Bundesamt
für Strahlenschutz

Ressortforschungsberichte zum Strahlenschutz

Möglichkeiten und Grenzen des Einsatzes von künstlicher Intelligenz in der behördlichen Kommunikation von Strahlenschutzinhalten am Beispiel EMF

Vorhaben 3621EMF101

PricewaterhouseCoopers GmbH Wirtschaftsprüfungsgesellschaft
Deutsches Forschungszentrum für Künstliche Intelligenz

Das Vorhaben wurde mit Mitteln des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV) und im Auftrag des Bundesamtes für Strahlenschutz (BfS) durchgeführt.

Dieser Band enthält einen Ergebnisbericht eines vom Bundesamt für Strahlenschutz im Rahmen der Ressortforschung des BMUV (Ressortforschungsplan) in Auftrag gegebenen Untersuchungsvorhabens. Verantwortlich für den Inhalt sind allein die Autoren. Das BfS übernimmt keine Gewähr für die Richtigkeit, die Genauigkeit und Vollständigkeit der Angaben sowie die Beachtung privater Rechte Dritter. Der Auftraggeber behält sich alle Rechte vor. Insbesondere darf dieser Bericht nur mit seiner Zustimmung ganz oder teilweise vervielfältigt werden.

Der Bericht gibt die Auffassung und Meinung des Auftragnehmers wieder und muss nicht mit der des BfS übereinstimmen.

Impressum

Bundesamt für Strahlenschutz
Postfach 10 01 49
38201 Salzgitter

Tel.: +49 30 18333-0
Fax: +49 30 18333-1885
E-Mail: ePost@bfs.de
De-Mail: epost@bfs.de-mail.de

www.bfs.de

BfS-RESFOR-237/24

Bitte beziehen Sie sich beim Zitieren dieses Dokumentes immer auf folgende URN: urn:nbn:de:0221-2024111248442

Salzgitter, November 2024

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

1 Künstliche Intelligenz in der Kommunikation

Die Bedeutung künstlicher Intelligenz (KI) in unserer heutigen Welt, insbesondere im Bereich der Kommunikation, nimmt kontinuierlich zu. Ein Beispiel dafür sind die auf vielen Webseiten bereits integrierten Chatbots, die den Nutzern bei der Navigation behilflich sind. Diese Chatbots können voll automatisiert Fragen der Besucher beantworten. Die Anzahl von Anfragen, die manuell per E-Mail oder Telefon beantwortet werden müssen, wird stark reduziert, wodurch Mitarbeiter entlastet werden. Neben den Chancen birgt der Einsatz von Chatbots auch Risiken, wenn diese irreführende Falschaussagen produzieren.

Für die Realisierung von Chatbots können sogenannte „Large Language Models" (LLM, großes Sprachmodell) genutzt werden. Zu diesen LLM zählt beispielsweise GPT-4, welches die technische Grundlage für ChatGPT bildet, das in den letzten Monaten für großes mediales Aufsehen gesorgt hat. LLMs sind neuronale Netze, die meist auf der sogenannten Transformer-Architektur (Vaswani et al., 2017) basieren. Durch immer leistungsfähigere Hardware ist es heutzutage möglich, diese Modelle mit riesigen Textmengen aus unterschiedlichen Quellen zu trainieren. Dies ermöglicht den Einsatz solcher Modelle für vielfältige Aufgaben wie das Schreiben von Text, das Beantworten von Fragen, die Übersetzung von Texten und sogar das Generieren von Programmcode (Krupp et al., 2023).

Der Einsatz von Chatbots beschränkt sich nicht nur auf Webseiten, sondern findet auch im Bildungsbereich Anwendung (Krupp et al., 2023). Allerdings besteht die Herausforderung darin, den unkontrollierten Einsatz dieser Technologien, wie zum Beispiel zur automatischen Bearbeitung von Hausaufgaben, zu verhindern. Solche Systeme, wie ChatGPT, weisen Schwächen im logischen Schlussfolgern auf, was zu potenziell irreführenden Informationen führen kann (Krupp et al., 2023). Dennoch könnten sie in der Lehre eingesetzt werden, um Lehrkräfte bei der Aufgabenerstellung zu unterstützen oder Schülern mit Beeinträchtigungen zu helfen.

In den letzten Jahren wurden erhebliche Fortschritte in der Forschung und Entwicklung künstlicher Intelligenz verzeichnet. Ein Beispiel hierfür ist der PageRank-Algorithmus, der durch Google bekannt wurde. Dieser Algorithmus bewertet Webseiten basierend auf ihrer Wichtigkeit, wobei die Anzahl und Qualität der eingehenden Links relevant sind (Page et al., 1998).

Der PageRank-Algorithmus hat in den letzten Jahren aus verschiedenen Gründen an Bedeutung verloren. Einerseits gab es viele Möglichkeiten der Manipulation, anderseits gibt es heutzutage viele weitere Faktoren, die die Relevanz einer Webseite bestimmen und vom ursprünglichen Algorithmus nicht berücksichtigt wurden. Alternativen zum PageRank-Algorithmus sind der TrustRank- (Gyöngyi et al., 2004) und Hilltop-Algorithmus (Bharat & Mihaila, 2000). Der TrustRank-Algorithmus ähnelt dem PageRank-Algorithmus, mit dem Unterschied, dass eine kleine Anzahl vertrauenswürdiger und nicht vertrauenswürdiger Webseiten manuell ausgewählt wird. Dagegen erreicht eine Seite bei dem Hilltop-Algorithmus eine hohe Wertung, wenn die Seite viele eingehende Links sogenannter Experten-Seiten hat. Ein Beispiel für eine Experten-Seite kann ein gut geführtes Verzeichnis darstellen. Somit bestimmt hier auch die Qualität der eingehenden Links die Ergebnisse.

Google ersetzte 2013 den PageRank- durch den Hummingbird-Algorithmus (Singhal, 2013), welcher auch semantische Informationen aus der Suchanfrage verwenden kann.

Eine weitere Möglichkeit, um Suchmaschinen zu verbessern, bietet der Einsatz von Embeddings (Mikolov et al., 2013). Diese Technologie ermöglicht es, semantisch ähnliche anstelle von syntaktisch ähnlichen Ergebnissen zu einer Suchanfrage zu liefern. Zum Beispiel könnten durch die Verwendung solcher Embeddings bei einer Suche nach dem Stichwort „Sonnenbrand" auch Ergebnisse zum Thema UV-Strahlung angezeigt werden. Eine primitive Suche, die auf einfachem String-Matching basiert, würde zu keinen Treffern führen, da sich beide Begriffe syntaktisch stark unterscheiden.

Ein weiteres bedeutendes Anwendungsgebiet künstlicher Intelligenz ist die maschinelle Übersetzung. Die ersten maschinellen Übersetzungsprogramme folgten regelbasierten Ansätzen (Wang et al., 2021). Diesen Systemen mangelte es jedoch an Skalierbarkeit, da sie auf manuellen Schreibregeln beruhten, die für jede Sprache definiert werden mussten. Mit dem Aufkommen von zweisprachigen Textkorpora um die Jahrtausendwende wurden Korpus-basierte Technologien wie „statistical machine translation" und „neural machine

translation“ immer bedeutsamer (Wang et al., 2021). Diese statistischen Methoden lernen das Übersetzen von einer Quell- in eine Zielsprache aus einer großen Menge an Daten, ohne von menschlichem Expertenwissen abhängig zu sein. Durch Fortschritte im Bereich des Deep Learning konnte die Qualität der Übersetzungen immer weiter verbessert werden. Dabei wird die Ausgangssprache zunächst in eine semantische Darstellung überführt, um daraus die Übersetzung zu generieren (Wang et al., 2021).

In den letzten zwei Jahren haben die Fortschritte in der KI eine beeindruckende Geschwindigkeit erreicht. Getrieben wurde diese Entwicklung durch mehrere Schlüsselfaktoren, die sowohl die technischen Möglichkeiten als auch die praktische Anwendung dieser Technologien erweitert haben. Zu den markantesten Entwicklungen zählt die Einführung und Weiterentwicklung neuer Architekturen, insbesondere der Transformer-Modelle. Ursprünglich für die Verarbeitung natürlicher Sprache konzipiert, haben sie sich als äußerst effektiv für eine Vielzahl von Anwendungen, einschließlich Bildverarbeitung und multimodale Aufgaben, erwiesen. Diese Architekturen haben durch ihre Fähigkeit, komplexe Datenmuster zu erkennen und zu interpretieren, signifikant zur Leistungssteigerung von KI-Systemen beigetragen.

Parallel dazu hat die zunehmende Verfügbarkeit und Leistungsfähigkeit von spezialisierter Hardware wie Grafikprozessoren (GPUs) und Tensor Processing Units (TPUs) das Training immer größerer Modelle ermöglicht. Diese Modelle können aus umfangreichen Datenmengen lernen, wodurch die Genauigkeit und Anwendungsbreite der KI erheblich verbessert wurden. Auch die Datenlandschaft hat sich verändert. Die Verfügbarkeit großer und diversifizierter Datensätze hat es KI-Modellen ermöglicht, präzisere und nuanciertere Muster zu erkennen. Diese Entwicklung hat die Effektivität von KI in verschiedenen Anwendungsfällen gesteigert und ihre Generalisierungsfähigkeit verbessert.

Nicht zu übersehen sind die Fortschritte in den Lernalgorithmen selbst. Techniken wie das selbstüberwachte Lernen und das Transferlernen haben das Training effizienter gemacht, was zu schnelleren Trainingszeiten und einer besseren Anpassungsfähigkeit der Modelle geführt hat. Diese Verbesserungen unterstützen eine schnellere Entwicklung und breitere Anwendung von KI-Lösungen. Zusätzlich hat die Demokratisierung von KI-Technologien durch die Verfügbarkeit von Open-Source-Tools und Frameworks eine breitere Gemeinschaft von Forschern und Entwicklern dazu befähigt, auf dem Gebiet der KI Innovation voranzutreiben. Dies hat nicht nur die Forschung vorangetrieben, sondern auch die kommerzielle Nutzung von KI in verschiedenen Sektoren beschleunigt.

Die als Ergebnis dieses Forschungsprojektes konzipierte Retrieval-Augmented Generation Architektur existierte noch nicht zu Beginn des Projektes im Juni 2023, um ein anschauliches Beispiel für die rasante Entwicklung auf dem Gebiet der (generativen) KI zu nennen.

2 Herausforderungen und Potenziale behördlicher Kommunikation zum Strahlenschutz

Die Nutzung von KI in der Kommunikation von EMF-Themen birgt sowohl Herausforderungen als auch Potenziale. Zu den Herausforderungen gehört die Gewährleistung der Genauigkeit und Zuverlässigkeit von KI-gesteuerten Systemen, um Fehlinformationen zu vermeiden. Dies erfordert eine sorgfältige Überwachung und kontinuierliche Anpassung der KI. Der Einsatz von KI bietet aber auch die Möglichkeit, Effizienz zu steigern und die Kommunikation zu personalisieren. Im Fokus dieses Kapitels stehen die Besonderheiten, Strategien und Zukunftsaussichten der Behördenkommunikation im Kontext von EMF.

2.1 Spezielle Anforderungen an die Kommunikation zu EMF

Neben den allgemeinen Anforderungen an die Benutzerfreundlichkeit von Informations-Webseiten gibt es auch spezielle Anforderungen an die Kommunikation im Kontext von elektromagnetischen Feldern. Diese ist anspruchsvoller, da Sender und -empfänger von Informationen oftmals einen unterschiedlichen Wissensstand aufweisen. Es besteht die Gefahr, dass der Sender ein bestimmtes Wissen beim Empfänger voraussetzt, wodurch es zu einer Misskommunikation zwischen den Kommunikationspartnern kommen kann. Um potenzielle Fehlerquellen in der Kommunikation zwischen dem BfS und den Bürgern zu minimieren, werden die kommunikativen Besonderheiten im Kontext von elektromagnetischen Feldern erläutert.

Im Allgemeinen erfordert die behördliche Kommunikation im Zusammenhang mit EMF ein hohes Maß an Transparenz und Fachwissen, um die Öffentlichkeit entsprechend zu informieren. Dazu ist das Kompetenzzentrum Elektromagnetische Felder (KEMF) des BfS eingerichtet.

Konkret lassen sich folgende spezielle Anforderungen der behördlichen Kommunikation im Kontext der EMF ableiten:

- **Komplexität der Materie:** Die EMF gehören zu einem komplexen wissenschaftlichen Thema, das Fachkenntnisse erfordert. Bei der Vermittlung von Informationen muss deswegen auf leicht verständliche Sprache geachtet werden, ohne die wissenschaftliche Grundlagen von EMF vollkommen auszuklammern.
- **Gesundheitsbedenken:** Bei manchen Personen kann eine Exposition gegenüber EMF, insbesondere im Zusammenhang mit Mobilfunk, WLAN und anderen drahtlosen Kommunikationstechnologien zu Gesundheitsbedenken führen. Es ist wichtig, diese in der Kommunikation ernst zu nehmen und über wissenschaftliche Erkenntnisse und über mögliche Maßnahmen zur Expositionsverringerung zu berichten.
- **Beteiligung der Öffentlichkeit:** Da elektromagnetische Felder die breite Öffentlichkeit betreffen und vereinzelte Gruppen bereits Verschwörungsmythen im Zusammenhang mit EMF verbreiten, sollte eine möglichst hohe Zahl an Bürgern mit vertrauenswürdigen Informationen versorgt werden. Um das zu gewährleisten, ist eine öffentliche Beteiligung zum Beispiel durch Feedback und Rückmeldungen zu den Informationen empfehlenswert.
- **Medienarbeit:** Die Medien sind ein wichtiger Kanal für die Kommunikation über EMF. Vereinzelt kann es auch zur Verbreitung von gefälschten oder unrichtigen Nachrichten kommen. Behörden sollten deswegen mit Medienanbietern kooperieren und sicherstellen, dass Informationen korrekt präsentiert werden.
- **Risikokommunikation:** Die verständliche Kommunikation über potenzielle Risiken im Zusammenhang mit EMF ist entscheidend, um das Vertrauen der Bevölkerung zu gewinnen. Insbesondere das BfS sollte daher klar darlegen können, welche Risiken bestehen und wie diese minimiert werden können.
- **Monitoring von Fortschritten in der Forschung:** Zur Nutzung von Technologien mit einem EMF wird umfassend und laufend geforscht. Eine Kommunikation des aktuellen Stands der Forschung in die Bevölkerung hinein ist ein wichtiger Bestandteil.

2.2 Interviews mit Fokus auf der Kommunikation des BfS

In der Behördenkommunikation wird künstlicher Intelligenz mit ihren Auswirkungen und Nutzungsmöglichkeiten großes Potenzial zugeschrieben. Zwei Interviews mit Experten im Bereich elektromagnetische Felder und für Bürgerkommunikation bieten eingehende Einblicke in die Potenziale und Herausforderungen dieser Technologie für eine effiziente und transparente Kommunikation in diesem spezifischen Kontext. Im Folgenden werden die gewonnenen Einblicke aus den Interviews in thematischen Bausteinen unterteilt dargestellt.

Interview 1: Presse- und Öffentlichkeitsarbeit des BfS

Im ersten Interview wurden die bisherigen Vorgehensweisen der Presse- und Öffentlichkeitsarbeit des BfS und Einsatzmöglichkeiten von KI in der Behördenkommunikation für die Bereiche Strahlenschutz und elektromagnetische Felder besprochen.

Einleitung

Es wurden die großen Potenziale deutlich, die KI für die Verbesserung der Effizienz und Transparenz in der Kommunikation mit der Öffentlichkeit, also Bürgern und Unternehmen in Bezug auf Fragen zur Strahlung bietet.

Pressearbeit und Internetveranstaltungen

Die Kommunikationsabteilung hat in den letzten Jahren verstärkt Informationen über Strahlungen und deren Wirkung in die Öffentlichkeit getragen. Eine entsprechende Pressearbeit und Online-Veranstaltungen mit entsprechenden Formaten wurden umgesetzt und als entscheidend eingestuft. Diese Maßnahmen haben dazu beigetragen, das Bewusstsein für das Thema zu schärfen und die Verbreitung von Informationen zu verbessern.

Bürgerkommunikation und Anfragen von Unternehmen

Ein großer Teil der Kommunikationsarbeit besteht in der Bearbeitung von Bürgeranfragen und Anfragen von Unternehmen. Im Jahr 2020 wurden über 5.100 Anfragen telefonisch und 1.421 schriftlich eingereicht. Interessanterweise bevorzugen sowohl die Bürger als auch die Mitarbeiter der Kommunikationsabteilung telefonische Anfragen, da über diesen Kommunikationskanal Fragen und Rückfragen einfacher geklärt werden können.

Weiterleitung von Anfragen und Chatbot

Um die kompetente Beantwortung von allen Anfragen sicherzustellen, werden Anfragen, die nicht von der Kommunikationsabteilung beantwortet werden können, konsequent an die entsprechenden Fachstellen weitergeleitet. Zudem wird die Einrichtung eines dialogorientierten Chatbots zur automatischen Beantwortung von Fragen rund um elektromagnetische Felder präferiert. Dieser Schritt zielt darauf ab, die öffentliche Wahrnehmung von EMF-Risiken zu klären und fundierte Informationen bereitzustellen.

Bearbeitungszeiten und Webseiten-Optimierung

Die festgelegte maximale Bearbeitungszeit von zehn Arbeitstagen für Anfragen wird von einigen Beteiligten als zu lang empfunden. Lediglich bei seltenen fachspezifischen Anfragen wird diese ausgereizt. Obwohl die Webseite informationsreich ist, wird sie von Laien oft als schwer verständlich wahrgenommen. Es wurde dringend empfohlen, die Webseite benutzerfreundlicher zu gestalten und die Informationen für die breite Öffentlichkeit zugänglicher zu machen. Dies könnte dazu beitragen, Bürger und Unternehmen besser zu informieren und ihre Anfragen zu reduzieren.

Interview 2: Kompetenzzentrum Elektromagnetische Felder

In einem zweiten Interview mit dem Co-Leiter des Informations- und Kommunikations-Teams des KEMF wurden Einsatzbereiche von KI in der Kommunikation des KEMF besprochen.

Einleitung

In dem Interview wurde die Bedeutung der künstlichen Intelligenz in der Kommunikation im Bereich elektromagnetische Felder erörtert. Das Gespräch verdeutlichte die großen Potenziale, die KI für die Verbesserung der Effizienz und Transparenz in der Kommunikation in Bezug auf Bürgeranfragen bieten.

Fokusbereiche in der Kommunikation von Inhalten zu EMF

Anders als bei sonstigen Themen zu Strahlung, bei denen es eindeutig wissenschaftlich nachgewiesene Wirkungen gibt, gibt es bei EMF eine weitgehende Konsensmeinung, dass unterhalb der empfohlenen Grenzwerte keine gesundheitsrelevanten Wirkungen nachgewiesen sind. Es muss daher zwischen „Care Communication“ und „Consensus Communication“ unterschieden werden. Kommunikation muss Vertrauen und Akzeptanz bei der Empfängergruppe insbesondere Bürgern herstellen.

Auf Bürgeranfragen muss daher stets zielgruppen- und bedarfsorientiert reagiert werden. Es werden durch das BfS verschiedene Zielgruppen identifiziert. Dies sind insbesondere Personen, die Informationen von einer autorisierten Quelle, in diesem Fall dem BfS, erhalten möchten. Des Weiteren ist eine Personengruppe mit wissenschaftlichem Interesse vorhanden. Eine dritte Personengruppe hat Bedenken zu bestimmten Themen, und setzt jedoch trotzdem Vertrauen in die Autorität des BfS zur Beantwortung ihrer Fragen. Eine vierte Gruppe stellen die Personen, die die Autorität kritisch hinterfragen und solche, die das BfS grundsätzlich ablehnen. Letztere Gruppe betreibt teils politischen Aktivismus.

Parameter der Kommunikation

Es existieren behördliche Rahmenbedingungen und klare Erwartungen an die Art und Weise, wie mit Menschen kommuniziert werden soll. Dabei ist es entscheidend, sicherzustellen, dass nur relevante und korrekte Informationen so verständlich wie möglich weitergegeben werden. Dabei sollten die Antworten auf wissenschaftlichen Erkenntnissen und Fakten basieren und transparent vermittelt werden. Diese wissenschaftlichen Inhalte sollten für Laien verständlich und prägnant vermittelt werden. Ein weiterer Parameter der Kommunikation ist das Prinzip „one message, many voices“, was bedeutet, dass alle Mitarbeiter identische Kerninhalte wiedergeben sollen. Jegliche Abweichung in den Antworten führt zu Verwirrung bei den Bürgern und bietet mehr Angriffsfläche für Kritiker. Ein weiterer wichtiger Aspekt bei der Kommunikation ist die Neutralität der Behörde. Die Behörde sieht es nicht als ihre Aufgabe, Akzeptanz für den Einsatz von Technologien zu schaffen.

Einsatz von KI zur Beantwortung von Fragen von Bürgern mit Gesundheitsbedenken bei Strahlung

Es ist von großer Bedeutung, dass sämtliche Kommunikationskanäle (Telefon, Post, E-Mail, Webseite) offenbleiben, um sicherzustellen, dass Menschen mit unterschiedlichen Hintergründen grundlegend versorgt werden können. Die Entwicklung eines Chatbots für (voll)automatisierte Antworten ist in Betracht zu ziehen, insbesondere bei weniger komplexen Themen. Es sollte jedoch ein Ablenkungsmechanismus integriert werden, der es ermöglicht, bei nicht zufriedenstellenden Antworten einen Mitarbeiter des BfS hinzuzuziehen. Bei der erstmaligen Implementierung von KI ist eine Überprüfung der Antworten durch Verantwortliche zwingend erforderlich, um deren Richtigkeit zu gewährleisten. Nach sorgfältiger Evaluierung könnte laut des Ansprechpartners die Verwendung von KI in risikoärmeren Bereichen in Betracht gezogen werden. Im Falle eines KI-Einsatzes muss dies dem Nutzer unbedingt transparent vermittelt werden. Es ist wichtig, bei jeder KI-Antwort Quellen und Referenzen anzugeben, damit der Nutzer sich weiter informieren oder die Antwort selbst überprüfen kann. Der Einsatz von KI könnte die Behörde möglicherweise angreifbarer machen.

Einschätzung zur Suchfunktion der BfS-Webseite

Vonseiten des Ansprechpartners wird darauf hingewiesen, dass trotz der Tatsache, dass für fast alle der gestellten Fragen Antworten auf der Webseite zu finden sind, die meisten Bürger dennoch den direkten Kontakt zum BfS suchen. Es wird vonseiten der Ansprechpartner des BfS als Hinweis gedeutet, dass die Suchfunktion auf der Webseite für die Besucher nicht zufriedenstellend ist. Der Ansprechpartner aufseiten des BfS könnte sich vorstellen, dass die Webseite in Zukunft, ähnlich wie die ChatGPT-Webseite von OpenAI, gestaltet wird, auf der Suchende direkt Fragen stellen können, anstatt auf der Seite nach Informationen suchen zu müssen.

Ausblick auf die Zukunft

Vonseiten des Ansprechpartners wird für die Zukunft angestrebt, dass etwa 80 Prozent der Fragen automatisiert beantwortet werden können, um sich verstärkt auf die 20 Prozent komplexeren Themen zu konzentrieren. Des Weiteren wird sich eine systematischere Auswertung der Kommunikationseingänge gewünscht, um herauszufinden, was für die Bürger von Relevanz ist. Bürger sollten die Möglichkeit haben, anzugeben, ob sie die Antwort lieber schnell von einer KI oder von einem Experten erhalten möchten, wobei die Antwortzeit entsprechend variiert. Zudem sollte es möglich sein, Antworten in andere Sprachen zu übersetzen, um alle Bürger einzubeziehen. Als letztes Wunschszenario wurde aufgenommen, dass durch den Einsatz von Gamification und KI die Thematik für die Bürger anschaulicher und verständlicher erklärt werden könnte. Eine Möglichkeit könnte auch sein, die Inhalte der Webseite algorithmisch zu restrukturieren.

Die Kommunikation mittels KI sollte sich auf Wissenschaftskommunikation mit dem Ziel beschränken, Informationen laienverständlich und komprimiert inklusive wissenschaftlicher Begründung zu liefern. Sorgen sollten dabei nicht adressiert werden. Die Kommunikation sollte sachlich bleiben.

Zukünftige unerwünschte Entwicklungen

Es ist entscheidend, dass die Anfälligkeit gegenüber „Bad Faith Actors“ durch den Einsatz von KI nicht zunimmt. Diese könnten ungenaue Antworten der KI als politisches Mittel nutzen. Es muss auch in Zukunft verhindert werden, dass Inhalte durch die KI falsch verstanden werden. Mit dem Einsatz von KI ist es schwieriger zu überprüfen, ob die andere Seite die Information richtig verstanden hat, im Vergleich zu beispielsweise einem Telefonat. Ein weiterer wichtiger Aspekt ist, dass die Bürger nicht den Eindruck haben sollten, durch KI schlechtere Informationen als bisher zu erhalten. Allerdings wäre etwa 5 Prozent mehr Unzufriedenheit akzeptabel, wenn dies im Hinblick auf den Nutzen der KI vertretbar ist.

Kernaussagen der Interviews

Die Interviews unterstreichen die entscheidende Rolle der künstlichen Intelligenz in der Kommunikation des BfS zu elektromagnetischen Feldern. Die potenzielle Implementierung eines Chatbots, die Überprüfung der Bearbeitungszeiten und die Verbesserung der Webseite sind als positive Schritte in Richtung transparenter und effektiver Kommunikation hervorgehoben. Die Notwendigkeit differenzierter Kommunikation, klare Parameter für wissenschaftsbasierte Antworten und die betonten Herausforderungen beim Einsatz von KI werden als Schlüsselfaktoren für die Zukunft angesehen. Insgesamt bieten die Interviews eine optimistische Perspektive auf die Nutzung von KI, um Ängste zu zerstreuen und gleichzeitig umfassende Informationen bereitzustellen. Gleichzeitig wird darauf geachtet, möglichen Missbrauch zu verhindern und die Wahrnehmung der KI als Bereicherung für die Bürger zu gewährleisten.

Deutlich wird damit, dass insbesondere die verschiedenen Zielgruppen und die hohen Anforderungen an eine wissenschaftsbasierte, transparente und verständliche Kommunikation erfasst werden müssen. Möglichkeiten der methodischen Vorbereitung auf eine erfolgreiche zielgruppenspezifische Kommunikation bildet der Einsatz von Personas.

2.3 Vorstellung relevanter Personas

Im folgenden Kapitel werden fiktive Charaktere vorgestellt, die für die Kommunikation des BfS relevant sind. Schlussfolgernd aus den Interviews wurden für die relevanten Anforderungen verschiedene Personas erstellt. Mit diesen Personas werden die Kommunikationsbedürfnisse der dort aufgenommenen Zielgruppen ausgearbeitet. Diese **Personas** aus Mitarbeiter- und Bürgersicht geben einen **Einblick in die Bedürfnisse, Herausforderungen und Anforderungen von ebendiesen Kommunikationspartnern**.

Alle im Folgenden beschriebenen Personas mit Namen und Eigenschaften sind frei ausgearbeitet. Ähnlichkeiten oder eine Identifizierung von lebenden/toten Personen und Persönlichkeitseigenschaften sind zufällig und nicht intendiert.

Name	Anna Meier
Rolle	Webdesignerin beim BfS/ITZBund
Zugehörigkeit und Tätigkeit	Anna Meier ist für den Aufbau und die Instandhaltung der Webseite des BfS zuständig. Es ist ihre Aufgabe zu bestimmen, wie die Webseite aussieht. Sie leitet auch alle Projekte, die sich mit dem strukturellen Aufbau der Webseite befassen.
Bedürfnisse	Anna Meier würde gerne wissen, wie sie die Webseite so umbauen kann, dass die Bürger so viele Informationen wie möglich so einfach wie möglich erhalten. Ihr Ziel ist es auch, die Anzahl der schriftlichen und telefonischen Anfragen durch passende Informationen auf der Webseite zu verringern.
Herausforderungen	Anna Meier weiß nicht, was die Bürger am meisten beschäftigt. Sie kann somit schwer einschätzen, womit sie sich als erstes befassen muss, um die Webseite anzupassen.
Anforderungen	Anna Meier benötigt statistische Informationen über die Art der Anfragen und die Themen, die die Bürger am meisten interessieren. Klassifikationen von Anfragen, Daten über die meistbesuchten Themenbereiche auf der Webseite, Verbesserungsvorschläge seitens der Bürger und Erklärungen von Artikeln in leichter Sprache könnten ihr helfen.

Tabelle 1: Persona 1 - Mitarbeiterin des BfS/ITZBund

Name	Emma Schneider
Rolle	Mutter
Zugehörigkeit und Tätigkeit	Emma Schneider ist Mitglied im Elternverein und will sich und ihre Kinder vor möglichen gesundheitlichen Gefahren schützen. Gleichzeitig will sie die anderen Mitglieder im Elternverein sowie Freunde und Bekannte davor warnen, damit diese sich vor Gefahren schützen können.
Bedürfnisse	Emma Schneider möchte gut fundierte und wissenschaftlich belegte Informationen und Artikel über die Arten und möglichen gesundheitlichen Auswirkungen von elektromagnetischen Strahlungen.
Herausforderungen	Emma Schneider benötigt klare, vertrauenswürdige Antworten und die dazugehörigen Quellen/Studien. Damit möchte sie sicherstellen, dass sie auf dem aktuellen Stand ist und gleichzeitig auch die anderen davon überzeugen kann, ohne Falschinformationen zu verbreiten.
Anforderungen	Emma Schneider benötigt auf der Webseite die verwendeten Studien und Quellen leicht zugänglich, sodass sie sich selbst von den Artikeln überzeugen kann.

Tabelle 2: Persona 2 – Mutter und Mitglied im Elternverein

Name	Familie Müller
Rolle	Gesundheitsbewusste Familie
Zugehörigkeit und Tätigkeit	Familie Müller ist eine gesundheitsbewusste Familie, die sich mit praktischen Tipps vor der Exposition gegenüber elektromagnetischer Strahlung schützen will. Die Familienmitglieder wissen, dass keine große Gefahr besteht, wollen aber trotzdem, falls möglich, auf der sicheren Seite sein, um komplett sorgenfrei zu bleiben.
Bedürfnisse	Familie Müller benötigt alltagstaugliche und einfache Tipps, wie die Eltern sich und ihre Kinder schützen können. Die Familie braucht keine große Menge an Informationen.
Herausforderungen	Wenn Familie Müller die Webseite öffnet, sieht sie zuerst einen Berg an neuen Begriffen und Informationen. Das sorgt dafür, dass die Familie überwältigt ist und sich eventuell mehr Sorgen um Strahlung macht, da Begriffe und Informationen äußerst kompliziert erscheinen.
Anforderungen	Familie Müller benötigt einen Artikel, der leicht umsetzbare Tipps und kurze, verständliche Erklärungen dazu enthält. Damit erhält die Familie Müller ein Verständnis für das Thema und kann selbst etwas unternehmen, um sich zu schützen.

Tabelle 3: Persona 3 - Gesundheitsbewusste Familie

Name	Heinrich Weber
Rolle	Privatperson
Zugehörigkeit und Tätigkeit	Heinrich Weber hat keine besondere Tätigkeit und ist ein normaler Bürger, der sich einfach über die elektromagnetische Felder informieren will. Er ist ein Laie und hat keine besonderen Kenntnisse zum Thema.
Bedürfnisse	Heinrich Weber benötigt unkomplizierte und leicht zu verstehende Informationen über allgemeine Themen. Dazu gehören zum Beispiel die Erklärungen, was EMF sind, welche Arten von Strahlung es gibt, wo Strahlung auftritt, oder worauf zu achten ist.
Herausforderungen	Für Heinrich Weber ist die große Komplexität der Webseite eine Herausforderung. Des Weiteren setzen die Einträge Wissen voraus. Die Seite ist auch nicht besonders leicht zu navigieren. Heinrich Weber findet nur kompliziertere Artikel, die zwar hilfreich, aber schwer zu verstehen sind.
Anforderungen	Heinrich Weber benötigt eine neu strukturierte Webseite mit einer leichten Darstellung der Informationen, FAQs und Einsteigerinformationen/Artikeln.

Tabelle 4: Persona 4 – Privatperson

Name	Martin Vorsicht
Rolle	Bürger mit Gesundheitsbedenken bei Strahlung
Zugehörigkeit und Tätigkeit	Martin Vorsicht hat im Internet gelesen, dass elektromagnetische Strahlungen sehr gefährlich sind. Er hat Angst, seiner Gesundheit zu schaden, wenn er sein Handy nah am Kopf hält oder wenn er sich in der Nähe von Sendemasten befindet.
Bedürfnisse	Martin Vorsicht kennt sich mit elektromagnetischer Strahlung nicht aus und die meisten Begriffe sind ihm unbekannt. Er will Beweise und Studien in leichter Sprache, die ihm die Angst nehmen und ihm zeigen, dass unterhalb der empfohlenen Grenzwerte keine gesundheitsrelevanten Wirkungen nachgewiesen sind.
Herausforderungen	Die Webseite ist Martin Vorsicht zu kompliziert und die Begriffe sind ihm unbekannt. Die Studien helfen ihm nicht, da er die vielen komplizierten Fachbegriffe nicht versteht und sich in der Folge auch nicht von den komplexen Beschreibungen im Text angesprochen fühlt.
Anforderungen	Martin Vorsicht braucht Antworten auf seine Sorgen in leichter Sprache. Ein FAQ-Bereich mit den meisten Irrtümern wäre eine Idee. Das Gespräch mit einem Chatbot, der gezielt auf seine Fragen antwortet, könnte ihm auch helfen. Falls er nicht auf Maschinen hören will, kann er telefonisch Kontakt aufnehmen, um mit einem Experten zu reden, der ihm alles in einfacher Sprache erklärt.

Tabelle 5: Persona 5 – Bürger mit Gesundheitsbedenken bei Strahlung

Name	Peter Schmidt
Rolle	Mitarbeiter der Presse- und Öffentlichkeitsarbeit beim BfS
Zugehörigkeit und Tätigkeit	Peter Schmidt ist für alle eingehenden Anfragen verantwortlich (schriftlich und telefonisch). Er beantwortet Fragen von Bürgen, die sich mehr informieren wollen oder Fragen zu bestimmten Inhalten auf der Webseite haben.
Bedürfnisse	Peter Schmidt benötigt ein Tool/Funktion, womit er die Anfragen schneller bearbeiten kann oder wodurch keine Anfragen mehr kommen, da die Bürger schon im Voraus informiert werden.
Herausforderungen	Peter Schmidt stellt fest, dass es viele verschiedene Arten von Anfragen gibt. Abhängig von Fragesteller und Frage werden verschiedene Informationen benötigt. Der Umgang und die Kommunikation erfolgen daher sehr individuell. Peter Schmidt soll trotzdem vertrauenswürdig und verständnisvoll auf die Bürger wirken.
Anforderungen	Peter Schmidt benötigt einen vorformulierten Text für jede Kategorie von schriftlichen Anfragen, den er leicht bearbeiten kann, bevor er ihn an den Bürger schickt oder telefonisch Auskunft gibt. Zudem benötigt er einen ausführlichen und leicht zugänglichen Bereich für FAQ mit den meisten Fragen, die die Bürger stellen. Zudem benötigt er einen Chatbot, mit dem die Bürger kommunizieren können, bevor sie anrufen.

Tabelle 6: Persona 6 – Mitarbeiter der Presse- und Öffentlichkeitsarbeit beim BfS

Die Vorstellung der Personas zeigt, dass ein breites Spektrum an Personen in Kommunikation mit dem BfS tritt, beziehungsweise an der Kommunikation beteiligt ist. Es werden vielfältige Bedürfnisse, Herausforderungen und Anforderungen beschrieben, auf die bei der Kommunikation sorgfältig eingegangen werden muss.

Diese individuellen Kommunikationsanforderungen der Kommunikationspartner in und außerhalb des BfS stehen neben den allgemeinen Anforderungen an Informations-Webseiten.

2.4 Allgemeine Anforderungen an die Benutzerfreundlichkeit von Informations-Webseiten

Die Kommunikation mit den Bundesbehörden ist ein wichtiger Grundpfeiler in demokratischen Systemen und Gesellschaften. Insbesondere die Bereitstellung von Informationen erlaubt es den Bürgern, politische, gesellschaftliche und institutionelle Entwicklungen zu verstehen und richtig einzuordnen. Aus diesem Grund ist die laufende Verbesserung der bundesweiten Informations-Webseiten wie die des BfS eine wichtige Aufgabe, da sie oftmals als erste Anlaufstelle für Bürger fungiert.

Um als benutzerfreundliche Seite eingestuft und von Nutzern als solche wahrgenommen zu werden, gibt es einige Anforderungen, die eine Informations-Webseite erfüllen muss. Generell sollte die Benutzerfreundlichkeit sowohl durch die Aufbereitung und Strukturierung der Informationen als auch durch das Design der Webseite sichergestellt werden.

Konkret lassen sich folgende Anforderungen an die Benutzerfreundlichkeit von Informations-Webseiten ableiten:

- **Barrierefreiheit:** Eine barrierefreie Webseite stellt sicher, dass die präsentierten Informationen auch für Menschen mit Behinderungen zugänglich sind. Dies kann zum Beispiel die Verwendung von Alt-Texten für Bilder, klaren Kontrasten, Vorleseprogrammen und Tastaturzugänglichkeit einschließen.
- **Verständliche und klare Sprache:** Der Inhalt auf der Webseite sollte in einer verständlichen und klaren Sprache präsentiert werden. Eine umfassende Verwendung technischer und wissenschaftlicher Fachbegriffe verringert die Verständlichkeit von Texten.
- **Optimierung für Mobilgeräte:** Aufgrund der hohen Nutzung von Mobilgeräten sollte die Webseite insbesondere für Smartphones und Tablets optimiert sein und eine reaktionsschnelle Gestaltung aufweisen.
- **Leichte Navigation:** Eine Webseite sollte durch eine einfache und intuitive Navigation ausgezeichnet sein. Dafür sind insbesondere eine gut strukturierte Menüführung und klare Hierarchien wichtig, sodass die Nutzer schnell zwischen Informationen navigieren können.
- **Aktualität der Informationen:** Die zur Verfügung gestellten Informationen sollten laufend aktualisiert werden und stets auf dem neuesten Stand sein. Veraltete Informationen stiften Verwirrung und können in einem institutionellen Kontext das Vertrauen in eine Behörde verringern.
- **Feedback-Möglichkeiten:** Benutzer sollten die Möglichkeit haben, Feedback zu geben oder Schwierigkeiten zu melden. Dies ermöglicht es der Behörde, die Webseite auf die Wünsche der Nutzer auszurichten.
- **Suchfunktion:** Eine effektive Suchfunktion ist entscheidend, um Nutzern die Möglichkeit zu geben, gezielt nach Informationen zu suchen.
- **Transparenz:** Es ist wichtig, dass sich Bundesbehörden durch eine hohe Transparenz auszeichnen. Informationen über Entscheidungsprozesse, politische Maßnahmen und Finanzen sollten klar dargestellt und zugänglich sein.
- **Multilinguale Unterstützung:** Mehrsprachige Webseiten stellen sicher, dass Bürger mit wenigen Deutschkenntnissen Zugang zu Informationen erhalten.

Mit den allgemeinen Anforderungen werden insbesondere die Zugänglichkeit, Struktur und Informationsgehalt beschrieben. Deutlich wird, dass verschiedene Arbeitsbereiche bei der Erstellung von Kommunikationsmaterialien und deren Bereitstellung innerhalb des BfS zusammenwirken müssen.

2.5 Feststellungen zu Herausforderungen und Potenzialen

Die Analyse der behördlichen Kommunikation im Kontext EMF verdeutlicht die anspruchsvollen Anforderungen, die dieser Themenbereich an die Informationsvermittlung stellt. Besondere Herausforderungen ergeben sich aus der Komplexität der Materie, den Gesundheitsbedenken der Bevölkerung, der Beteiligung der Öffentlichkeit sowie der Medienarbeit. Zudem ist situations- oder ereignisbedingt die Bedeutung einer notwendigen Risikokommunikation hervorzuheben.

Die durchgeführten Interviews mit Experten betonen die entscheidende Chance des Einsatzes künstlicher Intelligenz in der Behördenkommunikation zu EMF. Die Potenziale von KI, insbesondere in Form eines Chatbots, werden als Möglichkeit zur effizienten und transparenten Kommunikation hervorgehoben. Die Notwendigkeit differenzierter Kommunikation, klarer Parameter für wissenschaftsbasierte Antworten und die Herausforderungen beim Einsatz von KI werden als Schlüsselfaktoren für die Zukunft betrachtet.

Die vorgestellten Personas veranschaulichen die Vielfalt der Zielgruppen bei der Kommunikation und deren unterschiedliche Bedürfnisse. Von Webdesignern über Bürger mit Gesundheitsbedenken bis hin zu Mitarbeitern der Presse- und Öffentlichkeitsarbeit sind spezifische Anforderungen an die Informationsgestaltung und -vermittlung zu berücksichtigen.

Generell sollten Informations-Webseiten im behördlichen Kontext bestimmte Anforderungen erfüllen, um als benutzerfreundlich wahrgenommen zu werden. Dazu gehören Barrierefreiheit, verständliche Sprache, Optimierung für Mobilgeräte, leichte Navigation, Aktualität der Informationen, Feedback-Möglichkeiten, effektive Suchfunktion, Transparenz und multilinguale Unterstützung.

Die Gesamtbetrachtung der behördlichen Kommunikation zu EMF zeigt die Notwendigkeit einer kontinuierlichen Anpassung und Verbesserung der Informationsstrategien, um den spezifischen Bedürfnissen der Zielgruppen gerecht zu werden und gleichzeitig Vertrauen und Transparenz zu schaffen.

3 Webdatenanalyse der BfS-Webseite

Im Folgenden werden die vom BfS zur Verfügung gestellten Webseitendaten analysiert. Diese umfassen die „Export_Seiten_URL_YYYY-MM.csv“, die „Export_Suchbegriffe_(interne_Suche)_YYYY-MM.csv“ von September 2019 bis August 2023 und Protokolle von realen Anfragen, die vom BfS bearbeitet wurden.

Wir beginnen mit einer Analyse der Suchbegriffe, dem folgt eine Analyse der Seiten URLs und abschließend die Analyse der Anfrageprotokolle.

3.1 Am häufigsten gestellte Suchanfragen

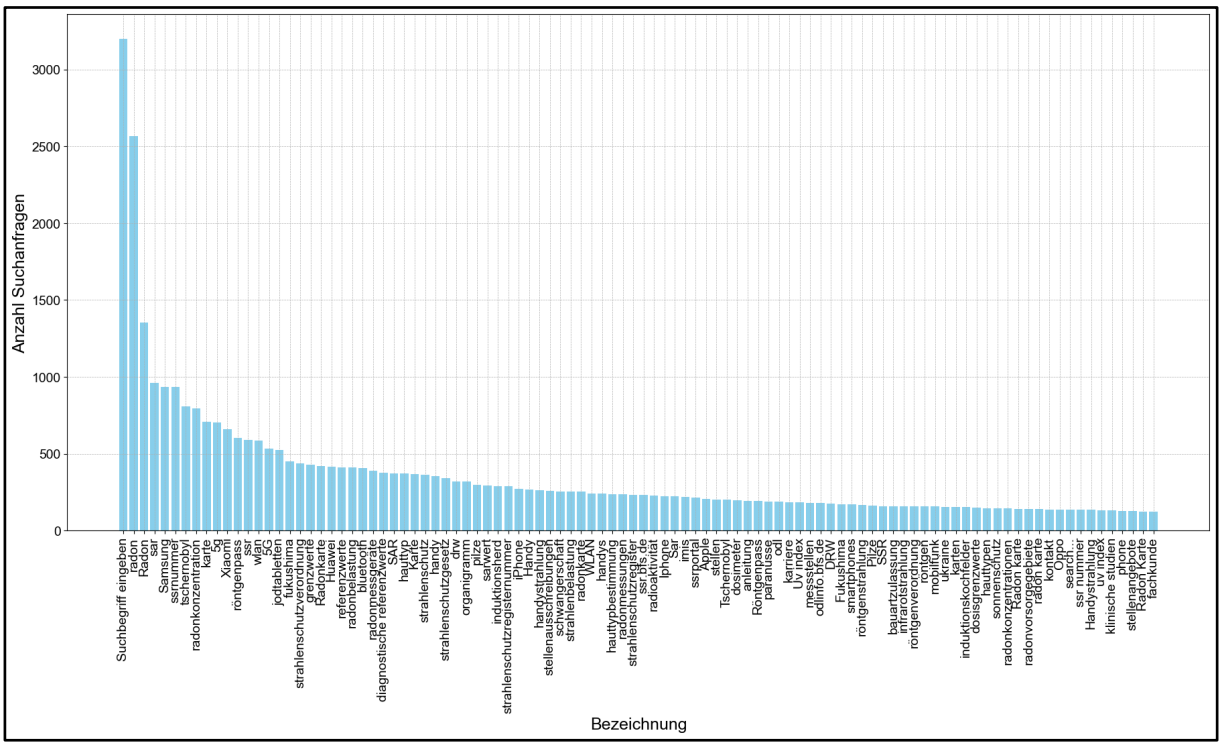

Abbildung 1: Gesamtzahl von Suchen aufgeschlüsselt nach Begriff

In Abbildung 1 sind die 100 am häufigsten gesuchten Begriffe und die Anzahl der getätigten Suchen mit diesem Begriff im Erhebungszeitraum (September 2019 – August 2023) aufgelistet. Dabei wurde der Sammelbegriff „Andere“ nicht berücksichtigt, da es sich dabei um einen Sammelbegriff handelt, der nicht genauer analysiert werden kann. Im Begriff „Andere“ sind über 88000 Suchanfragen gesammelt.

Bei genauer Betrachtung der Begriffe fallen Muster auf. Erfasst werden in den Suchanfragen beispielsweise verschiedenste Typen von Smartphones oder deren Hersteller. Da diese Geräte elektromagnetische Felder erzeugen, werden sie im Folgenden der Kategorie „EMF“ zugeordnet.

3.2 Kategorisierte Suchanfragen

Die Suchanfragen wurden auf Grund von semantischer Ähnlichkeit einer von vier Kategorien zugeordnet (siehe Abbildung 2). In der Kategorie „EMF“ befinden sich Suchbegriffe wie „Samsung“, „5g“ und „Handystrahlung“, während in der Kategorie „Radon“, Terme wie „Radonkarte“ und „Radonkonzentration“ enthalten sind. Unter der Kategorie „Andere Strahlung“ finden sich Begriffe wie „Jodtabletten“ und „Tschernobyl“ während Begriffe wie „Suchbegriff eingeben“ und „search“ in die Kategorie „Anderes“ fallen.

Anhand dieser Kategorien ist festzustellen, dass am häufigsten nach Themen im Bereich EMF gesucht wird. Die Kategorien „Radon“ und „Andere Strahlung“ haben jeweils etwa 30 Prozent weniger Suchanfragen verglichen mit „EMF“.

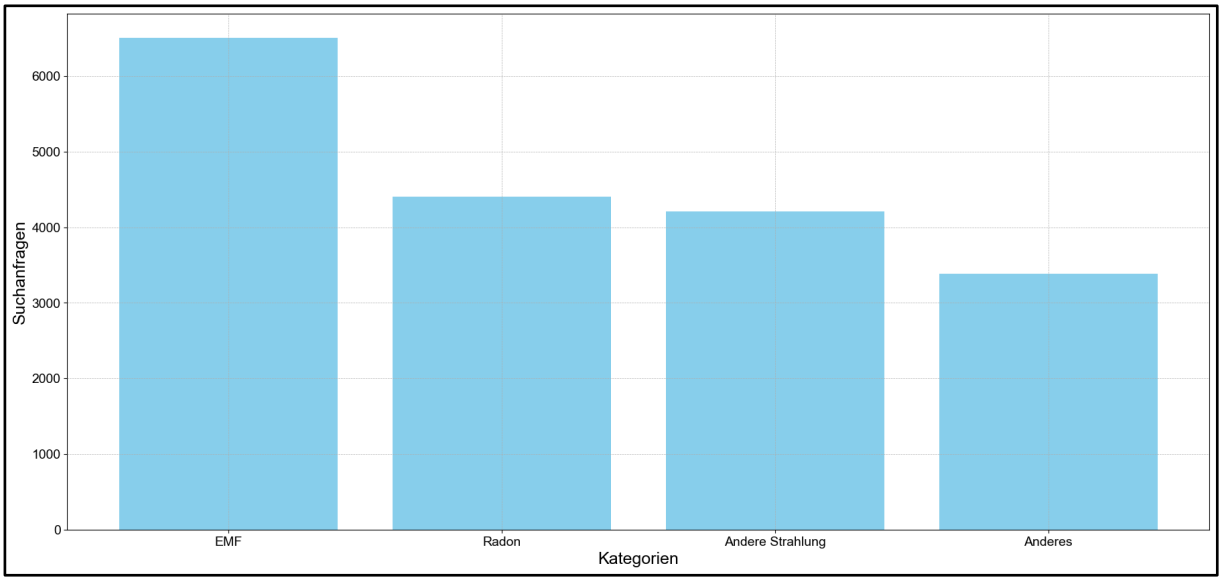

Abbildung 2: Kategorisierte Suchanfragen eingeteilt nach inhaltlichem Zusammenhang

3.3 Kategorisierte Suchanfragen pro Monat

Um einen genaueren Eindruck davon zu erhalten, wie sich die Relevanz der vier Kategorien im Lauf der Zeit verändert hat, wurde in Abbildung 3 die Menge der Suchanfragen pro Kategorie und pro Monat im Erhebungszeitraum berechnet. In diesem Kontext zeigen sich Aktivitätsspitzen vor allem zwischen Januar und Juli 2022. Insgesamt ist jedoch keine steigende Tendenz bezüglich der Anzahl der monatlichen Suchanfragen feststellbar. Insgesamt ist die Menge der monatlichen Suchanfragen gering, die Aktivitätsspitzen ausgenommen.

Auf Grundlage der bisher betrachteten Daten lassen sich einige Annahmen treffen. Zunächst wird die Suche hauptsächlich verwendet, um sich über aktuell wichtige Themen zu informieren und nicht als generelle Quelle von Informationen im Bereich Strahlung/Strahlenschutz. Diese Annahme würde die Aktivitätsspitzen erklären. Personen, die die Suchfunktion einmal genutzt haben, kehren nicht regelmäßig auf die Webseite zurück, da ansonsten eine positive Tendenz in der Menge der monatlichen Suchanfragen festgestellt werden würde). Alternativ könnte die Webseite allgemein einen niedrigen Bekanntheitsgrad aufweisen. Dies würde ebenfalls die konstant niedrigen Mengen an monatlichen Suchen erklären. Auf Grund dessen wird ein großes Steigerungspotenzial bezüglich der monatlichen Nutzung der Suche durch eine Verbesserung der Suchfunktion angenommen.

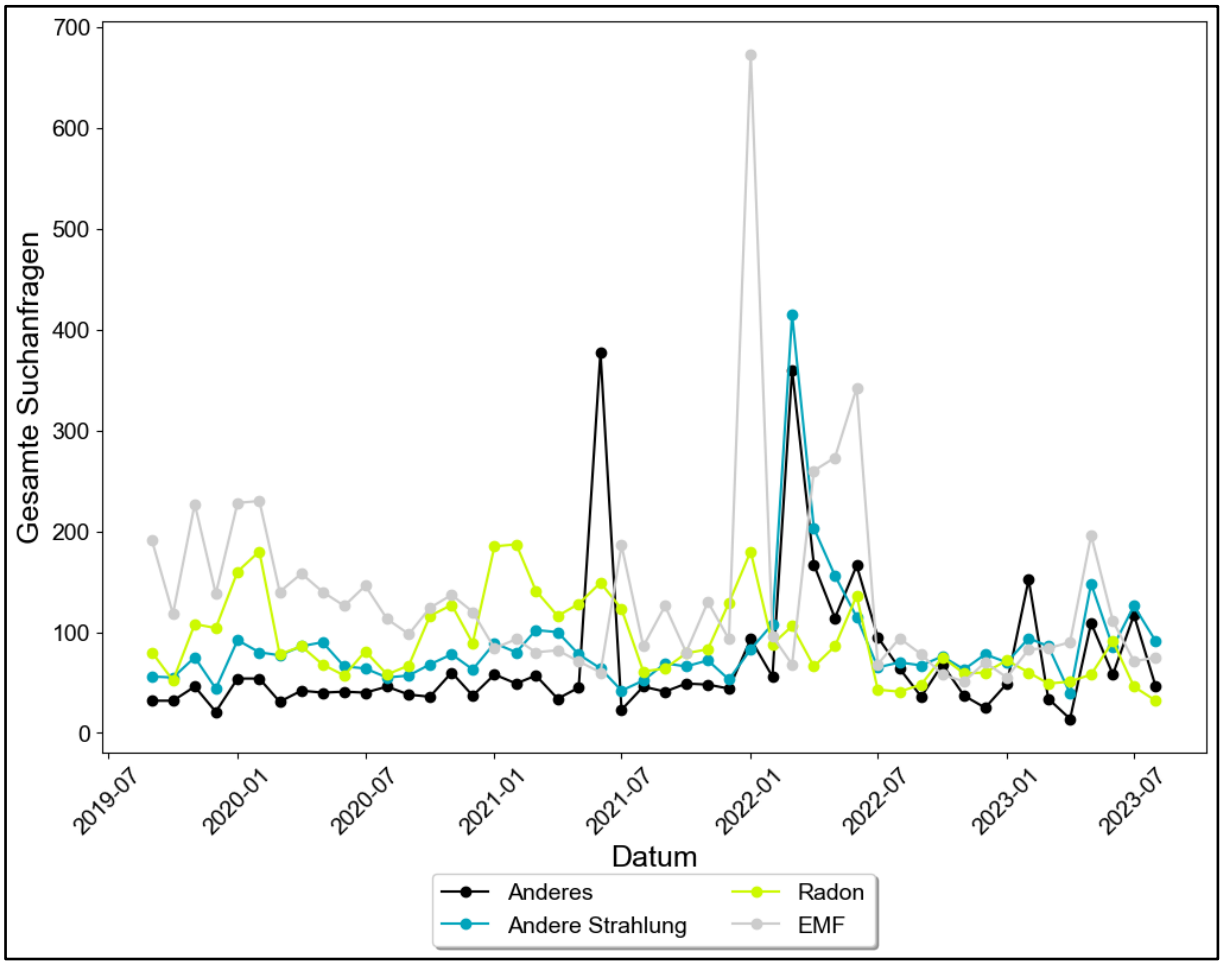

Abbildung 3: Gesamte Suchanfragen pro Kategorie und pro Monat

3.4 Häufigkeit von Aufrufen je Seite pro Monat

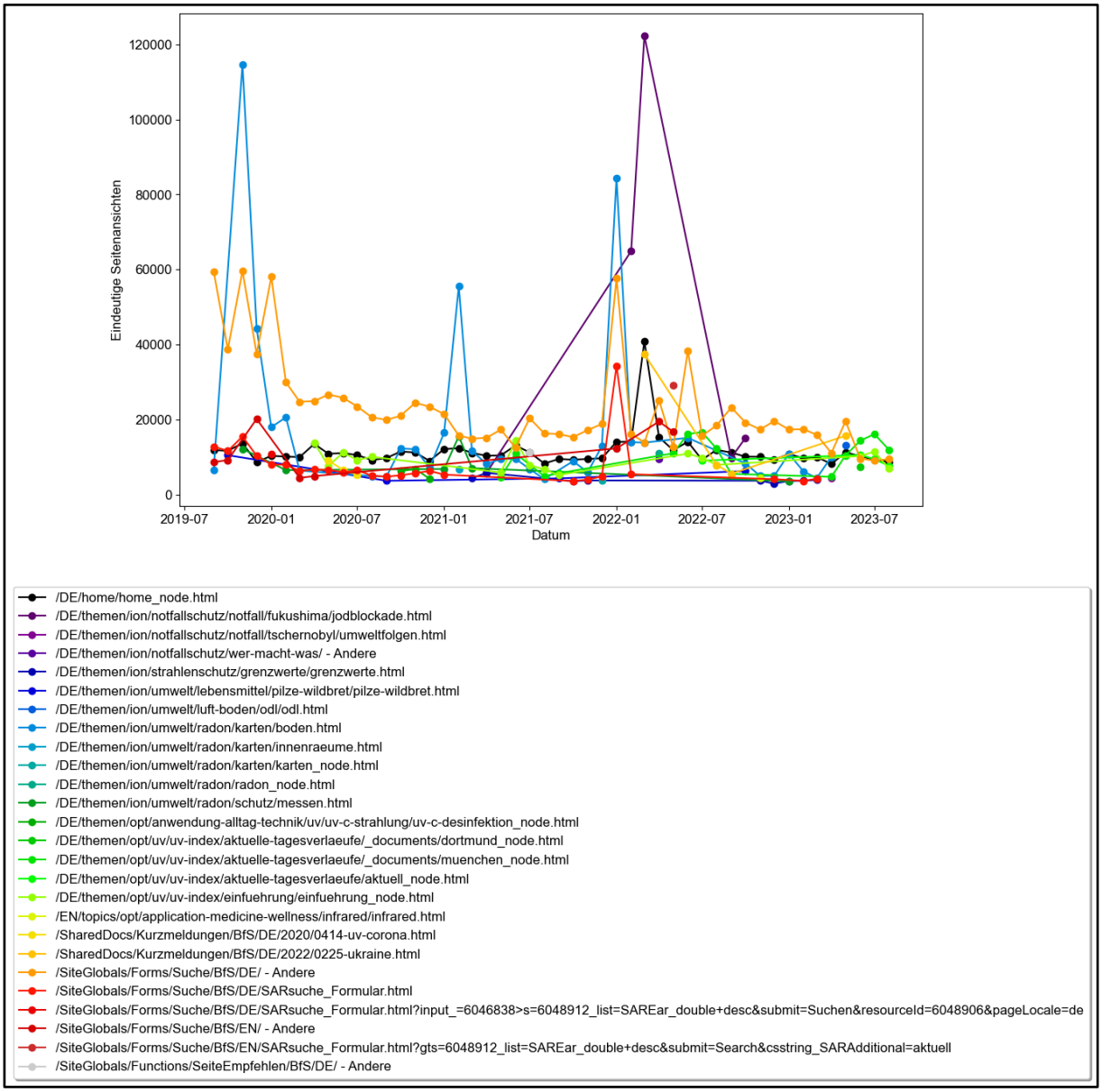

Abbildung 4: Top fünf der meistbesuchten Seiten pro Monat

Im Folgenden wird die gesamte Webseite, die Relevanz verschiedener Teile anhand mehrerer Metriken, angefangen mit Abbildung 4 untersucht. Es wird die Anzahl an monatlichen Seitenansichten im Erhebungszeitraum für die fünf in diesem Monat am häufigsten aufgerufenen Seiten ausgewertet.

Dabei ist festzustellen, dass eine oder mehrere Suchfunktionen (dargestellt durch das orange-rote Farbspektrum) in jedem Monat zu den fünf meistbesuchten Seiten zählen. Das ist insofern relevant, da es verdeutlicht, dass die Suchfunktion im Kontext der Gesamtseitenaufrufe häufig und regelmäßig genutzt wird. Außerdem finden sich auch Suchseiten in englischer Sprache unter den Top 5 wieder. Es besteht also auch unter Personen, deren Muttersprache nicht Deutsch ist und die nicht gut Deutsch sprechen, Interesse, die Seite zu benutzen. An dieser Grafik lassen sich ebenfalls Aktivitätsspitzen erkennen. Beispielsweise das Thema „Fukushima-Jodblockade“ (dunkles Violett) wird um März 2022 relevant und viel besucht, nachdem es ein Erdbeben und einen Tsunami in Fukushima gab.

3.5 Dauer von Aufrufen je Seite pro Monat

Untersucht man die Zeit in Stunden, die von Besuchern auf Seiten verbracht wurde, anstatt der Menge an eindeutigen Seitenansichten, so ergibt sich ein ähnliches Bild (siehe Abbildung 5). Auch hier lassen sich Aktivitätsspitzen feststellen, allerdings sind diese weniger ausgeprägt, was darauf hindeutet, dass viele der Besucher, die auf Grund aktuell relevanter Themen eine Seite besuchen, nur eine sehr kurze Zeit auf dieser verbringen. Dies kann mehrere Gründe haben. Entweder wird die gesuchte Information sehr schnell gefunden oder die Besucher sind nach Aufrufen der Seite überfordert, sodass sie die Suche nach für sie relevanten Informationen schnell aufgeben.

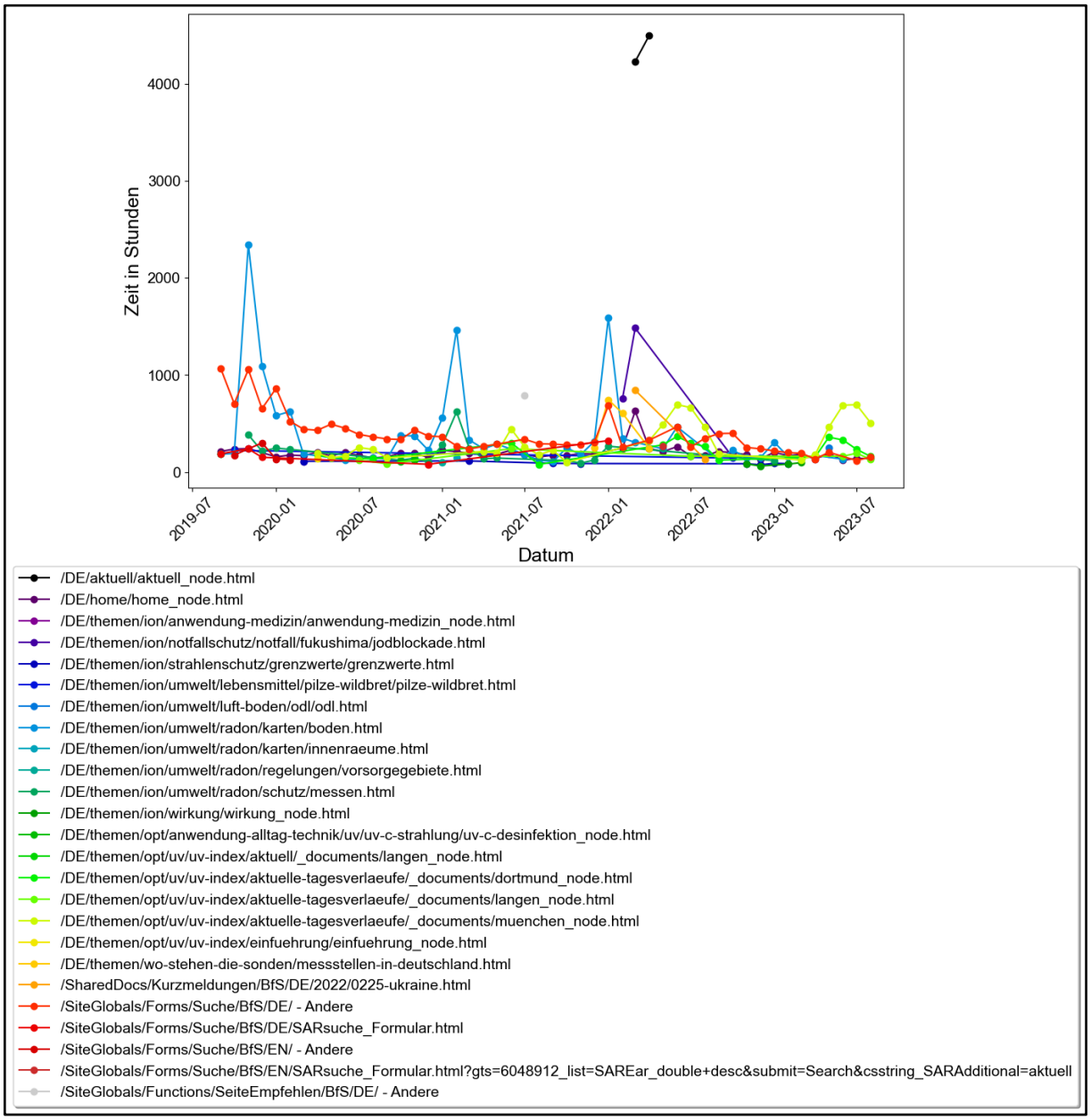

Abbildung 5: Die fünf am meisten aufgerufenen Seiten pro Monat, gemessen an dort verbrachter Zeit in Stunden

3.6 Gesamtzahl der Seitenaufrufe pro Monat

Abbildung 6 stellt einen Überblick über die gesamten monatlichen Seitenansichten im Erfassungszeitraum dar. Auch hier zeigen sich wieder starke Schwankungen, aber kein positiver Trend bezüglich des Wachstums der Webseite. Ein möglicher Grund dafür könnte sein, dass ein Teil der Bevölkerung nicht weiß, dass die Webseite existiert.

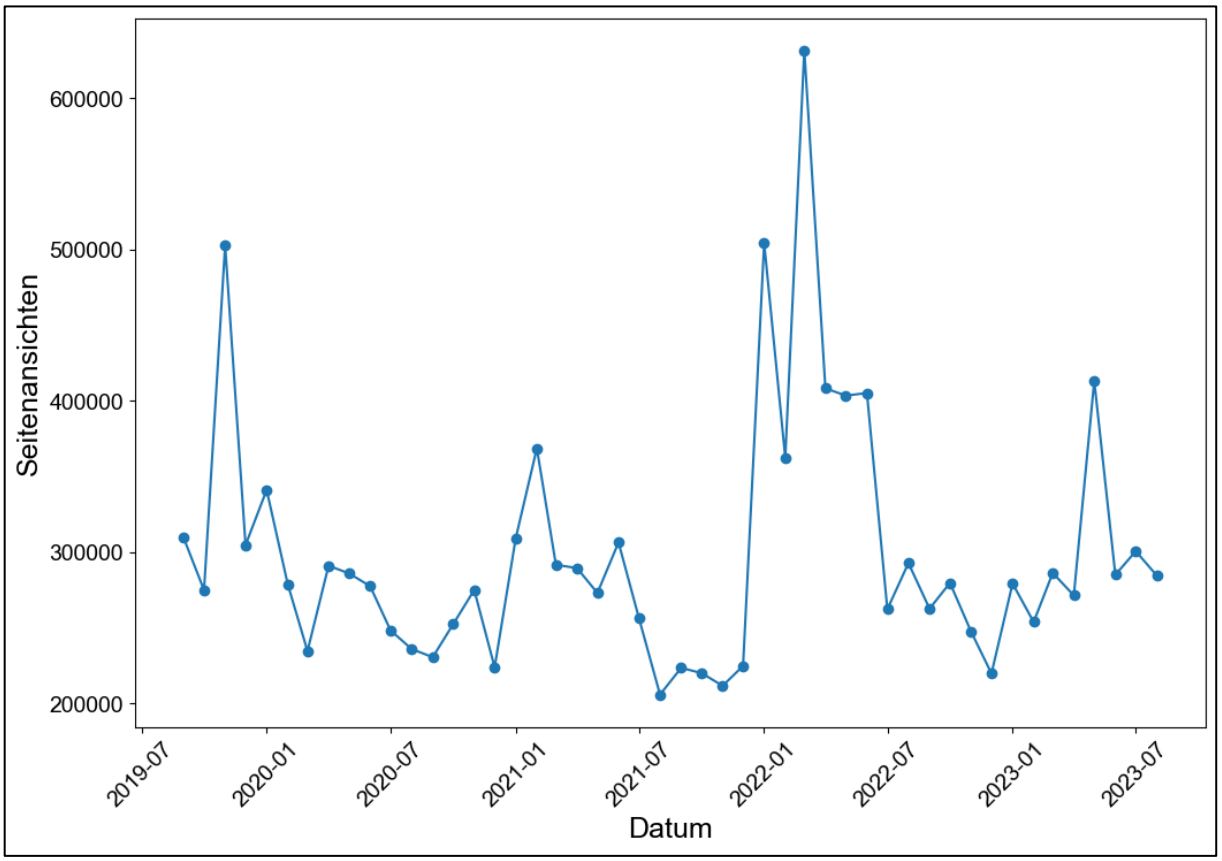

Abbildung 6: Die Gesamtmenge der monatlich angesehenen Seiten

3.7 Besuchte Seiten in Folge einer Suche pro Monat

In Abbildung 7 wird die durchschnittliche Anzahl an Seiten abgebildet, die pro Monat in Folge einer Suchanfrage im Erfassungszeitraum besucht wurden. Es wird ein Trend in Richtung 1,20 an durchschnittlich besuchten Ergebnisseiten sichtbar. Es wäre ideal, wenn Nutzer der Suchfunktion nach einer Suche auf Anhieb die richtige Seite finden, was einem Wert von 1.0 entspräche. Dieser Trend kann theoretisch zwei Gründe haben. Erstens könnte es daran liegen, dass mehr Nutzer als zu Beginn des Erhebungszeitraums relevante Informationen direkt auf Anhieb durch die Suchfunktion finden. Zweitens könnte der Trend schneller in Richtung 1,0 laufen, wenn die Anzahl an Nutzern gestiegen ist, die nach einer Suchanfrage keine Seite besuchen, da die Suche nicht die erwarteten Informationen zurückgegeben hat. Welcher der beiden Fälle zutrifft, ist aus den verfügbaren Daten nicht abzuleiten.

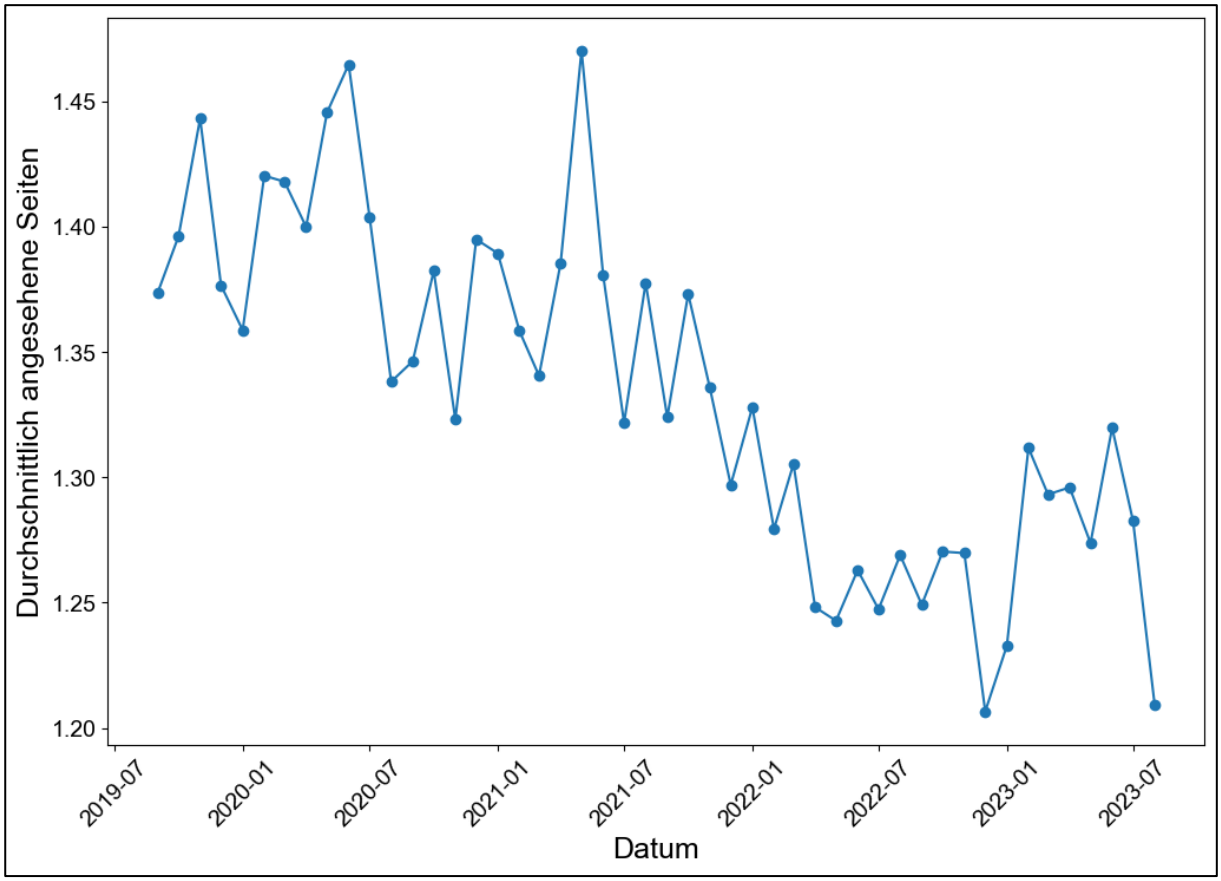

Abbildung 7: Die Menge der durchschnittlich nach einer Suche angesehenen Seiten pro Monat

4 KI-Use Cases in der Kommunikation des BfS

Basierend auf der umfassenden Datenanalyse und den Schlussfolgerungen aus den durchgeführten Interviews werden im Folgenden potenzielle Use Cases präsentiert, die zur Steigerung der Effizienz in der Behördenkommunikation eingesetzt werden können. Kapitel 6.3.1 bietet einen allgemeinen Überblick über die Möglichkeiten der Integration künstlicher Intelligenz in die Behördenkommunikation. Im darauffolgenden Kapitel wird eine vertiefte Betrachtung von drei im vorherigen Abschnitt genannten Use Cases vorgenommen, da diese für eine priorisierte Umsetzung empfohlen werden.

4.1 KI-Use-Cases Überblick

Im Folgenden werden verschiedene KI-Use Cases vorgestellt, die sich in Bezug auf die Behördenkommunikation eignen. Dabei werden Methoden wie das „finetuning" eines Large Language Model, die Übersetzung in andere Sprachen, eine Semantische Suche mittels Embeddings sowie die Vorklassifizierung von Bürgeranfragen vorgestellt.

4.1.1 LLM-Finetuning für Conversational Chatbots

Ein „large language model" ist ein künstlicher intelligenter Algorithmus, der darauf trainiert wurde, menschenähnliche Texte zu generieren, zu verstehen und darauf zu reagieren. Diese Modelle basieren auf maschinellem Lernen und Verwenden eine große Menge an Textdaten, um natürliche Sprachverarbeitungsaufgaben zu bewältigen.

LLM-Finetuning für Conversational Chatbots bezieht sich auf den Prozess, bei dem ein bereits trainiertes großes Sprachmodell, wie beispielsweise GPT-3, weiter verfeinert oder angepasst wird, um es speziell für die Aufgabe des Chatbot-Betriebes und der Konversation mit Benutzern zu optimieren. Dieser Feinabstimmungsprozess ist wichtig, um sicherzustellen, dass das Modell in der Lage ist, menschenähnliche und sinnvolle Dialoge zu erstellen und auf Anfragen in natürlicher Sprache angemessen zu reagieren.

Die Nutzung im BfS ermöglicht es den Menschen, sich mit einem Bot zu unterhalten, der das Thema gut kennt und sich anpassen kann, um die spezifischen Fragen zu beantworten, die ein Benutzer haben könnte. Dies könnte Mitarbeitende entlasten und zu einer schnelleren Bearbeitung von Anfragen führen.

Dafür kann beispielsweise Falcon 180B[1] genutzt werden. Für das Finetuning werden themenspezifische Daten benötigt. Danach kann mit dem Chatbot ähnlich interagiert werden, wie in einem Nachrichtenaustausch mit einer natürlichen Person.

Funktionsweise:
Personen können eine geschriebene Konversation mit dem Chatbot führen und ihre Fragen durch den Chatbot klären lassen.

Geplantes Einsatzgebiet beim BfS:
Der Conversational Chatbot kann als Assistenzsystem für die gesamte Seite dienen und zukünftig einen Großteil der Bürgeranfragen obsolet machen.

Bisherige Vorgehensweise beim BfS:
FAQs, Artikel, E-Mails, Briefe und Telefonate.

[1] https://huggingface.co/tiiuae/falcon-180B-chat)

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Wenn das LLM im Haus gefinetuned und betrieben werden soll, ist es notwendig, eine Serverarchitektur aufzubauen, die den zu erwartenden Ansprüchen der LLMs in Sachen Inferenz und Training gewachsen ist. Inferenz im Kontext der künstlichen Intelligenz und Informationstechnologie bezieht sich auf den Prozess, bei dem ein KI-System oder ein Algorithmus Schlussfolgerungen oder Vorhersagen auf Grundlage von vorhandenen Daten, Modellen oder Wissen zieht. Dieser Prozess kann auf verschiedene Arten durchgeführt werden und ist in vielen KI-Anwendungen von entscheidender Bedeutung. Zum Vergleich: die auf 2 Bit quantisierte Version von „Falcon 180B“ benötigt ungefähr 70GB vRAM für Inferenz und zusätzlichen Speicher für den Kontext. Eine Nvidia A100 80GB Grafikkarte kostet ungefähr 14 000 Euro. Es würden wahrscheinlich mehrere dieser Grafikkarten benötigt, um akzeptable Antwortzeiten und parallele Nutzerzahlen zu erlauben, die Details sind jedoch modellabhängig. Um „Falcon 180B“ nicht-quantisiert für Inferenz betreiben zu können, werden ungefähr 8 Nvidia A100 80GB Grafikkarten benötigt. Eine andere Option ist es, die dafür benötigte Serverleistung zu mieten. Des Weiteren wird Software benötigt. Hier bieten sich Python 3.10 und eine Reihe kostenloser Bibliotheken an.
Organisatorische Voraussetzungen	Für das initiale Finetuning werden Experten benötigt, die dafür sorgen, dass das LLM das gewünschte Verhalten erreicht. Des Weiteren wird mindestens ein Experte benötigt, um den Chatbot zu pflegen, zu überwachen und auf dem neuesten Stand zu halten.
Weitere Voraussetzungen	Für das Finetuning werden auf der Webseite enthaltene Informationen und zusätzliche vertrauenswürdige Ressourcen in einem maschinen-lesbaren Format benötigt.

Vorläufige Analyse des Use Cases:

Chancen	Ein Großteil der Standardfragen, die vorher per Mail beantwortet werden mussten, können vom System in Echtzeit automatisch beantwortet werden.
Risiken	LLMs können halluzinieren, was zu irreführenden oder falschen Ausgaben führen kann. Des Weiteren könnten „Bad Actors“ versuchen den Chatbot dazu zu bringen falsche Aussagen zu formulieren und dies dann als Aussage des BfS zu verkaufen.
Aufwand für die Implementierung	Hoch bis sehr hoch

Anmerkungen und Kommentare:
Der Chatbot muss kontinuierlich neu trainiert werden, um ihn an die aktuelle Version der Wissensdatenbank anzupassen.

4.1.2 Automatische Übersetzung (Fremdsprachen)

Mit Hilfe von LLMs und Natural-Language-Processing-Techniken (NLP-Techniken) ist es möglich, die Webseite automatisch in weitere andere Sprachen übersetzen zu lassen. Dies wäre hilfreich für alle, deren Muttersprache weder Deutsch noch Englisch ist. Nachdem die Übersetzung einmal generiert ist, kann die neue Seite genauso angelegt und verwaltet werden wie die bereits vorhandene englische Version der Seite.

Funktionsweise:
Mit Hilfe von Software werden zunächst alle aktuell auf der Webseite zugänglichen Texte automatisch übersetzt. Diese werden anschließend von einem Experten in die Webseite eingebunden. Danach werden alle neuen Texte mit Hilfe der Software übersetzt und in die Webseite eingebunden.

Geplantes Einsatzgebiet beim BfS:
Gesamte Webseite.

Bisherige Vorgehensweise beim BfS:
Unbekannt. Übersetzung ins Englische vorhanden.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Es wird genug Serverleistung benötigt, um die automatische Übersetzungssoftware regelmäßig durchzuführen. Die genauen Voraussetzungen richten sich dabei stark nach der verwendeten Software. Als Basis könnte Python 3.10 dienen. Mögliche Übersetzer wären beispielsweise Metas NLLB-200 600M (eigener Betrieb mit einer Grafikkarte mit mindestens 16GB VRAM) oder GPT-4 (für diesen Übersetzer würde kein Server benötigt, sondern ausschließlich ein API Zugang). Im Falle, dass PDF-Dokumente übersetzt werden müssen, ist zusätzliche Software nötig, um aus diesen Dokumenten Texte zu extrahieren.
Organisatorische Voraussetzungen	Es wird mindestens ein Data Scientist benötigt, um die Software zu entwickeln, in Stand zu halten und die Seite regelmäßig zu updaten.
Weitere Voraussetzungen	Um die Artikel übersetzen zu können, wäre der Zugang zu einem anderen Format als PDF vorzuziehen, da dieses die Extraktion von Textinhalten erschwert.

Vorläufige Analyse des Use Cases:

Chancen	Die Seite kann automatisch in mehrere, aktuell nicht unterstützten Sprachen übersetzt werden. Dies würde dazu führen, dass sie für mehr Menschen zugänglich wird.
Risiken	Für „low-resource“ Sprachen, ist die gewünschte Übersetzungsgenauigkeit möglicherweise nicht gegeben. Low-resource-Sprachen, auch als ressourcenarme Sprachen bezeichnet, sind Sprachen, für die nur begrenzte oder eingeschränkte linguistische Ressourcen wie Textkorpora, linguistische Tools, Lehrmaterialien und maschinelle Übersetzungsmodelle verfügbar sind. Diese Sprachen sind oft in Bezug auf ihre Verwendung in der Informationstechnologie und der künstlichen Intelligenz unterrepräsentiert. Das Fehlen von Ressourcen für diese Sprachen stellt Herausforderungen bei der Entwicklung von Anwendungen und Systemen dar, die natürliche Sprache verarbeiten, insbesondere im Kontext von maschinellem Lernen und KI. Dadurch könnten unter anderem auch Grafiken nicht automatisch übersetzt werden.
Aufwand für die Implementierung	Gering bis mittel

Anmerkungen und Kommentare:
Zunächst wäre es wichtig, herauszufinden, welche Sprachen potenzielle Nutzer der Seite sprechen, um das bestmögliche Übersetzungsprogramm auswählen zu können.

4.1.3 Automatische Übersetzung (einfache Sprache)

Mit Hilfe von Prompt Engineering erlauben es LLMs, Texte automatisch in einfache Sprache zu übersetzen. Es kann sich dieser Vorgang ebenfalls als eine Art Übersetzung vorgestellt werden, bei der die Sprache gleichbleibt und sich nur die Ausdrucksweise ändert. Eine bessere Verfügbarkeit der Seiteninhalte in einfacher Sprache würde die Nutzerfreundlichkeit der Webseite erhöhen.

Funktionsweise:
Mit Hilfe eines Programms können alle auf der Seite angezeigten Texte und alle neu hinzukommenden Texte automatisch übersetzt werden.

Geplantes Einsatzgebiet beim BfS:
Gesamte Webseite.

Bisherige Vorgehensweise beim BfS:
Unbekannt. Leichte Sprache teilweise auf Deutsch vorhanden.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Es wird genug Serverleistung benötigt, um die automatische Übersetzungssoftware regelmäßig laufen zu lassen. Die genauen Voraussetzungen richten sich dabei stark nach der verwendeten Software. Als Basis könnte Python 3.10 dienen. Mögliche Übersetzer wären beispielsweise GPT-4 (für diesen Übersetzer würde kein Server benötigt, sondern ausschließlich ein API-Zugang). Im Falle davon, dass PDF-Dokumente übersetzt werden müssen, ist zusätzliche Software nötig, um aus diesen Dokumenten Texte zu extrahieren.
Organisatorische Voraussetzungen	Es wird mindestens ein Data Scientist/Prompt Engineer benötigt, um die Software zu entwickeln, in Stand zu halten und die Seite regelmäßig upzudaten.
Weitere Voraussetzungen	Um die Artikel in leichte Sprache übersetzen zu können wäre Zugang zu einem anderen Format als PDF vorzuziehen, da dieses die Extraktion von Textinhalten erschwert.

Vorläufige Analyse des Use Cases:

Chancen	Die Übersetzung in einfache Sprache ermöglicht es mehr Menschen, die auf der Webseite präsentierten Informationen zu verstehen.
Risiken	Durch das Vereinfachen der Sprache könnten wichtige Informationen verloren gehen.
Aufwand für die Implementierung	Gering bis mittel

Anmerkungen und Kommentare:
Der Use Case „Automatische Übersetzung (einfache Sprache)“ kann theoretisch mit dem vorherigen Use Case „Automatische Übersetzung (Fremdsprache)“ kombiniert werden, um für mehrere Sprachen Texte in einfacher Sprache zu generieren.

4.1.4 Integration eines Vorlesewerkzeugs

Auf der Webseite wird ein Vorlesewerkzeug unter den Texten angeboten, mit dem beliebige Inhalte der Webseite angehört werden können, anstatt sie zu lesen.

Funktionsweise:
Mit Hilfe von „Text-to-Speech“-Software können längere Artikel automatisch zu Audiodateien konvertiert werden. Diese können dann in die Webseite eingebunden werden, um Nutzern zu erlauben sich Texte vorlesen zu lassen, anstatt sie lesen zu müssen. Dies würde es Menschen mit Sehbehinderung erleichtern, Zugang zu den auf der Seite enthaltenen Informationen bieten.

Geplantes Einsatzgebiet beim BfS:
Diverse Artikel auf der BfS-Webseite.

Bisherige Vorgehensweise beim BfS:
Unbekannt. Einige Videos sind verfügbar.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Es wird ein PC mit ausreichenden Kapazitäten benötigt, um die notwendige Software ausführen zu können. Ein führendes Modell ist aktuell Massively Multilingual Speech (MMS) von Meta[2], das in der Lage ist, für über 1000 Sprachen Texte in gesprochene Sprache zu übersetzen. Außerdem würde Python 3.10 und eine Reihe weiterer Bibliotheken benötigt.
Organisatorische Voraussetzungen	Benötigt werden Data Scientists oder NLP-Experten, um das richtige Modell auszuwählen, aufzusetzen, die Integration vorzunehmen sowie die Pflege/Wartung sicherzustellen.
Weitere Voraussetzungen	Keine.

Vorläufige Analyse des Use Cases:

Chancen	Einfacherer Zugang zu Inhalten sowohl für Personen mit Sehbehinderung als auch für Personen, die sich nebenbei über die angebotenen Themen informieren wollen.
Risiken	Die Komplexität einiger Fachbegriffe könnte zu einigen fehlerhaften Aussprachen führen.
Aufwand für die Implementierung	Gering bis Mittel. Dies ist abhängig von einer Reihe von Faktoren, die den Implementierungsaufwand wesentlich vergrößern könnten.

Anmerkungen und Kommentare:
Es existieren bereits Bibliotheken die „Text zu Audio“ konvertieren. Wenn es möglich ist, eine dieser Bibliotheken zu nutzen, ist der Aufwand gering. Soll die „Stimme“ verändert werden, erhöht sich der Aufwand.

[2] https://ai.meta.com/blog/multilingual-model-speech-recognition/

4.1.5 Semantische Suchfunktion mittels Embeddings

Text Embeddings („Einbettungen“) sind Vektoren in einem hochdimensionalen Vektorraum. Sie können dazu verwendet werden, semantische Ähnlichkeit zwischen Wörtern und Texten zu quantifizieren, indem sie die Bedeutung eines Wortes oder Textes aus dessen Kontext ermitteln. Diese Technologie ermöglicht es, zu einer Suchanfrage semantisch relevante, anstatt syntaktisch ähnlicher Ergebnisse zu liefern.

Funktionsweise:
Embeddings werden für die auf der BfS-Webseite zur Verfügung gestellten Informationen/Dokumente vorab erstellt. Für eine eingehende Suchanfrage wird das Embedding effizient zur Laufzeit berechnet. Nun wird die Ähnlichkeit zwischen der Suchanfrage und den Dokumenten berechnet und die semantisch relevantesten Ergebnisse werden zurückgegeben.

Es können vortrainierte Embedding Modelle verwendet werden, es kann darüber hinaus ein Finetuning auf den Dokumenten der BfS-Webseite stattfinden.

Geplantes Einsatzgebiet beim BfS:
Suchfunktion der BfS-Webseite.

Bisherige Vorgehensweise beim BfS:
Unbekannt, vermutlich einfaches String-Matching auf rein syntaktischer Ebene oder in der Datenbank hinterlegte Keywords.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Zugang zu Textdaten der BfS-Webseite: Die Verfügbarkeit von Textinhalten auf der Webseite ist entscheidend, um Text Embeddings zu erstellen. Text Embedding-Modelle: Auswahl von geeigneten Embedding-Modellen, die für die semantische Suche verwendet werden können. Rechenkapazitäten: Die Implementierung erfordert ausreichende Rechenleistung. Die Verwendung von Hardwarebeschleunigung ist sowohl für das Training/das Feintuning des Modells als auch für die Inferenz erforderlich. Die benötigte Rechenleistung variiert stark je nach dem verwendeten Embedding-Modell. Kleinere Modelle wie (S)BERT (Devlin et al., 2018; Nils et al., 2019) können problemlos auf handelsüblichen GPUs wie der NVIDIA GeForce RTX 3090 ausgeführt werden. Für größere Modelle ist jedoch leistungsstarke Server-Hardware erforderlich oder es muss auf Cloud-Computing-Ressourcen zurückgegriffen werden. Alternativ bietet sich die Möglichkeit, zum Beispiel das ADA-Embedding-Modell von OpenAI[3] über die API zu verwenden, was unter Umständen jedoch zu einer höheren Latenz führen kann.
Organisatorische Voraussetzungen	Es werden Data Scientists oder NLP-Experten mit Erfahrung in Natural Language Processing benötigt, um Embedding-Modelle auszuwählen, zu trainieren und zu integrieren.

[3] https://openai.com/blog/new-and-improved-embedding-model

Weitere Voraussetzungen	Keine.

Vorläufige Analyse des Use Cases:

Chancen	Die Implementierung von Text Embeddings kann zu genaueren und semantisch relevanten Suchergebnissen führen. Nutzer erhalten schnellere und relevantere Informationen, was die Zufriedenheit steigern kann. Darüber hinaus können multilinguale vortrainierte Embeddings verwendet werden, wodurch auch für Anfragen in Fremdsprachen, relevante Ergebnisse geliefert werden können.
Risiken	1. Komplexität: Die Integration von Embeddings kann technisch anspruchsvoll sein und zusätzliche Hardwareressourcen erfordern. 2. Datenqualität: Die Qualität der Suchergebnisse hängt von der Qualität der verfügbaren Textdaten ab. Unstrukturierte oder unvollständige Daten können die Ergebnisse beeinträchtigen.
Aufwand für die Implementierung	Gering bis Mittel. Dies ist abhängig vor allem von der Modellauswahl, der Datenverfügbarkeit und den erforderlichen Ressourcen. Die Implementierung könnte mehrere Monate in Anspruch nehmen, abhängig von der Komplexität und den vorhandenen Ressourcen.

Anmerkungen und Kommentare

Es ist wichtig, Benutzerfeedback zu sammeln und die Suche kontinuierlich zu überwachen und zu optimieren, um sicherzustellen, dass die Embedding-basierte Suche den Erwartungen der Benutzer entspricht. Die Zusammenarbeit zwischen verschiedenen Teams, einschließlich Datenwissenschaftlern, Entwicklern und Inhaltsexperten, ist entscheidend für den Erfolg des Projekts. Es sollte auch geprüft werden, ob die Vorteile der Embedding-Implementierung die damit verbundenen Kosten und den Aufwand rechtfertigen.

4.1.6 Verbesserte Suchfunktion durch Multimodalität

Die Integration von Multimodalität in die Suchfunktion ermöglicht es, nicht nur textuelle Informationen, sondern auch visuelle Inhalte wie Bilder in die Suchanfragen beziehungsweise Suchergebnisse einzubeziehen. Dadurch werden die Suchergebnisse genauer und vielfältiger, was die Benutzererfahrung erheblich verbessern kann.

Funktionsweise:
Die Suchfunktion wird durch die Integration einer Bilderkennungs- und Analysekomponente mithilfe von Machine-Learning und Computer Vision-Technologien und/oder Verwendung multimodaler Embeddings erweitert. Multimodale Suchanfragen, bei denen Benutzer sowohl Text als auch Bilder verwenden können, zum Beispiel Suche nach "Strahlenbelastung" mit hochgeladenem Bild der eigenen Mikrowelle, verbessern die Nutzererfahrung und führen dazu, schneller die gewünschten Informationen zu erhalten. Die Kombination von textuellen und visuellen Ergebnissen sorgt für umfassendere und relevantere Informationspräsentation.

Geplantes Einsatzgebiet beim BfS:
Suchfunktion der Webseite des BfS.

Bisherige Vorgehensweise beim BfS:
Die bisherige Suchfunktion auf der BfS-Webseite beschränkte sich auf textbasierte Suchanfragen und kann keine visuellen Inhalte berücksichtigen.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Zugang zu Text- und Bilddaten der BfS-Webseite: Die Verfügbarkeit von sowohl textuellen als auch visuellen Inhalten ist entscheidend, um eine multimodale Suche zu ermöglichen. Text- und Bildverarbeitungsmodelle: Auswahl und Implementierung von geeigneten Modellen und Algorithmen für die Text- und Bilderkennung sowie -analyse. Rechenkapazitäten: Die Implementierung erfordert ausreichende Rechenleistung, um Text- und Bildverarbeitung in Echtzeit durchzuführen (Server mit leistungsstarker(n) GPU(s)).
Organisatorische Voraussetzungen	Data Scientists, Computer Vision-Experten und NLP-Experten werden benötigt, um die Modelle zu entwickeln, zu trainieren und zu integrieren.
Weitere Voraussetzungen	Keine.

Vorläufige Analyse des Use Cases:

Chancen	Die Integration von Multimodalität in die Suchfunktion bietet die Chance auf präzisere und vielfältigere Suchergebnisse, was die Benutzererfahrung verbessern und die Zufriedenheit steigern kann.
Risiken	1. Komplexität: Die Implementierung von Bilderkennung und -analyse sowie die Kombination von textuellen und visuellen Ergebnissen ist technisch anspruchsvoll, setzt hochqualifiziertes Personal voraus und erfordert zusätzliche Hardware-Ressourcen. 2. Datenqualität: Die Qualität der visuellen Ergebnisse hängt von der Genauigkeit der Bilderkennung ab und unklare oder unvollständige Bilddaten können die Ergebnisse beeinträchtigen.
Aufwand für die Implementierung	Mittel bis Hoch. Dies ist abhängig vor allem von der Modellauswahl, der Datenverfügbarkeit und den erforderlichen Ressourcen. Die Implementierung könnte mehrere Monate in Anspruch nehmen, abhängig von der Komplexität und den vorhandenen Ressourcen.

Anmerkungen und Kommentare:
Die Integration von Multimodalität eröffnet neue Möglichkeiten für die Benutzer, relevante Informationen auf der Webseite zu finden. Die Qualität der Suchergebnisse hängt von der Genauigkeit der Bilderkennung und -analyse ab, daher ist eine sorgfältige Modellauswahl und -entwicklung entscheidend. Benutzerfeedback

sollte kontinuierlich gesammelt werden, um die Suche weiter zu optimieren und sicherzustellen, dass sie den Erwartungen der Benutzer entspricht. Generell sollte jedoch zunächst geprüft werden, ob die Vorteile der Multimodalitäts-Implementierung die damit verbundenen Kosten und den Aufwand rechtfertigen.

4.1.7 Automatische Vorklassifizierung von Bürgeranfragen

Die automatische Vorklassifizierung von schriftlichen Bürgeranfragen ermöglicht eine effiziente Zuordnung der Anfragen zu den entsprechenden Abteilungen des BfS, basierend auf ihrer fachlichen Tiefe und dem erforderlichen Aufwand.

Funktionsweise:

1. Anfrageanalyse: Die schriftlichen Bürgeranfragen werden automatisch analysiert, um ihren Inhalt und ihre fachliche Komplexität zu bewerten.
2. Klassifizierungsalgorithmen: Anhand von Textanalyse und maschinellem Lernen (ML) werden Klassifizierungsalgorithmen eingesetzt, um die Anfragen in Kategorien oder Abteilungen zu unterteilen.
3. Zuordnung: Die vorklassifizierten Anfragen werden automatisch den zuständigen Abteilungen zugewiesen, um die Bearbeitung zu optimieren.

Geplantes Einsatzgebiet beim BfS:
Alle Kanäle, über welche Bürgeranfragen schriftlich eingereicht werden können.

Bisherige Vorgehensweise beim BfS:
Bürgeranfragen werden manuell von Mitarbeitern geprüft und an die verantwortliche Abteilung weitergeleitet.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Handelsüblicher Rechner mit Grafikkarte ausreichend.
Organisatorische Voraussetzungen	NLP/ML Experten für Implementierung, Wartung und Weiterentwicklung.
Weitere Voraussetzungen	Anpassungsbedarf: Die Algorithmen müssen kontinuierlich aktualisiert und angepasst werden, um Veränderungen in den Anfragen zu berücksichtigen.

Vorläufige Analyse des Use Cases:

Chancen	Effizienzsteigerung: Die automatische Vorklassifizierung reduziert den manuellen Aufwand und beschleunigt die Bearbeitung von Bürgeranfragen. Verbesserte Kundenzufriedenheit: Bürger erhalten schnellere und präzisere Antworten auf ihre Anfragen.
Risiken	Fehlerhaft klassifizierte Anfragen müssen manuell an die verantwortliche Abteilung übermittelt werden (wie bisher).

Aufwand für die Implementierung	Niedrig bis mittel

4.1.8 Generierung von Antwortvorschlägen für schriftliche Bürgeranfragen

Die (semi-)automatische Beantwortung von schriftlichen Bürgeranfragen ermöglicht eine schnellere und genauere Reaktion auf Anfragen, indem relevante Informationen extrahiert und teilweise vorab in eine Antwortvorlage eingefügt werden. Dies führt zu einer Reduzierung des manuellen Arbeitsaufwands und einer beschleunigten Bearbeitung der Anfragen.

Funktionsweise:

1. Anfrageerfassung: Schriftliche Bürgeranfragen werden automatisch erfasst und in ein verarbeitbares Format umgewandelt.
2. Anfrageverstehen: Natürliche Sprachverarbeitung (NLP) und maschinelles Lernen (ML) werden eingesetzt, um den Inhalt und die Absicht der Anfragen zu verstehen.
3. Informationen extrahieren: Relevante Informationen aus der Anfrage werden extrahiert, z. B. Name des Antragstellers, Anliegen, Aktennummer, etc.
4. Wissensdatenbank: Eine umfangreiche Wissensdatenbank wird gepflegt, die Antworten auf häufig gestellte Fragen und Anliegen enthält.
5. Antwortgenerierung: Basierend auf dem Verständnis der Anfrage, den extrahierten Informationen und unter Berücksichtigung der Wissensdatenbank werden automatisch passende Antwortvorlagen generiert.
6. (Semi)-Automatische Antwortvorbereitung: Die Antwortvorlagen werden teilweise mit den extrahierten Informationen vorausgefüllt.
7. Manuelle Überprüfung und Anpassung: Die teilweise vorab gefüllten Antworten werden den zuständigen Mitarbeitern zur manuellen Überprüfung und gegebenenfalls Anpassung vorgelegt.
8. Antwortübermittlung: Die überprüften und angepassten Antworten werden über den gewählten Kommunikationskanal (zum Beispiel E-Mail, Webseite, Chatbot) an die Bürger gesendet.

Geplantes Einsatzgebiet beim BfS:
Alle schriftlichen Bürgeranfragen über verschiedene Kommunikationskanäle.

Bisherige Vorgehensweise beim BfS:
Bürgeranfragen werden manuell von Mitarbeitern geprüft und individuell beantwortet.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Je nach Komplexität der zugrundeliegenden ML-Modelle, einen leistungsstarken Rechner.
Organisatorische Voraussetzungen	Es werden NLP/ML-Experten für die Entwicklung und Pflege des Systems sowie Content-Manager für die Aktualisierung der Wissensdatenbank benötigt.
Weitere Voraussetzungen	Regelmäßige Aktualisierung der Wissensdatenbank, um aktuelle Informationen und Lösungen zu gewährleisten.

Vorläufige Analyse des Use Cases:

Chancen	Effizienzsteigerung: Die (semi-)automatische Beantwortung reduziert den manuellen Arbeitsaufwand erheblich und beschleunigt die Reaktion auf Bürgeranfragen. Verbesserte Kundenzufriedenheit: Bürger erhalten schnelle und präzise Antworten. Konsistenz: Durch die teilweise automatisierte Antwortvorbereitung wird eine konsistente Kommunikation gewährleistet.
Risiken	Komplexität der Anfragen: Einige Anfragen könnten sehr komplex sein und eine manuelle Intervention erfordern. Wissensdatenbankpflege: Die Aktualisierung und Pflege der Wissensdatenbank erfordern kontinuierlichen Aufwand.
Aufwand für die Implementierung	Mittel bis hoch

Anmerkungen und Kommentare:
Eine klare Schnittstelle zwischen der automatisierten Antwortvorlage und der manuellen Überprüfung ist entscheidend, um sicherzustellen, dass komplexe Anfragen angemessen behandelt werden und die Qualität der Antworten gewährleistet ist. Die Qualität der Wissensdatenbank und das Training der NLP-Modelle sind ebenfalls entscheidend für den Erfolg dieses Use Cases. Es sollte vorab geprüft werden, ob sich der Aufwand eines solchen Systems lohnt.

4.1.9 Interaktive FAQ

Effiziente Informationsbereitstellung und Verbesserung des Wissens über elektromagnetische Felder durch interaktive FAQ. Dieser Use Case zielt darauf ab, Bürgeranfragen im Zusammenhang mit elektromagnetischen Feldern durch die Bereitstellung von interaktiven FAQ effizienter zu beantworten. Das System verwendet maschinelles Lernen, Multimedia-Inhalte, Nutzerfeedback und natürliche Sprachverarbeitung, um relevante und kontextbezogene Informationen automatisch bereitzustellen und kontinuierlich zu verbessern.

Funktionsweise

1. Vorhersage von Fragen

Das System verwendet maschinelles Lernen, um die Fragen vorherzusagen, die Bürger wahrscheinlich zu elektromagnetischen Feldern stellen werden. Dies basiert auf dem Nutzerverhalten und vergangenen Abfragen.

2. Multimedia-FAQs

Die FAQs werden um multimediale Inhalte wie Videos, Infografiken und interaktive Simulationen erweitert, um komplexe Konzepte im Zusammenhang mit elektromagnetischen Feldern effektiver zu erklären.

3. Integration von Nutzerfeedback

Nutzer haben die Möglichkeit zu bewerten, wie hilfreich die FAQs waren und Feedback zu geben. KI analysiert dieses Feedback, um den Inhalt kontinuierlich zu verbessern und Bereiche mit detaillierteren Informationen zu identifizieren.

4. Natürliche Sprachverarbeitung

Das interaktive FAQ-System ist mit Fähigkeiten zur natürlichen Sprachverarbeitung ausgestattet, um eine Vielzahl von Nutzeranfragen zu verstehen und darauf zu antworten, einschließlich Anfragen mit unterschiedlichen technischen Details.

5. Kontextbezogene Antworten

Das System bietet kontextbezogene Antworten. Zum Beispiel kann es bei einer Anfrage zur Sicherheit elektromagnetischer Felder in der Nähe von Schulen lokalisierte und relevante Informationen basierend auf dem Standort des Nutzers bereitstellen.

Geplantes Einsatzgebiet beim BfS:
FAQs der Webseite des BfS.

Bisherige Vorgehensweise beim BfS:
FAQs in textueller Form mit Filterfunktion.

Voraussetzungen für einen Einsatz beim BfS:

Technische Voraussetzungen	Eine skalierbare Infrastruktur für das FAQ-System und ausreichende Bandbreite für Multimedia-Inhalte.
Organisatorische Voraussetzungen	Es werden Experten für maschinelles Lernen, Content-Ersteller für Multimedia-Inhalte, NLP-Spezialisten für die Sprachverarbeitung und Datenanalysten zur Auswertung von Nutzerfeedback benötigt.
Weitere Voraussetzungen	Regelmäßige Aktualisierung der FAQs und Integration von Nutzerfeedback in den Verbesserungsprozess.

Vorläufige Analyse des Use Cases:

Chancen	Verbesserte Kundenzufriedenheit: Bürger erhalten schnelle, präzise und multimediale Antworten auf ihre Fragen. Kontinuierliche Verbesserung: Das System kann durch die Integration von Nutzerfeedback und maschinellem Lernen kontinuierlich optimiert werden. Besseres Verständnis: Bürger können komplexe Konzepte besser verstehen, da multimediale Inhalte zur Verfügung stehen.
Risiken	**Komplexität der Anfragen**: Einige Fragen könnten sehr komplex sein und nicht durch das interaktive FAQ gelöst werden. **Content-Pflege**: Die Aktualisierung und Pflege der FAQs und Multimedia-Inhalte erfordert kontinuierlichen Aufwand.
Aufwand für die Implementierung	hoch

Anmerkungen und Kommentare:
Die Qualität der Multimedia-Inhalte und die Integration von Nutzerfeedback sind entscheidend für den Erfolg dieses Use Cases.

4.2 Priorisierung spezifischer KI-Use Cases

Im nachfolgenden Abschnitt werden drei KI-Anwendungsfälle detailliert beschrieben, die wir priorisiert für die Implementierung und Umsetzung vorschlagen. Wir haben die Maßnahmen als aussichtsreiche Chancen identifiziert, um die Kommunikation des BfS nachhaltig zu unterstützen. Die Auswahl dieser Use Cases wird durch eine eingehende Analyse ihrer besonderen Eignung begründet. Dabei werden die technischen Hintergründe ausführlich erläutert. Weiterhin wird aufgezeigt, ob die jeweilige Lösung kurz-, mittel- oder langfristig geplant ist.

4.2.1 Methoden zur Verbesserung der Suchfunktion der BfS-Webseite

Die Gewährleistung einer effizienten Suche nach relevanten Informationen auf der Webseite des BfS ist von entscheidender Bedeutung, um Bürgerinnen und Bürgern qualitativ hochwertige und genaue Informationen im Bereich „Strahlenschutz“, aber auch zu vielen weiteren Themenfeldern, zugänglich zu machen. Unsere internen Interviews und Befragungen haben ergeben, dass die bisherige Qualität der Suchfunktion überwiegend negativ bewertet wird, was sich auch in der insgesamt geringen Nutzung der Suchfunktion widerspiegelt.

Ein konkretes Beispiel für die mangelnde Qualität der bisherigen Suchfunktion zeigt sich bei der Suche nach dem Begriff „Strahlenschutz“. Diese Suche führt zu 1.154 Treffern. Hingegen ergibt eine Suche nach „Strahlen Schutz“, also mit einem Leerzeichen zwischen „Strahlen“ und „Schutz“, lediglich 67 Ergebnisse (Stand: 20. November 2023) (siehe Abbildung 8). Eine Suche nach „Ionisierung“ und „ionisierend“ liefert jedoch jeweils identische Ergebnisse (siehe Abbildung 9). Es sind somit bereits einige einfache Normalisierungsschritte in der Suche implementiert, wie beispielsweise Case Sensitivity (Unterscheidung von Groß- und Kleinschreibung) oder Stemming (Verfahren in der Textverarbeitung und der Computerlinguistik, bei dem Wörter auf ihren sogenannten Stamm reduziert werden) beziehungsweise Lemmatization (ein Prozess in der natürlichen Sprachverarbeitung, bei dem verschiedene Flexionsformen eines Wortes auf ihre Grundform, das sogenannte Lemma, reduziert werden). Dennoch handelt es sich bislang um eine rein „Keyword“-basierte Suche, was die Suche stark einschränkt. Konkrete und sprachlich komplexe Fragen wie „Wie viele Meter beträgt der gesetzlich vorgeschriebene Mindestabstand zwischen Hochspannungsleitungen und Wohngebieten?“ führen zu keinen relevanten Ergebnissen. Ebenso liefern Suchanfragen nach Synonymen oder semantisch ähnlichen Begriffen völlig unterschiedliche Ergebnisse: Die Suche nach „5G“ ergibt 71 Treffer, während „5. Mobilfunkgeneration“ zu 18 Ergebnissen führt, und „Fünfte Mobilfunkgeneration“ lediglich drei Artikel und eine Pressemitteilung liefert (Stand: 20. November 2023).

Darüber hinaus funktioniert die thematische Filterfunktion nur bedingt und basiert unserer Einschätzung nach auf einer manuellen thematischen Zuordnung der Webseiteninhalte und Dokumente zu vordefinierten Kategorien. Diese Kategorisierung ist jedoch unvollständig. Beispielsweise ist der Artikel „Die neue Mobilfunkgeneration 5G“ vom 4. Oktober 2019 der Kategorie „Elektromagnetische Felder“ zugeordnet und lässt sich somit über den entsprechenden Filter finden. Der Artikel „Die nächste Generation im Mobilfunk: 5G“ vom 17. März 2021 ist jedoch keinem Thema zugeordnet und erscheint daher nicht, wenn die Filterfunktion genutzt wird.

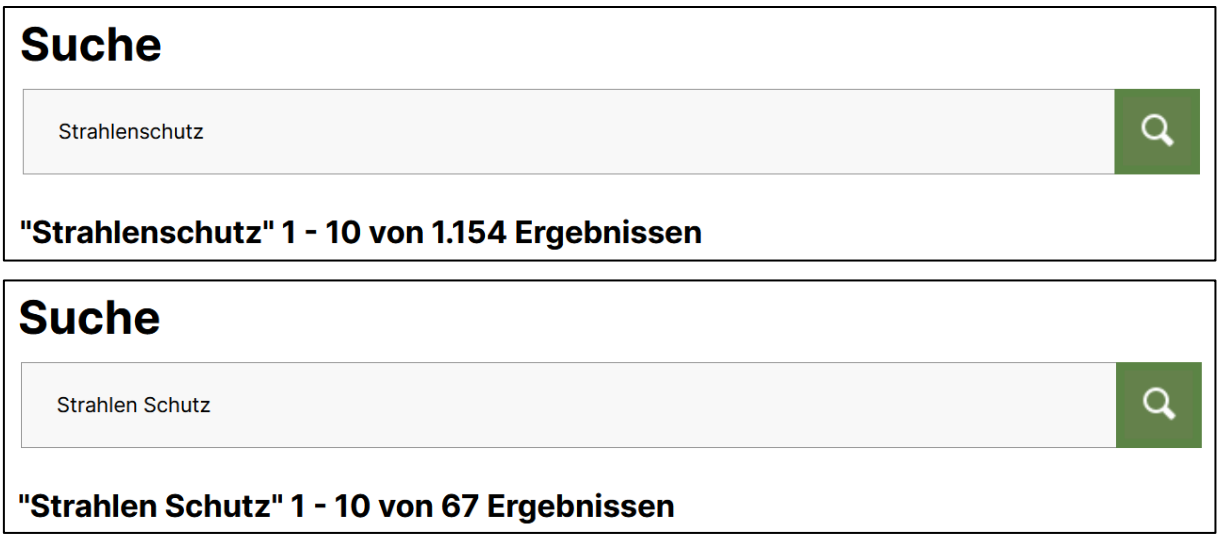

Abbildung 8: Unterschiedliche Ergebnisse bei Eingabe von „Strahlenschutz“ und „Strahlen Schutz“ in der Suchmaske der BfS-Webseite (Stand: 20. November 2023)

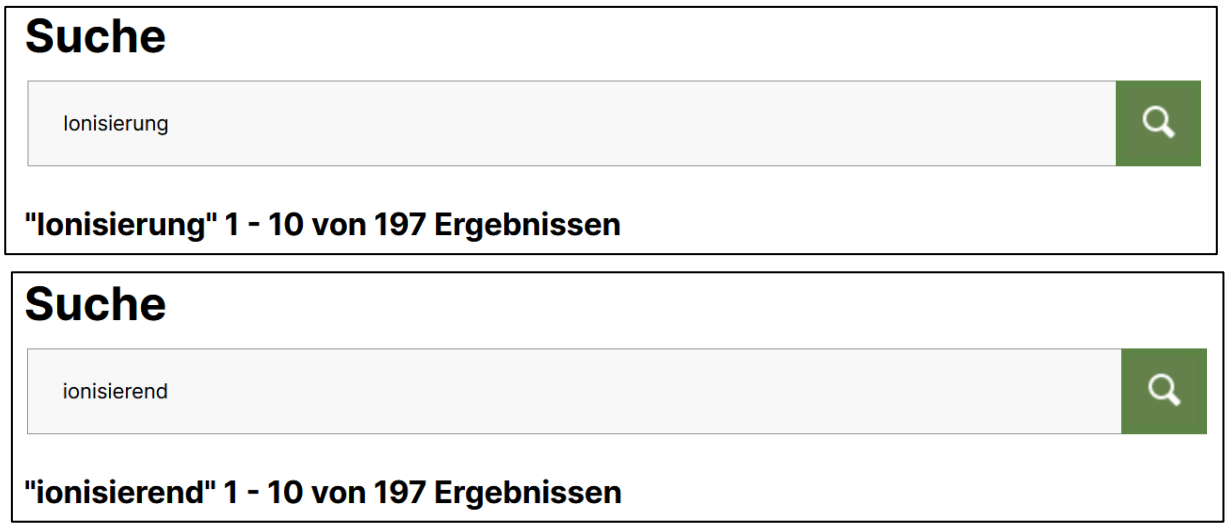

Abbildung 9: Die Suche nach „Ionisierung“ und „ionisierend“ führt zu den gleichen Suchtreffern in der Suchmaske der BfS-Webseite (Stand: 20. November 2023)

Der zunehmende Fortschritt leistungsfähiger KI-Systeme, insbesondere im Bereich des maschinellen Lernens, eröffnet vielversprechende Möglichkeiten zur Optimierung der bestehenden Suchfunktionalität der Webseite. Die Integration fortschrittlicher KI-Methoden zur Wissensabfrage und intelligenten Durchsuchung der Webseiteninhalte könnte die Schwächen der aktuellen Suchfunktion beheben und somit die Genauigkeit der Suchergebnisse verbessern. Darüber hinaus bietet die KI die Option, Inhalte automatisch thematisch zu kategorisieren. Im Folgenden präsentieren wir einen konkreten Implementierungsvorschlag.

Intelligente Suchfunktion

Eine Möglichkeit besteht darin, ein generatives Sprachmodell in die Webseite zu integrieren, um dem Nutzer der Suchfunktionalität maßgeschneiderte Antworten zu liefern. Dabei müssen jedoch potenzielle Herausforderungen berücksichtigt werden, die mit der Verwendung großer Sprachmodelle (LLMs) einhergehen. Ein Beispiel hierfür ist das sogenannte „Halluzinieren“ solcher Modelle, das dazu führen kann, dass falsche, aber überzeugende Antworten generiert werden. Dies birgt die Gefahr, dass unwissende Nutzer solche Informationen irrtümlicherweise für wahr halten.

Insbesondere bei sicherheitsrelevanten Fragen wie „Wie repariere ich meine defekte Mikrowelle?“ können fehlerhafte Antworten katastrophale Folgen haben und müssen ausgeschlossen werden können. Diesem Problem kann begegnet werden, indem dem LLM zusätzlicher Kontext (zum Beispiel der Inhalt der Webseite) zur Verfügung gestellt wird, auf dessen Basis die Antwort generiert wird. Das Modell kann zudem durch Prompting angewiesen werden, auf besonders sensible Themen keine Antwort zu generieren, zu welchem kein Wissen vorliegt.

Darüber hinaus sind Datenschutzvorgaben zu beachten, wenn ein solches Modell eingesetzt wird. Viele LLMs sind cloudbasiert und werden von ausländischen, oft amerikanischen Unternehmen betrieben, ohne Berücksichtigung der hierzulande geltenden Datenschutzvorgaben. Alternativ existieren jedoch Open-Source-Modelle, die auf eigener Serverinfrastruktur betrieben werden können – unter der Voraussetzung ausreichender Rechen- und Speicherkapazitäten.

Systemarchitektur und Pipeline

Um die intelligente Suche zu realisieren, schlagen wir vor, eine Pipeline zu nutzen, die ein generatives Sprachmodell mit einer Wissensdatenbank (Vektordatenbank) kombiniert, wie in Abbildung 10 dargestellt. Die Architektur besteht im Wesentlichen aus vier Hauptkomponenten: einem Embedding-Modell, einer Vektordatenbank, einer semantischen Suchfunktion und dem generativen Sprachmodell.

Im ersten Schritt werden die Inhalte der Webseite, einschließlich aller Artikel und Dokumente, indiziert, in Blöcke (Chunks) aufgeteilt und dem Embedding-Modell zugeführt. Dieses Modell berechnet für die Inhaltsblöcke Vektoren in einem hochdimensionalen Vektorraum. Diese Vektoren erfassen den semantischen Inhalt der Chunks und ermöglichen eine präzise Repräsentation der Informationen. Die generierten Vektoren werden in einer Vektordatenbank gespeichert. Dabei besteht die Möglichkeit, Elemente durch das Hinzufügen von Metadaten wie Links, Publikationsdatum des Inhaltes und anderen relevanten Informationen zu erweitern. Diese Vektordatenbank bildet die Grundlage für die spätere Suchfunktion.

Wenn ein Nutzer eine Suchanfrage stellt, wird diese ebenfalls durch das Embedding-Modell vektorisiert. Die semantische Suche verwendet Matching-Techniken wie zum Beispiel die Cosine Similarity, um in der Vektordatenbank nach relevanten Inhalten zu suchen. Dadurch erfolgt eine präzise und schnelle Identifikation von Informationen, die semantisch relevant zu der Anfrage des Nutzers sind. Für die Umsetzung der intelligenten Suche ist ein generatives Sprachmodell nicht zwingend notwendig, lässt sich aber dadurch nutzerfreundlicher gestalten. Zudem kann die aufgebaute Architektur für weitere Use Cases (vgl. 4.2.3) genutzt werden.

Darüber hinaus wird die Suchanfrage dem generativen Sprachmodell als Prompt übergeben. Die Ergebnisse der semantischen Suche werden dem Sprachmodell ebenfalls als Kontext präsentiert. Auf diese Weise kann das Modell die Ausgabe generieren, wobei eine umfassende und kontextsensitive Darstellung der Suchergebnisse gewährleistet wird.

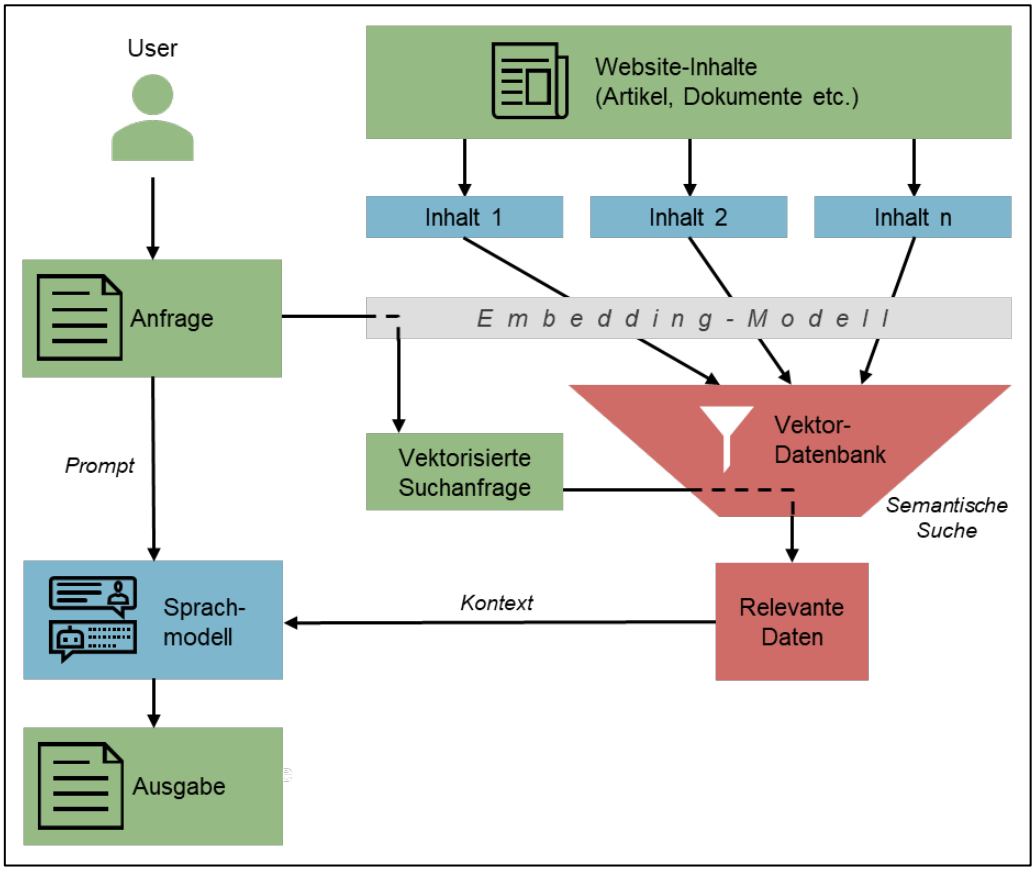

Abbildung 10: Pipeline der semantischen Suche mittels Kombination eines Embedding-Modells, einer semantischen Suche und eines Sprachmodells

Die Integration dieser Architektur schafft eine leistungsfähige und maßgeschneiderte Suchfunktion, die auf fortschrittlichen KI-Technologien basiert und die Qualität der Suchergebnisse auf der Webseite des BfS erheblich verbessern kann.

Die einzelnen Komponenten werden im Folgenden detailliert beleuchtet.

1. **Webseiteninhalte:**
   Hierbei handelt es sich um sämtliche textuelle Inhalte der Webseite, die Nutzer durchsuchen können. Diese können sowohl die HTML-Inhalte als auch andere Dokumente (zum Beispiel PDF-Dokumente) umfassen.

2. **Suchanfrage des Nutzers:**
   Der Nutzer kann über eine Suchmaske Suchanfragen an das System stellen. Diese kann entweder aus einem einzelnen Begriff (zum Beispiel „5G“, „WLAN“) bestehen oder auch eine komplexere und spezifischere Frage beinhalten (zum Beispiel „In welchem Frequenzbereich arbeitet 5G?“).

3. **Chunking:**
   „Chunking“ ist ein Schritt, der darauf abzielt, den Text in sinnvolle Blöcke oder Gruppen von Wörtern beziehungsweise Textpassagen zu unterteilen. Die Wahl der optimalen Chunk-Größe für die Inhalte und Dokumente auf der Webseite ist entscheidend, um genaue und relevante Suchergebnisse zu gewährleisten. Die Chunk-Größe definiert dabei die Menge an Kontext, die dem Embedding-Modell übergeben wird. Eine zu große Chunk-Größe, beispielsweise ein ganzer Absatz oder ein gesamtes Dokument, führt dazu, dass die vom Embedding-Modell generierte Vektorrepräsentation das allgemeine Thema des Eingabetextes erfasst, die Bedeutung einzelner Phrasen oder Sätze jedoch vernachlässigt. Dies erschwert die Generierung präziser Ergebnisse für eine Suchanfrage. Außerdem kann eine zu große Chunk-Größe dazu führen, dass für die eingehende Anfrage irrelevante Informationen als Kontext an das generative Sprachmodell übergeben werden. Eine zu kleine Chunk-Größe hingegen stellt dem Sprachmodell zu wenig

Kontextinformation zu. Sowohl eine zu große als auch eine zu klein gewählte Chunk-Größe können somit die Qualität der generierten Ausgabe negativ beeinflussen.

Die Anzahl der Tokens, die dem Sprachmodell übergeben wird, beeinflusst zudem die Performance und kann zu unnötig hohen Kosten führen, insbesondere wenn auf ein externes Modell (zum Beispiel GPT-4 von OpenAI) zurückgegriffen wird und zu viel Kontext übergeben wurde. Es muss außerdem berücksichtigt werden, dass, abhängig vom gewählten Embedding-Modell, die optimale Chunk-Größe variiert.

Im Gegensatz zum simplen Ansatz, die Eingabetexte in Chunks fester vordefinierter Länge zu unterteilen, können auch kontextsensitive Methoden angewendet werden. Diese Ansätze zielen darauf ab, die Eingabe in semantisch zusammenhängende Einheiten zu zerlegen. Zum Beispiel besteht die Möglichkeit, die Struktur und Hierarchie der Eingabe zu analysieren und das Aufteilen basierend auf dieser Analyse durchzuführen.

Eine alternative Strategie besteht darin, die Eingabe in Chunks unterschiedlicher Größen zu unterteilen. Obwohl dies zu Redundanzen führt, bietet es den Vorteil, sowohl den allgemeinen Kontext als auch detaillierte Inhalte in den Embeddings zu erfassen, was zu verbesserten Ergebnissen bei Suchanfragen führen kann.

4. **Embedding-Modell:**
   Das Embedding-Modell ermöglicht es, semantische Beziehungen zwischen Daten zu verstehen und zu lernen. Es wird dazu verwendet, Wörter oder ganze Textpassagen in hochdimensionale numerische Vektoren umzuwandeln, wobei semantisch ähnliche Elemente im Vektorraum nahe beieinander liegen. Diese Repräsentation ermöglicht der Suchfunktion semantisch relevante, anstatt syntaktisch ähnlicher, Ergebnisse zu liefern.

   Während frühe Embedding-Modelle wie Word2Vec (Mikolov et al., 2013) und GloVe (Pennington et al., 2014) lediglich die Semantik auf Wortebene erfassen konnten, sind modernere und leistungsfähigere Systeme, die auf der Transformer-Architektur basieren (Vaswani et al., 2017), dazu imstande, feine semantische Beziehungen in längeren Textsequenzen zu erlernen. Diese Modelle haben alle gemeinsam, dass sie auf einer großen Menge an Textdaten trainiert wurden und statistische Zusammenhänge in den Daten lernen. Die Semantik eines Wortes oder einer Textphrase wird somit durch den Kontext definiert, in welchem der Text in den Trainingsdaten erscheint.

   In den letzten Jahren haben sich zahlreiche vortrainierte Modelle auf Basis der Transformer-Klasse für diverse Anwendungsszenarien etabliert. Einige davon werden kommerziell vertrieben, wie zum Beispiel das ada-Modell von OpenAI oder die Embedding Modelle von Cohere. Viele andere werden hingegen kostenlos über HuggingFace zur Verfügung gestellt, wie beispielsweise BERT (Devlin et al., 2018), SBERT (Reimers & Gurevych, 2019) oder T5 (Raffel et al., 2020) basierte Embedding-Modelle. Viele dieser Modelle liefern bereits ohne Finetuning, also einer Feinabstimmung der Modellparameter für den individuellen Anwendungsfall, ausgezeichnete Ergebnisse und lassen sich leicht integrieren.

5. **Vektordatenbank:**
   Die Vektordatenbank bildet das Fundament der intelligenten Suchfunktion. Sie speichert die vom Embedding-Modell generierten Vektoren der indizierten Webseiteninhalte. Diese Vektoren ermöglichen eine effiziente Repräsentation des semantischen Inhaltes der Inhaltsblöcke. Die Vektordatenbank bietet die Möglichkeit, Metadaten wie Links (zum Beispiel Verweise auf wissenschaftliche Quellen) und andere relevante Informationen hinzuzufügen, um die Suche weiter zu verfeinern.

6. **Semantische Suche:**
   Die semantische Suche extrahiert bei einer gegebenen Suchanfrage semantisch relevante Inhalte aus der Vektordatenbank. Sobald ein Benutzer eine Suchanfrage stellt, erfolgt die Umwandlung durch dasselbe Embedding-Modell in eine Vektorrepräsentation, mit dem zuvor die Webseiteninhalte vektorisiert

wurden. Die semantische Suche nutzt dann Matching-Techniken, wie etwa die Cosine Similarity Metrik, um in der Vektordatenbank nach relevanten Inhalten zu suchen. Dieser Ansatz ermöglicht eine präzise Identifikation von Informationen, die semantisch mit der Anfrage des Benutzers korrelieren. Somit gewährleistet die semantische Suche eine effektive Filterung und Auswahl von relevanten Inhalten aus der Vektordatenbank.

7. **Generatives Sprachmodell:**
   Bei dem Sprachmodell handelt es sich um ein leistungsstarkes LLM, das in der Lage ist, komplexe natürliche Sprachanfragen zu verstehen und passende Antworten zu generieren. Die Suchanfrage des Benutzers wird dem LLM als Prompt übergeben, während die Ergebnisse der semantischen Suche als Kontextinformationen dienen. Dieser integrative Ansatz ermöglicht es, auf komplexe Fragen genaue und präzise Antworten zu liefern, die aus den verfügbaren Informationen auf der Webseite stammen. Bei der Integration eines LLMs müssen potenzielle Herausforderungen wie zum Beispiel das „Halluzinieren" berücksichtigt werden. Das Risiko hierfür kann ebenfalls durch das Bereitstellen des Webseite-Inhaltes als Kontext verringert werden, da hierdurch sichergestellt wird, dass die generierten Antworten auf validen Informationen basieren.

Weitere Aspekte

1. **Intelligente Suchfunktion im Kontext EMF:**
   Die beschriebene Suchfunktion ist grundsätzlich auf sämtliche Inhalte der Webseite anwendbar. Im Bereich der EMF wird jedoch ein besonders hoher Anspruch an die Qualität der Ergebnisse gestellt. Es muss garantiert werden können, dass diese wissenschaftlich fundiert und korrekt sind. Unpräzise oder fehlerhafte Informationen könnten nicht nur zu einem Vertrauensverlust seitens der Bürgerinnen und Bürger gegenüber dem BfS führen, sondern auch gezielte Angriffe auf das System provozieren. Solche Angriffe könnten darauf abzielen, das Sprachmodell dazu zu verleiten, falsche Antworten zu generieren, die dann als vermeintlicher Beleg für ein wissenschaftsfeindliches oder verschwörungstheoretisches Weltbild dienen könnten. Insbesondere im Kontext von EMF sind wir uns bewusst, dass dieses Themenfeld äußerst anfällig für Verschwörungstheorien ist.

   Wie bereits oben erläutert, können die identifizierten Risiken durch die Integration relevanter Kontextinformationen erheblich minimiert werden. Darüber hinaus besteht die Möglichkeit, bei jeder Suchanfrage das Prompt durch zusätzliche Informationen zu erweitern. Beispielsweise könnte die Anweisung an das Sprachmodell lauten, eine für Laien verständliche Sprache zu verwenden, auch bei emotional aufgeladenen Nutzeranfragen sachlich zu bleiben und sich ausschließlich auf die im Kontext verfügbaren Informationen zu beziehen. Darüber hinaus können generierte Antworten um Verweise auf wissenschaftliche Quellen ergänzt werden, sofern diese als Metadaten in der Datenbank verfügbar sind. Dieser Ansatz stellt sicher, dass die bereitgestellten Antworten nicht nur präzise und verständlich sind, sondern auch auf fundierten wissenschaftlichen Erkenntnissen beruhen, was die Glaubwürdigkeit der Webseite und des BfS insgesamt stärkt.

2. **Automatische Generierung von Kategorien für den Suchfilter:**
   Wie zu Beginn dargelegt, verfügt die Suchfunktion der BfS-Webseite derzeit lediglich über eine einfache Filterfunktion für die Webseiteninhalte, die auf manueller Zuordnung zu vordefinierten Kategorien, wie zum Beispiel „Elektromagnetische Felder", basiert und teilweise fehleranfällig ist. Um diesen Prozess zu optimieren, bietet sich eine Automatisierung der Kategorisierung an. Durch Einsatz von KI-Technologien können die Zuordnungen präziser, effizienter und fehlerfrei gestaltet werden, was die Gesamtgenauigkeit und Benutzerfreundlichkeit der Filterfunktion verbessert.

   Die automatische Generierung von Kategorien für den Suchfilter kann ebenfalls mithilfe eines Embedding-Modells oder eines generativen Sprachmodells realisiert werden. Im Folgenden werden beide Ansätze im Detail erläutert.

- **Embedding-Modell:**
  Für die Nutzung des Embedding-Modells zur automatischen Generierung von Kategorien erfolgt zunächst die manuelle Festlegung vordefinierter relevanter Kategorien. Diese dienen als Grundlage für die Kategorisierung der Webseiteninhalte. Anschließend werden die Embeddings dieser vordefinierten Kategorien berechnet, wodurch sie im Vektorraum repräsentiert werden.

  Für sämtliche Inhalte (insbesondere Artikel, Publikationen, Pressemitteilungen) auf der BfS-Webseite werden ebenfalls Embeddings berechnet. Die Zuordnung zu den vordefinierten Kategorien erfolgt anschließend auf Grundlage des besten Matches im Vektorraum. Bei Hinzufügen neuer Inhalte zur Webseite müssen zunächst deren Embeddings berechnet und anschließend einer passenden Kategorie zugeordnet werden. Änderungen an den vordefinierten Kategorien, sei es durch Hinzufügen, Umbenennen oder Löschen, erfordern eine Neukategorisierung aller Inhalte.

- **LLM (generatives Sprachmodell):**
  Auch ein generatives Sprachmodell kann verwendet werden, indem ihm die Webseiteninhalte als Eingabe übergeben werden. Durch geeignetes Prompting wird das LLM angeregt, die Inhalte zu kategorisieren. Hierbei kann entweder auf vordefinierte Kategorien zurückgegriffen werden, oder das Modell erstellt eigene Kategorien.

  Für die Verarbeitung neuer Inhalte genügt es, das LLM auf diese anzuwenden, um sie automatisch einer passenden Kategorie zuzuordnen oder eine neue Kategorie zu generieren. Bei der automatischen Generierung von Kategorien könnte es jedoch notwendig sein, ein Postprocessing durchzuführen, um semantisch ähnliche Kategorien zu reduzieren. Ein möglicher Ansatz hierfür wäre das Clustering der generierten Kategorien, beispielsweise mithilfe von Embeddings. Dieses Clustering könnte auf maximal zehn bis maximal zwanzig Kategorien beschränkt werden, um die Übersichtlichkeit und Benutzerfreundlichkeit zu wahren.

3. **Multimodalität:**
   Moderne leistungsfähige Modelle wie GPT-4 bieten heutzutage die Möglichkeit, sowohl Text- als auch Bildinformationen zu verarbeiten. Diese Fortschritte eröffnen die Möglichkeit, solche Modelle in eine multimodale Suche zu integrieren. Dadurch können beispielsweise zu einer Suchanfrage passende Infografiken präsentiert werden. Der grundlegende Aufbau der Suche bleibt dem textbasierten Ansatz ähnlich, jedoch erfordert die multimodale Integration, dass sowohl das Embedding- als auch das generative Sprachmodell multimodale Funktionalitäten aufweisen.

4.2.2 Automatische Übersetzung der Webseiteninhalte für Mehrsprachigkeit

Um sicherzustellen, dass die Webseite des BfS inklusiv und für Sprecher verschiedener Sprachen zugänglich ist, kann eine automatische Übersetzungsfunktion implementiert werden. Diese Funktion ermöglicht es, die Webseiteninhalte in verschiedene Sprachen zu übersetzen, um eine breitere Nutzerbasis anzusprechen.

Die automatische Übersetzung wird durch fortschrittliche maschinelle Übersetzungsmodelle realisiert, die auf KI basieren. Diese Modelle, oft auf Transformer-Architekturen wie dem Google Translate Modell oder vergleichbaren Technologien aufbauend, können präzise und kontextsensitive Übersetzungen liefern. Durch die Integration von maschinellen Übersetzungstechnologien wird die Webseite des BfS in der Lage sein, ihre Inhalte in mehreren Sprachen anzubieten, darunter Englisch, Französisch, Spanisch und anderen wichtige Sprachen.

Die mehrsprachige Webseite trägt nicht nur zur Barrierefreiheit bei, sondern fördert auch die Verbreitung von rein wissenschaftlichen Informationen über Strahlenschutz auf globaler Ebene. Dies ist besonders relevant, da der Themenkomplex des Strahlenschutzes internationale Relevanz hat. Nutzer aus verschiedenen Ländern können somit auf die Inhalte der Webseite zugreifen und von den umfassenden Informationen profitieren, die das BfS bereitstellt. Informationen im Kontext von Gesetzgebung, Regulierung, Verordnung,

Aufsicht, Überwachung und politischer Kontextualisierung, wovon der Großteil der Informationen auf der BfS-Webseite betroffen ist, sind hingegen eher für Deutschland relevant.

Die automatische Übersetzungsfunktion wird kontinuierlich optimiert, um eine hohe Qualität und Genauigkeit der Übersetzungen sicherzustellen. Dabei werden auch kulturelle Nuancen berücksichtigt, um sicherzustellen, dass die Botschaften und Informationen der Webseite angemessen und präzise in verschiedenen Sprachen vermittelt werden. Durch diese Initiative könnte das BfS seine Rolle als Informationsquelle im Bereich des Strahlenschutzes stärken.

4.2.3 Verarbeitung von Bürgeranfragen

Eine wichtige Aufgabe der Mitarbeiter des BfS besteht darin, Anfragen von Bürgern zu beantworten. Diese Anfragen stammen teilweise von Technologiekritikern, die sich mit dem Thema Strahlenschutz auseinandersetzen. Eine Vielzahl dieser Anfragen kann telefonisch beantwortet werden, jedoch erfordert ein erheblicher Anteil eine schriftliche Ausführung. Dies stellt einen beträchtlichen zeitlichen Aufwand für die Mitarbeiter dar. Insbesondere liegt ein hoher Aufwand vor, da sich viele der gestellten Fragen auf ähnliche Themen wie „Mikrowellenstrahlung“ oder „5G“ beziehen und folglich ähnliche Antworten erfordern.

Ein besonderes Augenmerk gilt den Fragen von Systemkritikern, da vermeintliche Unstimmigkeit oder jeder vermeintliche Widerspruch politisch gegen die Organisation oder den Staat instrumentalisiert werden könnte. Daher ist die Bearbeitung derartiger Anfragen eine verstärkte Aufmerksamkeit. Mit dem Ziel, mehr Aufmerksamkeit auf kritische oder komplexe Fragestellungen zu lenken, wird die Integration von Large Language Models in Betracht gezogen. Diese können dazu beitragen, wiederkehrende Fragen automatisiert zu beantworten, indem das Modell mit Fakten aus wissenschaftlichen Studien und den bisherigen Antworten gefüttert wird. Der strukturelle Aufbau vieler Antworten kann dabei von einem LLM beibehalten werden.

Um die Richtigkeit der Antworten zu gewährleisten, erfolgt eine initiale Evaluierung des Modells anhand von Probedaten (Phase 1). Bei zufriedenstellenden Ergebnissen könnten die generierten Texte den Mitarbeitern des BfS zur Verfügung gestellt werden, wobei diese die Möglichkeit haben, die Texte nach Bedarf zu modifizieren oder zu löschen (Phase 2). Bei zufriedenstellenden Modellantworten könnten die Mitarbeiter diese an den Auftraggeber weiterleiten. In einer abschließenden Phase (Phase 3) wäre sogar eine vollautomatisierte Antwort direkt an den Bürger denkbar, jedoch erfordert dies eine fortlaufende Re-Evaluierung und Prüfung.

Die Implementierung in Phase 2 könnte laut Aussagen von Interviewpartnern des BfS die Mitarbeiter um bis zu 80 Prozent entlasten und ihnen die Möglichkeit geben, sich verstärkt auf komplexere Fragestellungen zu fokussieren.

5 Technisches Detailkonzept

Im Folgenden wird genauer beschrieben, welche Use-Cases sich ergeben haben und wie diese technische umgesetzt werden können. Zuerst werden Möglichkeiten zur Übersetzung von Inhalten beschrieben. Dabei wird auf die statische- sowie auf die Echtzeitübersetzung eingegangen. Anschließend werden die technischen Aspekte beim Use-Case der automatisierten Beantwortung von Bürgeranfragen und die Verbesserungen der Suchfunktion auf der Webseite des BfS erläutert. Da sich beide Use-Cases in ihrer technischen Umsetzung überschneiden, werden zuerst die allgemeinen technischen Umsetzungsdetails beschrieben. Anschließend wird auf den jeweilig speziellen Use-Case genauer eingegangen.

5.1 Übersetzungssoftware

Um sicherzustellen, dass die Webseite des BfS inklusiv und für Sprecher verschiedener Sprachen zugänglich ist, kann eine automatische Übersetzungsfunktion implementiert werden. Diese Funktion ermöglicht es, die Webseiteninhalte in verschiedene Sprachen zu übersetzen, um eine breitere Nutzerbasis anzusprechen. Es gibt beispielsweise zwei Möglichkeiten, um die Übersetzungskomponente erfolgreich zu implementieren. Zum einen können alle Inhalte der BfS-Webseite heruntergeladen werden, übersetzt und erneut in einer anderssprachigen Version hochgeladen werden. Zum anderen können die Text in Echtzeit übersetzt werden. Beide Möglichkeiten werden im Folgenden beschrieben.

5.1.1 Statische Übersetzung

Zuerst muss der Text, der übersetzt werden soll, aus der Webseite extrahiert werden. Die Webseiteninhalte sollten dem BfS bereit vorliegen. Falls dies nicht der Fall ist, kann die Extraktion mit Web-Scraping-Tools wie BeautifulSoup4 oder Scrapy in Python erfolgen. Nachdem der Text extrahiert wurde, kann er mit einer Übersetzungsbibliothek oder API in die Zielsprache übersetzt werden. Es gibt verschiedene Übersetzungsdienste, darunter:

- **Google Cloud Translation API[4]:** Ermöglicht es, Text in Echtzeit in über 100 Sprachen zu übersetzen. Es ist eine leistungsfähige und einfach zu verwendende API, erfordert jedoch ein Google Cloud-Konto und ist kostenpflichtig.
- **Microsoft Translator Text API[5]**: Ähnlich wie die Google API ermöglicht auch Microsofts Dienst das Übersetzen von Text in verschiedene Sprachen.
- **DeepL API[6]:** Bietet hochwertige Übersetzungen durch die Verwendung von künstlicher Intelligenz und maschinellem Lernen. DeepL ist bekannt für seine Genauigkeit und natürliche Übersetzungen.
- **Open-Source-Bibliotheken[7]:** Bibliotheken wie “translate” bieten eine einfache Schnittstelle für die Übersetzung, die verschiedene Übersetzungsdienste als Backend nutzen kann.
- **Finetuning von Transformer Modellen[8]:** Es können auch eigene Transformer-Modelle nachtrainiert werden, die beispielsweise über “Huggingface” bereitgestellt werden. Durch das Nachtrainieren beziehungsweise Finetuning mittels eigener Fachtexte des BfS kann speziell bei komplexen Themen eine solide Übersetzungsleistung gewährleistet werden. Ein beliebtes multilinguales Übersetzungsmodell ist

[4] https://cloud.google.com/translate/docs/reference/rest

[5] https://www.microsoft.com/en-us/translator/business/translator-api/

[6] https://www.deepl.com

[7] https://pypi.org/project/translate/

[8] https://huggingface.co/models

"facebook/mbart-large-50-many-to-many-mmt", ein Modell, mit dem über 50 verschiedene Sprachen übersetzt werden können. Für diesen Fall würden Trainingsdaten benötigt werden mit Quell- und Zielsprache für jede Zielsprache, in die Übersetzt werden soll. Das Erstellen eines eigenen Trainigsdatensatzes kann einen erheblichen Mehraufwand mit sich bringen.

Nachdem der Text übersetzt wurde, muss er an der richtigen Stelle in die Webseite eingefügt werden. Dies kann manuell oder durch ein Skript erfolgen, das den ursprünglichen Text durch die Übersetzung ersetzt.

Zusammenfassend beinhaltet der Prozess der automatisierten Übersetzung von Webseiteninhalten die Übersetzung durch verschiedene Dienste oder APIs in die gewünschte Sprache. Abschließend wird der übersetzte Text wieder in die Webseite eingefügt. Dieser Vorgang nutzt etablierte Übersetzungsdienste sowie die Möglichkeit, spezifische Übersetzungsmodelle für verbesserte Genauigkeit anzupassen. Bei kritischen Inhalten auf der Webseite, für die das BfS bei einer fehlerhaften Übersetzung haften würde, macht es Sinn, die jeweiligen Übersetzungen nochmal manuell zu prüfen.

5.1.2 Echtzeitübersetzung

Für die Echtzeitübersetzung von Webseiteninhalten gibt es mehrere Ansätze, die auf der Integration von Übersetzungs-APIs und dynamischen Webtechnologien basieren. Eine solche Lösung erfordert im Allgemeinen die Kombination von Client- und Serverseitentechnologien, um eine nahtlose und sofortige Übersetzung zu ermöglichen, ohne die Seite neu laden zu müssen.

Zuerst wird eine Übersetzungs-API ausgewählt, die Echtzeitübersetzungen unterstützt. Beliebte Optionen sind die bereits genannten Google Cloud Translation API, Microsoft Translator Text API und DeepL API. Diese APIs bieten robuste Unterstützung für eine Vielzahl von Sprachen und werden häufig aufgrund ihrer hohen Qualität und Skalierbarkeit gewählt.

Auf der Clientseite (im Browser des Benutzers) verwendet man JavaScript, um die Interaktion mit der Übersetzungs-API zu steuern. Moderne Frontend-Frameworks wie React, Vue.js oder Angular können verwendet werden, um eine dynamische Benutzeroberfläche zu erstellen, die die Echtzeitübersetzung unterstützt. Ein Beispielprozess könnte wie folgt aussehen:

- **Sprachauswahl:** Eine Dropdown-Liste oder ein anderes Auswahlwerkzeug wird angeboten, damit der Benutzer die Zielsprache auswählen kann.
- **Erfassen von Textänderungen:** Event-Listener werden verwendet, um Änderungen im DOM oder spezifische Benutzeraktionen zu erfassen, die eine Übersetzung erfordern könnten.
- **API-Anfragen:** Der zu übersetzende Text wird mittels AJAX/Fetch-Anfragen an die ausgewählte Übersetzungs-API gesendet. Dies kann direkt von der Clientseite aus erfolgen oder über einen Server-Proxy, um API-Schlüssel zu schützen und die Sicherheit zu erhöhen.
- **Aktualisierung der Benutzeroberfläche:** Die Übersetzung von der API wird empfangen und die Benutzeroberfläche entsprechend aktualisiert, indem der ursprüngliche Text durch die Übersetzung ersetzt wird.

Obwohl viele Übersetzungsanfragen direkt von der Clientseite aus erfolgen können, ist es oft sinnvoll, serverseitige Unterstützung für die Verwaltung von API-Anfragen und -Antworten zu implementieren. Dies kann helfen, die Sicherheit (z.B. durch Verbergen von API-Schlüsseln) zu erhöhen und die Last auf dem Client zu verringern.

Beispielimplementierung:

- **Proxy-Endpoint:** Ein Proxy-Endpoint wird auf dem Server erstellt, der Anfragen von der Clientseite entgegennimmt, sie an die Übersetzungs-API weiterleitet und dann die Antwort an den Client zurückgibt.
- **Caching:** Caching-Strategien werden auf dem Server implementiert, um häufig angeforderte Übersetzungen zu speichern und so die Antwortzeiten zu verbessern und die API-Kosten zu reduzieren.

Die Implementierung einer Echtzeitübersetzungslösung für Webseiteninhalte erfordert eine sorgfältige Planung und die Auswahl geeigneter Technologien und Dienste. Durch die Kombination von client- und serverseitigen Technologien kann eine nahtlose, schnelle und benutzerfreundliche Übersetzungserfahrung geschaffen werden, die es Benutzern ermöglicht, Inhalte in ihrer bevorzugten Sprache zu konsumieren, ohne die Seite neu laden zu müssen.

5.1.3 Zusammenfassung

Zusammenfassend eignen sich beide vorgestellte Übersetzungsmethoden, um die Inhalte auf der Webseite des BfS multilingualer zu gestalten. Jeder Ansatz kommt mit Vor- und Nachteilen. Bei der statischen Übersetzung liegt der Vorteil darin, dass kritische Texte bereits im Vorhinein überprüft werden können. Jedoch ist die statische Methode unflexibler, da alle Inhalte bei einer neuen Version der Übersetzungssoftware neu übersetzt werden müssen. Die Echtzeitübersetzung gestaltet sich als flexibler, da sich neue Softwareversionen und Sprachen einfach integrieren lassen. Jedoch bringt dieser Ansatz eine erhöhte Latenz mit, da die Übersetzungen dynamisch zur Laufzeit berechnet werden und es gibt keine Möglichkeit, Übersetzungsfehler zu korrigieren.

5.2 Suchfunktion & Bürgeranfragen: Technische Aspekte

Die Optimierung der Suchfunktionalität und die Verarbeitung von Bürgeranfragen weisen eine beträchtliche technische Schnittmenge auf. Dies ermöglicht eine effiziente Synergie beider KI-basierten Anwendungsfälle, indem einheitliche Algorithmen und Verarbeitungsmodelle zum Einsatz kommen. Die Konvergenz in den technischen Anforderungen und Methodiken resultiert in einer Ressourceneffizienz, bei der mit einer einzigen, ausgeklügelten Lösung multiple operative Aufgaben bewältigt werden. Diese strategische Bündelung fördert nicht nur die Kosteneffizienz, sondern trägt auch zur Konsistenz und Skalierbarkeit der KI-Systeme bei.

Auf die detaillierten technischen Punkte wird im Folgenden eingegangen. Anschließend wird für beide KI-Anwendungsfälle genauer spezifiziert, wie die Methode in der Praxis implementiert werden kann.

5.2.1 LLM-Finetuning

Eine Möglichkeit, um beide KI-Anwendungsfälle zu lösen ist das Finetuning eines Sprachmodells mit Daten des BfS. Finetuning eines Sprachmodells ist ein Prozess, bei dem ein vortrainiertes Sprachmodell weiter trainiert wird, um es auf eine spezifische Aufgabe oder einen spezifischen Datensatz anzupassen. Während des Finetunings wird das Modell mit einem kleineren, spezifischen Datensatz trainiert, der auf die jeweilige Aufgabe oder den Anwendungsbereich zugeschnitten ist. Ziel ist es, die Leistung des Modells zu verbessern, indem es die Besonderheiten und Nuancen des speziellen Datensatzes lernt. Dies ermöglicht es dem Modell, präzisere Vorhersagen oder Generierungen für die gewünschte Aufgabe zu machen, wie z.B. Textklassifikation, Frage-Antwort-Systeme, Textgenerierung in einem bestimmten Stil oder Kontextverständnis in einem Fachgebiet. Finetuning ist eine effektive Methode, um die Nützlichkeit und Anwendbarkeit von generellen Sprachmodellen auf spezialisierte Aufgaben und Branchen zu erweitern.

LoRA (Low Rank Adaptation) ergänzt diesen Prozess, indem es eine Technik darstellt, die es ermöglicht, nur einen kleinen Teil der Parameter eines vortrainierten Modells anzupassen, anstatt das gesamte Modell neu zu trainieren. Bei LoRA werden Änderungen an den Gewichten des Modells durch die Anwendung niedrigdimensionaler Anpassungen vorgenommen, was zu einer effizienteren und ressourcenschonenderen Anpassung führt. Dies reduziert den Aufwand für das Training und die Speicheranforderungen erheblich, ohne dabei signifikant an Modellleistung einzubüßen.

Insgesamt bleibt der Finetuning-Prozess, sowie die Erstellung eines hochqualitativen Datensatzes jedoch ressourcenaufwändig. Zudem bringt es eine gewisse Inflexibilität mit sich, da bei neuen, oder sich ändernden Dokumenten in der Datenbank das Modell erneut auf die Daten nachtrainiert werden muss.

5.2.2 Retrieval Augmented Generation

Ein flexiblerer und ressourcenschonender Ansatz für beide KI-Anwendungsfälle nennt sich "Retrieval Augmented Generation" (RAG). Die Retrieval Augmented Generation (RAG) ist ein fortschrittlicher Ansatz in der künstlichen Intelligenz, insbesondere im Bereich des Natural Language Processing (NLP), der die Vorteile von Information Retrieval (IR) und generativen Sprachmodellen kombiniert, um Antworten zu generieren, die sowohl genau als auch inhaltlich reichhaltig sind. Im RAG-Prozess ist es ohne erheblichen Aufwand möglich, neue Dokumente als Datengrundlage hinzuzufügen. Folgende Schritte werden bei der RAG-Methode vollzogen (siehe Abbildung 10):

1. Datenvorbereitung und Indexierung
2. Transformation in Embeddings
3. Aufbau der Vektordatenbank
4. Nutzeranfrage und Vektorisierung
5. Retrieval von Informationen
6. Antwortgenerierung
7. Refinement und Feedback-Integration
8. Kontinuierliches Lernen

Datenvorbereitung und Indexierung

Der erste Schritt in der RAG-Pipeline ist die Sammlung und Vorbereitung von Daten, die als Grundlage für das Information Retrieval dienen. Inhalte werden aus verschiedenen Quellen aggregiert. Das kann beispielsweise Webseiten, wissenschaftliche Artikel, Datenbanken oder auch interne Dokumente umfassen. Diese werden dann indiziert, d.h., es wird ein systematisches Verzeichnis aller Schlüsselinformationen erstellt, was schnellen Zugriff und effiziente Suchfunktionalität ermöglicht.

Transformation in Embeddings

Die indizierten Daten werden anschließend durch ein Embedding-Modell verarbeitet, das Text in hochdimensionale Vektoren umwandelt. Dieses Modell, häufig ein neuronales Netzwerk, erfasst die Bedeutung und den Kontext von Wörtern oder Textabschnitten, indem es sie in Punkte innerhalb eines Vektorraums transformiert, wobei ähnliche Bedeutungen nahe beieinander liegen.

Aufbau der Vektordatenbank

Die generierten Embeddings werden in einer Vektordatenbank gespeichert, die oft mit zusätzlichen Metadaten angereichert wird, um eine detailliertere und kontextualisierte Suche zu ermöglichen. Die Vektordatenbank ermöglicht es, Abfragen in Echtzeit durchzuführen und die relevantesten Inhalte schnell zu identifizieren.

Nutzeranfrage und Vektorisierung

Wenn ein Nutzer eine Anfrage stellt, wird diese Anfrage auch in einen Vektor umgewandelt, indem das gleiche Embedding-Modell verwendet wird. Dies ermöglicht es dem System, die Anfrage mit den in der Vektordatenbank gespeicherten Informationen zu vergleichen.

Retrieval von Informationen

Im nächsten Schritt verwendet die RAG-Methode Algorithmen wie die Cosine Similarity, um in der Vektordatenbank nach den Embeddings zu suchen, die der Vektorrepräsentation der Nutzeranfrage am ähnlichsten sind. Diese ausgewählten Informationen werden dann als Kontext für das generative Modell bereitgestellt. Vereinfacht kann man sich dies mit der Sortierung in einem Supermarkt vorstellen. Sucht man beispielsweise nach Softdrinks, wird man in einem Regal ähnliche Getränke finden, wie eine Coca-Cola oder Pepsi. Diese befinden sich an einer ähnlichen Stelle im Supermarkt, da sie ähnliche Eigenschaften aufweisen. Übertragen auf das Retrieval von Informationen im Vektorraum bedeutet dies, dass Absätze, also Chunks, die eine ähnliche Bedeutung haben, sich im selben Raum befinden. Wie viele relevante Chunks zurückgegeben werden kann programmatisch eingestellt werden sowie ein "Threshold" um nur Dokumente zurückgegeben werden, die mindestens n Ähnlichkeit zur Query haben.

Antwortgenerierung

Das generative Sprachmodell, oft ein Transformer-basiertes Modell, empfängt die Anfrage zusammen mit den aus der Vektordatenbank abgerufenen Kontextinformationen. Das Modell generiert daraufhin eine Antwort, die auf der originalen Anfrage und dem zusätzlichen Kontext basiert. Ein möglicher Prompt vor und während der Generierung ist in den folgenden beiden Abbildungen zu sehen. Die erste Abbildung 11 zeigt den Prompt mit den Platzhaltern {context} und {question}. In dieser Prompt-Vorlage wird das Modell darauf angewiesen, nur die gefundenen Kontextinformationen zu verwenden, um die Frage des Nutzers zu beantworten. Sätze wie "Erfinde keine Antwort!" gelten als effektives Mittel, um Halluzinationen reduzieren zu können. Die Frage oder Aufgabe des Nutzers wird in den Frage-Platzhalter eingespeist. Daraufhin wird die Frage ebenfalls von dem Embeddingmodell vektorisiert und die ähnlichsten Dokumente zur Frage werden zurückgegeben. Die zurückgegebenen Dokumente werden daraufhin in den Kontext-Platzhalter eingefügt. Ein weiterer Vorteil dieser Methode ist, dass die gefundenen Informationen in der Antwort des LLMs als Quelle verlinkt werden können, sodass ein Nutzer die Informationen auf der Webseite des BfS selbst überprüfen kann. Dies stellt eine transparente Kommunikation sicher.

```
prompt = """
Du bist ein hilfsbereiter Assistent.
Verwende die folgenden Kontext Informationen, um die Frage des Nutzers zu beantworten.
Wenn du die Antwort nicht kennst, sag einfach, dass du es nicht weißt.
Erfinde keine Antwort!!

Hier die Kontext Informationen:
### KONTEXT ###
{kontext}

Hier die Frage des Nutzers:
### FRAGE ###
{frage}

### Antwort ###
"""
```

Abbildung 11: Initiales Prompt

Ein Praxisbeispiel findet sich in Abbildung 12. Hier stellt der Nutzer die Frage, wie das BfS die 5G Technologie einschätzt. Zur Simulation der Suche wurde ein Beitrag zu 5G auf der Webseite des BfS als Kontext in das Systemprompt des LLMs eingespeist. Das LLM, in diesem Fall ein open-source Modell namens Mixtral 8x7B, antwortet daraufhin auf die Frage anhand der Informationen aus dem Kontext.

```
prompt = """
Du bist ein hilfsbereiter Assistent.
Verwende die folgenden Kontext Informationen, um die Frage des Nutzers zu beantworten.
Wenn du die Antwort nicht kennst, sag einfach, dass du es nicht weißt.
Erfinde keine Antwort!!

Hier die Kontext Informationen:
### KONTEXT ###
Seit 2020 wird die nächste Mobilfunkgeneration 5G eingeführt. Selbstfahrende Autos,
sprachgesteuerte Assistenten und intelligente Kühlschränke sind nur einige Beispiele dafür,
was die höheren Datenübertragungsraten der neuen Mobilfunktechnologie unterstützen werden könnten.
Es gibt jedoch auch Bedenken. Dazu gehört insbesondere die Frage, ob der 5G-Ausbau auch
ein gesundheitliches Risiko nach sich zieht. Das Bundesamt für Strahlenschutz (BfS) geht
nach derzeitigem wissenschaftlichen Kenntnisstand nicht von negativen gesundheitlichen
Auswirkungen aus, sieht aber auch noch offene Fragen. Neben den Grundlagen zur 5. Mobilfunkgeneration
werden in diesem Artikel, in dem Keine ausführliche Bewertung dieser Technologie aus Strahlenschutzsicht
und in unserer Schriftenreihe der Strahlenschutzstandpunkt.

Hier die Frage des Nutzers:
### FRAGE ###
Wie schätzt das BfS 5G ein?

### Antwort ###
Das Bundesamt für Strahlenschutz (BfS) geht nach derzeitigem wissenschaftlichen
Kenntnisstand nicht von negativen gesundheitlichen Auswirkungen aus, sieht aber auch noch
offene Fragen.
"""
```

Abbildung 12: Prompt mit eingefügten Inhalten

Refinement und Feedback-Integration

Die generierte Antwort kann einem optionalen Verfeinerungsprozess unterzogen werden, bei dem menschliche Prüfer oder zusätzliche Algorithmen die Qualität sicherstellen und die Antwort weiter verbessern können. Nutzerfeedback wird ebenfalls gesammelt und kann verwendet werden, um das System kontinuierlich zu optimieren.

Kontinuierliches Lernen

In einem iterativen Prozess wird das System kontinuierlich mit neuen Daten gefüttert, und die Modelle werden feinjustiert, um die Genauigkeit und Relevanz der Antworten zu verbessern. Dieses kontinuierliche Lernen ist entscheidend, um die Leistung der RAG-Methode aufrechtzuerhalten und zu steigern.

Die RAG-Methode bietet somit eine dynamische und kontextbewusste Antworterstellung, die nicht nur die direkte Anfrage adressiert, sondern den Informationsgehalt durch Hinzunahme und Synthese relevanter Wissensfragmente erweitert. Sie stellt einen signifikanten Fortschritt gegenüber herkömmlichen Chatbots oder Suchmaschinen dar, indem sie Nutzern Antworten liefert, die sowohl genauer als auch informativer sind.

5.3 Technische Umsetzung

Um die oben beschriebenen Punkte in der Praxis umzusetzen, werden im Folgenden Methoden und Frameworks beschrieben, die angewendet werden können, um am Ende automatisiert Bürgeranfragen beantworten zu können.

5.3.1 Vektordatenbank erstellen

Um eine Vektordatenbank zu erstellen, müssen die Dokumente zuerst in Embeddings umgewandelt werden, um sie im Vektorraum darzustellen. Es gibt eine Vielzahl an Embedding-Modellen, welche sich von Use-Case zu Use-Case unterscheiden. Die beliebtesten Embedding-Modelle für die deutsche Sprache sind (Stand Januar 2024):

- Intfloat Multilingual E5: ein multilinguales, open-source state-of-the-art Embedding Modell
- T-Systems RoBERTa: ein Englisch-Deutsches Embedding Modell

Sobald ein Embedding-Modell ausgewählt wurde, kann eine Vektordatenbank erstellt werden. Dafür gibt es ebenfalls eine Vielzahl an Frameworks, wie zum Beispiel:

- ChromaDB[9]
- Langchain FAISS[10]
- Pinecone[11]
- Weaviate[12]
- PGVector[13]

Sobald ein Vektordatenbank-Framework ausgewählt wurde, können sowohl die bereits beantwortenden Bürgeranfragen als auch Inhalte der Webseite des BfS in den Vektorraum eingespeist werden. Da viele Sprachmodelle mit ihrer Kontextlänge begrenzt sind, müssen die Daten vorher mittels Chunking aufgeteilt werden.

Die Anzahl der Tokens, die dem Sprachmodell übergeben wird, beeinflusst zudem die Performance und kann zu unnötig hohen Kosten führen, insbesondere wenn auf ein externes Modell (zum Beispiel GPT-4 von OpenAI) zurückgegriffen wird und zu viel Kontext übergeben wurde. Es muss außerdem berücksichtigt werden, dass, abhängig vom gewählten Embedding-Modell, die optimale Chunk-Größe variiert.

Im Gegensatz zum simplen Ansatz, die Eingabetexte in Chunks fester vordefinierter Länge zu unterteilen, können auch kontextsensitive Methoden angewendet werden. Diese Ansätze zielen darauf ab, die Eingabe in semantisch zusammenhängende Einheiten zu zerlegen. Zum Beispiel besteht die Möglichkeit, die Struktur und Hierarchie der Eingabe zu analysieren und das Aufteilen basierend auf dieser Analyse durchzuführen.

Sind diese Schritte erledigt wurde die Wissensdatenbank, die als Grundstütze des Projekts gilt, erfolgreich erstellt. Als nächster wichtiger Punkt muss ein geeignetes Sprachmodell ausgewählt werden.

5.3.2 Auswahl des Sprachmodells

Bei dem Sprachmodell handelt es sich um ein leistungsstarkes LLM, das in der Lage ist, komplexe natürliche Sprachanfragen zu verstehen und passende Antworten zu generieren. Die Suchanfrage des Benutzers wird dem LLM als Prompt übergeben, während die Ergebnisse der semantischen Suche als Kontextinformationen dienen. Dieser integrative Ansatz ermöglicht es, auf komplexe Fragen genaue und präzise Antworten zu liefern, die aus den verfügbaren Informationen auf der Webseite stammen. Bei der Integration eines LLMs müssen potenzielle Herausforderungen wie zum Beispiel das „Halluzinieren“ berücksichtigt werden. Das Risiko hierfür kann ebenfalls durch das Bereitstellen des Webseite-Inhalts als Kontext verringert werden, da hierdurch sichergestellt wird, dass die generierten Antworten auf validen Informationen basieren.

Bei der Auswahl des Sprachmodells stehen zwei Optionen zur Verfügung. Zum einen kann auf das derzeit beste Sprachmodell von OpenAI (GPT-4, GPT-3.5) zurückgegriffen werden. Die Modelle von OpenAI überzeugen durch ihre Qualität und sind verhältnismäßig kostengünstig in der Nutzung. Kosten entstehen abhängig von der Anzahl Anfragen bzw. wie viel Text in der Anfrage an das Modell geschickt wird und wie viel Text als

[9] https://www.trychroma.com/

[10] https://python.langchain.com/docs/integrations/vectorstores/faiss

[11] https://www.pinecone.io/

[12] https://weaviate.io/

[13] https://github.com/pgvector/pgvector

Antwort generiert wird. Ein nicht unbedeutender Nachteil ist jedoch, dass alle Daten an die Server von OpenAI übersendet werden, was eine DSGVO konforme Nutzung schwierig gestaltet.

Als Alternative existieren verschiedene Open-Source Modelle, die ebenfalls durch ihre Qualität überzeugen und kommerziell genutzt werden können. Es fallen jedoch Kosten für das Hosten auf Servern an, da diese Modelle rechenintensiv sind. Dies bietet jedoch den großen Vorteil, dass die verarbeiteten Daten vollständig auf den eigenen Servern verbleiben. Beliebte Open-Source Modelle für die deutsche Sprache sind:

- Modelle von MistralAI (Mistral und Mixtral 8x7B)[14]
- Modelle von Meta (Llama 2)[15]

Mistral und Llama 2 Modelle, welche auf die deutsche Sprache nachtrainiert wurden:

- Modelle von LeoLM (Leo-Hessian-AI)[16]
- Modelle von EM German (EM-German-Mistral)[17]
- Modelle von Vago Solutions (SauerkrautLM-Mixtral-8x7B)[18]

Es gibt kein universales Modell, dass sich für jeden Use-Case am besten eignet. Vielmehr muss evaluiert werden, mit welchem Modell die Ergebnisse am zufriedenstellen. Verschiedenste Faktoren beeinflussen die Ausgabequalität eines Modells. Diese sind zum einen modellspezifische Eigenschaften, die nicht vom Nutzer beeinflusst werden. Hierunter fallen z.B. die Modellarchitektur, die Größe des Modells (Anzahl Parameter), die Trainingsalgorithmen sowie die Trainingsdaten und deren Qualität. Einen großen Einfluss auf die Qualität der generierten Antwort hat jedoch auch die gewählte Prompt-Strategie, die vom Nutzer definiert werden kann. Durch das Prompt werden dem Modell Anweisungen gegeben, wie es antworten soll und welche Informationen es für die Antwort verwendet.

Dennoch ist das Modell Mixtral 8x7B hervorzuheben. Mistral AI hat Mixtral 8x7B eingeführt, ein äußerst effizientes Modell für spärliche Mischungen von Experten (MoE) mit offenen Gewichten, lizenziert unter Apache 2.0. Dieses Modell zeichnet sich durch seine schnelle Inferenz aus und ist sechsmal schneller als Llama 2 70B. Es überzeugt durch ausgezeichnete Kosten-/Leistungsverhältnisse und konkurriert oder übertrifft GPT-3.5 in den meisten Standard-Benchmarks, was es zu einem führenden Modell mit offenen Gewichten und einer großzügigen Lizenz macht.[19]

Mixtral verwendet eine ähnliche Architektur wie Mistral 7B und verfügt über eine robuste Fähigkeit, einen Kontext von 32k Tokens zu verarbeiten. Es unterstützt Englisch, Französisch, Italienisch, Deutsch und Spanisch und zeigt eine starke Leistung bei der Codegenerierung. Mit insgesamt 46,7B Parametern arbeitet Mixtral mit der Effizienz und den Kosten eines 12,9B-Modells. Allerdings erfordert es eine hohe VRAM für den Betrieb, wodurch es für Setups mit erheblichen VRAM-Ressourcen besser geeignet ist.

Neben dem Basismodell hat Mistral AI Mixtral 8x7B Instruct veröffentlicht. Diese Variante wurde speziell für präzises Anweisungsverfolgen feinabgestimmt und erreicht eine Qualität der Ausgabe die vergleichbar ist mit GPT-3.5 (Stand: Januar 2024). Mixtral kann auch dazu aufgefordert werden, bestimmte Ausgaben für Anwendungen mit hohem Moderationsniveau einzuschränken. Ohne spezifische Aufforderungen folgt es jedoch den gegebenen Anweisungen.

---

[14] https://mistral.ai/

[15] https://llama.meta.com/

[16] https://laion.ai/blog/leo-lm/

[17] https://github.com/jphme/EM_German

[18] https://vago-solutions.ai/technologie/

[19] https://mistral.ai/news/mixtral-of-experts/

5.3.3 Hardwareanforderungen

Die Hardwareanforderungen hängen maßgeblich von der Größe des Modells, also der Anzahl der Parameter ab. Mixtral hat beispielsweise 45 Milliarden Parameter. Die Parameter werden standardmäßig in half-precision, also als 16 Bit Gleitkommazahlen gespeichert, woraus sich eine Speicheranforderung von mindestens 90 GB errechnen lässt.

Üblicherweise werden große Sprachmodelle aus Performanzgründen auf GPUs ausgeführt und müssen daher in den VRAM geladen werden. Prinzipiell lässt sich jedes Sprachmodell auch im Arbeitsspeicher halten und man kann die Berechnungen mittels der CPU durchführen, was jedoch die Berechnungszeit maßgeblich negativ beeinflusst und für den praktischen Einsatz nicht empfehlenswert ist. Um einen Kompromiss zwischen Performanz und Hardwarekosten zu finden, ist es auch möglich, nur einen Teil des Modells in eine kleinere Grafikkarte zu laden und den Rest im RAM zu halten.

Es besteht jedoch auch die Möglichkeit eine sogenannte Post-Training Quantization durchzuführen, also eine Quantifizierung der Modellparameter zur Laufzeit. Dies bedeutet, dass die Parameter beispielsweise in 8 oder 4 Bit komprimiert werden, wodurch sich die Speicheranforderung im Vergleich zur half-precision halbiert bzw. viertelt. Üblicherweise führt dies bei 8 Bit zu einem kaum messbaren negativen Einfluss auf die Ausgabequalität, bei 4 Bit lediglich zu einem geringen. Erst die Reduzierung auf 3 oder weniger Bit führt, laut aktuellen Studien, zu einem deutlichen Qualitätsverlust des Modells.[20] Mittels 4 Bit Quantifizierung lässt sich das Mixtral Modell beispielsweise mit zwei GeForce RTX 3090 (je 16 GB) oder einer V100-32GB Grafikarte effizient betreiben. Da die Speicheranforderung jedoch in Abhängigkeit von der Länge der dem Modell übergebenen Eingabe steigt (in etwa 1GB/1000 Tokens), ist es empfehlenswert, mindestens zwei RTX 4090 (je 24 GB) oder eine A100-40GB Hardwarebeschleuniger Karte zum Betreiben des Modells zu nutzen, um auch längere Abgaben verarbeiten zu können. Darüber hinaus sollten dem System mindestens 64 GB Hauptspeicher zur Verfügung stehen.

Schließlich sei darauf hingewiesen, dass es sich hierbei um Empfehlungen handelt und die tatsächliche Leistung von verschiedenen Faktoren abhängig ist, darunter die spezifische Aufgabe, die Implementierung des Modells und andere Systemprozesse. Tabelle 7 zeigt mögliche Kosten, die für verschiedene Modelle bei der Anschaffung der geeigneten Hardware anfallen könnten (Stand März 2024, Durchschnittswerte).

Preis	Hardware	Modell
Bis ca. 4.000 Euro	- 2x GeForce RTX 4090 (je 24GB) - 1x V100 32GB	- Mixtral 8x7B 4 Bit - Llama 2 13B
Bis ca. 40.000 Euro	- 3x A100 (je 40GB)	- Mixtral 8x7B

Tabelle 7: Hardware Kostenüberblick, Stand März 2024

[20] Dettmers, T., & Zettlemoyer, L. (2023, July). The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning* (pp. 7750-7774). PMLR.

5.3.4 Zusammenführung von Vektordatenbank und Sprachmodell

Um eine performante Anwendung zu erstellen, muss die Vektordatenbank und das Sprachmodell sinnvoll zusammengeführt werden. Dafür gibt es bereits open-source Frameworks, die den Prozess vereinfachen:

- LlamaIndex[21]
- Langchain[22]

Beide Frameworks sind spezialisiert um Retrieval Augmented Generation (RAG) umzusetzen. RAG ist die Technologie, die verwendet wird, um Bürgeranfragen automatisiert beantworten zu können. Mithilfe von Verarbeitungsketten (Chains) werden die Schritte aus Abbildung 10 nach und nach ausgeführt. Dabei stellt sich die Herausforderung, wie viele Dokumente bei der Vektorsuche dem Sprachmodell übergeben werden. Zum einen sollen ausreichend Informationen zur Verfügung gestellt werden, zum anderen können zu viele Suchergebnisse im Prompt die Kontextlänge des Modells überschreiten, was zu einer verschlechterten Performanz führen könnte. Hier gibt es ebenfalls keinen Wert, der allgemein am besten funktioniert – es muss evaluiert werden, wie viele gefundene Suchergebnisse in den Prompt eingespeist werden. Ein gängiger Startwert sind vier Suchergebnisse.

Um die Qualität der Antworten zu verbessern, stellen beide Frameworks ebenfalls fortgeschrittene RAG-Methoden zur Verfügung:

- **Parent Document Retriever:** Dokument wird in größere Chunks aufgeteilt (Parent Chunks). Diese größeren Chunks werden wiederum in kleinere Chunks aufgeteilt (Child Chunks). Nun wird die Suche auf die Child Chunks angewendet, und falls die relevante Information sich im Child Chunk befindet, wird der Parent Chunk als Ergebnis zurückgegeben, um dem Modell mehr Kontext zu geben
- **Ensemble Retriever:** Kombination mehrerer Retriever, z.B. Cosine Similarity Suche kombiniert mit einer Maximal Marginal Relevance (MMR) Suche
- **Contextual Compressor:** gefundenen Kontext auf die relevanten Informationen komprimieren.
- **Finetuning von Embeddings:** Erstellung von eigenen Embeddings, um Dokumente besser finden zu können.

5.3.5 Evaluierung

Ein sehr wichtiger Punkt ist es, das Modell zu evaluieren, bevor es zum Einsatz kommt. Dadurch kann die Qualität der Antworten eingesehen werden und eingeschätzt werden, ob das Modell verlässlich eingesetzt werden kann. Für die Evaluation gibt es verschiedene Ansätze.

Zum einen können Antworten auf häufig gestellte Bürgeranfragen manuell überprüft werden. Zum anderen gibt es open-source Frameworks, die die Qualität der Antworten automatisiert evaluieren (z.B. RAGAS Framework). Hier sind die Schlüsselmetriken, die für die Evaluierung von RAG-Systemen verwendet werden können:

Retriever:

- context_precision: Misst, wie relevant der abgerufene Kontext in Bezug auf die gestellte Frage ist. Dies gibt Aufschluss über die Qualität der Retrieval-Pipeline.
- context_recall: Misst die Fähigkeit des Retrievers, alle notwendigen Informationen abzurufen, die zur Beantwortung der Frage benötigt werden.

[21] https://www.llamaindex.ai/

[22] https://www.langchain.com/

Generator (LLM):

- faithfulness: Misst die faktische Konsistenz der Antwort im Verhältnis zum Kontext basierend auf der gestellten Frage, insbesondere hinsichtlich der Vermeidung von Halluzinationen.
- answer_relevancy: Misst, wie zielgerichtet und relevant die Antwort im Verhältnis zur gestellten Frage ist.

Das harmonische Mittel dieser vier Aspekte ergibt den RAGAS-Score, der als einzelnes Maß für die Leistung des QA-Systems über alle wichtigen Aspekte hinweg dient.

Diese Metriken ermöglichen es, das RAG-System umfassend zu evaluieren, indem sowohl die Qualität des Retrieval-Prozesses als auch die Güte, der vom Sprachmodell generierten Antworten beurteilt werden. Als Basis wird hierfür ein Testdatensatz benötigt. Dafür können die bereits vorhandenen Antworten auf Bürgeranfragen verwendet werden, denn diese enthalten bereits eine Frage und eine richtige Antwort auf die Frage. Anschließend lässt man das Modell die gleiche Frage beantworten und vergleicht die richtige sowie die generierte Antwort auf ihre Richtigkeit und Ähnlichkeit. Daraus ergeben sich Scores, die es einfach machen, das beste Sprachmodell, die beste RAG-Methode und die beste Vektordatenbank festzulegen.

6 Auswahl von Use-Cases

Bei der Entwicklung von Technologielösungen für die behördliche Kommunikation, insbesondere unter Einsatz von KI, steht man oft vor der Herausforderung, aus einer Vielzahl möglicher Use-Cases denjenigen auszuwählen, der den größten Mehrwert bietet und technisch realisierbar ist. Diese Auswahl ist entscheidend für den Erfolg des Prototyps und die spätere Implementierung in realen Anwendungsszenarien. Im Folgenden soll erläutert werden, wie dieser Prozess der Auswahl eines einzelnen Use-Cases aus mehreren Möglichkeiten gestaltet werden kann, um eine effektive und zielgerichtete prototypische Entwicklung im Bereich der KI-gestützten behördlichen Kommunikation zu gewährleisten.

Die initiale Aufstellung von neun Use-Cases zur Verwendung von KI wurde im weiteren Projektverlauf bewertet. Die Bewertung und Priorisierung dieser Use-Cases basierte auf verschiedenen Kriterien. Zu diesen Kriterien gehören u.a. der erwartete Nutzen und die Auswirkungen auf die Stakeholder, die technische Machbarkeit, die Verfügbarkeit von Daten zur Unterstützung der KI-Modelle, die Kosten für die Entwicklung und Implementierung sowie die Übereinstimmung mit den strategischen Zielen der Behörde. Damit wurde eine ganzheitliche Sicht auf die Potenziale und Herausforderungen jedes Use-Cases möglich. Passende Use-Cases wurden weiterverfolgt und ggf. konsolidiert, sodass zum Schluss drei Use-Cases für verschiedene Umsetzungszeitspannen, kurz-, mittel- und langfristig, übrig-blieben.

6.1 Automatische Übersetzung

Einer Verwendung automatisierter Übersetzung im BfS steht nur der Risikoappetit der Organisation entgegen. Die Vorteile der automatisierten Übersetzung sind die schnelle Verfügbarkeit von Inhalten für eine sprachlich vielfältige Zielgruppe. Nachteilig ist, dass die dadurch generierten Inhalte von der Güte der verwendeten Lösung abhängen und möglicherweise validiert werden müssen, da es sich um offizielle Kommunikation einer Bundesbehörde handelt. Hierzu bräuchte man Dolmetscher mit entsprechender Fachkenntnis, um die Übersetzungen auf ihre Korrektheit zu prüfen. Dies würden die Effekte der KI-Nutzung negieren. Daher ist der Nutzen direkt von der Akzeptanz von Risiken abhängig, welche entstehen, wenn Inhalte durch eine KI übersetzt und so direkt verfügbar gemacht werden, ohne viel Aufwand für deren Validierung zu investieren.

Da das Thema elektromagnetische Felder in manchen gesellschaftlichen Gruppen eine gewisse Brisanz hat, besteht ein Risiko, dass fachliche Fehler unabsichtlich durch Ungenauigkeiten in der automatischen Übersetzung publiziert werden.

Der Aufwand für die Umsetzung inkl. manuelle Qualitätssicherung steht in einem zu hohen Verhältnis zum Nutzen. Die automatische Übersetzung eignet sich daher sicher für weniger risikoreiche Inhalte, jedoch, nach interner Rücksprache im KEMF, nicht für Inhalte zu elektromagnetischen Feldern.

6.2 Verbesserung der Suchfunktion der BfS-Webseite

Die intelligente Suchfunktion, speziell ausgerichtet auf den Bereich der elektromagnetischen Felder, erfordert eine Präzision und wissenschaftliche Fundierung der Suchergebnisse, die über herkömmliche Standards hinausgehen. Dies ist essenziell, um nicht nur das Vertrauen der Nutzer in das BfS zu wahren, sondern auch um mögliche gezielte Desinformationskampagnen abzuwehren, die darauf abzielen könnten, durch die Generierung falscher Antworten, die Glaubwürdigkeit der Institution zu untergraben. In Anbetracht der Sensibilität des Themas EMF und seiner Anfälligkeit für Verschwörungstheorien ist eine besonders sorgfältige Gestaltung der Suchfunktion vonnöten.

Durch die Integration relevanter Kontextinformationen und die Erweiterung der Suchanfragen um spezifische Anweisungen kann das Sprachmodell dazu angehalten werden, ausschließlich auf Basis der verfügbaren Informationen zu antworten. Es kann zudem angeleitet werden, auch bei emotional aufgeladenen Anfragen eine sachliche und für Laien verständliche Sprache zu wahren. Die Möglichkeit, Antworten mit Verweisen auf

wissenschaftliche Quellen zu versehen, sofern diese als Metadaten vorhanden sind, stärkt die Glaubwürdigkeit und wissenschaftliche Stichhaltigkeit der generierten Inhalte.

Zur Verbesserung der Suchfilterfunktion auf der BfS-Webseite bietet sich die Automatisierung der Kategorisierung mithilfe von KI-Technologien an, um die Genauigkeit und Benutzerfreundlichkeit zu steigern. Mit einem Embedding-Modell könnten vordefinierte Kategorien im Vektorraum repräsentiert und sämtliche Webseiteninhalte diesen zugeordnet werden, basierend auf dem besten Match im Vektorraum. Änderungen an den Kategorien würden eine Neukategorisierung erfordern.

Ein alternativer Ansatz unter Verwendung eines generativen Sprachmodells würde es ermöglichen, Inhalte zu kategorisieren oder neue Kategorien zu erstellen, basierend auf dem Eingangstext der Webseite. Dies könnte durch ein geeignetes Prompting unterstützt werden. Die automatische Generierung von Kategorien erfordert jedoch eventuell ein Clustering, um die Übersichtlichkeit zu gewährleisten und redundante Kategorien zu vermeiden.

Die Einbeziehung von Multimodalität stellt einen weiteren Innovationsschritt dar. Modelle wie GPT-4, die Text- und Bildinformationen verarbeiten können, ermöglichen es, Suchanfragen, um passende Infografiken oder Bildmaterial zu ergänzen. Für eine erfolgreiche multimodale Integration müssen sowohl das Embedding- als auch das Sprachmodell über die entsprechenden Fähigkeiten verfügen, um eine nahtlose Verarbeitung und Darstellung multimodaler Inhalte zu gewährleisten.

Der Use-Case „verbesserte Suchfunktion“ lässt sich kurzfristig umsetzen mit einem mittleren zu erwartenden Nutzen für die Webseitenbenutzer. Die zu schaffenden technischen Voraussetzungen sind allerdings etwas aufwändiger, Kosten und Nutzen müssen daher genauestens abgewogen werden. Da jedoch der Use-Case „Automatisierte Verarbeitung von Bürgeranfragen“ auf derselben technologischen Grundlage aufbaut, können langfristig Synergieeffekte genutzt werden. Der letztgenannte Use-Case bewirkt eine deutliche Entlastung der Mitarbeiter im BfS und führt zu einer schnelleren Beantwortung von Bürgeranfragen. Daher wird die automatisierte Verarbeitung der Bürgeranfragen priorisiert. Die verbesserte Suchfunktion kann nachgelagert umgesetzt werden.

6.3 Automatische Verarbeitung von Bürgeranfragen

Ein weiterer KI-Anwendungsfall, den wir priorisiert für die Implementierung und Umsetzung vorschlagen, ist die automatisierte Verarbeitung von Bürgeranfragen. Eine wichtige Aufgabe der Mitarbeiter des BfS besteht darin, Anfragen von Bürgern zu beantworten. Diese Anfragen stammen teilweise von Technologiekritikern, die sich mit dem Thema Strahlenschutz auseinandersetzen. Eine Vielzahl dieser Anfragen kann telefonisch beantwortet werden, jedoch erfordert ein erheblicher Anteil eine schriftliche Ausführung. Dies stellt einen beträchtlichen zeitlichen Aufwand für die Mitarbeiter dar. Insbesondere liegt ein hoher Aufwand vor, da sich viele der gestellten Fragen auf ähnliche Themen wie „Mikrowellenstrahlung“ oder „5G“ beziehen und folglich ähnliche Antworten erfordern.

Ein besonderes Augenmerk gilt den Fragen von Systemkritikern, da jede wahrgenommene Unstimmigkeit oder jeder wahrgenommene Widerspruch politisch gegen die Organisation oder den Staat instrumentalisiert werden könnte. Daher erfordern derartige Anfragen eine verstärkte Aufmerksamkeit. Mit dem Ziel, mehr Aufmerksamkeit auf kritische oder komplexe Fragestellungen zu lenken, wird die Integration von Large Language Models und speziell die Integration der RAG-Methode in Betracht gezogen. Diese können dazu beitragen, wiederkehrende Fragen automatisiert zu beantworten, indem das Modell von einem Vektorretriever mit Fakten aus wissenschaftlichen Studien und den bisherigen Antworten kontextualisiert wird.

Zum Beispiel könnte die RAG-Methode angewendet werden, um relevante Informationen zu finden, und auf Basis der gefundenen Inhalte eine Antwort zu generieren Als Beispiel könnten bisher gestellte Fragen und E-Mail-Antworten in eine Vektordatenbank indexiert werden. Zusätzlich können Webseiteninhalte oder Publikationen aus dem DORIS-System hinzugefügt werden, um die Wissensbasis zu erweitern. Stellt nun ein Nutzer eine Frage, die in der Vergangenheit bereits gestellt und beantwortet wurde, kann die jeweilig richtige Antwort als Kontext in das Modell-Prompt gegeben werden und eine neue Antwort generiert werden, die

sich auf die Bedürfnisse des Nutzers anpasst. Durch die Angabe im Prompt, dass ausschließlich die gefundenen Inhalte zur Beantwortung der Frage benutzt werden sollen, können Halluzinationen reduziert und die Richtigkeit der Antwort sichergestellt werden.

Ein weiterer wichtiger Punkt ist, dass die Antworten stets zielgruppen- und bedarfsorientiert angefertigt werden müssen. Um dies zu gewährleisten, kann das LLM, das für die Beantwortung der Bürgeranfragen verwendet wird, weiter verfeinert (Finetuning) werden. Durch Finetuning kann dem Sprachmodell ein gewisser Antwortstil beigebracht werden. Hierfür wird das Sprachmodell mit bereits beantworteten Bürgeranfragen gefüttert und ist so in der Lage konsistent auf die jeweilige Zielgruppe einzugehen, wie es bisher Mitarbeiter des BfS getan haben.

Um die Qualität der Antworten evaluieren zu können, wird ein “Roll-out" in 3 Phasen empfohlen:

- **Phase 1:** Initiale Evaluierung des Modells anhand von Probedaten
- **Phase 2:** Bei zufriedenstellenden Ergebnissen in Phase 1 könnten die generierten Texte den Mitarbeitern des BfS zur Verfügung gestellt werden, wobei diese die Möglichkeit haben, die Texte nach Bedarf zu modifizieren oder zu löschen (Phase 2). Bei zufriedenstellenden Modellantworten könnten die Mitarbeiter diese an den Auftraggeber weiterleiten
- **Phase 3:** In einer abschließenden Phase wäre eine vollautomatisierte Antwort direkt an den Bürger denkbar, jedoch erfordert dies eine fortlaufende Re-Evaluierung und Prüfung

Soll-Prozess Bürgeranfrage

Folgende Abbildung stellt den Prozess ab Phase 2 dar. Die Anfrage des Bürgers wird zunächst mit dem Embedding-Modell umgewandelt. Anschließend werden in der Vektordatenbank die ähnlichsten Embeddings zur Anfrage gesucht und zurückgegeben. Falls es zur Anfrage keine relevanten Dokumente gibt, wird die Anfrage wie gehabt zu einem Mitarbeiter des BfS weitergeleitet und dieser beantwortet die Anfrage manuell. Falls es aber schon mal eine ähnliche Anfrage gab, oder falls in der Vektordatenbank andere relevante Dokumente gefunden werden (z.B. indexierte Paper), werden die gefundenen Inhalte an das LLM übergeben. Dieses fertigt dann, basierend auf den gefundenen Inhalten, eine Antwort an. Anschließend wird die LLM-Antwort, samt Quellenangaben zur besseren Transparenz, an einen Mitarbeiter gesendet. Dieser kann nun prüfen, ob die Antwort richtig und zufriedenstellen ist oder nicht. Hier hat der Mitarbeiter die Möglichkeit, die Antwort bei Bedarf zu modifizieren. Zu guter Letzt wird die generierte oder korrigierte Antwort an den Nutzer zurückgesendet.

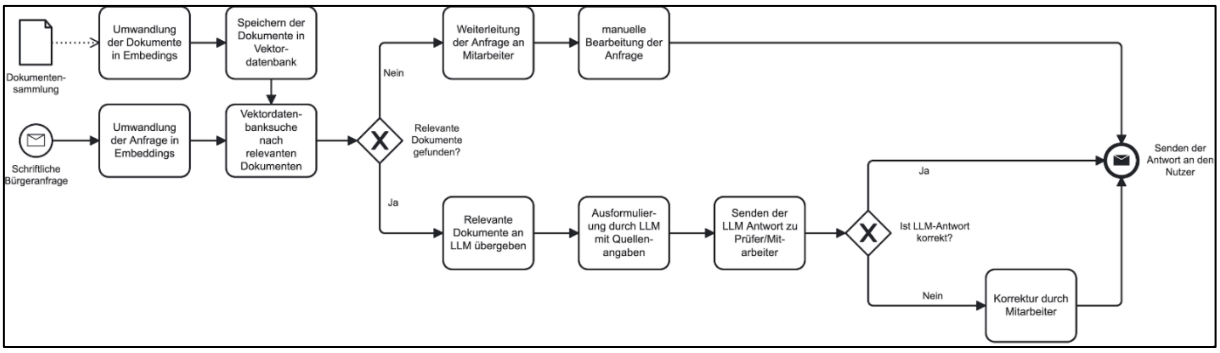

Abbildung 13: Soll-Prozess Bürgeranfrage

Da, wie bereits beschrieben, die Umsetzung der automatisierten Verarbeitung von Bürgeranfragen intern wie extern großes Potential birgt, sollte man sich zuerst auf diesen Use-Case konzentrieren. Die dadurch geschaffene technische Infrastruktur kann für weitere Projekte, wie de verbesserte Suchfunktion genutzt werden.

Neben der reinen Entwicklung des Sprachmodells und der Vektordatenbank wird für eine Verwendung in der Presse und Öffentlichkeitsarbeit eine grafische Benutzeroberfläche benötigt. Die Benutzeroberfläche soll den

Status der Anfrage anzeigen und den BfS-Mitarbeiter in die weitere Bearbeitung einbinden, sollte dies notwendig sein. Hierzu sollte eine Art Workflow-System genutzt werden. Mittels Schnittstellenaufrufen kommuniziert das KI-System mit der Oberfläche und umgekehrt und bindet ebenfalls BfS-spezifische Anwendungen wie E-Mail-Eingang und Vorgangsbearbeitungssystem im Hintergrund ein.

7 Handlungsempfehlungen

Im Folgenden werden Handlungsempfehlungen gegeben, auf welche Implementierung sich das BfS spezialisieren sollte, um die Webseitenerfahrung sowie interne Prozesse, wie die Beantwortung von Bürgeranfragen, zu verbessern und effizienter zu gestalten. Im technischen Detailkonzept wurde bereits auf verschiedene Implementationsmöglichkeiten eingegangen. In den Handlungsempfehlungen werden diese Implementationsmöglichkeiten aufgegriffen erörtert, welche sich am besten eignen. Um unsere Empfehlungen zu rechtfertigen, gehen wir dabei auf State-of-the-art Prozesse in der Wissenschaft sowie auf konkrete Praxiserfahrungen ein.

7.1 Organisatorische Handlungsempfehlungen

Organisatorische Handlungsempfehlungen zielen auf nichttechnische Maßnahmen ab, welche die Umsetzung von KI-Projekten nachhaltig ermöglichen bzw. diese fördern und in einen gesamtstrategischen Kontext einbetten. Unserer Erfahrung nach haben sich die folgenden organisatorischen Maßnahmen als wirkungsvoll erwiesen, die Nutzung von KI in Organisationen zu etablieren.

7.1.1 Einrichtung eines KI-Steuerungsteams

Für die erfolgreiche Integration von Künstlicher Intelligenz in die behördliche Arbeit ist die Bildung eines interdisziplinären KI-Steuerungsteams essenziell. Dieses Team setzt sich aus Experten verschiedener Fachbereiche zusammen, einschließlich IT, Kommunikation, Recht und Datenschutz, um eine ganzheitliche Betrachtung der KI-Anwendungen zu gewährleisten. Die Hauptaufgabe des Teams liegt in der Koordination und Überwachung aller KI-bezogenen Projekte, von der Konzeption über die Implementierung bis hin zur fortlaufenden Optimierung. Durch regelmäßige Schulungen bleibt das Team auf dem neuesten Stand der Technik und kann proaktiv auf Herausforderungen reagieren. Die Einrichtung eines solchen Teams schafft eine zentrale Anlaufstelle für alle KI-relevanten Fragen und fördert eine konsistente Strategie im Umgang mit KI-Technologien. Dies ermöglicht es der Behörde, innovativ und effizient zu arbeiten, Risiken zu minimieren und den Nutzen von KI voll auszuschöpfen.

Wichtig ist hierbei zu beachten, dass keine Parallelstrukturen im BfS aber auch im Ressortbereich entstehen. Das BfS könnte KI-Steuerungsteams für die Kompetenzzentren einrichten, welche aber unter sich gut vernetzt und im regelmäßigen Austausch stehen. Somit können domänenspezifische Anwendungen verfolgt, aber auch übergeordnete Ziele erreicht und Synergieeffekte genutzt werden.

7.1.2 Schulung und Weiterbildung

Um den erfolgreichen Einsatz von KI im BfS zu gewährleisten, ist eine fundierte Schulung der Mitarbeiter unerlässlich. Durch gezielte Weiterbildungsmaßnahmen erlangen sie ein tiefgreifendes Verständnis für die Funktionsweise, Möglichkeiten und ethischen Aspekte der KI. Besonders wichtig ist, dass die Mitarbeiter lernen, wie KI-gestützte Systeme die Kommunikation und den Informationsaustausch mit der Öffentlichkeit verbessern können. Schulungen sollten sowohl grundlegende Einführungen in die KI umfassen als auch spezifische Anwendungen in der behördlichen Praxis. Der Fokus liegt dabei auf der Vermittlung von Kompetenzen, die es den Mitarbeitern ermöglichen, KI-Tools effektiv zu nutzen, zu verwalten und weiterzuentwickeln. Durch regelmäßige Fortbildungen bleiben die Mitarbeiter zudem über aktuelle Entwicklungen und Best Practices informiert. Eine gut ausgebildete Belegschaft ist der Schlüssel zu einer innovativen und zukunftsfähigen Verwaltung, die KI verantwortungsvoll und zum Wohle der Gesellschaft einsetzt.

7.1.3 Verbesserung der Datenqualität und -verfügbarkeit

Die Qualität und Verfügbarkeit von Daten spielen eine entscheidende Rolle für den erfolgreichen Einsatz von KI in der Behördenkommunikation, aber auch bspw. im Wissensmanagement. Eine solide Datenbasis ermöglicht es KI-Systemen, präzise und relevante Informationen zu generieren, die den Bürgern zur Verfügung gestellt werden können. Um dies zu erreichen, müssen vorhandene Datenbestände systematisch erfasst, aufbereitet und für die Verwendung durch KI-Anwendungen optimiert werden. Dies schließt die Standardisierung von Datenformaten, die Verbesserung von Datenstrukturen und die Sicherstellung der Datenaktualität mit ein. Zudem ist eine enge Zusammenarbeit zwischen den Fachabteilungen erforderlich, um eine durchgängige Datenqualität zu gewährleisten. Die Implementierung von Richtlinien für das Datenmanagement und regelmäßige Qualitätskontrollen sind weitere wichtige Schritte. Durch die Verbesserung der Datenqualität und -verfügbarkeit können KI-Anwendungen effizienter gestaltet und die Informationsbedürfnisse der Öffentlichkeit besser bedient werden.

7.1.4 Implementierung eines Feedback-Systems

Ein effektives Feedback-System ist unerlässlich, um den Erfolg von KI-gestützten Kommunikationsangeboten zu messen und kontinuierlich zu verbessern. Durch das Sammeln und Analysieren von Rückmeldungen der Nutzer kann das BfS verstehen, wie ihre KI-Anwendungen wahrgenommen werden und wo Verbesserungspotenzial besteht. Ein solches System ermöglicht es nicht nur, die Benutzerfreundlichkeit und Relevanz der angebotenen Informationen zu erhöhen, sondern fördert auch das Vertrauen und die Akzeptanz bei den Bürgern.

Feedback kann über verschiedene Kanäle eingeholt werden, beispielsweise durch Online-Umfragen oder Kommentarfunktionen. Die gewonnenen Erkenntnisse sollten systematisch ausgewertet und in die Weiterentwicklung der KI-Strategie einbezogen werden. Ein proaktiver Umgang mit Nutzerfeedback trägt dazu bei, die Qualität und Effektivität der KI-gestützten Kommunikation kontinuierlich zu steigern.

7.1.5 Datenschutz und Informationssicherheit

Beim Einsatz von KI beim BfS muss ein besonderes Augenmerk auf Datenschutz und Datensicherheit gelegt werden. KI-Systeme verarbeiten, je nach Anwendungsfall, häufig große Mengen sensibler Daten, wie bspw. die Kontaktdaten von Anfragenden, was sie zu einem kritischen Punkt in Bezug auf Datenschutz und Sicherheitsrisiken macht. Um diese Risiken zu minimieren, müssen strenge Datenschutzrichtlinien implementiert und konsequent eingehalten werden. Dies beinhaltet die Einrichtung sicherer Datenverarbeitungsprozesse, die regelmäßige Überprüfung von Sicherheitsprotokollen und die Schulung von Mitarbeitern in Datenschutzbestimmungen. Zudem sollte die Transparenz gegenüber den Bürgern erhöht werden, indem klar kommuniziert wird, wie ihre Daten verwendet werden. Durch die Einhaltung höchster Datenschutz- und Sicherheitsstandards können Behörden das Vertrauen der Öffentlichkeit in den Einsatz von KI stärken und gleichzeitig die Integrität und Vertraulichkeit der verarbeiteten Informationen gewährleisten.

Bei der Integration von KI in die Arbeitsabläufe des BfS ist es essenziell, dass KI-gestützte Prozesse nahtlos in das bestehende Informationssicherheitsmanagementsystem (ISMS) eingebettet werden. Dies erfordert eine sorgfältige Planung und Implementierung verschiedener Maßnahmen.

Zunächst ist eine umfassende Risikoanalyse (z.B. nach BSI-Standard 200-3) spezifisch für den Einsatz von KI vonnöten. Diese Analyse dient dazu, potenzielle Risiken für die Informationssicherheit zu identifizieren und zu bewerten. Insbesondere die einzigartigen Schwachstellen von KI-Systemen, wie die Anfälligkeit für manipulative Angriffe, müssen erkannt und in das ISMS integriert werden. Auf der Basis dieser Risikoanalyse sind spezifische Sicherheitsrichtlinien für KI-Anwendungen zu entwickeln, die den gesamten Lebenszyklus der KI-Systeme umfassen – von der Entwicklung über die Implementierung bis zum Betrieb.

Für den Entwicklungsprozess der KI ist es unabdingbar, Sicherheitsaspekte von Beginn an zu berücksichtigen (Security by Design). Dies beinhaltet die Auswahl sicherer Entwicklungswerkzeuge, die Anwendung von Verschlüsselungsverfahren und eine regelmäßige Überprüfung des Codes auf Sicherheitslücken.

Des Weiteren sind strikte Zugriffskontrollen und ein effektives Berechtigungsmanagement für KI-Systeme zu implementieren. Nur autorisiertes Personal sollte Zugang zu den Systemen und den verarbeiteten Daten haben. Die Protokollierung und Überwachung aller relevanten Aktivitäten rund um die KI-Systeme ist unerlässlich, um ungewöhnliche Verhaltensmuster oder Sicherheitsvorfälle frühzeitig erkennen zu können.

Für den Fall eines Sicherheitsvorfalles müssen spezifische Reaktionspläne für KI-Systeme vorhanden sein. Diese sollten detaillierte Schritte zur Eindämmung des Vorfalles, zur Untersuchung der Ursachen und zur Wiederherstellung der Systemintegrität enthalten. Zudem ist eine kontinuierliche Verbesserung des ISMS hinsichtlich KI-Sicherheitsmaßnahmen unverzichtbar. Regelmäßige Sicherheitsaudits und die Anpassung der Sicherheitsstrategie an neue technologische Entwicklungen und Bedrohungsszenarien sollten als fester Bestandteil in das ISMS aufgenommen werden.

7.1.6 Entwicklung von Richtlinien

Für den verantwortungsvollen Umgang mit KI ist die Entwicklung interner Richtlinien notwendig. Diese Richtlinien sollten ethische Prinzipien, rechtliche Rahmenbedingungen und fachliche Standards für den Einsatz von KI umfassen. Sie dienen als Leitfaden für Entwickler, Anwender und Entscheidungsträger, um die Integrität und Transparenz der KI-Nutzung sicherzustellen. Die Richtlinien müssen regelmäßig überprüft und an neue Erkenntnisse und gesellschaftliche Anforderungen angepasst werden. Durch die Einbindung von Stakeholdern in den Entwicklungsprozess können vielfältige Perspektiven berücksichtigt und das Vertrauen in KI-basierte Systeme gestärkt werden. Die klar definierten Rahmenbedingungen fördern nicht nur die ethische Verantwortung und Rechtskonformität, sondern auch die Innovationskraft und Wettbewerbsfähigkeit der Organisation.

7.1.7 Transparente Kommunikation

Eine offene und transparente Kommunikation über den Einsatz von KI schafft eine Vertrauensbasis zwischen Nutzern und Anbietern. Sie dient dazu, über Funktionsweisen, Ziele und den Umgang mit Nutzerdaten aufzuklären. Transparente Kommunikation muss auch die Grenzen und Herausforderungen von KI-Technologien ansprechen, um realistische Erwartungen zu setzen und potenzielle Missverständnisse zu vermeiden. Durch regelmäßige Berichterstattung über Fortschritte, Anpassungen und die Erfolge von KI-Projekten kann eine positive Wahrnehmung in der Öffentlichkeit gefördert werden. Dies erfordert klare, verständliche Informationen, die auch für Laien nachvollziehbar sind. Zudem sollten Nutzer die Möglichkeit haben, Rückfragen zu stellen und Feedback zu geben, um einen dialogorientierten Austausch zu fördern. Eine solche Kommunikationsstrategie hilft nicht nur, Ängste und Vorbehalte abzubauen, sondern stärkt auch das Verantwortungsbewusstsein der Organisationen im Umgang mit KI-Technologien.

7.1.8 Evaluation und kontinuierliche Verbesserung

Die systematische Evaluation von KI-Anwendungen ist entscheidend, um deren Effektivität und Nutzerakzeptanz zu gewährleisten. Durch regelmäßige Überprüfungen können Schwachstellen identifiziert, Anpassungen vorgenommen und die Leistung der Systeme kontinuierlich verbessert werden. Dies umfasst nicht nur technische Aspekte, sondern auch die Nutzererfahrung und den Beitrag der KI zur Erreichung organisatorischer Ziele. Evaluationsprozesse sollten dabei verschiedene Perspektiven berücksichtigen, einschließlich Feedback von Endnutzern, Leistungsdaten der KI-Systeme und Vergleiche mit alternativen Lösungen. Die gewonnenen Erkenntnisse dienen als Grundlage für strategische Entscheidungen und die Weiterentwicklung der KI-Strategie. Eine Kultur der kontinuierlichen Verbesserung, die Innovationen fördert und gleichzeitig Risiken minimiert, ist für den langfristigen Erfolg von KI-Projekten unerlässlich.

7.1.9 Förderung der Kollaboration

Kollaboration spielt eine entscheidende Rolle beim Einsatz von KI, indem sie den Austausch von Wissen, Erfahrungen und Ressourcen zwischen verschiedenen Akteuren ermöglicht. Das BfS sollten aktiv den Dialog und die Zusammenarbeit mit anderen Behörden, Forschungseinrichtungen, der Industrie und der Zivilgesellschaft suchen. Durch gemeinsame Projekte, Arbeitsgruppen und Netzwerke können Synergien geschaffen und innovative Lösungen schneller entwickelt werden. Eine offene Kollaborationskultur trägt dazu bei, Standards zu setzen, Best Practices zu teilen und gemeinsame Herausforderungen effektiver zu bewältigen. Die Förderung von Partnerschaften kann auch dazu beitragen, die Entwicklung ethischer Richtlinien für den Einsatz von KI voranzutreiben und einen breiten gesellschaftlichen Konsens über die Nutzung dieser Technologien zu erreichen.

Als Initiator hierfür eignet sich das unter 7.1.1 beschriebene KI-Steuerungsteam. Welches auch bspw. den regelmäßigen Austausch mit dem KI-Lab im Umweltbundesamt pflegt.

7.2 Technische Handlungsempfehlungen

Primär sollte das BfS die Use-Cases zur Verbesserung der Suchfunktion sowie zur automatisierten Beantwortung von Bürgeranfragen implementieren. Da sich beide Anwendungsfälle in ihrer technischen Ausführung überschneiden, ermöglicht eine einzige Implementierung das Erreichen von zwei Verbesserungen. Dies führt zu effizienten Synergien zwischen den KI-basierten Anwendungsfällen durch den Einsatz einheitlicher Algorithmen und Verarbeitungsmodelle.

Die Übereinstimmung in den technischen Anforderungen und Methoden führt zu einer Ressourceneffizienz, bei der mit einer durchdachten Lösung mehrere operative Aufgaben bewältigt werden. Diese strategische Konzentration fördert nicht nur die Kosteneffizienz, sondern trägt auch zur Konsistenz und Skalierbarkeit der KI-Systeme bei.

Das bedeutet jedoch nicht, dass der Use-Case der Übersetzungssoftware generell vernachlässigt werden sollte. Vielmehr ermöglicht der Einsatz eines Large Language Modells (LLM) beim priorisierten Use-Case als zusätzlichen Vorteil die Übersetzung von Inhalten für Nutzer auf der Webseite in andere Sprachen.

Als technische Grundlage für beide Problemstellungen sollte das BfS den „Retrieval Augmented Generation"-Ansatz verfolgen. Eine detaillierte Beschreibung des Prozesses findet sich in der Detailformulierung. Hier eine kurze Wiederholung der wichtigsten erforderlichen Schritte für den erfolgreichen Einsatz dieser Technologie:

1. Datenvorbereitung
2. Transformation in Embeddings
3. Aufbau der Vektordatenbank
4. Retrieval von Informationen
5. Antwortgenerierung
6. Evaluation

Im weiteren Verlauf werden diese Schritte nochmals betrachtet und klare Empfehlungen zur technischen Umsetzung für die Verbesserung der Suchfunktion und für die automatisierte Beantwortung von Bürgeranfragen ausgesprochen.

7.2.1 Datenvorbereitung

In der ersten Phase des RAG-Prozesses werden die abzurufenden Daten gesammelt und aufbereitet. Da die intelligente Suchfunktion sich primär auf den Bereich der elektromagnetischen Felder (EMF) konzentrieren soll, ist eine wissenschaftliche Fundierung der Suchergebnisse essenziell. Es wird empfohlen, relevante wissenschaftliche Publikationen zu sammeln und der Datenbank hinzuzufügen. Zusätzlich ist es ratsam sämtliche Inhalte der BfS-Webseite zu extrahieren und diese Informationen ebenfalls in die Datenbank zu integrieren. Für beide Anwendungsfälle sollte jeweils eine dedizierte Datenbasis erstellt werden.

Im Kontext der automatisierten Beantwortung von Bürgeranfragen wird zudem die Integration bereits beantworteter Anfragen in die Datenbasis vorgeschlagen. Vor der Transformation der Dokumente in numerische Vektoren ist eine sinnvolle Segmentierung notwendig. Dieser als „Chunking" bekannte Prozess ist aufgrund der Kontextlängenbegrenzung des Embedding-Modells und des LLMs unabdingbar. Die direkte Eingabe kompletter Dokumente würde schnell an diese Grenzen stoßen, weshalb eine Aufteilung in kleinere Einheiten erfolgen muss. Es wird die Verwendung des Open-Source-Frameworks „Unstructured.io" nahegelegt, welches sich insbesondere für die Zerlegung von PDF-Dokumenten und allgemein unstrukturierten Daten eignet. Dabei kommt häufig die OCR-Technologie zum Einsatz, um Texte, Bilder oder Tabellen effektiv aus PDF-Dateien zu extrahieren. Nach Abschluss des Chunking-Prozesses für jedes Dokument kann dann mit dem nächsten Schritt fortgefahren werden.

7.2.2 Transformation in Embeddings

Um eine effiziente Auffindbarkeit der segmentierten Dokumente im Vektorraum zu gewährleisten, ist es erforderlich, die einzelnen Segmente, die sogenannten „Chunks", numerisch im Vektorraum darzustellen. An dieser Stelle werden Embedding-Modelle eingesetzt. In der ausführlichen Beschreibung wurden bereits zwei Modelle präsentiert, die sich besonders für die deutsche Sprache eignen. Basierend auf der Präzision, die in vorherigen Projekten bei PwC erzielt wurde, sollte die Wahl des BfS auf das „Intfloat-Multilingual-E5-Instruct-Modell"[23] fallen. Die Genauigkeit dieses Embedding-Modells ist von signifikanter Bedeutung für die gesamte Anwendung, da es die Grundlage für die Suchfunktionalität im Vektorraum bildet.

In einem weiteren Schritt werden auch die Anfragen der Nutzer – sei es eine Suchanfrage auf der Webseite oder eine Bürgeranfrage – mit demselben Embedding-Modell in numerische Vektoren umgewandelt.

7.2.3 Aufbau der Vektordatenbank

Die vorher erzeugten Embeddings müssen in einer Datenbank gespeichert werden, die es ermöglicht, relevante Informationen schnell und verlässlich finden zu können. In der Detailbeschreibung wurden bereits verschiedene Vektordatenbanken-Frameworks vorgestellt. Dem BfS wird klar empfohlen, eine Open-Source Lösung zu verwenden, da diese kostenlos sind und in ihrer Performanz ähnlich wie Closed-Source Dienste abschneiden. In diesem Kontext wird die ChromaDB-Library[24] empfohlen, da diese als eine der performantesten Open-Source Vektordatenbank gilt, die sich speziell für KI-Use-Cases eignen. Eine gleichwertige Alternative ist PGVector[25], eine PostgreSQL-Erweiterung für die Vektor-Ähnlichkeitssuche.

[23] https://huggingface.co/intfloat/multilingual-e5-large-instruct

[24] https://www.trychroma.com/

[25] https://github.com/pgvector/pgvector

7.2.4 Retrieval von Informationen

Ein weiterer wesentlicher Aspekt ist das Retrieval, die Suche nach relevanten Inhalten in der Vektordatenbank, um die Anfragen der Nutzer optimal zu beantworten. Üblicherweise nutzt RAG die Kosinus-Ähnlichkeit, um Inhalte zu finden, die dem Nutzereingaben am nächsten stehen. Es sollte jedoch eine fortgeschrittenere RAG-Technik, das sogenannte "Ensemble Retrieval", implementiert werden. Diese kombiniert Kosinus-Ähnlichkeitssuche mit einer Stichwortsuche, wie etwa BM25, und hat in vorangegangenen Projekten zu überlegenen Suchergebnissen geführt. Sollten genügend Ressourcen zur Verfügung stehen, wird empfohlen, diverse fortgeschrittene Retrieval-Methoden zu evaluieren, da die optimale Lösung je nach Anwendungsfall variieren kann. Weitere fortgeschrittene Methoden umfassen:

- Parent-Document-Retrieval: Bei dieser Methode werden Dokumente in kleinere (Child-Chunks) und größere Einheiten (Parent-Chunks) unterteilt. Bei der Retrieval-Suche wird zunächst nach den kleinen Chunks gesucht, und falls diese zur Anfrage passend sind, wird der Übergeordnete Parent-Chunk für einen breiteren Kontext an das LLM übergeben.
- Kombination aus Kosinus-Ähnlichkeit, Keyword-Suche und Maximal Marginal Relevance (MMR): Hier wird das Ensemble-Retrieval um die MMR-Suche erweitert, welche die Relevanz und Diversität der zurückgegebenen Dokumente steigern.
- Identifikation von Top-K und Festlegung eines Score-Thresholds: Top-K definiert im Retrieval-Kontext die Anzahl der zurückzugebenden Suchergebnisse. Mit einem Score-Threshold lässt sich ein Mindestwert festlegen, der bei der Suche erreicht werden muss, damit ein Inhalt als relevant eingestuft wird.

7.2.5 Antwortgenerierung

Wie bereits in der Detailbeschreibung beschrieben wird ein LLM eingesetzt, um die gefundenen Inhalte aus der Vektordatenbank in eine natürliche Sprache, angepasst an die Frage des Nutzers, auszuformulieren. Das LLM kann bei der Suche nach Inhalten die Nutzererfahrung verbessern, in dem es auf die Suchanfrage des Nutzers eine Antwort in natürlicher Sprache anfertigt und zudem noch die relevanten Quellen angibt. Bei der automatisierten Beantwortung der Bürgeranfragen hilft das LLM dem Mitarbeiter, eine passende Antwort auf die Anfrage des Bürgers zu generieren. Dabei wird nach bereits ähnlichen Fragen und Antworten in der Vergangenheit gesucht. Dies wird kombiniert mit der Suche nach relevanten Inhalten auf der Webseite oder den wissenschaftlichen Artikeln, die im Punkt „Datenvorbereitung“ gesammelt werden.

Dem BfS wird klar empfohlen, auf ein Open-Source Sprachmodell zurückzugreifen, da damit nur beim Hosting des Modells auf den eigenen Server kosten anfallen (z.B. Energiekosten). Bei Closed-Source Diensten wie OpenAI fallen pro verarbeiteten und generierten Token kosten an, was schwer skalierbar ist. Ein weiterer großer Vorteil des Self-Hostings ist, dass keine der verwendeten Daten an Drittanbieter gehen, wenn der gesamte Prozess auf einer eigenen Serverinfrastruktur läuft.

Als Open-Source Modell wird das SauerkrautLM-Mixtral-8x7B[26] empfohlen, da dieses bei PwC-internen Evaluationen am besten für die deutsche Sprache abgeschnitten hat. Dieses Modell basiert auf dem state-of-the-art Open-Source Modell von MistralAI, welches von VagoSolutions weiter auf deutsche Texte nachtrainiert wurde, und sich somit am besten für die deutsche Sprache eignet. Im Übersetzungskontext ist dieses Modell ebenfalls der Englischen, Französischen, Spanischen und Italienischen Sprache mächtig.

[26] https://huggingface.co/VAGOsolutions/SauerkrautLM-Mixtral-8x7B-Instruct

7.2.6 Evaluation

Ein wichtiger Aspekt ist, das RAG-Modell und seine generierten Antworten, sei es bei der intelligenten Suche oder bei der automatisierten Beantwortung von Bürgeranfragen, zu evaluieren. Um Antworten evaluieren zu können, werden Frage-Antwort Beispiele benötigt, die als Grundwahrheit gelten. Im Fall der automatisierten Beantwortung von Bürgeranfragen gibt es den praktischen Nebeneffekt, dass bereits manuell beantwortete Fragen vorliegen. Diese können als Grundwahrheit angenommen werden. Dadurch können die Antworten des Modells und die Grundwahrheit-Antworten verglichen und evaluiert werden. Als Framework wird hier „RAGAS“ [27] empfohlen, womit der Retrieval- sowie der Generierungsprozess effektiv evaluiert werden kann. Die Evaluierung endet mit einem RAGAS-Score, und je höher dieser ist, umso besser schneidet die Methode ab.

7.2.7 Hardwareanforderungen

Um das RAG-Modell für die beiden Use-Cases zu hosten, gibt es bestimmte Hardwareanforderungen, die erfüllt werden müssen. In der Detailbeschreibung wurden bereits Empfehlungen für Sprachmodelle von verschiedenen Größen gegeben. Da das SauerkrautLM-Mixtral-8x7B ein sehr großes Modell mit 46,7 Milliarden Parametern ist, ist viel Rechenleistung nötig, um das Modell auf eignen Servern zu hosten. Aus diesem Grund wird empfohlen, eine 4 Bit komprimierte Version des Modells zu verwenden, was bereits mit 2 GeForce RTX 4090 (je 24GB) umsetzbar ist. Das jeweilige 4 Bit Modell weist, wie im technischen Detailkonzept bereits erläutert, einen guten Trade-off zwischen Performanz und Rechenleistung auf.

7.2.8 Zusammenfassung

In der Zusammenfassung, dargestellt in Tabelle 8, werden alle relevanten Empfehlungen aufgeführt.

Implementierungsschritt	Empfehlung
Datenvorbereitung	Chunking mit Unstructured.io
Embedding-Transformation	intfloat/multilingual-e5-large-instruct
Vektordatenbank	ChromaDB/ PGVector
Retrieval	Ensemble Retrieval mit Kosinus-Ähnlichkeit kombiniert mit BM25-Keywordsuche
Antwortgenerierung/ LLM	SauerkrautLM-Mixtral-8x7B
Evaluation	RAGAS-Framework
Hardware	2x GeForce RTX 4090 (je 24GB)

Tabelle 8: Zusammenfassung der technischen Handlungsempfehlungen

[27] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Für die Datenvorbereitung ist das Aufteilen der Dokumente in kleinere Segmente (Chunks) mittels des Unstructured-Frameworks ratsam, das sich besonders für die Extraktion von Inhalten aus unstrukturierten oder komplexen Dokumenten wie beispielsweise PDF-Dateien eignet. Als Embedding-Modell wird „intfloat/multilingual-e5-large-instruct“ empfohlen, welches sich für die Darstellung deutscher Inhalte als numerische Vektoren bewährt hat. Die Speicherung der generierten Embeddings aus den Chunks sollte vorzugsweise in ChromaDB, einem Open-Source Vektordatenbank-Framework, erfolgen. Im Retrieval-Prozess wird ein Ensemble-Retrieval-Ansatz nahegelegt, der die Kosinusähnlichkeitssuche mit einer Stichwortsuche kombiniert und dadurch die Suchgenauigkeit verbessert. Als das leistungsfähigste Sprachmodell (LLM) für die Generierung notwendiger Inhalte hat sich das „SauerkrautLM-Mixtral-8x7B“ von VagoSolutions herausgestellt, eine auf deutsche Texte spezialisierte Adaption des Mixtral-8x7B Modells von MistralAI. Zur Evaluation der Qualität und Korrektheit der generierten Antworten ist das RAGAS-Framework zu verwenden, das eine separate Bewertung des Retrieval- sowie des Generierungsprozesses ermöglicht. Für das Hosting des Modells auf eigenen Servern sind 2 GeForce RTX 4090 Grafikkarten zu empfehlen. Grundsätzlich wird empfohlen auf komprimierte Versionen des SauerkrautLM-Mixtral-8x7B-Modells zurückzugreifen.

8 Fazit

Die Ergebnisse des Projektes haben das Potenzial der digitalen Transformation innerhalb des BfS eindrucksvoll demonstriert. Durch die Automatisierung der Bearbeitung von Bürgeranfragen könnte nicht nur die interne Effizienz gesteigert werden, sondern auch eine neue Art der Interaktion mit den Bürgern etabliert werden, die schneller und zugänglicher ist als traditionelle Methoden. Darüber hinaus kann das System zur Entlastung der Mitarbeiter des KEMF beigetragen, welche sich nun verstärkt komplexeren Anfragen widmen kann. Zudem wurden Einblicke in die Bedürfnisse und Anliegen der Bürger geliefert. Die Vorteile werden auch im KEMF gesehen. Die befragten Mitarbeiter stehen der Technologie sehr offen gegenüber und erhoffen sich schon jetzt eine baldige Implementierung und damit Unterstützung bis hin zur Entlastung.

Aus dem Projekt ergeben sich zahlreiche Lernergebnisse und Best Practices, die für die Planung und Durchführung zukünftiger digitaler Initiativen von großer Bedeutung sind. Die agile Projektmanagementmethode erwies sich als äußerst effektiv, um flexibel auf Herausforderungen zu reagieren und die Lösung kontinuierlich auszugestalten und mit technischen Details zu versehen. Die Wichtigkeit einer soliden Datenbasis und die Notwendigkeit, Transparenz und Datenschutz in den Mittelpunkt aller KI-basierten Projekte zu stellen, wurden ebenfalls deutlich.

Blickt man in die Zukunft, so bietet das Projekt eine solide Grundlage für weitere Innovationen und Verbesserungen. Die Möglichkeit, das KI-System durch die Einbeziehung von Spracherkennungstechnologien zu erweitern, um auch telefonisch eingereichte Anfragen automatisiert bearbeiten zu können, sowie die kontinuierliche Analyse der bisher gesammelten Daten, bieten enorme Chancen- So kann die Bürgerkommunikation des KEMF, aber auch die des gesamten BfS, weiter verbessert und noch besser auf die Bedürfnisse der Bürger eingegangen werden.

Zusammenfassend lässt sich sagen, dass das Projekt die Grundlagen für technologischen Fortschritt in der digitalen Transformation der öffentlichen Verwaltung darstellt. Es hat nicht nur gezeigt, dass durch den Einsatz moderner Technologien interne Prozesse optimiert und die Servicequalität für die Bürger verbessert, werden können, sondern auch wertvolle Einblicke für ähnliche Initiativen in anderen Bereichen geliefert. Die fortlaufende Entwicklung und Anpassung des Systems, die Berücksichtigung ethischer und rechtlicher Aspekte sowie die Förderung von Transparenz und Bürgerbeteiligung werden entscheidend sein, um das volle Potenzial der KI in der öffentlichen Verwaltung auszuschöpfen und das Vertrauen der Öffentlichkeit in diese Technologien zu stärken.

Referenzen

- Bharat, K., & Mihaila, A. (2000). Hilltop: A search engine based on expert documents [Online]. University of Toronto. https://ftp.cs.toronto.edu/pub/reports/csrg/405/hilltop.html (Zugegriffen am: November 24, 2023).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Gyöngyi, Z., Garcia-Molina, H., & Pedersen, J. (2004, August). Combating web spam with trustrank. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (pp. 576-587).
- Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., ... & Karolus, J. (2023). Unreflected Acceptance--Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education. arXiv preprint arXiv:2309.03087.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bring order to the web. Technical report, stanford University.
- Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485-5551.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084
- Singhal, A. (2013). Fifteen years on–and we just getting started [Online]. Google Blog. https://search.googleblog.com/2013/09/fifteen-years-onand-were-just-getting.html (Zugegriffen am: November 24, 2023).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
- Wang, H., Wu, H., He, Z., Huang, L., & Church K. W. (2021). Progress in Machine Translation. Engineering 18 (2022) 143-153.