP):** z.B. Natürliche und konversationsähnliche Interaktion mit Kunden, Sprach- und Textanalyse für Kundenfeedback und -absichten, Unterstützung mehrerer Sprachen, Generierung von Inhalten wie Produktbeschreibungen oder Marketingmaterialien
- **Hybride Chatbots (Kombination aus regelbasiert und KI):** Abdeckung einer breiten Palette von Aufgaben, einschließlich Kundensupport, Vertrieb und Marketing, Lernfähigkeit und Anpassungsfähigkeit an Benutzerpräferenzen, Fähigkeit, komplexe Probleme mit einem regelbasierten Ansatz zu lösen

Regelbasierte Chatbots sind ideal für einfache, wiederkehrende Aufgaben und häufig gestellte Fragen. Sie sind kostengünstig und leicht zu implementieren. Chatbots, die auf maschinellem

Lernen basieren, können komplexere Interaktionen verarbeiten und mit der Zeit lernen und sich anpassen. Dadurch eignen sie sich besonders für den Kundenservice. Darüber hinaus können KMU auch von vorgefertigten, branchenspezifischen Chatbots profitieren, die bereits mit relevantem Fachwissen trainiert wurden. Diese Chatbots können schnell implementiert werden und bieten sofortige Unterstützung in Bereichen wie Einzelhandel, Gastgewerbe oder Finanzdienstleistungen.

2.3. Wie gehen KMU bei der Entwicklung eines KI-Chatbots vor?

Die Entwicklung eines KI-Chatbots erfordert eine sorgfältige Planung und die Bewältigung einiger Herausforderungen. Unternehmen sollten zunächst die spezifischen Ziele und Anforderungen ihres Chatbots definieren, sei es die Verbesserung der Kundenakquise, des Kundenservice oder die Rationalisierung interner Prozesse. Dann müssen sie eine KI-Chatbot-Plattform auswählen, die den gegebenen technischen Möglichkeiten entspricht und zum Budget passt. Die Datensammlung und -aufbereitung ist ein nächster kritischer Schritt, da die Qualität der Trainingsdaten die Leistung des Chatbots beeinflusst. Unternehmen sollten relevante Daten sammeln, häufig gestellte Fragen ermitteln und diese entsprechend strukturieren.

Schritt für Schritt zum eigenen KI-Chatbot

1. Ziele und Anforderungen des Unternehmens definieren, z. B. Kundenservice, Kundenakquise oder Verbesserung interner Prozesse.
2. Budget für die Entwicklung, den Betrieb und die Wartung des KI-Chatbots planen, und dabei mögliche langfristige Einsparungen durch die Automatisierung berücksichtigen.
3. Hochwertige und relevante Daten für das Training des Chatbots aufbereiten. Dazu gehören häufig gestellte Fragen, Kundengespräche, Produktinformationen.
4. Anforderungsgerechte KI-Chatbot-Plattform auswählen, dabei Faktoren wie Anpassbarkeit, Integration, Datensicherheit und Skalierbarkeit berücksichtigen.
5. Orientiert an den Bedürfnissen der Nutzenden realistische und effektive Gesprächsabläufe entwerfen und die Interaktion der eigenen Marke entsprechend personalisieren.
6. Testversionen des Chatbots orientiert am Feedback von Nutzenden optimieren.
7. Den KI-Chatbot in bestehende Systeme und Prozesse, z. B. Website, Software zum Customer-Relationship-Management (CRM) oder Kundendienstplattform integrieren.
8. Die Leistung des Chatbots kontinuierlich überwachen und bei Bedarf optimieren sowie die regelmäßige Wartung planen.
9. Datenschutzbestimmungen und ethische Überlegungen bei der Verarbeitung sensibler Kundendaten berücksichtigen.
10. Durch Schulung und Sensibilisierung bei Nutzenden die Akzeptanz fördern.
11. Erfolge durch Key Performance Indicators (KIPs) messen und Verbesserungsmöglichkeiten identifizieren.

In den Entwicklungsprozess sollten möglichst verschiedene Abteilungen des Unternehmens einbezogen werden. Für eine erfolgreiche Umsetzung ist die Expertise aus verschiedenen Abteilungen gefragt von der IT-Abteilung über die Marketingabteilung bis hin zur Personalabteilung. Die Entwicklung von Gesprächsabläufen erfordert Kreativität und ein tiefes Verständnis der Zielgruppe. Die Antworten des Chatbots sollten natürlich und ansprechend sein und die Marke des Unternehmens widerspiegeln. Technische Herausforderungen können insbesondere bei der Integration in bestehende Systeme und Schnittstellen auftreten. Hier können Unternehmen von der Zusammenarbeit mit erfahrenen Dienstleistern profitieren, die über die notwendigen Kompetenzen verfügen.

Sorgfältiges Testen und Optimieren sind unerlässlich, um die Leistungsfähigkeit des KI-Chatbots zu gewährleisten. Die Rückmeldung von den Nutzenden und die laufende Überwachung der Leistung helfen bei der Feinabstimmung der Antworten und der Verbesserung der Treffsicherheit. Darüber hinaus erfordert die Wartung und Aktualisierung des Chatbots regelmäßige Aufmerksamkeit, um sicherzustellen, dass er über die neuesten Funktionen verfügt. Unternehmen sollten sich auch der Datenschutz- und Sicherheitsaspekte bewusst sein und gewährleisten, dass der KI-Chatbot den geltenden Vorschriften entspricht.

Zusammenfassend lässt sich sagen, dass Unternehmen ihre eigenen KI-Chatbots entwickeln können, indem sie ihre Ziele definieren, die richtige Plattform auswählen, qualitativ hochwertige Daten sammeln und Gesprächsabläufe entwerfen, die auf ihre Zielgruppe zugeschnitten sind. Technische Herausforderungen können durch die Zusammenarbeit mit externen Dienstleistern gemeistert werden, und regelmäßige Wartung und Optimierung sorgen dafür, dass ein KI-Chatbot effizient arbeitet (vgl. Kasten oben).

2.4. Hintergrund: Aufbau und Funktionsweise eines KI-Chatbots

KI-Chatbots bestehen im Wesentlichen aus mehreren Komponenten, die zusammenwirken, um eine natürliche und ansprechende Konversation mit den Nutzenden zu ermöglichen. Im Folgenden wird ein Überblick über ihren Aufbau und ihre Funktionsweise gegeben:

1. Die Nutzenden interagieren mit einem KI-Chatbot über eine textbasierte **Schnittstelle**, wie z. B. ein Chat-Fenster auf einer Website oder einer mobilen App, oder über eine sprachbasierte Schnittstelle, die Sprachbefehle und -antworten ermöglicht.

2. Die **Verarbeitung natürlicher Sprache** *(Natural Language Processing, siehe Glossar)* erlaubt es dem KI-Chatbot, die Eingaben der Nutzenden zu verstehen.

3. Um die Absicht oder den Zweck der Eingabe der Nutzenden zu interpretieren, verwendet der KI-Chatbot **maschinelles Lernen**. Das bedeutet, dass die verarbeiteten Wörter und Kontexte mit einem vordefinierten Satz von Kategorien oder Absichten verglichen werden, um zu bestimmen, worum es den Nutzenden geht.

4. Basierend auf der interpretierten Absicht folgt der KI-Chatbot einer Reihe von **logischen Regeln und Entscheidungsbäumen**. Diese bestimmen eine angemessene Reaktion oder

Aktion des Chatbots, wie z. B. das Abrufen einer vordefinierten Antwort, die Berechnung eines Werts oder das Auslösen eines komplexeren Workflows.

5. Der KI-Chatbot generiert daraufhin eine **Antwort**, die auf der interpretierten Absicht und allen verfügbaren relevanten Daten basiert. Die Antwort kann in Form von Text, Audio oder sogar interaktiver Elemente wie Schaltflächen oder Bildern ausgegeben werden. Ziel ist es, eine auf die Benutzeranfrage zugeschnittene Antwort zu geben.
6. KI-Chatbots nutzen maschinelles Lernen, um aus Interaktionen mit Benutzern zu lernen und sich im Laufe der Zeit anzupassen. Sie analysieren Muster in Benutzerbeiträgen, bewerten die Effektivität ihrer Antworten und passen sich an, um die Genauigkeit und Relevanz ihrer Reaktionen zu verbessern. Dieser Lernaspekt ermöglicht es Chatbots, sich weiterzuentwickeln und komplexere Anfragen im Laufe der Zeit zu bewältigen. Dadurch besteht aber auch das Risiko, dass Chatbots manipuliert werden können.

Unternehmen können eigene KI-Chatbots entwickeln oder vorgefertigte Lösungen nutzen, wobei sie die oben genannten Komponenten und deren Integration berücksichtigen müssen. Durch die sorgfältige Planung und Anpassung an ihre spezifischen Anforderungen können KMU effektive KI-Chatbots implementieren.

2.5. Beispielprompts für die eigene Recherche

Für unsere Recherche zu Mehrwert und Anwendungsfällen von KI-Chatbots in KMU haben wir unter anderem auch HuggingChat (vgl. https://huggingface.co/chat/) verwendet. Anregungen, wie KMU vorgehen können, wenn sie selbst Open Source KI-Chatbots wie HuggingChat für solche Zwecke einsetzen möchten, finden sich in der Praxishilfe „Beschäftigte mit Open Educational Resources (OER) und Künstlicher Intelligenz (KI) gezielt fördern!“ des Zukunftszentrum Süd unter: https://www.f-bb.de/fileadmin/PDFs-Publikationen/240924_ZZSUE_f-bb-online_Praxishilfe_final.pdf.

Die folgenden Beispielprompts helfen, sich Wissen anzueignen, das man haben sollte, bevor man sich intensiver mit dem Thema beschäftigt und bieten einen guten Ausgangspunkt für eigene Recherchen:

- *„Warum brauchen Unternehmen, respektive KMU eigene Chatbots? Kann nicht HuggingChat einfach alle Fragen beantworten? Antworte in Stichpunkten.“*
- *„Wozu braucht ein KMU einen KI-Chatbot? Antworte als Unternehmensberater in wenigen kurzen Sätzen. Berücksichtige dabei alle Einsatzmöglichkeiten von KI-Chatbots.“*
- *„Welche Einsatzmöglichkeiten gibt es in KMU für KI-Chatbots? Antworte in Stichpunkten.“*
- *„Welche KI-Chatbots passen zu KMU? Antworte als Unternehmensberater in wenigen kurzen Sätzen. Berücksichtige dabei alle Arten von KI-Chatbots und zeige auf wie diese arbeiten.“*
- *„Welche KI-Chatbot eignet sich für welche Aufgabe? Antworte in Stichpunkten.“*
- *„Wie gehen KMU bei der Entwicklung eines eigenen KI-Chatbots vor? Antworte als Unternehmensberater in einem kurzen Fließtext. Gehe dabei auch auf mögliche Herausforderungen und deren Lösung ein.“*

- *„Wie können KMU bei der Implementierung eines KI-Chatbots vorgehen? Was gilt es zu beachten? Antworte in Stichpunkten!“*
- *„Wie wird/ist ein KI-Chatbot aufgebaut und wie funktioniert er? Antworte als Unternehmensberater in einem Fließtext. Gehe dabei auch auf technische Details ein.“*

3. Empfehlungen aus eigener Erfahrung

Bei der Entwicklung der Chatbots für den KI-Experimentierraum (vgl. https://zukunftszentrum-sued.de/ki-experimentierraum/) hat das Zukunftszentrum Süd in der Zusammenarbeit mit dem externen Dienstleister wertvolle Erfahrungen gesammelt, von denen KMU, die selbst einen KI-Chatbot entwickeln möchten, profitieren können.

Eine zentrale Erkenntnis war, dass nicht nur die technischen Anforderungen und Funktionen zu berücksichtigen sind, sondern auch der **Charakter des Chatbots**. Dies betrifft insbesondere die Art und Weise, wie der Chatbot mit Nutzenden interagiert und welche Rolle er innerhalb der Kommunikationsstrategie im Unternehmen spielt.

Die Praxistauglichkeit eines Chatbots für den jeweiligen Einsatzzweck basiert aber nicht nur auf dem Charakterdesign, sondern auch auf intensiven Testphasen. Nur durch kontinuierliches **Testen und Optimieren** kann sichergestellt werden, dass Chatbots mit der Zeit immer fehlerfreier arbeiten. Gerade beim Einsatz von *LLM (Large Lange Models) (vgl. Glossar)* ist das Testen von entscheidender Bedeutung, da solche Sprachmodelle dazu neigen, zu halluzinieren – also Antworten zu geben, die auf falschen Annahmen oder Informationen beruhen.

Bei der Entwicklung unserer KI-Chatbots (Max und Sophie) wurde bewusst auf die ***Open-Source-Sprachmodelle*** *(vgl. Glossar)* zurückgegriffen. Diese bieten die Flexibilität, die Modelle an spezifische Anforderungen anzupassen, und ermöglichen es, die Datensicherheit zu gewährleisten. Trotz der vielen Vorteile, die Open-Source-Sprachmodelle bieten, ist es wichtig, ihre Grenzen zu kennen, insbesondere im Hinblick auf den unbeabsichtigten Sprachwechsel von Deutsch auf Englisch und mögliche Inkonsistenzen. Regelmäßige Aktualisierung und die Einbeziehung von Feedback sind entscheidend, um ihre Leistung langfristig zu verbessern.

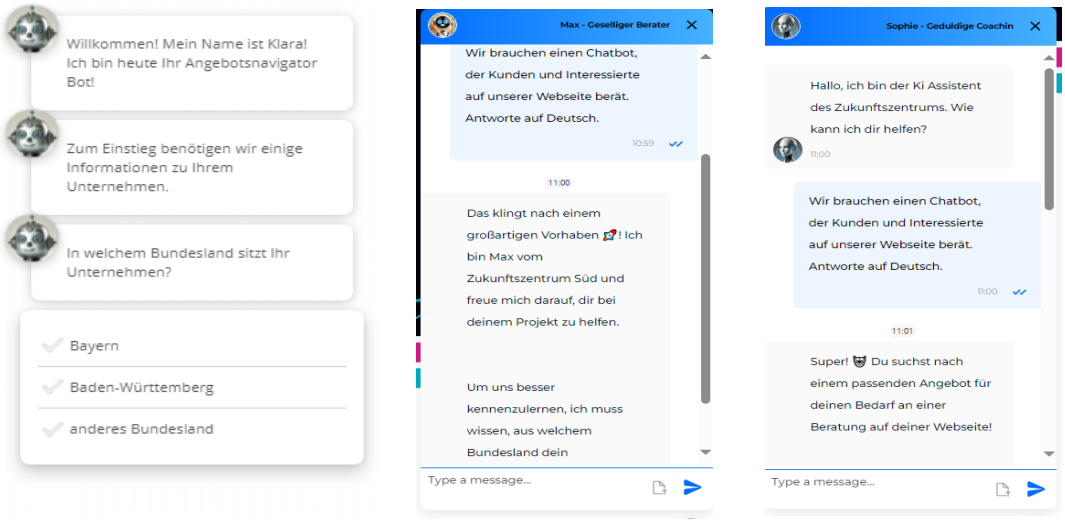
Eine echte Herausforderung war, dass für die Entwicklung und den Betrieb der KI-Chatbots nur ein schmales **Budget** zur Verfügung stand, gleichzeitig sollten hohe Anforderungen an Datenschutz und Sicherheit gewährleistet sein. Diese Faktoren haben den Gestaltungsspielraum und die Möglichkeiten für die technische Umsetzung stark eingeschränkt. Denn Chatbots, die auf großen Sprachmodellen basieren, benötigen erhebliche Rechenressourcen, um nahezu in Echtzeit auf komplexe Anfragen reagieren zu können. Da die notwendigen Serverkapazitäten nicht bereitgestellt werden können, nahmen wir Einbußen bei der Performance bewusst in Kauf, da sonst die Realisierung überhaupt nicht möglich gewesen wäre. Es ist wichtig, Nutzenden das zu kommunizieren, um unrealistische Erwartungen zu vermeiden. Unser Ziel war es, den Nutzenden den Mehrwert eines maßgeschneiderten KI-Chatbots zu

veranschaulichen, der auf ihre spezifischen Bedürfnisse zugeschnitten ist, auch wenn dies manchmal auf Kosten der Antwortgeschwindigkeit geht.

Die **Leistungsfähigkeit** unserer KI-Chatbots kann mit der marktführender KI-Systeme wie ChatGPT von OpenAI nicht mithalten. ChatGPT wird durch einen enormen finanziellen Aufwand und eine hochentwickelte Infrastruktur getragen. Unternehmen wie OpenAI verfügen über spezialisierte Rechenzentren, hochoptimierte Server und enorme Ressourcen, die es ihnen ermöglichen, Anfragen nahezu in Echtzeit zu bearbeiten. Diese immensen finanziellen Mittel führen zu einer außergewöhnlichen Geschwindigkeit und Genauigkeit bei der Beantwortung von Nutzeranfragen.

Entscheidend für die erfolgreiche Implementierung einer KI-gestützten Chatbot-Plattform ist die **Gestaltung des Benutzerinterfaces**. Ein gut durchdachtes *Chatwidget (siehe Glossar)* spielt dabei eine zentrale Rolle, da es die visuelle Schnittstelle bildet, über die die Nutzenden mit der Chatbot-Plattform in Echtzeit kommunizieren können. Die Wahl des richtigen Benutzerinterfaces und die nahtlose Anbindung an bestehende Systeme sind ebenso wichtig wie die Implementierung der Chatbot-Plattform selbst (vgl. Abbildung unten).

Abbildung 1: Chatwidgets mit Beispielanfragen zur textbasierten Spracheingabe bei KI-Chatbots



4. Praxistipps für die agile Zusammenarbeit

Eine enge und effiziente Zusammenarbeit zwischen Dienstleistern und Kunden ist entscheidend für den Erfolg eines Projekts, insbesondere wenn es um komplexe technologische Entwicklungen wie KI-Chatbots geht. Agiles Projektmanagement kann helfen, mit dem oft hohen Grad an Dringlichkeit, Komplexität und Unsicherheit bei der Umsetzung umzugehen. Der iterative Ansatz und der Einsatz agiler Methoden ermöglicht die bei der Entwicklung von KI-Chatbots notwendige flexible Anpassung des Produkts an sich verändernde Anforderungen: Über Experimentieren kommt man zu erprobten Lösungen!

Unternehmen, die wenig Erfahrung mit der Zusammenarbeit in agilen Projekten haben, können einige grundlegende Ansätze und bewährte Praktiken anwenden, um die Zusammenarbeit mit Dienstleistern zu verbessern. Hintergrundinformationen zum agilen Projektmanagement erhalten Unternehmen zum Beispiel über die Plattform „Projekte leicht gemacht“ (vgl. https://projekte-leicht-gemacht.de/projektmanagement/agiles-projektmanagement/#Methoden-im-agilen-Projektmanagement). Das Zukunftszentrum Süd gibt über ein innovatives Qualifizierungsangebot eine erste Einführung in agile Projektarbeit.

Die folgenden Tipps speisen sich aus unseren Erfahrungen bei der Entwicklung des KI-Experimentierraums und richten sich an Unternehmen, die lernen möchten, agiler zu arbeiten und die Zusammenarbeit mit externen Partnern zu optimieren.

1. **Klare Zieldefinition und Kommunikation:** Der erste Schritt zu einer erfolgreichen Zusammenarbeit ist eine klare Definition der Ziele des Projekts. Dazu gehören nicht nur die technischen Anforderungen, sondern auch die erwarteten Ergebnisse und der Zeitrahmen.

Tipp: Schaffen Sie zu Beginn des Projekts ein gemeinsames Verständnis der Ziele. Regelmäßige Besprechungen, in denen die Erwartungen beider Seiten diskutiert und angepasst werden, helfen Missverständnisse zu vermeiden. Ein gut strukturierter Projektplan mit Meilensteinen und klaren Verantwortlichkeiten erleichtert die Zusammenarbeit und stellt sicher, dass alle Beteiligten auf das gleiche Ziel hinarbeiten.

2. **Agile Methoden nutzen:** *Agile Methoden (siehe Glossar)*, wie **Scrum** oder **Kanban**, eignen sich hervorragend für die Zusammenarbeit mit Dienstleistern, da sie Flexibilität ermöglichen und den Fokus auf die kontinuierliche Lieferung von Ergebnissen legen. Durch iterative Zyklen („Sprints“) und regelmäßige Feedbackschleifen können Unternehmen und Dienstleister frühzeitig Verbesserungen vornehmen und auf veränderte Anforderungen reagieren.

Tipp: Führen Sie wöchentliche Sprints oder Sprint-Meetings ein, in denen die Fortschritte besprochen und Aktivitäten priorisiert werden. Ein laufend aktualisiertes Produkt-Backlog hilft, die nächsten Schritte klar zu priorisieren und die Ressourcen effizient einzusetzen. Nutzen Sie Tools wie **Trello, Jira oder Asana**, um Aufgaben zu verteilen und den Fortschritt in Echtzeit zu verfolgen.

3. **Transparenz und offenes Feedback:** Ein weiterer zentraler Aspekt der agilen Zusammenarbeit ist der offene Umgang mit Problemen und Fortschritten. Dienstleister und Kunden sollten eine Kultur der Transparenz pflegen, die es erlaubt, Fehler anzusprechen und daraus zu lernen.

Tipp: Organisieren Sie regelmäßige **Retrospektiven**, um zu evaluieren, was gut läuft und wo Verbesserungen notwendig sind. Diese Treffen schaffen ein Umfeld, in dem Kritik konstruktiv geäußert und Prozesse verbessert werden können. Offenes und schnelles Feedback, auch

zwischen den regulären Treffen, stellt sicher, dass Probleme sofort erkannt und angegangen werden können.

4. **Verantwortung teilen:** Ein häufiger Fehler bei der Zusammenarbeit mit Dienstleistern besteht darin, dass das Projekt entweder komplett an den Dienstleister abgegeben wird oder der Kunde zu stark eingreift. Für eine erfolgreiche agile Zusammenarbeit ist es wichtig, dass beide Seiten Verantwortung übernehmen.

Tipp: Definieren Sie klare Verantwortlichkeiten. Dienstleister sollten für die technische Umsetzung verantwortlich sein und fachliche Expertise bereitstellen, während der Kunde die geschäftlichen Anforderungen und strategischen Entscheidungen festlegt. Gemeinsame Entscheidungsfindung und offene Diskussionen fördern das Vertrauen und schaffen ein ausgewogenes Verhältnis zwischen Kontrolle und Flexibilität.

5. **Schnelle Entscheidungsfindung:** Eine der größten Hürden in agilen Projekten sind langwierige Entscheidungsprozesse. Unternehmen, die die Zusammenarbeit mit Dienstleistern verbessern wollen, sollten darauf achten, die Entscheidungswege so kurz wie möglich zu halten. Lange Wartezeiten auf Feedback oder Entscheidungen verlangsamen das Projekt und können die Produktivität des Dienstleisters beeinträchtigen.

Tipp: Stellen Sie sicher, dass es immer eine verantwortliche Person gibt, die schnell Entscheidungen treffen kann, sei es auf Kundenseite oder beim Dienstleister. **Daily Standups** helfen, kleinere Entscheidungen sofort zu treffen, und stellen sicher, dass alle Beteiligten auf dem gleichen Stand sind. So werden unnötige Wartezeiten vermieden und ein reibungsloser Projektverlauf gewährleistet.

6. **Dokumentation und Wissensaustausch**: Obwohl agile Projekte oft wenig formale Dokumentation erfordern, ist es dennoch wichtig, dass der Wissensaustausch zwischen Dienstleister und Kunde kontinuierlich und klar erfolgt. Gerade in komplexen Projekten wie der Entwicklung von LLM-Chatbots müssen technische Spezifikationen und Entscheidungsprozesse nachvollziehbar bleiben.

Tipp: Erstellen Sie eine gemeinsame Wissensbasis, zum Beispiel in Form eines **Wikis** oder einer **Cloud-basierten Dokumentation**. Hier sollten alle relevanten Informationen zu Projektanforderungen, Projektfortschritt und technischen Details festgehalten werden. So können alle Beteiligten auch bei Personalwechseln oder längeren Pausen im Projekt schnell wieder einsteigen.

7. **Flexibilität und Anpassungsfähigkeit:** Auch bei klar definierten Zielen kann es vorkommen, dass sich die Anforderungen während des Projekts ändern oder neue Erkenntnisse hinzukommen. Unternehmen sollten als Kunden offen für solche Änderungen sein und die Dienstleister bei den notwendigen Anpassungen unterstützen.

Tipp: Stellen Sie sicher, dass der Projektplan flexibel genug ist, um auf Veränderungen reagieren zu können. Planen Sie Pufferzeiten für Unvorhergesehenes ein und fördern Sie eine Kultur der **Anpassungsfähigkeit**. Agile Projekte leben von der Fähigkeit, schnell auf neue Herausforderungen zu reagieren und sich entsprechend neu auszurichten.

8. **Einbeziehung des Kunden:** Für eine erfolgreiche Zusammenarbeit ist es entscheidend, dass der Kunde in den gesamten Entwicklungsprozess einbezogen ist. Nur so können Anforderungen präzise vermittelt und Missverständnisse vermieden werden. Auch der regelmäßige Austausch von Feedback hilft dabei, das Projekt zielgerichtet zu steuern.

Tipp: Schulung und Onboarding des Kunden-Teams sind entscheidend, damit alle auf demselben Wissenstand sind. Regelmäßige Status-Updates und die Beteiligung des Kunden an Sprint-Reviews oder Retrospektiven stellen sicher, dass das Projekt den Erwartungen entspricht und der Kunde ein aktiver Partner im Prozess bleibt.

5. KI-Chatbots selbst entwickeln – so sind wir vorgegangen

Im folgenden Erfahrungsbericht zeigen wir auf, wie wir bei der Entwicklung der Chatbots für den KI-Experimentierraum des Zukunftszentrums Süd vorgegangen sind. Wir geben detaillierte Einblicke von ersten Designüberlegungen über die technischen Spezifikationen bis hin zur Server- und Infrastrukturplanung sowie möglichen Stolpersteinen und den Umgang damit. Ein besonderes Augenmerk legen wir dabei auf Rahmenbedingungen wie Anforderungen des Datenschutzes und die Kostenplanung. Diese plastische Darstellung soll Unternehmen, die selbst vor der Entscheidung stehen, einen KI-Chatbot zu entwickeln, Hinweise geben worauf dabei insbesondere zu achten ist.

5.1. Charakterdesign

Als Beratungslotsen für das Zukunftszentrum Süd haben wir drei unterschiedliche Charaktere entwickelt, die auf verschiedene Bedürfnisse und Szenarien zugeschnitten sind:

Klara – die Logikexpertin steht für Zuverlässigkeit und Struktur. Sie arbeitet einen klar definierten Prozess ab und bietet eine standardisierte Form der Beratung an. Ihre Stärke liegt in der Lotsenberatung. Algorithmisch kann sie Unternehmen auf Unterstützungsangebote verschiedener öffentlich geförderter Initiativen verweisen. Klara führt Nutzende durch vordefinierte Fragebäume und gibt eindeutige Antworten zu festgelegten Themen.

Max – der gesellige Berater ist ein KI-Chatbot, der auf einem leistungsfähigen Open-Source-Sprachmodell basiert. Seine Stärke liegt in der sozialen Interaktion. Er kann schnell nützliche Ratschläge geben. Max wurde entwickelt, um eine entspannte und freundliche Atmosphäre zu schaffen. Er ist gesellig, manchmal vielleicht etwas zu gesellig, und neigt dazu, sich bei längeren Gesprächen zu wiederholen oder ins Englische zu wechseln.

Sophie – die geduldige Coachin ist ebenfalls ein KI-Chatbot, der auf einem Open-Source-Sprachmodell (LLM) basiert. Bei ihr wird das LLM-Modell jedoch mit der Retrieval-Augmented-Generation *(RAG siehe Glossar)* kombiniert. Das bedeutet, dass sie Informationen aus vorgegebenen Quellen abrufen kann, um genauere und hilfreichere Antworten zu geben. Dies macht sie zu einer erfahrenen und geduldigen Beraterin, deren Schwerpunkt auf einer gründlichen Problemanalyse und dem Abrufen fundierter Informationen aus einer umfangreichen Datenbank liegt. Sie nimmt sich Zeit für ihre Antworten und eignet sich besonders für Situationen, in denen komplexe und detaillierte Informationen benötigt werden.

Jeder dieser Chatbots simuliert eine einzigartige Persönlichkeit und Funktion, die durch Designentscheidungen und spezifische Technologien unterstützt werden (vgl. Tabelle unten). Jeder Chatbot wurde auf Benutzerfreundlichkeit, Konsistenz und Performance getestet.

Tabelle 1: Technische Merkmale und Anwendungsfälle der Chatbots im Vergleich

Art des Chatbots	Technische Merkmale	Anwendungsfall
 **Klara - Die Logikexpertin** (Regelbasierter Chatbot)	Klara ist ein regelbasierter Chatbot, der konsistent arbeitet, d.h. auf gleiche Eingaben immer die gleiche Ausgabe liefert. Sie greift auf Informationen aus einer Datenbank zu und erzeugt über vordefinierte Prozesse Antworten. Klara verwendet keine freien Sprachmodelle *(Open Source LLMs)* und kann freie Texteingaben nicht verarbeiten. Dies schränkt ihre Flexibilität ein, sorgt aber für eine hohe Verbindlichkeit der Antworten.	Regelbasierte Chatbots wie Klara sind ideal für Szenarien, in denen Anwendende klar definierte Fragen haben und strukturierte Antworten benötigen, z.B. im IT-Support oder bei standardisierten Prozessen.
 **Max - Der gesellige Berater** (LLM-basierter Chatbot)	Max basiert auf einem (Open Source) LLM-Modell. Dadurch kann er flexibel auf unterschiedliche Anfragen reagieren, allerdings auf Kosten der Konsistenz. Gelegentlich „vergisst" er Details oder wiederholt sich. Da er nicht auf eine definierte Datenbank zurückgreift, gibt er auch falsche oder irreführende Antworten. Max neigt dazu, bei längeren Gesprächen ins Englische zu wechseln. Dies ist auf eine technische Einschränkung des Sprachmodells zurückzuführen.	LLM-basierte Chatbots wie Max eignen sich hervorragend für interaktive und dialogische Anwendungen, bei denen eine freundliche, informelle Beratung gefragt ist, wie z.B. im Verkauf oder bei der Produktberatung.
 **Sophie - Die geduldige Coachin** (LLM-basierter Chatbot mit RAG-System)	Sophie verwendet eine Kombination aus einem LLM und einem RAG-System, das es ihr ermöglicht, in Echtzeit auf eine externe Datenbank zuzugreifen und fundierte Informationen zu liefern. Manchmal springt sie zwischen verschiedenen Datensätzen hin und her, was gelegentlich zu Inkonsistenzen führt. Da sie auf in einer Datenbank hinterlegte Informationen zugreift, antwortet sie zwar langsamer, aber wesentlich präziser und zuverlässiger als Max.	Durch die Verbindung eines freien Sprachmodells (LLM) mit Datenbankwissen (RAG-System) kann ein KI-Chatbot wie Sophie zu einer Vielzahl von Themen detailliert beraten.

5.2. Technische Umsetzung

Die Einbindung von Large Language Models (LLMs) in unsere Anwendungen und Arbeitsabläufe ist für uns entscheidend, um maßgeschneiderte KI-Lösungen zu entwickeln. Wir setzen dabei auf *Ollama (siehe Glossar)*, eine Plattform, die eine einfache Möglichkeit bietet, Sprachmodelle lokal zu hosten und effizient in unsere Systeme zu integrieren. Diese Plattform ermöglicht es, Modelldateien herunterzuladen und direkt auf unseren eigenen Servern zu betreiben, ohne auf externe Cloud-Dienste angewiesen zu sein. Dadurch haben wir die volle Kontrolle über unsere Daten und können die Modelle an unsere spezifischen Anforderungen anpassen. Wir schätzen insbesondere die Möglichkeit, Ollama über Python *(siehe Glossar)* anzusteuern, da dies eine direkte Integration in unsere bestehenden Software-Stacks ermöglicht. Über die *API-Schnittstellen (siehe Glossar*) von Ollama können wir LLMs einfach in Python-basierte Anwendungen einbinden. Dies erleichtert uns die Entwicklung von Anwendungen, die von der Textgenerierung über Datenanalyse bis hin zur Automatisierung komplexer Abläufe reichen. Wir haben die Sprachmodelle Llama 3.1, Llama 2, Mistral, phi3, zephyr und command-r getestet. Die Auswahl der Modelle, die für uns sinnvoll einsetzbar waren, reduzierte sich schnell. Einige Modelle waren zu ressourcenintensiv (command-r) und andere lieferten teilweise unnatürliche Dialoge, weil sie z. B. nicht ausreichend auf Deutsch trainiert waren (Llama 2, phi3 und zephyr). Somit blieben nur die Modelle von Mistral und Llama 3.1 übrig. Beide Modelle bieten gute Leistungen bei der Verarbeitung natürlicher Sprache und könnten theoretisch unsere Anforderungen abdecken, wobei Mistral attraktiv ist, wenn es darum geht, spezifische Anwendungsfälle mit geringen Hardwareanforderungen zu unterstützen. Es ist flexibel anpassbar und eignet sich gut für gezielte Feinabstimmungen auf bestimmte Datensätze und Sprachen. Llama wiederum bietet mehrere Modellgrößen (7B, 13B, 30B), die es uns ermöglichen, das Modell auf unsere Hardware abzustimmen. Diese Vielfalt stellt sicher, dass wir auch auf weniger leistungsstarken Maschinen eine relativ stabile und zuverlässige KI-Leistung erzielen können. In unseren Tests hat sich Llama als robust und vielseitig erwiesen, so dass wir es in vielen Bereichen einsetzen können, ohne ständig Anpassungen vornehmen zu müssen. Aus diesem Grund haben wir uns letztendlich für den Einsatz von Llama entschieden. Es bleibt noch zu erwähnen, dass wir uns auch mit der Integration von Sprachmodellen beschäftigt haben, die über *HuggingFace (siehe Glossar)* bereitgestellt werden. HuggingFace bietet eine umfassende Plattform mit einer großen Auswahl an Modellen. Die Modelle können in der Cloud trainiert werden und Nutzende profitieren von der einfachen Bereitstellung über APIs. Die aus Datenschutzgründen notwendige lokale Bereitstellung von huggingface-Modellen erwies sich jedoch als zu ressourcenintensiv. Ein sinnvoller Betrieb auf unserer Hardware war daher nicht möglich.

5.3. Budget und Kostenfaktoren

LLM-basierte Chatbots benötigen erhebliche Rechenressourcen, um nahezu in Echtzeit auf komplexe Anfragen reagieren zu können (vgl. Kasten unten). Für die Realisierung des KI-Experimentierraums stand aber nur ein sehr begrenztes Budget für Serverkosten zu Verfügung.

Dies stellte eine erhebliche Hürde dar, da moderne LLM-basierte Chatbots in der Regel auf Servern mit deutlich höheren Kosten betrieben werden müssen. Je nach Komplexität des Modells, der Menge an zu verarbeitenden Anfragen und der notwendigen Rechenleistung liegen die Kosten aktuell typischerweise zwischen 600 € und 2.000 € pro Monat.

Technische Daten zur Server-Konfiguration:

- CPU: 4-8 virtuelle Kerne
- GPU: Keine dedizierte GPU (bei Max und Sophie mussten wir auf optimierte Modelle zurückgreifen, damit sie überhaupt auf CPU-basierter Infrastruktur laufen konnten)
- RAM: 16-32 GB
- Speicherplatz: 500 GB SSD
- Netzwerk: Geteilte Gigabit-Leitungen

Mit nur 130 € Budget für Serverkosten pro Monat mussten wir kreative Lösungen finden, um die KI-Chatbots überhaupt lauffähig zu machen. Das bedeutete, dass wir auf kostengünstigere Server zurückgreifen mussten, die in der Regel weniger Rechenleistung und Speicherplatz bieten. Konkret haben wir uns für kleinere VPS (Virtual Private Server) ohne GPU-Unterstützung entschieden (technische Daten siehe Kasten).

Typische Serveranforderungen für LLM-Chatbots:

CPU vs. GPU *(siehe Glossar)*: Während CPUs für einfache Aufgaben wie regelbasierte Chatbots (wie Klara) ausreichen, benötigen LLMs (wie Max und Sophie) leistungsfähige GPUs, um die Berechnungen für die Sprachverarbeitung und die Antwortgenerierung in akzeptabler Geschwindigkeit zu leisten (typische Server für LLMs verwenden GPUs wie **NVIDIA`s A100** oder **V100**, die für das Training und die Ausführung von KI-Modellen optimiert sind).

Speicherbedarf: LLMs benötigen oft große Mengen an Arbeitsspeicher (RAM), um die erforderlichen Datenmengen zu verarbeiten und die komplexen Berechnungen zu bewältigen. Für unsere Anwendungsfälle wären mindestens **64 GB RAM** pro Server erforderlich.

Speicherplatz: Die Modelle selbst und die Daten, die sie verarbeiten, benötigen ebenfalls viel Speicherplatz. Für einfache Sprachmodelle sind **4TB SSD-Speicher** keine Seltenheit.

Netzwerkbandbreite: Um eine reibungslose Benutzererfahrung zu gewährleisten, ist eine gute Netzwerkbandbreite erforderlich, insbesondere wenn der Chatbot viele parallele Anfragen bearbeiten muss.

Während diese Server für die weniger anspruchsvollen Aufgaben von Klara ausreichen, stoßen sie bei komplexeren Anfragen wie denen von Max und Sophie an ihre Grenzen. Diese beschränkte Serverkapazität führte zu einer Reihe von Kompromissen in Bezug auf Geschwindigkeit und der Parallelisierung: Die Verarbeitungszeit, insbesondere bei Sophie, ist aufgrund der fehlenden GPU-Unterstützung deutlich langsamer als idealerweise gewünscht. Die Anzahl der gleichzeitigen Benutzerinteraktionen ist stark eingeschränkt. Während ein dedizierter Server mit GPUs problemlos Dutzende von gleichzeitigen Anfragen verarbeiten kann, stößt unser Server bereits bei wenigen gleichzeitigen Anfragen an seine Leistungsgrenze. Ohne die

Möglichkeit, auf teure GPU-unterstützte Server oder hochoptimierte Cloud-Infrastrukturen zurückzugreifen, kann die Geschwindigkeit und Flexibilität des Chatbots nicht mit der eines ChatGPT mithalten. Es ist daher zu erwarten, dass Nutzende, die an die Reaktionsgeschwindigkeit von ChatGPT gewöhnt sind, eine verzögerte Antwortzeit und gelegentlich weniger genaue Ergebnisse bei den eigenen Chatbots nicht tolerieren werden.

5.4. Benutzerinterface und Integration: Anbindung an bestehende Systeme

Für eine erfolgreiche Implementierung einer KI-gestützten Chatbot-Plattform ist die Gestaltung des Benutzerinterfaces entscheidend. Bei der Entwicklung, der Integration und dem Betrieb einer Chatbot-Lösung haben wir insbesondere folgende Aspekte berücksichtigt:

1. Benutzerfreundlichkeit (User Experience)

Das Chatwidget sollte einfach und intuitiv zu bedienen sein, um die Akzeptanz der Nutzenden zu fördern. Dies umfasst:

- *Responsive Design:* Das Chatwidget sollte auf verschiedenen Geräten (Desktop, Tablet, Smartphone) optimal funktionieren, um eine konsistente Nutzererfahrung zu gewährleisten.
- *Klarer Einstiegspunkt:* Das Chatwidget sollte leicht auffindbar und zugänglich sein, z. B. durch ein gut platziertes Icon auf der Website oder in der App.
- *Visuelle Gestaltung:* Eine ansprechende, zum Markenauftritt passende Gestaltung schafft Vertrauen und motiviert zur Interaktion.

2. Anbindung an bestehende Systeme

Die Integration der Chatbot-Plattform in bestehende IT-Systeme und Systeme des *Customer Relationship Management (CRM siehe Glossar)* ist entscheidend, um eine konsistente und effiziente Nutzererfahrung zu gewährleisten. Hier sind die wesentlichen Elemente der Integration:

- *Website oder App:* Das Chatwidget ist die Benutzeroberfläche, über die Nutzende mit der Chatbot-Plattform kommunizieren. Es sollte nahtlos in bestehende Web- oder App-Infrastrukturen eingebettet werden. Dies kann durch JavaScript-Widgets, Plugins oder API-basierte Einbindungen erfolgen, sodass die Nutzer auf einfache Weise Zugang zur Chatbot-Funktionalität erhalten.
- *CRM-Systeme und Datenbanken:* Damit die Chatbot-Plattform personalisierte Antworten geben kann, ist es wichtig, dass sie auf Daten aus bereits genutzten CRM-Systemen oder internen Datenbanken des Unternehmens zugreifen kann. Das Chatwidget übermittelt die Nutzereingaben an die Plattform, die auf Basis der CRM-Daten individuell angepasste Antworten generiert. Die Datenverarbeitung und -integration erfolgt hierbei auf der Plattform, während das Chatwidget als Vermittler fungiert.
- *Support- und Ticketsysteme:* In Fällen, in denen die Chatbot-Plattform nicht in der Lage ist, eine Anfrage vollständig zu bearbeiten, sollte sie Anfragen an menschliche Mitarbeiter oder ein Support-Ticketsystem weiterleiten können. Dies stellt sicher, dass

komplexe Anliegen nicht unbeantwortet bleiben und eine nahtlose Übergabe an den menschlichen Support erfolgt.

3. API-Anbindung zwischen Chatwidget und LLM

Die Kommunikation zwischen dem Chatwidget und einem Large Language Model (LLM) erfolgt häufig über eine API, die eine reibungslose Datenübertragung ermöglicht. Die API-Verbindung zwischen dem Chatwidget und dem LLM kann mithilfe von Python realisiert werden, beispielsweise über einen *Flask-Server (siehe Glossar)*. Je nach Anwendungsfall können verschiedene Parameter wie Modelltyp oder Konversationskontext angepasst werden, um die Antworten der Chatbot-Plattform besser auf den Kontext der Nutzenden abzustimmen.

Ablauf der Kommunikation: Das Chatwidget sendet die Nutzereingaben an die LLM-API, empfängt die Antwort und stellt sie den Nutzenden zur Verfügung. Dies sorgt für eine natürliche und flüssige Konversation, da die Nutzenden direkt auf ihre Eingaben eine Antwort erhalten.

4. Technische Anforderungen und Sicherheit

Neben der Funktionalität muss die gesamte Chatbot-Integration sicher gestaltet sein, insbesondere wenn sie in Umgebungen eingesetzt wird, in denen sensible Daten verarbeitet werden. Hier sind die wichtigsten Aspekte:

- *Datenschutzkonformität:* Die Chatbot-Plattform und das Chatwidget müssen sicherstellen, dass alle Daten gemäß den geltenden Datenschutzgesetzen (wie der DSGVO) verarbeitet und übertragen werden. Das Chatwidget als Schnittstelle sollte Daten nur verschlüsselt übertragen, während die Plattform die Datenverarbeitung übernimmt.
- *Sicherheitsprotokolle:* Die Verbindung zwischen den Nutzenden und der Chatbot-Plattform sollte durch *TLS-Verschlüsselung (Transport Layer Security siehe Glossar)* abgesichert werden, um eine sichere Kommunikation zu gewährleisten. Dies gilt sowohl für die Übertragung der Daten vom Chatwidget zur Plattform als auch für die Kommunikation innerhalb der Backend-Systeme.
- *Skalierbarkeit:* Die Architektur der Chatbot-Plattform und des Chatwidgets sollte so ausgelegt sein, dass sie bei steigenden Nutzendenzahlen skalierbar ist, ohne an Performance zu verlieren. Das bedeutet, dass das Chatwidget auch bei hohem Datenaufkommen reibungslos mit der Plattform kommunizieren kann.

5. Anpassungsfähigkeit und Flexibilität

Das Chatwidget sollte sich flexibel an die Bedürfnisse des Unternehmens anpassen lassen, um eine einheitliche Nutzererfahrung zu bieten:

- *Branding:* Farben, Logos und Schriftarten des Chatwidgets sollten an das Corporate Design des Unternehmens angepasst werden können, um eine konsistente Markenpräsenz zu gewährleisten.

- *Funktionsumfang:* Je nach Anwendungsfall kann das Chatwidget so konfiguriert werden, dass es bestimmte Aufgaben priorisiert (z. B. Support, Verkauf, Beratung) und entsprechende Funktionen anbietet (z. B. Buttons, Auswahlfelder, Textfelder). Die Anpassung erfolgt häufig über Konfigurationen, die in der Chatbot-Plattform festgelegt und über das Chatwidget sichtbar gemacht werden.

5.5. Datenschutz und Compliance

Neben den technischen und budgetären Einschränkungen spielte der Datenschutz eine zentrale Rolle in der Entscheidung für unsere Serverinfrastruktur. Da wir personenbezogene Daten von Nutzenden und Unternehmen verarbeiten, mussten wir sicherstellen, dass alle rechtlichen Anforderungen – insbesondere die **Datenschutz-Grundverordnung (DSGVO)** – vollständig eingehalten werden. Dies schloss den Einsatz von Servern in unsicheren Drittstaaten kategorisch aus (vgl. Kasten unten).

Anforderungen an den Datenschutz

- **Serverstandort:** Server müssen sich innerhalb der **Europäischen Union** befinden, um den strengen Richtlinien der DSGVO zu entsprechen. Cloud- oder Serverdienste aus Drittstaaten, insbesondere den USA, kamen wegen möglicher Sicherheitsrisiken und der nicht garantierten Einhaltung europäischer Datenschutzstandards nicht in Frage.
- **Verschlüsselung:** Die Daten, die zwischen den Chatbots und den Nutzenden ausgetauscht werden, müssen durchgehend verschlüsselt werden. Wir haben **TLS 1.3** für die Übertragung verwendet, um sicherzustellen, dass alle Kommunikationskanäle sicher sind.
- **Speicherung personenbezogener Daten:** In Übereinstimmung mit der DSGVO mussten wir sicherstellen, dass keine personenbezogenen Daten länger als notwendig gespeichert werden. Hier haben wir strikte Richtlinien implementiert, um die Datenverarbeitung auf das Nötigste zu beschränken.

Technische Maßnahmen zur Einhaltung der Compliance

- **Datenminimierung:** Wir haben darauf geachtet, dass die Chatbots nur die Daten erheben, die für den jeweiligen Anwendungsfall absolut notwendig sind. So wurden z.B. Standortdaten und branchenspezifische Informationen von Unternehmen von Max und Sophie nur auf explizite Eingabe der Nutzenden erfasst.
- **Anonymisierung und Pseudonymisierung:** Wo immer möglich, wurden die erfassten Daten anonymisiert oder pseudonymisiert, um das Risiko bei einem möglichen Datenleck zu minimieren.
- **Sicherheitsüberprüfungen:** Regelmäßige Sicherheitsüberprüfungen der Serverstruktur sorgen dafür, dass keine Schwachstellen in der Infrastruktur ausgenutzt werden können.

5.6. Lessons Learned

Die Serverinfrastruktur erwies sich als eine der größten Herausforderungen des Projekts. Trotz erheblicher Budgeteinschränkungen konnten wir durch kreative Optimierungen und eine sorgfältige Auswahl der Server eine funktionierende Lösung schaffen. Der größte Lerneffekt war, dass die Hardwareauswahl entscheidend für den langfristigen Erfolg eines LLM-basierten Chatbots ist. Die Grenzen der Rechenleistung und die fehlende GPU-Unterstützung führten zu Leistungseinbußen und damit zu Verzögerungen bei der Reaktionszeit der Antworten, was in zukünftigen Projekten berücksichtigt werden muss.

6. Fazit

Dieser Erfahrungsbericht macht deutlich, dass der Weg zu einem erfolgreichen KI-Chatbot nicht nur technisches Know-how erfordert, sondern auch eine klare Strategie, kreative Lösungen und eine starke Zusammenarbeit zwischen allen Beteiligten. Von der ersten Designüberlegung bis zur finalen Implementierung gibt es viele Stolpersteine, die jedoch mit einem agilen Ansatz und den richtigen Tools gemeistert werden können.

Trotz eines streng limitierten Budgets und den hohen Anforderungen an Datenschutz und Serverleistung ist es gelungen, mit Klara, Max und Sophie drei einzigartige Chatbot-Charaktere zu entwickeln, die unterschiedliche Bedürfnisse abdecken. Sie sind nicht nur Beispiele dafür, wie man mit frei verfügbaren Sprachmodellen und kreativen Serverlösungen ein leistungsfähiges System aufbaut, sondern auch, wie man mit flexiblen, agilen Methoden ein Projekt erfolgreich vorantreibt. Diese Chatbot-Charaktere können nun ggf. zu einem späteren Zeitpunkt bei ausreichendem Budget mit wenig Aufwand auf leistungsfähigere Server übertragen werden, um dann auch die Antwortgeschwindigkeit zu verbessern. Dies ist derzeit noch eine der größten Hürden.

Für Unternehmen, die erste Schritte im Bereich KI-Chatbots gehen oder ihre bestehenden Prozesse optimieren möchten, bietet der Erfahrungsbericht wertvolle Anregungen. Er zeigt, dass auch unter schwierigen Bedingungen hochwertige Lösungen entwickelt werden können, wenn die Zusammenarbeit zwischen Kunden und Dienstleistern auf Transparenz, Flexibilität und einer klaren Zieldefinition basiert. Am Ende entscheidet nicht allein die Technologie über den Erfolg eines KI-Chatbots – es ist das Zusammenspiel von Mensch, Maschine und Methodik. Unternehmen, die diese Prinzipien umsetzen, schaffen die Grundlage für nachhaltigen Erfolg in der Kundeninteraktion und Effizienzsteigerung im Alltag.

7. Gut beraten durch das Zukunftszentrum Süd

Das Zukunftszentrum Süd berät KMU, wie sie Künstliche Intelligenz (KI) gewinnbringend einsetzen können. Wir begleiten Unternehmen in der Transformation und zeigen ihnen

Entwicklungsmöglichkeiten auf; ein möglicher Anwendungsfall ist die Unterstützung der Kommunikation mit Kunden oder auch unternehmensintern durch KI-Chatbots.

Für Unternehmen, die erste Erfahrungen mit KI sammeln möchten, stehen verschiedene Qualifizierungsangebote zur Verfügung. In unserem interaktiven KI-Planspiel erfahren Unternehmen, wie sie KI strategisch implementieren können; auch hier und in weiteren Qualifizierungen sind KI-Chatbots und ihre Einsatzmöglichkeiten im Marketing, in der Produktion, im Personalbereich und der unternehmensinternen Kommunikation ein Thema. Eingegangen wird dabei auf die Möglichkeit der Eigenentwicklung, aber auch auf Standardlösungen. Im Workshop „Generative KI verstehen und einsetzen“ können Unternehmen beispielsweise den Umgang mit generativen KI-Systemen, u.a. mit ChatGPT, üben.

Das Zukunftszentrum Süd wird im Rahmen des Programms „Zukunftszentren“ durch das Bundesministerium für Arbeit und Soziales (BMAS) und die Europäische Union über den Europäischen Sozialfonds Plus (ESF Plus) sowie anteilig durch die Landesministerien für Wirtschaft in Bayern und Baden-Württemberg gefördert. Umgesetzt wird das Zukunftszentrum Süd durch das Forschungsinstitut Betriebliche Bildung (f-bb) am Standort Nürnberg im Verbund mit dem Bildungswerk der Bayerischen Wirtschaft (bbw), dem Bildungswerk der Baden-Württembergischen Wirtschaft (BIWE) und der Technischen Hochschule Deggendorf. Unser öffentlich gefördertes Angebot richtet sich an Unternehmen mit Sitz in Bayern oder Baden-Württemberg. Weitere Informationen dazu finden Sie auf der **Website unter https://zukunftszentrum-sued.de**.

Ziel der Europäischen Union ist es, dass alle Menschen eine berufliche Perspektive erhalten. Der Europäische Sozialfonds Plus (ESF Plus) trägt zu einem sozialeren Europa bei und setzt die Europäische Säule sozialer Rechte in die Praxis um. Er investiert vor Ort in Maßnahmen, um Menschen bei der Bewältigung wirtschaftlicher und sozialer Herausforderungen zu unterstützen und ihre Beschäftigungschancen zu verbessern. Der ESF Plus unterstützt die Menschen durch Ausbildung und Qualifizierung und trägt zum Abbau von Benachteiligungen auf dem Arbeitsmarkt bei. Er fördert Gründer*innen und hilft kleinen und mittleren Unternehmen (KMU) bei der Fachkräftesicherung. Mehr zum ESF unter: www.esf.de.

8. Glossar

Agile Methoden (Scrum, Kanban): Agile Methoden sind Ansätze im Projektmanagement, die durch kurze Entwicklungszyklen und regelmäßige Anpassungen an Veränderungen geprägt sind. Sie ermöglichen eine flexible und reaktionsschnelle Arbeitsweise (Scrum und Kanban sind zwei populäre Frameworks für die agile Projektverwaltung und Softwareentwicklung).

API (Application Programming Interface): Eine Schnittstelle, über die verschiedene Softwaresysteme miteinander kommunizieren können. In diesem Projekt wird die API verwendet, um den Chatbot mit anderen Anwendungen oder Datenbanken zu verknüpfen.

Chatwidget: Ein Chatwidget ist das visuelle Interface (Benutzeroberfläche), über das Nutzer mit einem Chatbot interagieren können. Es erscheint typischerweise als kleines Chatfenster auf einer Website oder in einer App und ermöglicht eine direkte, textbasierte oder sprachgesteuerte Kommunikation zwischen Nutzer und Chatbot.

CRM (Customer Relationship Management): Ein System zur Verwaltung von Kundenbeziehungen. Durch die Integration von Chatbots mit einem CRM-System können Unternehmen personalisierte Kundendaten nutzen, um gezielte und individuelle Antworten zu geben.

Datenpseudonymisierung: Ein Verfahren, bei dem personenbezogene Daten so verarbeitet werden, dass sie ohne zusätzliche Informationen keiner bestimmten Person zugeordnet werden können. Dies dient dem Schutz der Privatsphäre.

DSGVO (Datenschutz-Grundverordnung): Eine Verordnung der Europäischen Union, die den Schutz personenbezogener Daten und die Privatsphäre von Nutzern regelt. Unternehmen müssen sicherstellen, dass alle Systeme, einschließlich Chatbots, DSGVO-konform sind.

CPU (Central Processing Unit): CPU steht für "Central Processing Unit" und bedeutet übersetzt so viel wie zentrale Verarbeitungseinheit. Die CPU ist der Hauptprozessor eines Computers und damit das Herzstück eines Rechners.

Flask-Server: Flask ist ein Microframework für Webanwendungen, das es Entwicklern ermöglicht, schnell und einfach Webdienste zu erstellen, die auf Python basieren und flexibel einsetzbar sind. Ein Flask-Server ist ein Server, der auf diesem Framework basiert und für den Betrieb von Webanwendungen sorgt.

GPU (Graphics Processing Unit): Eine GPU (Graphics Processing Unit) ist ein für parallele Berechnungen optimierter Prozessor, der häufig für komplexe KI-Modelle und maschinelles Lernen eingesetzt wird. GPUs sind für die schnelle Verarbeitung großer Datenmengen in LLMs unerlässlich.

HuggingFace: HuggingFace ist eine Open-Source-Plattform für maschinelles Lernen, die eine umfangreiche Sammlung von vortrainierten Modellen und Tools für natürliche Sprachverarbeitung (NLP) und künstliche Intelligenz (KI) bereitstellt. Die Plattform ermöglicht Entwicklern den einfachen Zugriff und die Integration von leistungsfähigen KI-Modellen in ihre eigenen Anwendungen.

KI-Chatbot (Künstliche Intelligenz Chatbot): Ein Computerprogramm, das mittels künstlicher Intelligenz in der Lage ist, natürliche Sprache zu verstehen und darauf zu reagieren. Es simuliert eine menschliche Konversation, um Nutzern bei verschiedenen Anfragen zu helfen.

LLM (Large Language Model): Ein Sprachmodell, das auf einer großen Menge von Textdaten trainiert wurde und komplexe Sprachverarbeitung ermöglicht. LLMs wie GPT-3 oder LLaMA können menschliche Sprache verstehen, Texte generieren und auf Fragen antworten.

NLP (Natural Language Processing): Natürliche Sprachverarbeitung ist ein Teilbereich der künstlichen Intelligenz, der sich damit beschäftigt, wie Computer menschliche Sprache verstehen und verarbeiten können. Es ist die Grundlage, auf der KI-Chatbots aufgebaut sind.

Ollama: Ollama ist eine Plattform, die es ermöglicht große Sprachmodelle direkt auf den eigenen Computer herunterzuladen und zu hosten. Die Plattform unterstützt eine Vielzahl von LLMs, darunter Metas / Facebooks Llama 3, Microsofts Phi 3, Mistral von minstral.ai, Googles Gemma und vielen weiteren.

Open-Source: Software oder Modelle, deren Quellcode öffentlich zugänglich ist, sodass sie frei genutzt, verändert und verteilt werden können. In diesem Kontext wurden Open-Source-LLMs für die Entwicklung der Chatbots verwendet.

Python: Python ist eine hochentwickelte Programmiersprache, die für die Entwicklung von Software, Skripten und Anwendungen verwendet wird und sich durch ihre Lesbarkeit, Flexibilität und Vielseitigkeit auszeichnet. Sie ist eine der beliebtesten Programmiersprachen und wird in vielen Bereichen wie Webentwicklung, Datenanalyse, künstlicher Intelligenz und maschinellem Lernen eingesetzt.

RAG (Retrieval-Augmented Generation): Eine Methode, bei der die Antwort eines Sprachmodells nicht nur auf seinen trainierten Daten basiert, sondern auch Informationen aus externen Datenquellen in Echtzeit abgerufen werden, um genauere und aktuellere Antworten zu liefern.

Sprints: Kurze Arbeitsphasen, typischerweise im Rahmen von Scrum, in denen Teams bestimmte Aufgaben in einem festgelegten Zeitraum bearbeiten und danach ihre Fortschritte überprüfen.

TLS-Verschlüsselung (Transport Layer Security): TLS (Transport Layer Security) ist ein Protokoll für die sichere Kommunikation im Internet, das die Verschlüsselung von Daten ermöglicht, die zwischen einem Client (z. B. einem Webbrowser) und einem Server ausgetauscht werden. Durch TLS wird sichergestellt, dass Daten während der Übertragung vor unbefugtem Zugriff und Manipulation geschützt sind.

VPS (Virtual Private Server): Ein virtueller Server, der eine isolierte Umgebung bietet, um Anwendungen zu hosten. Es handelt sich um eine kostengünstige Alternative zu einem dedizierten , d.h. physikalisch eigenständigen Server mit fester IP-Adresse, die in diesem Projekt aufgrund des Budgets verwendet wurde.

Zu den Autor*innen

Dominique Dauser ist wissenschaftliche Mitarbeiterin am Forschungsinstitut Betriebliche Bildung (f-bb). Das f-bb ist Konsortialleitung des Zukunftszentrum Süd.

Markus Utomo ist Gründer und Geschäftsführer des Designstudios Markus Utomo Design und Experte für Webdesign, Spielzeugdesign, 3D-Druck und Gestaltung.

Im vorliegenden Whitepaper beschreiben Markus Utomo aus Sicht des externen Dienstleisters und Dominique Dauser aus Kundensicht ihre eigenen Erfahrungen und Erkenntnisse im Entwicklungsprozess der Chatbots für den KI-Experimentierraums des Zukunftszentraums Süd.

Außerdem zuletzt vom f-bb veröffentlicht

Bauer, P., Wittig, W., & Weber, H. (2024): *Stärkung der Ausbildungsbereitschaft von Betrieben: Wie der Transfer von Bildungsinnovationen gelingen kann. Arbeitshilfe für die Transferpraxis.* f-bb-online 02/24. https://www.f-bb.de/unsere-arbeit/publikationen/staerkung-der-ausbildungsbereitschaft-von-betrieben-wie-der-transfer-von-bildungsinnovationen-geling/

Berger, N., Baderschneider, A., & Drummer, K. (2023): *Beratungsleitfaden für eine klischeefreie Berufsorientierung. Leitfaden zur Gestaltung von Informations- und Beratungsangeboten unterschiedlicher Zielgruppen.* f-bb-online 02/2023. https://www.f-bb.de/unsere-arbeit/publikationen/beratungsleitfaden-fuer-eine-klischeefreie-berufsorientierung-leitfaden-zur-gestaltung-von-informati/

Dauser, D. (2024): *Beschäftigte mit Open Educational Resources (OER) und Künstlicher Intelligenz (KI) gezielt fördern! Eine Praxishilfe für die betriebliche Personalentwicklung im Mittelstand.* f-bb-online 04/24. https://www.f-bb.de/de/unsere-arbeit/publikationen/beschaeftigte-mit-open-educational-resources-oer-und-kuenstlicher-intelligenz-ki-gezielt-foerdern/

Fischer, A., Jöchner, A., Pabst, C., Lorenz, S., & Schley, T. (2023): *KI-basierte Personalisierung berufsbezogener Weiterbildung. Ein Praxisleitfaden für Bildungsanbieter.* f-bb-Reihe: Leitfaden für die Bildungspraxis (Bd. 73). Bielefeld: wbv Publikation.

Fischer, A., Jöchner, A., & Dauser, D. (2024). *Open Educational Resources (OER) und Künstliche Intelligenz (KI) – Entwicklungschancen für die berufliche Weiterbildung.* f-bb-online 03/24.

Pabst, C., Jöchner, A., Fischer, A., Lorenz, S., & Schley, T. (2023): *Modularisierung berufsbezogener Weiterbildung. Ein Praxisleitfaden für Bildungsanbieter.* f-bb-Reihe: Leitfaden für die Bildungspraxis (BD. 74). Bielefeld: wbv Publikation.

Pfeiffer, I., & Weber, H. (Hrsg.) (2023): *Zum Konzept der Nachhaltigkeit in Arbeit, Beruf und Bildung – Stand in Forschung und Praxis*. Bonn. https://www.f-bb.de/unsere-arbeit/publikationen/zum-konzept-der-nachhaltigkeit-in-arbeit-beruf-und-bildung-stand-in-forschung-und-praxis/

Richter, K., & Müller, J. (2023): *Berufliche Weiterbildung im Kontext der digitalen Transformation. Digitale Methoden und Medienformate zur Gestaltung beruflicher Bildungsinhalte.* f-bb-online 04/23. https://www.f-bb.de/unsere-arbeit/publikationen/berufliche-weiterbildung-im-kontext-der-digitalen-transformation-digitale-methoden-und-medienformat/