Lösungsansätze für die datenschutz-konforme Nutzung von GenAI in der öffentlichen Verwaltung

Studie im Auftrag des Datenschutzbeauftragten des Kantons Luzern

Jürgen Vogel, Annett Laube und Peter von Niederhäusern

Berner Fachhochschule (BFH)

Finale Version (11.04.2025)

Berner Fachhochschule

Institut IDAS

Inhaltsverzeichnis

1 Management Summary 3
2 Einleitung 4
2.1 Inhalt des Berichts und methodisches Vorgehen 5
2.2 Maschinelles Lernen und Generative Künstliche Intelligenz (GenAI) 6
2.3 Herausforderungen im Zusammenhang mit GenAI 7
2.4 Autoren der Studie 9
3 Praxisbeispiele und Fallstudien 11
3.1 Chatbots 11
3.2 Informationszugang und Suche 12
3.3 Office-Applikationen 13
3.4 Fach-Applikationen 14
3.5 Sprach-Erkennung und Protokollierung 14
3.6 Bilder und Videos 15
3.7 Europäische Kommission als Vorreiter für produktive Lösungen 15
3.8 Zusammenfassung 16
4 Ausgangslage 17
4.1 Strategie 17
4.2 Rechtliche und ethische Rahmenbedingungen 19
4.3 Exemplarische Anwendungsfälle 21
5 Lösungsansätze für den Einsatz von GenAI 23
5.1 Modellauswahl 23
5.2 Best Practices für IT-Sicherheit, Betrieb und Wartung 25
5.3 Aufbau und Betrieb 26
5.4 Prompt Management 28
6 Datenschutzkonforme Implementierung 29
6.1 Phase 1: Erhebung von Trainingsdaten 29
6.2 Phase 2: Vorbereitende Datenverarbeitung 30
6.3 Phase 3: Training 30
6.4 Phase 4: Operation - Verarbeitung von Prompts und Outputs 31
6.5 Phase 5: Training mit Prompts (optional) 32
6.6 Einordnung der Massnahmen 32
6.7 Beispielhafte Anwendung auf die exemplarischen Anwendungsfälle 33
7 Handlungsempfehlungen 36
7.1 Daten- und KI-Kultur 36
7.2 Weiterbildung und Rekrutierung von Mitarbeitenden 36
7.3 Gesamtbetrachtung der rechtlichen Rahmenbedingungen 37
7.4 Applikationen und Infrastruktur 37
7.5 Transparenz gegenüber der Öffentlichkeit und Politik 37
7.6 Risikomanagement 38
7.7 Bedarfsanalyse und Wirtschaftlichkeit 38
7.8 Vernetzung 39
8 Zusammenfassung und Ausblick 40
9 Tabellenverzeichnis 41
10 Glossar 42
11 Literaturverzeichnis 45

1 Management Summary

Der technologische Fortschritt auf dem Gebiet der Künstlichen Intelligenz (KI) hat in den letzten Jahren zahlreiche neue Anwendungen im Arbeitsalltag ermöglicht. Ansätze der sogenannten generativen KI (engl. Generative AI (GenAI)) wie Large Language Models (LLMs) können zunehmend auch komplexe Aufgaben lösen wie die automatisierte Beantwortung von Supportanfragen oder die Erstellung von fachlichen Texten. Gerade auch für öffentliche Verwaltungen ergibt sich somit ein grosses Potential zur Unterstützung der Mitarbeitenden im Arbeitsalltag, zur Optimierung der Verwaltungsprozesse und zur effizienten Interaktion mit Bürger/innen und anderen Institutionen. In dieser Studie werden rechtliche, organisatorische und technische Rahmenbedingungen und Herausforderungen untersucht, um speziell Lösungen auf Basis von GenAI und LLM für öffentliche Verwaltungen und andere Organe des öffentlichen Rechts praktisch nutzbar zu machen, unter besonderer Berücksichtigung einer datenschutzkonformen Umsetzung und der Sicherstellung der digitalen Souveränität. Die Studie stützt sich dabei auf Interviews mit Fachexperten und eine umfangreiche Recherche zu den Elementen einer Daten- und KI-Strategie im öffentlichen Sektor, zu den rechtlichen Rahmenbedingungen (Legitimierung, Begründungspflicht, Diskriminierungsverbot und Datenschutz) für den Einsatz von KI sowie zu aktuellen GenAI-Lösungen und -Anwendungen im Umfeld öffentlicher Verwaltungen. Anhand exemplarischer Anwendungsfälle (Suchmaschinen und Chatbots, KI-Assistenten für Office- und Fach-Applikationen, Spracherkennung) werden verschiedene Alternativen bezüglich der technischen Umsetzung auf Basis von kommerziellen oder Open Source LLM diskutiert und die jeweils umzusetzenden Datenschutz-Massnahmen analysiert. Weiterhin empfiehlt die Studie begleitende Massnahmen, die auf Basis der gewonnenen Erkenntnisse für die erfolgreiche Konzeptionierung, Entwicklung und Betrieb einer GenAI-Lösung umzusetzen sind: die Etablierung einer Daten- und KI-Kultur, bei der es zur gelebten Normalität wird, dass Mitarbeitende zielgerichtet und regelkonform mit Daten und KI umgehen und diese effektiv und effizient einsetzen können, eine Priorisierung der möglichen GenAI-Anwendungsfälle mit Hilfe einer Kosten-Nutzen-Analyse, den Aufbau zusätzlicher Infrastruktur und insbesondere personeller Ressourcen für Entwicklung und Betrieb von eigenen GenAI-Lösungen (durch Weiterbildung und Rekrutierung), vertrauensbildende Massnahmen und Transparenz gegenüber Mitarbeitenden und Betroffenen in Bezug auf die praktisch eingesetzten GenAI-Lösungen, ein systematischer Umgang mit allfälligen Restrisiken der eingesetzten Lösungen (Risikomanagement) sowie eine enge Zusammenarbeit mit externen Partnern wie kommerziellen Anbietern, Forschungsinstituten und anderen öffentlichen Institutionen zum allgemeinen Erfahrungsaustausch und der gemeinsamen Umsetzung von aufwändigen GenAI-Projekten.

2 Einleitung

Der technologische Fortschritt auf dem Gebiet der Künstlichen Intelligenz (KI) (engl. *Artificial Intelligence (AI)*), insbesondere beim zugrundeliegenden maschinellen Lernen (ML), hat in den letzten Jahren zahlreiche neue Anwendungen ermöglicht, die sich im Arbeitsalltag etabliert haben, beispielsweise bei der Vorhersage von Finanzzahlen [28], dem Informationszugriff via *Chatbot* [115], oder der (teil-)automatischen Bearbeitung von Geschäftsprozessen [1] und Verwaltungsfällen [29]. Ansätze der sogenannten generativen KI (engl. *Generative AI (GenAI)*) wie *Large Language Models (LLMs)* können zunehmend auch komplexe Aufgaben lösen wie die Beantwortung von Beratungsanfragen von Bürger/innen an eine öffentliche Verwaltung [61] oder die Erstellung von fachlichen Texten [12], beispielsweise zur Bearbeitung von juristischen Fällen [140]. Wissenschaftliche Studien haben dabei nachgewiesen, dass die Qualität der Ergebnisse, die von einer GenAI automatisch erzeugt werden, in den letzten Jahren in vielen Aufgabenbereichen kontinuierlich verbessert werden konnte und mittlerweile das Niveau einer menschlichen Fachkraft erreichen kann [21, 155]. Der zu beobachtende intensive Ressourceneinsatz in Forschung und Entwicklung im Bereich KI [100] lässt zudem vermuten, dass sich die Verfahren auch in den nächsten Jahren beständig weiterentwickeln und somit auch weitere, innovative Anwendungen ermöglichen werden. Gerade auch für öffentliche Verwaltungen ergibt sich somit ein grosses Potential zur Unterstützung der Mitarbeitenden im Arbeitsalltag, bei der Optimierung der Verwaltungsprozesse und der effizienteren Interaktion mit Bürger/innen und anderen Institutionen [40, 104, 152, 159].

Möglich wurde dieser Fortschritt durch das Zusammenspiel von verschiedenen Faktoren, die zur erfolgreichen Realisierung einer KI-Lösung notwendig sind [102]: der Entwicklung leistungsfähiger ML-Verfahren, der Verfügbarkeit einer grossen Menge an qualitativ hochwertigen Daten sowie einer leistungsfähigen IT-Infrastruktur zum Entwickeln (Trainieren) der ML-Modelle, ergänzt durch eine einfache Bereitstellung und Anwendung der ML-Modelle für konkrete Aufgaben, beispielsweise als Chatbot oder Dienst, und nicht zuletzt ein rentables Geschäftsmodell, das die kontinuierliche Weiterentwicklung einer KI-Lösung nachhaltig ermöglicht.

Diese Erfolgsfaktoren in konkreten Projekten auch umzusetzen, ist gleichzeitig eine grosse Herausforderung [102] und erfordert einen entsprechenden Einsatz an finanziellen, technischen und menschlichen Ressourcen. Als Folge lässt sich beobachten, dass sich in vielen Geschäftsbereichen eine klassische Rollenverteilung mit spezialisierten KI-Anbietern auf der einen und KI-Anwendern auf der anderen Seite ergeben hat. Insbesondere kleinere öffentliche Verwaltungen mit limitierten eigenen Kapazitäten im Bereich Entwicklung und Betrieb von IT-Infrastrukturen im Allgemeinen und KI-Lösungen im Speziellen würde man daher eher auf der Seite der reinen KI-Anwender sehen. Diese Rollenverteilung hat jedoch auch zentrale Nachteile: Zunächst sind die geltenden rechtlichen Rahmenbedingungen einzuhalten und gerade Verwaltungsprozesse verarbeiten häufig sensitive personenbezogene Daten und müssen entsprechend die geltenden Anforderungen zum Datenschutz einhalten [159]. In Bezug auf ML und KI können personenbezogene Daten sowohl beim Trainieren als auch beim Ausführen des ML-Modells Verwendung finden und sind entsprechend zu schützen, was naturgemäss schwierig ist, wenn Training und/oder Ausführung an einen externen Anbieter ausgelagert sind [95].

Ein weiterer zentraler Nachteil liegt in den Fragen von Handlungskompetenz und Kontrolle, die wichtige Grundlagen der notwendigen digitalen Souveränität [124] von öffentlichen Institutionen sind: Wenn KI-Lösungen zentrale Aufgaben in Verwaltungsprozessen unterstützen oder vollständig übernehmen, besteht generell ein gewisses Risiko von intransparenten, fehlerhaften oder unfairen Entscheidungen und Prozessen [101], welches durch die Auslagerung von Entwicklung und Betrieb an einen Drittanbieter tendenziell verstärkt wird [103, 106]. Mangelnde Transparenz wird entsprechend auch häufig in Bürgerumfragen als mögliche Gefahr beim Einsatz von KI-basierten Lösungen in der öffentlichen Verwaltung genannt [163].

2.1 Inhalt des Berichts und methodisches Vorgehen

KI-basierte Lösungen bieten ein grosses Potential für eine effiziente und zielgerichtete öffentliche Verwaltung. In dieser Studie werden rechtliche, organisatorische und technische Rahmenbedingungen und Herausforderungen untersucht, um speziell Lösungen auf Basis von GenAI für öffentliche Verwaltungen und andere Organe des öffentlichen Rechts praktisch nutzbar zu machen, unter besonderer Berücksichtigung einer datenschutzkonformen Umsetzung und der Sicherstellung der digitalen Souveränität. Ziel der Studie ist ausserdem, verschiedene Umsetzungsoptionen darzustellen und diesbezüglich Empfehlungen zu geben.

In Kapitel 2 geben wir zunächst einen allgemeinen Überblick zum Inhalt der Studie und eine Einführung zu ML und GenAI. In Kapitel 3 untersuchen wir aktuelle Forschungsarbeiten und relevante bereits umgesetzte KI-Lösungen mit Schwerpunkt auf Language Models (LMs). Kapitel 4 beleuchtet die zugrundeliegende Digitalisierungsstrategie öffentlicher Verwaltungen, rechtlichen Rahmenbedingungen für den Einsatz von GenAI und definiert exemplarische Anwendungsfälle, um die Diskussion in den nachfolgenden Kapiteln zu konkretisieren. Kapitel 5 gibt einen Überblick zu den technischen Optionen bei der Umsetzung von GenAI-Lösungen in diesen Anwendungsfällen. Notwendige Massnahmen, um diese GenAI-Lösungen datenschutzkonform umzusetzen, werden in Kapitel 6 näher ausgeführt. Kapitel 7 gibt Empfehlungen zur weiteren Vorgehensweise bei Entwicklung und Betrieb von GenAI-Lösungen. Kapitel 8 schliesst mit einer Zusammenfassung und einem Ausblick.

Die inhaltliche Ausarbeitung der Studie basiert dabei auf einer wissenschaftlichen Literaturrecherche zum aktuellen Stand der Forschung im Bereich KI-Lösungen mit Schwerpunkt auf Fachartikeln zu GenAI und LM aus den letzten zwei Jahren, um den schnellen Fortschritt in diesem Bereich zu berücksichtigen. Bei Recherchen zu Anwendungsfällen liegt der Fokus auf öffentlichen Verwaltungen und andere Organisationen des öffentlichen Rechts. Geographisch liegt der Schwerpunkt wegen der vergleichbaren rechtlichen und ethischen Rahmenbedingungen auf Europa (EU/CH). Für das Auffinden von Fachartikeln wurden dabei vor allem die spezialisierte Suchmaschine Google Scholar[1] sowie die proprietären Suchmaschinen der wichtigsten Wissenschaftsverlage und -organisationen wie Springer[2], Elsevier[3], ACM[4], IEEE[5] etc. verwendet.

Weiter stützt sich der vorliegende Bericht auf eine selbst durchgeführte Marktstudie zu technischen Komponenten für eine mögliche Umsetzung von GenAI-Lösungen, die Expertise und Erfahrungen der Autoren, und nicht zuletzt auf drei Interviews mit Experten aus dem Kanton Luzern:

- Daniel Spichty, amtierender Datenschutzbeauftragter des Kantons Luzern und Auftraggeber der Studie,
- Marco Strebel, Stab Dienststellenleiter der Dienstelle Informatik und Mitglied der «Community of Practice KI» des Kantons Luzern, und
- Dimitri Bucher, Leiter Business Engineering und Projektmanagement des Justiz- und Sicherheitsdepartements des Kantons Luzern und Organisations- und Informatikverantwortlicher des Departements.

Diese Experten wurden dabei in separaten, ca. 1-stündigen Interviews zu allenfalls schon realisierten und potenziellen zukünftigen Anwendungsfällen, der verfügbaren IT-Infrastruktur, relevanten Rahmenbedingungen inklusive verfügbarer finanzieller und personeller Ressourcen, und möglichen Herausforderungen bei der Umsetzung von KI-basierten Lösungen befragt. Die Erkenntnisse aus den Interviews wurden von den Autoren als Basis für die inhaltliche Ausgestaltung des Berichts verwendetet; auf wörtliche Zitate wurde dagegen bewusst verzichtet. Für die engagierte Mitarbeit der genannten Personen und die daraus gewonnen Einsichten bedanken wir uns an dieser Stelle herzlich.

[1] https://scholar.google.com
[2] https://link.springer.com
[3] https://www.sciencedirect.com
[4] https://dl.acm.org
[5] https://ieeexplore.ieee.org/

2.2 Maschinelles Lernen und Generative Künstliche Intelligenz (GenAI)

Mit Künstlicher Intelligenz (KI) (engl. *Artificial Intelligence (AI)*) bezeichnet man computergestützte Verfahren, die Aufgaben lösen, die ursprünglich menschliche Intelligenz erfordern, wie z.B. Sprachverarbeitung, Problemlösung und Entscheidungsfindung [41]. *Maschinelles Lernen (ML)* ist ein Teilbereich der KI und entwickelt Algorithmen, die es ermöglichen, aus einer gegebenen Datenmenge, den sogenannten Trainingsdaten, eine Lösung für ein gegebenes Problem zu lernen und in Form eines Modells so abzubilden, dass dieses auch für Situationen, die nicht explizit Teil des Trainings waren, qualitativ hochwertige Lösungen liefern kann [87].

Je nach Art der Problemstellung unterscheidet man dabei zwischen *Discriminative AI* und *Generative AI (GenAI)* [48]. Verfahren der Discriminative AI ordnen gegebene Datensätze in Klassen ein. Beispielsweise kommt in der Steuerverwaltung des Kanton Bern ein ML-basiertes Verfahren zum Einsatz, das eine eingereichte Steuererklärung dahingehend einteilt (klassifiziert), ob dieses korrigiert und damit manuell überprüft werden muss oder aber automatisch veranlagt werden kann[6]. Diese Klassifikation basiert auf verschiedenen Merkmalen (*Features*) der eingereichten Steuerklärung, wie den ausgefüllten Feldern der Steuerformulare, den beigelegten Dokumenten und dem Ergebnis der zurückliegenden Erklärungen. Der ML-Algorithmus stellt dann einen Zusammenhang zwischen der Problemlösung (hier: der Klassifikation) und diesen Merkmalen her, indem die vielen konkreten Beispielfälle der Trainingsdaten analysiert und die zugrundeliegenden Muster abstrahiert werden. Im vorliegenden Fall bestehen die Trainingsdaten aus der Sammlung der historischen Steuererklärungen, bei denen die Klassifikation in die Kategorien «Korrektur erforderlich» bzw. «Korrektur nicht erforderlich» manuell von den Fachpersonen durchgeführt wurde und bei denen man davon ausgeht, dass diese fehlerfrei erfolgte. Weil der Lernprozess hier anhand von Beispielfällen mit vorgegebener korrekter Lösung abläuft, spricht man von überwachtem Lernen. Die aus den Trainingsdaten abgeleiteten Muster und Zusammenhänge werden dann als Ergebnis des Lernprozesses in einem sogenannten ML-Modell abgespeichert. Je nach ML-Algorithmus beschreibt das Modell beispielsweise eine mathematische Funktion oder ein neuronales Netz [87]. Die Qualität des trainierten Modells wird anhand einer geeigneten Metrik (z.B. der Fehlerquote) überprüft, indem es auf Testdaten angewendet wird, also Daten mit bekannter richtiger Problemlösung, die aber nicht für das Training verwendet wurden. Die Qualitätsmetrik kann dann durch den Vergleich zwischen erwartetem und beobachteten Ergebnis berechnet werden. Typischerweise erfordert die Entwicklung einer ML-basierten Lösung einen iterativen Entwicklungsprozess, bei dem die Lösung z.B. durch Anpassung der Merkmale oder Anpassung der Parameter des ML-Algorithmus sukzessive verbessert wird, bis die Metrik die notwendige Qualität der Lösung für deren praktischen Einsatz bescheinigt [7]. Das erfolgreich trainierte Modell ist dann in der Lage, die gestellte Aufgabe für zuvor unbekannte Datensätze (hier: die neu eingereichten Steuerfälle) zu erfüllen. Im Fall der Steuerveranlagung des Kantons Bern ist die entwickelte ML-Lösung bereits in der Lage, 20% der eingereichten Erklärungen natürlicher Personen automatisch zu veranlagen (wobei die gemessene Fehlerquote aus naheliegenden Gründen nicht publiziert wurde).

GenAI-Verfahren werden dagegen verwendet, um aus den trainierten Modellen neue synthetische Daten zu erzeugen, die in ihren Merkmalen und deren Zusammenhängen den originalen Trainingsdaten ähneln [15]. Konkret lassen sich so beispielsweise aus einem Sprachmodell, das auf Basis von Textdokumenten wie Webseiten oder Büchern trainiert wurde, neue Texte generieren, die sich sprachlich und inhaltlich an den originalen Text orientieren. Das Sprachmodell wird dafür anhand der Trainingsdaten zunächst auf die Aufgabe trainiert, für eine gegebene Sequenz von Worten das nächste, wahrscheinlichste Wort vorherzusagen [76], beispielsweise dass das Wort «Herren» wahrscheinlich auf die Sequenz «Sehr geehrte Damen und» folgt. Wiederholt angewendet können so auch längere Texte erzeugt werden, um beispielsweise Zusammenfassungen oder Übersetzungen von Texten zu erstellen oder Antworten auf die Anfrage eines Menschen an einen Chatbot.

[6] https://www.beinfo.sites.be.ch/de/start/rubriken/Zoom/zoom-202203-KI.html

Während des Trainings analysiert der Lernalgorithmus dafür selbständig die in den Trainingsdokumenten vorkommenden Wortsequenzen, weswegen man auch von unüberwachtem (oder selbst-überwachtem) Lernen spricht. Um die Vielzahl der relevanten Merkmale und Muster dieser Wortsequenzen genügend gut abbilden zu können, setzen moderne GenAI-Verfahren dabei auf neuronale Netzwerke mit Milliarden von zu trainierenden Parametern (Neuronen) [116], die in speziellen Anordnungen (Architekturen) miteinander verbunden sind. Aufgrund ihrer Komplexität bezeichnet man solche neuronalen Netzwerke auch als *Deep Neural Networks,* die zugrundeliegenden Lern-Algorithmen als *Deep Learning* [48] und das resultierende Sprachmodell als *Language Model (LM)* [76]. Im Anschluss an dieses unüberwachte *Vortraining* bezeichnet man ein LM auch als *Foundation Model*, welches dann in einem weiteren überwachtem (oder bestärkendem) Training an eine spezifische Aufgabe wie die maschinelle Übersetzung oder die Beantwortung von Bürgeranfragen angepasst werden kann [108]. Beispielsweise verwendet der Chatbot der Stadt Luzern ein LM, um Bürgeranfragen auf Basis der Inhalte der öffentlichen Webseiten zu beantworten[7]. Andere Anwendungsmöglichkeiten von LMs sind das Erzeugen von Zusammenfassungen, Übersetzungen, Formularen oder Briefen [162].

Um ein LM in einem Chatbot oder einer anderen Applikation nutzen zu können, wird dieses typischerweise in einer speziellen Laufzeitumgebung ausgeführt und dann über eine technische Schnittstelle eingebunden [3]. Die eigentliche Interaktion der Anwendung mit dem LM geschieht über eine in natürlicher Sprache formulierte Anfrage, dem sogenannten *Prompt*. Das LM verarbeitet den gegebenen Prompt und generiert durch einen als *Inferenz* (engl. *Inference)* genannten Prozess die passende Antwort. Da der Prompt einen entscheidenden Anteil an der Qualität der Antwort hat, wurden in den letzten Jahren im Rahmen des sogenannten *Prompt Engineering* verschiedene Prompting-Strategien entwickelt [134]. In diesem Zusammenhang wird vor allem auch die Technik des *Retrieval Augmented Generation (RAG)* erfolgreich eingesetzt, bei der über den Prompt zusätzliche, spezialisierte Informationsquellen in die Anfrage mit einbezogen werden [93]. Weitere Details zur technischen Funktionsweise bzw. Umsetzung von LM-Lösungen finden sich in Kapitel 5.

2.3 Herausforderungen im Zusammenhang mit GenAI

Angelehnt an die oben beschriebene grundsätzliche Funktionsweise am Beispiel eines LLMs ergeben sich eine Reihe von **technischen Herausforderungen** bei der Umsetzung einer GenAI-Lösung: Zunächst erfordert die Entwicklung einer KI-Lösung wie oben beschrieben eine umfangreiche Sammlung von repräsentativen Trainings- und Testdaten in hoher Qualität. Je nach Lernalgorithmus ist dabei insbesondere die gewünschte korrekte Lösung (Klassifikation) ganz (alle Daten beim überwachten Lernen) oder zumindest teilweise (Testdaten beim unüberwachten Lernen bzw. bei Anpassung eines Foundation Models) in manueller Arbeit zu erstellen, was entsprechend aufwändig ist. Werden für die Erstellung der Datenbasis Daten aus verschiedenen Quellen verwendet, ist häufig eine Datenintegration [164] erforderlich. Gerade in öffentlichen Verwaltungen sind für die verschiedenen Verwaltungsgeschäfte oftmals eine Vielzahl unterschiedlicher Standard- und Fachanwendungen mit jeweils eigener Datenverwaltung im Einsatz. Die Integration dieser Datensilos [120] insbesondere zwischen verschiedenen Institutionen ist aus Gründen des Datenschutzes oftmals allerdings auch gar nicht ohne Weiteres zulässig. Um den (erlaubten) Datenaustausch in der Schweiz in verschiedenen Themengebieten (Landwirtschaft, Gesundheit etc.) kontrolliert und strukturiert zu ermöglichen, hat der Bund den Aufbau eines so genanntes Datenökosystems initiiert[8].

Eine weitere technische Herausforderung ist die eigentliche Entwicklung der KI-Lösung, die eine systematische Auswahl der am besten geeigneten Verarbeitungsschritte und ML-Algorithmen und -Modelle [70] erfordert, kombiniert mit einer genauen Anpassung an die jeweilige

[7] https://www.stadtluzern.ch/aktuelles/newslist/2298589
[8] https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-103816.html und https://www.bk.admin.ch/bk/de/home/digitale-transformation-ikt-lenkung/datenoekosystem_schweiz/anlaufstelle-datenoekosystem-schweiz.html

Aufgabenstellung, die verfügbare Datenbasis, und die vorhandene IT-Infrastruktur. Aufgrund der Vielzahl der möglichen Varianten werden KI-Lösungen wie oben beschrieben in der Regel iterativ entwickelt, wobei die insgesamt notwendige Entwicklungszeit oft nur auf Basis von Erfahrungswerten geschätzt werden kann. Ein Hauptziel dieser iterativen Entwicklung ist die qualitative Verbesserung der Lösung bzw. das Beheben von Fehlern. Beispielsweise ist das so genannte *Halluzinieren* ein verbreiteter Fehler von LMs, bei dem das LM zwar einen syntaktisch korrekten Text erzeugt, dieser aber inhaltlich falsch und für den unkritischen Menschen dennoch überzeugend formuliert ist [68]. Die Wahrscheinlichkeit, dass ein LM halluziniert, wird dadurch beeinflusst, wie das LM die Wahrscheinlichkeit berechnet mit der Wortsequenzen miteinander im Zusammenhang stehen (siehe oben), und kann auch vom Benutzenden via Prompt (über Angabe einer sogenannten Temperatur) beeinflusst werden [13]. Andere wichtige Fehlerarten bei LMs sind Missinterpretation des Prompts [77] und unethische, diskriminierende (siehe unten und Abschnitt 4.2) oder mit Gesetzen in Konflikt stehende [154] Antworten. Da moderne ML-Modelle aufgrund ihrer Komplexität nicht direkt interpretiert werden können, erfolgt die Erkennung solcher Fehler durch systematische Untersuchung des ML-Modells mit Hilfe von vordefinierten Testfällen (auch Testdaten oder Benchmark genannt). Solche Benchmarks sind teilweise auch öffentlich verfügbar [72] und können daher auch gut zum Vergleich verschiedener Lösungen verwendet werden. Doch selbst mit sorgfältigem Anpassen und Testen lässt sich nicht vermeiden, dass ein gewisser Anteil an Fehlern (oft gemessen als Fehlerquote (*Error Rate*)) im ML-Modell verbleibt und entweder akzeptiert (oder das Projekt erfolglos beendet) werden muss. Um dem verantwortlichen Projektteam und den späteren Benutzenden das Aufdecken von Fehlern in komplexen ML-Modellen zu erleichtern, ist der Einsatz von *Explainable AI (XAI)* möglich, d.h. die Bereitstellung einer für Menschen interpretierbaren Erklärung zusätzlich zum Ergebnis [58]. Ein LM kann beispielsweise mit der sogenannten *Chain of Thought*-Methode angewiesen werden, die schrittweise Herleitung der Antwort zu erklären [166]. XAI unterstützt daher auch die Vertrauensbildung der Benutzenden [2, 57] und ist eine wichtige Anforderung bei der Integration der ML-Modells in die Benutzerschnittstelle der eigentlichen Anwendung [34].

Nicht zuletzt sind bei der Umsetzung von GenAI-Lösungen auch verschiedene **rechtliche Rahmenbedingungen** einzuhalten [63]: Zunächst sind personenbezogene Daten, die bei Training und Ausführung eines ML-Modells zum Einsatz kommen könnten, gemäss Datenschutzgesetz zu handhaben [159]. Weiter müssen GenAI-Lösungen, die zur Entscheidungsunterstützung oder -automatisierung verwendet werden, gemäss Rechenschaftspflicht transparent, erklärbar und konsistent sein [64]. Schliesslich darf eine GenAI-Lösung Betroffene nicht diskriminieren, also ungerechtfertigt aufgrund von bestimmten Persönlichkeitsmerkmalen wie Geschlecht benachteiligen [132]. Rechtliche und **ethische Belange** bei Entwicklung und Einsatz von KI-Lösungen insbesondere im öffentlichen Sektor werden auch in Öffentlichkeit und Politik kritisch betrachtet. Beispielsweise hat AlgorithmWatch[9] in der Schweiz eine Petition[10] zur möglichen Diskriminierung durch KI eingebracht, die in einigen KI-Lösungen auch bereits nachweisbar war[11]. Auch deswegen ist es notwendig, dass sowohl die direkt Nutzenden wie auch die indirekt Betroffenen einer GenAI-Lösung dieser auch vertrauen [57, 71]. Auf die ethischen und rechtlichen Herausforderungen wird vertiefter in Abschnitt 4.2 eingegangen.

Schliesslich ist zu entscheiden, ob die fertige KI-Lösung ganz oder teilweise entweder auf der eigenen Infrastruktur, d.h., *on premise,* oder auf einer externen, Cloud-basierten Infrastruktur betrieben werden soll. Da die iterative Entwicklung nach der Produktivstellung der KI-Lösung fortgeführt werden sollte, um das Modell laufend zu beobachten und bei Bedarf an geänderte Daten oder Anforderungen anzupassen, wird weiter ein als **MLOps** bezeichnetes Betriebskonzept benötigt, welches Datenmanagement, (Weiter-)Entwicklung, und Betrieb ganzheitlich betrachtet und die notwendige Infrastruktur, Tools und Prozesse beinhaltet [105].

[9] https://algorithmwatch.ch/en/
[10] https://algorithmwatch.ch/de/ki-appell-ubergabe/
[11] https://algorithmwatch.ch/de/explainer-oeffentlicher-sektor/

Insgesamt ergibt sich durch diese vielfältigen Aspekte und Herausforderungen somit eine hohe **Komplexität** bei der Lösungsentwicklung. Entsprechend konnten Chen et al. in einer empirischen Studie zu Hilfsgesuchen und Diskussionen von Entwicklern auf einschlägigen Online Foren auch feststellen, dass diese mit einer Vielzahl an konzeptionellen und praktischen Fragestellungen konfrontiert sind [27]. Die allgemeine Empfehlungen für ML-Projekte, wonach die Lösung iterativ zu entwickeln [8] (und nach der Produktivstellung auch weiterzuentwickeln [85]) ist, gilt somit insbesondere auch für GenAI-Projekte. Ein erfolgreiches Projekt erfordert somit auch einen entsprechenden Einsatz an **Ressourcen** in Bezug auf Personal, Zeit und finanziellen Mitteln. Angesichts des allgemeinen Fachkräftemangels im ICT-Bereich[12] fällt es vielen Arbeitgebern jedoch schwer, Mitarbeitende, die KI-Lösungen umsetzen und betreiben können, zu rekrutieren[13], insbesondere auch im öffentlichen Sektor[14].

Angesichts dieser komplexen technischen, organisatorischen, wirtschaftlichen und rechtlichen Herausforderungen zeigen empirische Studien zum Erfolg von KI-Projekten auch einen hohen Anteil von gescheiterten Projekten, so ergab z.B. der «Data Science Survey» von Rexer Analytics aus dem Jahr 2023, dass die Mehrheit der entwickelten Modelle nicht produktiv eingesetzt wird[15], und Gartner geht davon aus, dass bis zu 85% der Projekte scheitern[16].

Insbesondere kleinere öffentliche Verwaltungen mit limitierten eigenen Kapazitäten im Bereich Entwicklung und Betrieb von IT-Infrastrukturen im Allgemeinen und KI im Speziellen sind heutzutage oft nicht in der Lage, eigene KI-Lösungen zu realisieren. Als Folge ergibt sich eine klassische Rollenverteilung zwischen externen KI-Anbietern auf der einen und KI-Anwendern auf der anderen Seite. Eine externe KI-Lösung kann entweder eine Standardapplikation oder das Ergebnis einer Auftragsentwicklung sein und wird heutzutage meist als Cloud-Anwendung realisiert. Diese Rollenverteilung gefährdet jedoch die **digitale Souveränität** von öffentlichen Verwaltungen [124], die bei einer allfälligen Auslagerung von fachlich zentralen Funktionen an eine von einem Drittanbieter entwickelte und/oder betriebene GenAI-Lösung tendenziell geschwächt wird [103, 106].

Als Grundlage für die systematische Umsetzung von GenAI-Lösungen bzw. für die systematische Nutzung von Daten im Allgemeinen braucht es eine organisationsübergreifende **Datenstrategie** [33], die den Umgang mit (internen und externen) Daten definiert und regelt, wie diese erfasst, aufbereitet, integriert, gespeichert, verwendet und verwaltet werden (siehe Abschnitt 4.1).

2.4 Autoren der Studie

Die vorliegende Studie wurde im Auftrag des amtierenden Datenschutzbeauftragten des Kantons Luzern, Schweiz[17], im Zeitraum vom 01.01.25 - 11.04.25 erstellt und durch diesen finanziert. Die Autoren der Studie sind Mitarbeitende am «Institute for Data Applications and Security (IDAS)»[18] an der Berner Fachhochschule (BFH), Schweiz[19]. Das IDAS vereint u.a. zwei spezialisierte Forschungsgruppen:

- Die Forschungsgruppe «Applied Machine Intelligence (AMI)»[20] konzentriert sich auf die praxisnahe Anwendung von Künstlicher Intelligenz, insbesondere im Bereich Maschinelles Lernen und Large Language Models (LLMs). Die Forschungsgruppe AMI betreibt mit dem «BFH Generative AI Lab»[21] auch eine eigene Deep Learning-Infrastruktur

[12] https://www.ict-berufsbildung.ch/resources/IWSB_ICT-Bildungsbedarf_20301.pdf
[13] https://www2.deloitte.com/content/dam/insights/us/articles/6546_talent-and-workforce-effects-in-the-age-of-ai/DI_Talent-and-workforce-effects-in-the-age-of-AI.pdf
[14] https://www.pwc.ch/de/insights/oeffentlicher-sektor/fachkraeftemangel-studie.html
[15] https://drive.google.com/file/d/1Mz3WmtcvUl-00qaT2XKCxdE5-pqbOOjz/view
[16] https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence
[17] https://datenschutz.lu.ch
[18] https://www.bfh.ch/de/forschung/forschungsbereiche/institute-for-data-applications-and-security-idas/
[19] https://www.bfh.ch
[20] https://www.bfh.ch/de/forschung/forschungsbereiche/applied-machine-intelligence/
[21] https://www.bfh.ch/ti/de/aktuell/generative-ai-lab/

bestehend aus mehreren Hochleistungs-Workstations und verfügt daher über die entsprechenden praktischen Erfahrungen.

- Die Forschungsgruppe «Identity and Access Management (IAM)»[22] widmet sich Sicherheitsaspekten wie Zugriffskontrolle, Schutz der Privatsphäre und der Anonymisierung personenbezogener Daten.

Folgende Mitarbeitende des IDAS haben als Autorinnen und Autoren der Studie mitgearbeitet:

- Prof. Dr. Annett Laube[23], Dozentin für Informatik, Leiterin IDAS und Forschungsgruppe IAM, Fachexpertin für Privacy und Zugriffskontrolle,
- Prof. Dr. Jürgen Vogel[24], Dozent für Informatik und Fachexperte für Data Engineering, ML, NLP und LM am IDAS, Forschungsgruppe AMI, und
- Peter von Niederhäusern[25], Dozent für Informatik und Fachexperte für Infrastruktur für KI und GenAI am IDAS, Forschungsgruppe AMI.

[22] https://www.bfh.ch/de/forschung/forschungsbereiche/identity-access-management-iam/
[23] https://www.bfh.ch/de/annett-laube
[24] https://www.bfh.ch/de/juergen-vogel
[25] https://www.bfh.ch/de/peter-alfred-von-niederhaeusern

3 Praxisbeispiele und Fallstudien

In diesem Kapitel werden ausgewählte Forschungsarbeiten und Lösungen im Bereich GenAI[26] vorgestellt, die im Rahmen der Literaturrecherche gefunden und auf Basis der durchgeführten Experteninterviews für diese Studie als relevant klassifiziert wurden. Schwerpunkt der Analyse ist GenAI mit Anwendungen im Bereich der Erstellung, Bearbeitung und Auswertung von Texten im Umfeld öffentlicher Verwaltungen und verwandter Organisationen[27]. Eine Übersicht zu Projekten mit KI-Bezug im Umfeld der Schweizerischen Bundesverwaltung findet sich beispielsweise beim Kompetenznetzwerk für künstliche Intelligenz (CNAI)[28].

3.1 Chatbots

Ein Chatbot ist ein autonomes, textbasiertes Dialogsystem, das von einem Benutzer oder einer Benutzerin zum Informationsausstauch mit einem Computersystem auf der Basis von natürlich-sprachigen Nachrichten verwendet wird, wobei die Text-Ein/Ausgabe auch per Audio erfolgen kann [89]. Heutzutage sind Chatbots bereits weit verbreitet und werden standardmässig im Web (z.B. Schweizerische Post[29]), auf mobilen Endgeräten (z.B. Apple Siri[30], Amazon Alexa[31]) und seit neuestem auch als direkte Erweiterung von Desktop-Betriebssystemen (z.B. MS Copilot[32]) als alternative Benutzerschnittstelle, für allgemeine Suchanfragen, zur Abwicklung von Einkäufen oder zur Handhabung von Supportfällen eingesetzt.

Als Alternative zu dieser Kommunikation zwischen Mensch und Maschine kann ein Chatbot aber auch als Assistent bei einer zwischenmenschlichen Kommunikation eingesetzt werden: In ihrer Studie in einem grossen Call-Center kamen die Autoren zum Ergebnis, dass Service-Mitarbeiter/innen, die in der Kundenkommunikation parallel durch einen Chatbot unterstützt werden, im Durchschnitt 14% mehr Service-Anfragen pro Stunde erfolgreich bearbeiten können, also signifikant produktiver sind [22].

Im Kontext der öffentlichen Verwaltung werden Chatbots meist als Informationsquelle für Bürger/innen eingesetzt und ermöglichen so einen vereinfachten alternativen Zugriff auf die öffentlich verfügbaren Web-Seiten und Dokumente [112, 139]. Der UN E-Government Survey gibt an, dass im Jahr 2024 bereits 27% der Mitgliedsstaaten Chatbots auf ihren Informationsportalen anbieten [151]. Der sogenannte «WienBot»[33] der Stadt Wien kann beispielsweise in mehreren Sprachen Auskunft zu Parkgebühren oder Öffnungszeiten geben, ein ähnliches Angebot gibt es auch in den Städten St. Gallen[34] und Zürich[35]. Im Kanton Luzern unterstützt der Chatbot «Wasi»[36] Bürger und Bürgerinnen bei Fragen und Anliegen rund um die Prämienverbilligung bei Krankenkassen, ähnliche Chatbots existieren auch in den Kantonen Aargau und St. Gallen. In Aargau konnten dadurch 30% der Telefonanfragen eingespart werden[37]. Die Eidgenössische Stiftungsaufsicht ESA hat einen Chatbot zur Beantwortung von allgemeinen Fragen zum Stiftungswesen in der Schweiz auf Basis des kommerziellen ChatGPT von OpenAI umgesetzt[38].

[26] GenAI wird im Umfeld des öffentlichen Sektors hauptsächlich zur Erzeugung und Analyse von Texten und Bildern verwendet. Beim viel diskutierten Anwendungsfall «Prozessautomatisierung» kommen dagegen vorrangig Discriminative AI-Verfahren zum Einsatz (vgl. Abschnitt 2.2) [119], beispielsweise bei der Steuerveranlagung, die im Rahmen dieser Studie nicht im Detail untersucht werden. Auch Vorhersagemodelle, wie sie bei Smart City-Anwendungsfällen eingesetzt werden (z.B. bei Smart City LuzernNord, siehe https://www.luzernnord.ch/smart-city/ziel) fallen in die Kategorie Discriminative AI.

[27] Eine generelle Übersicht zu möglichen GenAI-Anwendungen findet sich beispielsweise in [56].

[28] https://cnai.swiss/dienstleistungen/projektdatenbank/

[29] https://www.post.ch/de/hilfe-und-kontakt

[30] https://www.apple.com/siri/

[31] https://alexa.com

[32] https://copilot.microsoft.com/

[33] https://www.wien.gv.at/bot/

[34] https://www.stadt.sg.ch/home/verwaltung-politik/verwaltung-dienste/chatbot.html

[35] https://chat.zuericitygpt.ch

[36] https://www.was-luzern.ch/praemienverbilligung

[37] https://www.sva-aargau.ch/sites/default/files/media/document/Medienmitteiung%20Chatbot%20Maxi.pdf

[38] https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-99829.html und https://www.fragesi.ch

Ebenfalls möglich ist der Einsatz von Chatbots für interne Nutzende, also die Mitarbeitenden der Verwaltung oder Organisation [91]. Die wesentlichen Unterschiede liegen dabei in der möglichen Anpassung der Eingaben (Prompts) und Ergebnisdarstellung an die fachspezifischen Anforderungen der Nutzenden sowie der Anbindung zusätzlicher Datenquellen, die neben internen Web-Seiten und Dokumenten via RAG (siehe Abschnitt 2.2) grundsätzlich auch Datenbanken mit einschliesst [67] und damit einen integrierten Zugriff auf alle relevanten Informationsquellen ermöglichen kann. In ihrer Studie zum Einsatz von Chatbots in verschiedenen US-amerikanischen Behörden konnten die Autoren feststellen, dass der Einsatz von Chatbots tendenziell dazu führt, dass Verwaltungsmitarbeiter/innen von einfachen Informationsanfragen entlastet werden und die gewonnene Zeit für komplexere Fälle und Aufgaben nutzen können [26]. In [91] untersuchten die Autoren eines vor allem in Norwegen eingesetzten Chatbots, der auf der Plattform von boost.ai[39] implementiert wurde, anhand von Interviews sowohl mit Bürgern und Verwaltungsangestellten. Dabei nannten beide Seiten vor allem den vereinfachten, schnelleren und qualitativ besseren Informationszugang als zentrale Vorteile. Die Bürger und Bürgerinnen vertrauten den generierten Antworten mehrheitlich. Auf Seite der Angestellten wurde zusätzlich noch die Entlastung bzw. die ermöglichte Priorisierung der Arbeitszeit hin zu komplexeren und wertstiftenden Aufgaben positiv bewertet.

Während in der Fachliteratur neben der Informationsabfrage häufig auch andere Anwendungsfälle für Chatbots genannt werden [42], wie beispielsweise die Unterstützung beim Ausfüllen von Anträgen, dem Erstellen von Einträgen im Handelsregister[40], oder der Abfrage von öffentlich verfügbaren Datensätzen (*Open Data* [42]) des BFS mit Statbot.Swiss[41] [113], handelt es sich hier häufig um forschungsnahe Pilotstudien, die sich entweder noch in einem (frühen) Entwicklungsstadium befinden, im Stadium eines Forschungsprototypen geblieben sind, oder nach kurzem Praxiseinsatz aufgrund von Fehlern, mangelnder Benutzerakzeptanz oder unzureichender Finanzierung wieder deaktiviert wurden [35].

Die technische Umsetzung eines Chatbots erfolgt heutzutage meist auf Basis eines LM in Kombination mit RAG (siehe Abschnitt 2.2). Ein solcher Chatbot ist zur Zeit beispielsweise für die vereinfachte Abfrage von statistischen Daten des Schweizerischen Parlaments in Entwicklung[42]. Für einfachere Anwendungsfälle können aber auch mit einem regelbasierten Chatbot, der anhand vordefinierter Dialogverläufe agiert, gute Ergebnisse erzielt werden [60].

3.2 Informationszugang und Suche

Mit Chatbots technisch eng verwandt und ebenfalls applikationsübergreifend umgesetzt ist die klassische Suchmaschine, deren Benutzerschnittelle sich meist an der bekannten Web-Suche [24] mit Eingabe von Schlüsselworten oder Textpassagen als Suchanfrage und der absteigend sortierten Ergebnisliste als Antwort der Suchmaschine orientiert. Eine Suchmaschine kann dabei neben Web-Seiten auch andere Datenquellen wie Textdokumente oder auch Mediendateien verarbeiten. Um die vom Benutzenden eingegebenen Schlüsselworte in einer Frage flexibler mit den Inhalten der durchsuchten Dokumente abgleichen zu können, verwendet man meist eine sogenannte semantische Suche, die auf einen inhaltlichen Vergleich statt eines rein lexikalischen Vergleichs von Texten abzielt. Für deren algorithmische Umsetzung existieren hierfür zahlreiche Verfahren [25], beispielsweise auf der Basis von Word Embeddings [74] oder zunehmend auch durch Einsatz eines LM [161]. Zusätzlich erlaubt die natürlichsprachige Schnittstelle des Chatbots eine einfachere und flexiblere Verwendung [127].

In ihrer Studie zu Art und Komplexität der Suchanfragen, die Benutzer/innen der LM-unterstützten Suchmaschine Bing Copilot[43] absetzen, fanden die Autoren einen deutlich höheren Anteil von Anfragen, die auf komplexe oder kreative Aufgabenstellungen hinweisen,

[39] https://boost.ai
[40] https://labsoflatvia.com/en/news/registry-of-enterprises-virtual-assistant-una
[41] https://www.bfs.admin.ch/bfs/en/home/dscc/blog/2023-02-statbot.html
[42] https://cnai.swiss/wp-content/uploads/2024/12/CNAI_Projekte_D_10_0-2.pdf
[43] https://bing.com/chat

als dies bei der klassischen Web-Suche, bei der es primär um das Auffinden von faktischem Wissen geht, der Fall ist [145].

Neben generischen Lösungen im Umfeld Suche wie z.B. ElasticSearch[44] existieren auch domänenspezifische Lösungen, die in der Regel auf Struktur und Inhalt der verwendeten Texte angepasst sind und darüber hinaus nützliche Zusatzfunktionalitäten anbieten. Beispielsweise kombinieren Produkte wie DeepJudge[45] und legal-i[46] semantische Suche und Funktionalitäten zur Fallbearbeitung speziell für juristische Texte, z.B. eine eingebettete Dokumentenklassifikation [94], eine graphische Darstellung des zeitlichen Ablaufs eines Falls oder automatische Zusammenfassungen. Im Kanton Zürich ist momentan eine semantische Suche für Protokolle, Beschlüsse, Gesetzestexte etc. in Entwicklung[47]. Im Auftrag der Europäischen Kommission wird zur Zeit ebenfalls eine semantische Suche für das öffentliche Finanzierungsportal «EU Funding & Tenders Portal[48]» entwickelt[49].

3.3 Office-Applikationen

Mitarbeitende der öffentlichen Verwaltung greifen im Arbeitsalltag wie andere Sachbearbeitende auch häufig auf Standardapplikationen wie Email-Programme oder Office-Applikationen zurück, beispielsweise, um mit einer Textverarbeitung Briefe oder Berichte zu erstellen oder mit einer Tabellenkalkulation Projekte zu budgetieren [12]. Ein LM kann dabei den Mitarbeitenden unterstützen, indem Textbausteine erstellt, Textpassagen zusammengefasst [126], in einen anderen Stil (z.B. einfache Sprache [50]) umgeschrieben oder eine andere Sprache übersetzt [83] werden. Im Kanton Zürich ist z.B. momentan eine Lösung zur (teil-)automatischen Vereinfachung von Texten in Entwicklung[50].

Seit 2023 können Nutzende von Microsoft 365-Applikationen[51] hierfür den integrierten KI-Assistenten Copilot[52] verwenden [143]. In ihrer Studie zu dessen Einfluss auf den Arbeitsalltag schliessen die Autoren aus ihrer empirischen Beobachtung, dass Office-Anwender aufgrund der KI-Unterstützung zwischen 10%-20% mehr Dokumente bearbeiten können, auf eine entsprechende Steigerung der Produktivität [73]. Technisch kombiniert MS Copilot dabei die lokale Office-Applikation mit einem organisations-spezifischen Microsoft 365 Cloud-Service zur Bereitstellung der gemeinsam genutzten Datenquellen um dem zentralen, ebenfalls Cloud-basierten LM (GPT4 von OpenAI)[53]. In Situationen, in denen diese tiefe technische Integration entweder nicht möglich oder nicht wünschenswert ist, können GenAI-Lösungen alternativ als Plugin (z.B. RedInk[54]) oder Popup-Dialog (z.B. LLM-for-X [147]) zur Verfügung gestellt werden. Auch in der Schweiz setzen praktisch alle öffentlichen Verwaltungen auf MS Office 365, z.B. der Bund[55].

Gerade bei Applikationen, die (noch) keine integrierte Unterstützung für KI-Funktionalität bieten, kann häufig beobachtet werden, dass Nutzende, die darauf nicht verzichten möchten, separate Applikationen (hauptsächlich Chatbots) mit privatem Account und/oder Endgerät einsetzen. Dieses Phänomen wird als *Shadow AI* [11] oder auch als “Bring your own AI (BYOAI)” bezeichnet. Bei einer auf den öffentlichen Sektor fokussierten Studie in UK gaben immerhin 22% der Befragten an, GenAI zu verwenden [20]. Und eine grossangelegte Studie von Microsoft und LinkedIn kam zum Ergebnis, dass bereits 75% der evaluierten Mitarbeitenden auf KI-Funktionalität im Arbeitsalltag zugreifen, und von diesen geben 90% an, dass sie dadurch Zeit

[44] https://www.elastic.co
[45] https://www.deepjudge.ai/product
[46] https://legal-i.ch
[47] https://www.zh.ch/de/politik-staat/kanton/kantonale-verwaltung/digitale-verwaltung/kuenstliche-intelligenz.html
[48] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/home
[49] https://commission.europa.eu/publications/artificial-intelligence-european-commission-aiec-communication_en
[50] https://www.zh.ch/de/politik-staat/kanton/kantonale-verwaltung/digitale-verwaltung/kuenstliche-intelligenz.html
[51] https://www.office.com
[52] https://copilot.microsoft.com/
[53] https://learn.microsoft.com/de-de/copilot/microsoft-365/microsoft-365-copilot-architecture
[54] https://www.vischer.com/redink-de/
[55] https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-102770.html

sparen und 85%, dass sie in der Folge mehr Zeit für aus ihrer Sicht wichtigere Aufgaben haben[56]. Gleichzeitig sind viele Organisationen bei der Umsetzung der dafür notwendige KI-Funktionalität stark im Rückstand und fast 80% der evaluierten Nutzenden verwenden Shadow AI, mit entsprechend hohen Risiken in Bezug auf Arbeitsqualität, Security und Datenschutz.

3.4 Fach-Applikationen

Neben Standardapplikationen verwenden Mitarbeitende in öffentlichen Verwaltungen und anderen Organisationen häufig Fachapplikationen[57] für spezifische Aufgaben wie die Fallbearbeitung, beispielsweise im Sozialwesen [88], bei der Einbürgerung, bei Genehmigungsverfahren etc. Für die Fallbearbeitung verwenden und erfassen die Mitarbeitenden dabei neben strukturierten Daten zum Klienten (demographische Daten, Lebenssituation etc.) und Fall (Typisierung, Finanzen, zeitlicher Verlauf etc.) vor allem Texte und Dokumente (Gesprächsprotokolle, Akten, Briefe, Formulare etc.). In Schweizer Verwaltungen wird hierfür das Fallbearbeitungssystem GEVER[58] verwendet. Gerade bei komplexeren, lange andauernden, oder durch mehrere Mitarbeitende bearbeiteten Fälle, könnte ein LM eine automatische Fallzusammenfassung für die Mitarbeitenden erzeugen, die den aktuellen Stand und die wichtigsten fallspezifischen Details übersichtlich darstellt. Nelson et al. untersuchten in ihrer Studie beispielsweise die Unterstützung von Mitarbeitenden einer US-basierten NGO bei der Betreuung von Obdachlosen durch die Erzeugung von automatischen Fallzusammenfassungen [109]. Als Vorteile nannten die Mitarbeitenden vor allem Zeitersparnis und leichtere Bestimmung der weiteren Vorgehensweise bei der Fallbearbeitung. Problematisch war dagegen, dass die Zusammenfassung je nach Fall unterschiedliche Informationen erhalten sollte und es somit keinen einheitlichen Standard gibt und von vielen Studienteilnehmenden befürchtet wird, dass wichtige Informationen fehlen oder falsch wiedergegeben werden. Technisch basierte der Prototyp auf dem Open Source-Modell Llama2 (siehe Kapitel 5.1) und wurde vor allem durch Prompt Engineering umgesetzt. In einer Studie von Roberts et al. in UK wurden proprietäre, Cloud-basierte Modelle von Google und OpenAI (siehe Kapitel 5.1) verwendet, um Patientendossiers im Umfeld der stationären plastischen Chirurgie in Form eines für den Patienten verständlichen Austrittsbriefs zusammenzufassen [131]. Als zentralen Vorteil sehen die Autoren der Studie die verbesserte Kommunikation zwischen Arzt und Patient sowie die Zeitersparnis bei der Erstellung des Austrittsbriefs, zu der die Klinik gesetzlich verpflichtet ist[59]. Problematisch war, dass die Zusammenfassungen teils fachliche Fehler enthielten und sich die Erzeugung des korrekten Ausgabeformats als schwierig erwies. Die Autoren empfehlen daher eine menschliche Nachbearbeitung gemäss dem Human-in-the-Loop-Ansatz [167]. Wie in vielen anderen Fallstudien auch ignorieren die Autoren den hier nicht datenschutzkonformen Umgang mit sensiblen Patientendaten, abgesehen von einer Erwähnung in einem Nebensatz.

3.5 Sprach-Erkennung und Protokollierung

In vielen Gesprächssituationen im Umfeld des öffentlichen Sektors werden Protokolle zur Dokumentation, Veröffentlichung oder weiteren Analyse angefertigt, beispielsweise bei Beratungsgesprächen, Gerichtsanhörungen [96], Parlamentsdebatten [158] oder der Polizeiarbeit [62]. Angesichts des hohen manuellen Aufwands für wort-wörtliche oder zusammenfassende Protokolle ist der unterstützende oder vollautomatische Einsatz einer sogenannten Spracherkennung (*Automatic Speech Recognition ASR*), die ein Audiosignal in Text transkribiert [76], besonders lohnenswert. Moderne Spracherkennungssysteme basieren technologisch auf komplexen neuronalen Netzen, den sogenannten Transformern [80], die das eingehende Audiosignal auf eine Wortsequenz abbilden, also transkribieren. Zur Qualitätsverbesserung kann diese initiale Wortsequenz noch mit einem LM abgeglichen werden,

[56] https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part
[57] Eine allgemeine Übersicht für den Kanton Luzern findet sich beispielsweise hier: https://informatik.lu.ch/Aktuelle_Projekte/Digitaler_Kanton
[58] https://www.bk.admin.ch/bk/de/home/dokumentation/gever-bund.html und https://www.fabasoft.com/de/produkte/onegov-gever-geschaeftsverwaltung
[59] https://www.gmc-uk.org/professional-standards/learning-materials/blog---montgomery--judgement

welches verschiedene Varianten dieser Sequenz bezüglich ihrer tatsächlich beobachtbaren Auftrittshäufigkeit bewerten und somit die wahrscheinlichste Variante als Endergebnis ausgewählt werden kann [125]. Für Gesprächssituationen mit mehreren Sprecher/innen wie z.B. in Gerichtssälen ist als weitere Funktionalität neben der eigentlichen Erkennung eine zusätzliche Sprecher-Differenzierung [99] bzw. Sprecher-Erkennung [149] für die Zuordnung des jeweiligen Sprechers erforderlich.

Die erzielbare Ergebnisqualität dieser Systeme hängt dabei neben dem eingesetzten ML-Modell vor allem auch an der Qualität des Audiosignals (z.B. Rauschen), der Anzahl und Variabilität der Sprecher, der Sprache selbst, ob mehrere Sprachen oder Dialekt gesprochen werden, etc. [76]. In ihrer Publikation zum Transkribieren von italienischen und polnischen Gerichtsanhörungen ermittelten Lööf et al. beispielsweise hohe Wortfehlerraten von 30%-50% [96], wenn auch unter Einsatz mittlerweile veralteter Technologien, während Garneau und Bolduc mit Hilfe von kommerziellen, Cloud-basierten Spracherkennungssystemen bei französisch-sprachigen Anhörungen deutlich tiefere Wortfehlerraten im Bereich von 15%-18% erzielen konnten [51]. Ergebnisse auf vergleichbaren Niveau wurden von Harrington beim Transkribieren von polizeilichen Befragungen im Rahmen strafrechtlicher Ermittlungen erreicht, ebenfalls unter Einsatz kommerzieller, Cloud-basierter Systeme [62], wobei die insgesamt erzielte Qualität neben dem verwendeten System auch von der Audio-Qualität und dem Sprecher-Akzent beeinflusst wurde. Wirth und Peinl konnten durch spezielle Anpassung (Fine-tuning) von ASR-Modellen dagegen Wortfehlerraten von unter 10% bei der Transkription von Debatten des deutschen Bundestages erzielen [158]. Im Kanton Zürich ist momentan eine Lösung zur Transkription von Besprechungen in Entwicklung[60].

3.6 Bilder und Videos

Analog zu den oben beschriebenen Anwendungsfällen im Zusammenhang mit Sprache und Textdokumenten können entsprechend trainierte KI- oder GenAI-Lösungen auch für die Bearbeitung, Analyse oder Erzeugung von Bildern oder Videos eingesetzt werden [39, 41]. GenAI-Lösungen wie DALL-E[61] werden beispielsweise für die prompt-basierte Erzeugung von Bildern [165] verwendet. In einer öffentlichen Verwaltung könnte dies z.B. für die Visualisierung von Bauvorhaben [118] verwendet werden. Schwerpunktmässig finden sich Anwendungen aber eher in der KI-basierten Klassifikation von Bildern [97], beispielsweise bei der Dokumentation und Verfolgung von Ordnungswidrigkeiten (Littering, Falschparken etc.) auf Basis von Fotos [78], der Identifizierung von Tatverdächtigen via Gesichtserkennung [128] oder der Vorsortierung von potentiell strafrechtlich relevantem Bildmaterial[62] in der Polizeiarbeit, der Analyse des Energieerzeugungspotentials von Photovoltaik-Anlangen auf Gebäudedächern anhand von Satellitenbildern [144] oder der Kategorisierung der Flächennutzung, ebenfalls anhand von Satellitenbildern [16]. Letzteres wurde in der Schweiz zum Zweck der statistischen Raumbeobachtung vom BFS produktiv umgesetzt[63].

3.7 Europäische Kommission als Vorreiter für produktive Lösungen

Die Europäische Kommission (EC) hat mehrere KI-basierte Lösungen für die Erstellung, Bearbeitung und Analyse von Dokumenten mit Unterstützung der 24 Amtssprachen (plus teilweise auch andere wie z.B. Arabisch oder Chinesisch) entwickeln lassen, die sich bereits im praktischen Einsatz befinden. Einige davon sind auch öffentlich verfügbar[64] und können (nach vorheriger Registrierung) entweder als Web-Applikation direkt oder über eine entsprechende API in anderen Applikationen verwendet werden. Beispielsweise ist eTranslation[65] eine mit

[60] https://www.zh.ch/de/politik-staat/kanton/kantonale-verwaltung/digitale-verwaltung/kuenstliche-intelligenz.html
[61] https://openai.com/index/dall-e-2/
[62] https://www.presseportal.de/blaulicht/pm/105578/4620141
[63] https://www.experimental.bfs.admin.ch/expstat/de/home/projekte/adele.html
[64] https://language-tools.ec.europa.eu
[65] https://commission.europa.eu/resources/etranslation_en

aktueller Technologie realisierte maschinelle Übersetzung[66], eSummary ein Tool zur Textzusammenfassung und Speech-to-Text eines für die Spracherkennung. Eine umfangreiche Roadmap[67] gibt weiter eine Übersicht zu Systemen, die sich in der Entwicklung bzw. noch in der Planung befinden. Beispielsweise sollen Mitarbeiter/innen der EC GenAI bei der Erstellung von Briefing-Unterlagen und Berichten einsetzen können. Die entwickelten KI-Lösungen setzen dabei hauptsächlich auf Open Source-LM und -Softwarebibliotheken, wobei detaillierte Angaben zu verwendeten Technologien, Implementierungspartnern und erzielter Ergebnisqualität nicht gefunden werden konnten.

3.8 Zusammenfassung

Zusammenfassend lässt sich feststellen, dass die untersuchten KI- und GenAI-Lösungen im Umfeld öffentlicher Verwaltungen und Dienste darauf abzielen, administrative Arbeitsabläufe effizienter zu gestalten, die Mitarbeitenden in ihrem Arbeitsalltag zu unterstützen und zu entlasten, und die Kommunikation und den Zugang zu Informationen und Diensten für Bürger und Bürgerinnen zu vereinfachen, zu beschleunigen und qualitativ zu verbessern. Angesichts dessen, dass Mitarbeitende im öffentlichen Sektor einen beträchtlichen Teil ihrer Arbeitszeit mit der Bearbeitung von Dokumenten verbringen (laut einer Studie von Deloitte in UK kamen die Autoren auf ein Zeitanteil von 30%[68]), bieten LM-gestützte Funktionen entsprechend ein grosses Verbesserungs- und Einsparpotential [22]. Etliche Mitarbeitende setzen GenAI bereits in ihrem Arbeitsalltag ein [20], häufig allerdings innoffiziell als Shadow AI via Web-Dienst in Form von maschineller Übersetzung oder als Chatbot zur Informationsbeschaffung. Während auf Ebene EU und vereinzelt auch in Ländern und Gemeinden GenAI-Lösungen auch produktiv umgesetzt oder verwendet werden, sind viele Institutionen noch zurückhaltend und die gefundenen Projekte befinden sich häufig noch im Forschungsstadium. Abgesehen von den oben beschriebenen GenAI-Lösungen der EU setzen die Institutionen bei der produktiven Umsetzung häufig auf die Standardprodukte der grossen Anbieter (Microsoft, OpenAI etc.), die meist Cloud-basiert sind und wenig bis gar keine Kontrolle zur Handhabung der Daten bieten.

[66] Die verwendeten Trainingsdaten sind ebenfalls öffentlich: https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en
[67] https://commission.europa.eu/publications/artificial-intelligence-european-commission-aiec-communication_en
[68] https://www2.deloitte.com/content/dam/insights/us/articles/3834_How-much-time-and-money-can-AI-save-government/DUP_How-much-time-and-money-can-AI-save-government.pdf

4 Ausgangslage

In diesem Kapitel beleuchten wir in Abschnitt 4.1 zunächst die strategischen Grundlagen für die Nutzung von KI-Lösungen in öffentlichen Verwaltungen. Abschnitt 4.2 diskutiert die allgemeinen rechtlichen und ethischen Rahmenbedingungen inklusive Datenschutz, deren Einhaltung bei der Umsetzung einer KI-Lösung zwingend zu beachten sind. Anschliessend definieren wir in Abschnitt 4.3 auf Basis der in Kapitel 3 vorgestellten Praxisbeispiele vier exemplarische Anwendungsfälle, welche zur Konkretisierung der nachfolgenden Kapitel genutzt werden.

4.1 Strategie

In zahlreichen öffentlichen Verwaltungen wird die **Digitalisierung** [82] der Amtsgeschäfte mit grossem Einsatz vorangetrieben [9, 151]. In der Schweiz gibt es entsprechende Digitalisierungsstrategien sowohl beim Schweizerischen Bund[69] als auch in allen Kantonen[70]. Für die Bürgerinnen sichtbar wird dies beispielsweise durch umfangreiche Web-Auftritte oder Online-Schalter, so auch im Kanton Luzern[71]. Ein häufig genanntes Ziel im Zusammenhang mit der angestrebten Digitalisierung ist die sogenannte **digitale Souveränität**, die eine Abhängigkeit von externen Technologieanbietern verhindern sowie Kontrolle und Transparenz über Prozesse und Daten ermöglichen soll [124]. Wenn KI-Lösungen zentrale Aufgaben in Verwaltungsprozessen unterstützen oder vollständig übernehmen, besteht generell ein gewisses Risiko von intransparenten, fehlerhaften oder unfairen Entscheidungen und Prozessen [101], welches durch die Auslagerung von Entwicklung und Betrieb an einen Drittanbieter tendenziell verstärkt wird [103, 106]. Mangelnde Transparenz wird häufig auch in entsprechenden Bürgerumfragen als mögliche Gefahr beim Einsatz von KI-basierten Lösungen in der öffentlichen Verwaltung genannt [163].

Organisationen sollten allgemein ihre Daten als wertvolle und wertstiftende Ressource («Asset») betrachten und diese systematisch nutzen [81]. Grundlage dafür ist eine organisationsübergreifende **Datenstrategie**, die den Umgang mit (internen und externen) Daten definiert und regelt, wie diese erfasst, aufbereitet, integriert, gespeichert, verwendet und verwaltet werden, um die strategischen und operativen Ziele der Organisation zu erreichen [33]. Typische Ziele einer Datenstrategie sind dabei z.B. das Aufbrechen von Datensilos zur organisationsübergreifenden Datennutzung [121] oder die systematische Umsetzung von KI-Projekten [4, 86]. Beispielsweise formuliert die Datenstrategie des eidgenössisches Departements für Umwelt, Verkehr, Energie und Kommunikation (UVEK) die zentralen Ziele folgendermassen: «Unternehmerische Potenziale von Datenprodukten [zu] erkennen und nutzen … durch die Anwendung von Datenanalyse, Data Science, Modellierungstechniken und anderen Methoden sowie die Einbindung von Daten in Entscheidungsprozesse» und als Grundlagen hierfür «eine starke Datenkultur zu schaffen und die Datenkompetenz auf breiter Ebene zu verbessern» sowie «Daten [zu] harmonisieren um deren Nutzung und Wiederverwendung zu ermöglichen»[72]. Analoge Datenstrategien finden sich auch auf kantonaler Ebene, z.B. in den Kanton Zürich[73] und Luzern[74].

Ergänzt wird eine Datenstrategie in öffentlichen Verwaltungen zunehmend durch eine dedizierte **KI-Strategie**, die meist schwerpunktmässig auf die Anwendungsfelder Informationsabruf, Verwaltungsprozesse und Diensterbringung abzielt [66]. Eine entsprechende KI-Strategie wurde beispielsweise von der Schweizerischen Bundeskanzlei[75] und

[69] https://www.bk.admin.ch/bk/de/home/digitale-transformation-ikt-lenkung/digitale-bundesverwaltung.html
[70] https://www.digitale-verwaltung-schweiz.ch/Kantonale-Digitalisierungsstrategien und https://www.lu.ch/-/klu/ris/cdws/document?fileid=7ab6eb240b054e1ea928fe688a3cac46 als Beispiel für den Kanton Luzern
[71] https://my.lu.ch/
[72] https://www.uvek.admin.ch/dam/uvek/de/dokumente/dasuvek/datenstrategie-uvek-2024-2027.pdf.download.pdf/datenstrategie-uvek-2024-2027.pdf
[73] https://www.zh.ch/de/politik-staat/kanton/kantonale-verwaltung/digitale-verwaltung/strategische-initiativen.html
[74] https://www.lu.ch/-/klu/ris/cdws/document?fileid=7ab6eb240b054e1ea928fe688a3cac46
[75] https://www.bk.admin.ch/bk/de/home/digitale-transformation-ikt-lenkung/ikt-vorgaben/strategien-teilstrategien/sb021-strategie-einsatz-von-ki-systemen-in-der-bundesverwaltung.html

vom Kanton Zürich[76] herausgegeben. KI-Lösungen zeichnen sich im Allgemeinen durch sehr hohe Anforderungen an die Daten selbst (in Bezug auf den notwendigen Umfang, Detailierungsgrad und Qualität), die technische Infrastruktur (zum Speichern und Verarbeiten der Daten) und insbesondere auch die Qualifikation der Mitarbeitenden (in Bezug auf Entwicklung und Anwendung solcher Produkte). Eine Daten- und/oder KI-Strategie adressiert weiterhin neben den technischen Anforderungen auch die wirtschaftlichen und organisatorischen Rahmenbedingungen [153], die bei der Umsetzung zu beachten sind. Zusätzlich werden auch die rechtlichen und ethischen Leitlinien definiert, beispielsweise zum Umgang mit sensiblen personenbezogenen Daten [49]. Da es gerade hier in Bezug auf KI-Lösungen noch etliche offene Fragen bzw. neue regulatorische Entwicklungen gibt (siehe Abschnitt 4.2), hat beispielsweise der Kanton Zürich eine umfassende wissenschaftliche Studie ausarbeiten lassen [19].

Bei der Umsetzung einer Datenstrategie sind dann alle Mitarbeitenden gefordert, die ihre täglichen Aufgaben mit Hilfe der Daten besser, also mit höherer Qualität oder grösserer Effizienz erledigen sollen. Dafür braucht es eine sogenannte **Datenkultur**, bei der es zur gelebten Normalität wird, dass Mitarbeitende zielgerichtet mit Daten umgehen und diese effektiv und effizient einsetzen können [36]. Dies wiederum setzt voraus, dass die Mitarbeitenden auch über die notwendigen Kompetenzen verfügen bzw. sich diese aneignen können und möchten [37]. Zur Verankerung der Datenkultur sollten datenspezifische Aufgaben und Verantwortlichkeiten innerhalb der Organisation für alle transparent festgelegt werden, beispielsweise wer die Einhaltung der Richtlinien zum Datenschutz wie überprüft oder wie vorzugehen ist, wenn mangelnde Datenqualität beobachtet wird. Hierfür ist es empfehlenswert, dedizierte Rollen innerhalb einer Organisation zu etablieren, wie beispielsweise die des **Data Steward**s, der übergreifend für Aufgaben der Data Governance wie Datenqualität und Meta-Datenmanagement zuständig ist [123].

Um die Datenstrategie greifbar und schrittweise umsetzbar zu machen, sollte weiterhin eine technische **Datenarchitektur** definiert werden, welche die für die Umsetzung benötigten technischen Komponenten (wie Datenquellen, Datenschnittstellen, Datenspeicher, Tools für die Datenanalyse etc.) in einer Gesamtübersicht festlegt [31].

Bei einer erfolgreichen Umsetzung der Datenstrategie entstehen neben diesen strukturellen auch konkrete Ergebnisse, die einen unmittelbaren Mehrwert für die Mitarbeitenden bieten und entsprechend auch für deren Motivation zur aktiven Mitgestaltung von zentraler Bedeutung sind: Zunächst die Anwendungen (**Datenprodukte**) selbst, die den Mitarbeitenden mit Hilfe der Daten beispielsweise Informationen bereitstellen oder sie durch (teil-)automatisierte Prozesse entlasten. Im Kanton Zürich sind einige dieser Datenprodukte auch als Open Source verfügbar[77]. Als zentrale Dokumentation aller vorhandenen Daten und ihrer Beschreibung durch verschiedene Meta-Daten (wie Quelle, Format, Semantik, Verwendung, Zugriffsrechte bzw. Einschränkungen etc.) ist weiter ein **Datenkatalog** zu erstellen und zu pflegen [117], um das aus unserer Erfahrung weit verbreitete Problem zu verhindern, dass selbst die IT-Verantwortlichen häufig nicht (mehr) wissen, welche Daten für welchen Zweck in einer Datenbank oder Datei abgelegt sind. Behördenübergreifend verantwortet in der Schweiz das Bundesamt für Statistik (BFS) im Rahmen des Programms «Nationale Datenbewirtschaftung (NaDB)»[78] eine Harmonisierung der Datensätze der öffentlichen Verwaltung, die in einem zentralen Datenkatalog I14Y[79] gesammelt und verwaltet werden.

Eine Datenkultur beinhaltet neben der Verfügbarkeit dieser technischen Komponenten aber auch Formate für den regelmässigen Informationsaustausch (Interpretation) zu allen Datenaspekten, z.B. Meetings oder Newsletter zu Projekterfolgen oder neuen Entwicklungen.

[76] https://www.zh.ch/de/politik-staat/kanton/kantonale-verwaltung/digitale-verwaltung/kuenstliche-intelligenz.html
[77] https://github.com/machinelearningZH
[78] https://www.bfs.admin.ch/bfs/de/home/nadb/nadb.html
[79] https://www.i14y.admin.ch/en/home

4.2 Rechtliche und ethische Rahmenbedingungen

Aus rechtlicher Sicht sind für den Einsatz von KI in der öffentlichen Verwaltung eine Reihe von allgemeingültigen Grundsätzen relevant. Von zentraler Bedeutung sind der menschenzentrierte Einsatz von KI, die Verhinderung von Diskriminierung (Voreingenommenheit) und der Schutz anderer Menschenrechte, die Pflicht zur Begründung rechtlicher Entscheidungen sowie Privatsphäre und Datenschutzrechte. Im Rahmen des Untersuchungsgrundsatzes muss der Datensatz für das Training und die Nutzung von ML-Modellen vollständig und genau sein. **Transparenz**, Erklärbarkeit und Rechenschaftspflicht von ML-Algorithmen sind von grundlegender Bedeutung, um die Kontrollmöglichkeit KI-basierter (Verwaltungs-)Entscheidungen durch Justizbehörden und die Zivilgesellschaft zu gewährleisten [19, 55, 135]. Ausserdem haben Betroffene gemäss Bundesverfassung ein Recht auf vorgängige Äusserung und Mitwirkung in Verwaltungsverfahren, welches durch eine allfällige (Teil-)Automatisierung eingeschränkt sein könnte [19]. Zumindest wird empfohlen, Entscheidungen nur dann automatisiert durchzuführen, wenn sie im Sinne des Betroffenen positiv ausfallen [129].

Ein weiterer zentraler Rechtsbegriff ist der Ermessensspielraum von Behörden. Bei Ermessensentscheidungen gilt KI als unzulässig oder zumindest kritisch [17, 19, 53]. KI kann hingegen dort eingesetzt werden, wo Entscheidungen oder Teile davon einem faktischen Beweis zugänglich sind [122], typischerweise in der Massenverwaltung (z.B. Steuerverfahren oder Sozialversicherungsverfahren). Statt einer generellen Freigabe von KI ist zusätzlich noch das bereichsspezifische Verwaltungsrecht zu berücksichtigen [53, 148]. Die rechtlichen Voraussetzungen hängen somit massgeblich davon ab, in welchem Verwaltungsbereich KI eingesetzt werden soll, welche genaue Anwendung vorgesehen ist, und welchen Automatisierungsgrad die Lösung haben soll [63]. Insbesondere wenn schützenswerte Personendaten verwendet werden, braucht es nach Einschätzung von Rechtsexperten eine formelle, durch die Legislative erlassene Ermächtigungsgrundlage [18]. Dies würde beispielsweise auch einen Chatbot betreffen, der Auskunft zu personenbezogenen Daten gibt (z.B. bei Anfrage zum Bearbeitungsstatus eines eingereichten Anliegens) oder der beim Ausfüllen eines Antrags unterstützt.

Eine grundlegende Herausforderung beim Einsatz von KI-Lösungen entsteht dadurch, dass ein trainiertes ML-Modell praktisch immer eine gewisse Restfehlerquote aufweist, d.h., dass statistisch gesehen ein gewisser Prozentsatz der generierten Ergebnisse falsch ist (siehe Abschnitt 2.3). Bei einem Chatbot würde dies konkret heissen, dass eine Anfrage falsch beantwortet oder ein Formular fehlerhaft ausgefüllt würde. Problematisch ist dies, weil öffentliche Verwaltungen einer hohen Sorgfaltsflicht unterliegen und Bürgerinnen und Bürger behördlichen Auskünften und Entscheidungen grundsätzlich vertrauen können sollen (Vertrauensschutz) [59]. Somit gilt es, eine praktikable Lösung für den Umgang mit dem Restrisiko zu finden.

Je nach Anwendungsfall können bei Training und/oder Anwendung eines ML-Modells personenbezogene Daten zum Einsatz kommen [46], beispielsweise demographische Daten oder Beobachtungsdaten zum Benutzerverhalten [43]. Bei der Verarbeitung von personenbezogenen Daten greifen generell die jeweiligen Gesetze zum **Datenschutz**, also auf Ebene EU die DSGVO [43] bzw. in der Schweiz das dazu kompatible revidierte Datenschutzgesetz DSG [136], wobei es im Detail Unterschiede beispielsweise zur praktischen Umsetzung, zu Meldepflichten oder zu möglichen Sanktionen gibt[80]. Die wichtigsten Grundpflichten gemäss DSGVO sind, dass es für die Verarbeitung der personenbezogenen Daten eine Rechtsgrundlage gibt (z.B. durch Einwilligung) und dass die Verarbeitung nachweislich für den Betroffenen nachvollziehbar (transparent), zweckgebunden, auf den notwendigen Umfang inhaltlich und die notwendige Speicherdauer zeitlich beschränkt, und vor unbefugtem Zugriff geschützt zu verarbeiten sind [82]. Weiter gelten für die Umsetzung dieser Grundpflichten die Prinzipen «Privacy by Design», also dass relevante Funktionen bei der Lösungsentwicklung von Beginn an mit realisiert werden, und «Privacy by Default», also dass alle Funktionen ohne Eingreifen des Nutzenden in der geforderten Qualität zur Verfügung

[80] https://www.pwc.ch/de/insights/regulierung/dsg-vergleich-dsgvo.html und https://swissprivacy.law/55/

stehen [136]. Betroffene haben weiter das Recht auf Auskunft bezüglich Speicherung und Verarbeitung ihrer personenbezogenen Daten (Informationspflicht). Allgemein gilt weiterhin aufgrund des Untersuchungsgrundsatzes im Verwaltungsrecht, dass alle für eine Entscheidung verwendeten Daten korrekt, aktuell und vollständig sein müssen [19].

Bei der weitverbreiteten Anwendung von GenAI in Form eines Chatbots (siehe Kapitel 3) sind beispielsweise auch die Gesprächs- und Kommunikationsdaten (IP-Adresse, Geräteprofile etc.) personenbezogen und ggf. zusätzlich die Profildaten bei Verwendung eines entsprechenden Nutzerkontos [63]. Im Sinne der Datensparsamkeit dürfen personenbezogene Daten nur für den eigentlichen Zweck erhoben werden, bei reinen Informationsabfragen (z.B. nach Öffnungszeiten eines Amtes) braucht es beispielsweise keine demographischen Angaben, auch wenn eine personalisierte Anrede mit dem Namen aus Sicht des UI-Designs positiv sein könnte. Ausserdem sind die personenbezogenen Daten nach der Verwendung zu löschen – oder deren Weiterverwendung durch eine entsprechende gesetzliche Grundlage zu ermöglichen [19]. Der Benutzende ist zudem vorgängig über Funktionsweise des Chatbots und Verwendung der Daten zu informieren und muss dieser Verwendung auch zustimmen. Der Umgang mit personenbezogenen Daten ist weiter auch für das Training des zugrundeliegenden LM relevant (siehe Kapitel 6).

Verzichtet eine KI-Lösung dagegen auf die Verwendung von personenbezogenen Daten, greift der Datenschutz dagegen nicht. Technisch sinnvoll und möglich ist dies beispielsweise bei einem Chatbot, der rein für allgemeine Auskünfte wie Öffnungszeiten etc. zuständig ist. Alternativ kann der Personenbezug in Daten durch Anonymisierung [98] so entfernt werden, dass die Wiederherstellung des Bezugs nicht oder nur mit unverhältnismässigem Aufwand möglich ist [160].

Eine GenAI-Lösung darf Betroffene nicht diskriminieren, also ungerechtfertigt aufgrund von bestimmten Persönlichkeitsmerkmalen wie Herkunft, Alter oder Geschlecht benachteiligen [132]. Eine häufige Ursache von **Diskriminierung** durch KI ist, dass der zugrundeliegende Trainingsdatensatz falsch oder statistisch unausgewogen ist oder Beispiele für bewusste oder unbewusste Diskriminierung durch einen Menschen enthält, was sich dann auch im resultierenden ML-Modell widerspiegelt und daher in zukünftigen Vorhersagen reproduziert wird [114]. Solche Modelle werden als voreingenommen (*biased*) bezeichnet und verletzen das Diskriminierungsverbot. Ein Bias kann aber auch (bewusst oder unbewusst) durch die Systementwickler verursacht werden, die zahlreiche Detail-Entscheidungen treffen und Entwicklungsalternativen abzuwägen haben [84]. Aufgrund dieser komplexen Herausforderungen ist es nicht weiter verwunderlich, dass in etlichen Praxisbeispielen nach der Produktivstellung Verstösse gegen das Diskriminierungsverbot festgestellt wurden, z.B. durch Organisationen wie AlgorithmWatch[81]. In der Studie von Eubanks [44] wurden beispielsweise mehrere Fälle aufgedeckt, in denen die entwickelten KI-Lösungen unfair sind und besonders arme Bürger/innen diskriminieren. Und eine in Australien eingesetzte Spracherkennungssoftware, welche die Englischkenntnisse von Bewerbenden auf ein Arbeitsvisum beurteilte, wurde eine irisch-stämmige englischsprachige Muttersprachlerin als ungenügend eingestuft[82], offensichtlich wegen unzureichender Trainingsdaten.

Im Rahmen des EU-Gesetzes zur Künstlichen Intelligenz (EU AI Act[83]) werden GenAI-Lösungen auch aufgrund der gegebenen Diskriminierungsgefahr in verschiedene Risikoklassen mit darauf abgestuften Anforderungen eingeteilt [132].

Diskriminierung lässt sich nur vermeiden, wenn Trainingsdaten und resultierendes ML-Modell in Bezug auf Korrektheit und repräsentative Abdeckung sorgfältig ausgewählt und auf potenziell enthaltene Diskriminierung analysiert werden. Da es dennoch keinen vollständigen Schutz vor Diskriminierung (und anderen potenziellen Fehlern) gibt wird auch empfohlen, dass KI-Lösungen Verfahren nicht vollständig automatisieren, sondern der Unterstützung eines menschlichen Entscheidungsträgers dienen sollten, damit dieser allfällige Diskriminierungen

[81] https://algorithmwatch.ch/de/was-ist-algorithmische-diskriminierung/
[82] https://www.abc.net.au/news/2017-08-09/voice-recognition-computer-native-english-speaker-visa-limbo/8789076
[83] https://artificialintelligenceact.eu/de/

und sonstige Fehler entdecken und korrigieren kann [19]. Aber auch bei diesem sogenannten *Human-in-the-Loop*-Ansatz [167] besteht immer noch das Risiko, dass eine falsche Vorhersage von einem unerfahrenen oder allzu vertrauensseligen Entscheidungsträger/innen unentdeckt bleibt. Dieser sogenannte Automatisierungs-Bias [141] wurde auch für den öffentlichen Sektor berichtet [44]. Als Abhilfe wird *Explainable AI (XAI)* empfohlen, d.h. die Bereitstellung einer für Menschen interpretierbaren Erklärung zusätzlich zum Ergebnis [58], was insbesondere für nicht-technische Nutzende ohne tieferes Verständnis von KI wichtig ist [90] und auch die Vertrauensbildung im Allgemeinen unterstützt [2, 57]. Vertrauen sollte andererseits aber auch kein blindes Vertrauen auf Seiten des Anwendenden sein, sondern angesichts von potenziellen Fehlern des ML-Modells wie dem möglichen Halluzinieren von LM auch das kritische Auseinandersetzen mit dem maschinell erzeugten Ergebnis beinhalten, was wiederum XAI bedingt. Allerdings wurde in einer Fallstudie mit Polizeikräften festgestellt, dass auch modernste XAI-Verfahren einen Bestätigungs-Bias nicht verhindern konnten, also dass die Studienteilnehmende nur denjenigen ML-Ergebnissen vertrauten, die auch ihrer eigenen Einschätzung entsprachen [138].

Die rechtliche Anforderung der **Begründungspflicht** besagt, dass Betroffene einer Verwaltungsentscheidung das Recht auf eine sachliche Begründung haben, um die Entscheidung nachvollziehen und ggf. anfechten zu können [18]. Die Begründungspflicht betrifft die gesamte ML-Verarbeitungspipeline [6], einschliesslich der Erstellung des zugrunde liegenden Datensatzes [69], der Entwicklung und des Trainings des ML-Modells [30] und seiner Anwendung in einem Verwaltungsprozess [137]. Um der Begründungspflicht Genüge zu tun, wird empfohlen, dass ML-basierte Lösungen transparent (d. h. Entwicklungsprozess, Daten und Code offenlegen), interpretierbar (vgl. XAI) [133] und formal beaufsichtigt [142] sein sollten. Für Betroffene sollte damit die Entscheidungslogik, die verwendeten Daten, und allfällige Besonderheiten des konkreten Falls einsehbar sein [19].

Diese allgemeinen Richtlinien in der Praxis tatsächlich umzusetzen ist angesichts der Verwendung von umfangreichen Datensätzen mit hochdimensionalen Merkmalen und komplexen ML-Algorithmen wie LM eine grosse Herausforderung [54], entsprechend bezeichnet man diese Lösungen auch als *Black Box*. Ausserdem richten sich viele Lösungen eher an ML-Experten und Expertinnen als an nicht-technische Endnutzende in öffentlichen Verwaltungen [90, 101].

Bei der Betrachtung der rechtlichen Rahmenbedingungen für KI-Lösungen ist zu beachten, dass spezifische gesetzliche Regelungen gerade erst kürzlich verabschiedet wurden bzw. erst noch in der Diskussion sind. Der bereits erwähnte EU AI Act wurde beispielsweise erst 2024 verabschiedet und die praktischen Implikationen werden noch intensiv und kontrovers diskutiert [23, 52]. Im Schweizer Parlament sind in den letzten Jahren weit über 70 Vorstösse zum Thema KI eingereicht worden[84] und die weitgehende Übernahme des EU AI Act ist zwar geplant, steht aber noch aus[85].

4.3 Exemplarische Anwendungsfälle

Angesichts der vielfältigen Einsatzmöglichkeiten von ML und GenAI (vgl. Kapitel 3) und der damit verbundenen unterschiedlichen Anforderungen, fokussieren wir uns in der weiteren detaillierteren Betrachtung auf die folgenden exemplarischen Anwendungsfälle (*Use Cases*). Die Anwendungsfälle werden zunächst rein funktional beschrieben, ohne sich auf eine spezifische Umsetzung oder ein bestimmtes Modell festzulegen; die möglichen technischen Optionen hierzu werden in Kapitel 5 und die Implikationen für den Datenschutz in Kapitel 6 beschrieben.

UC1 «Chatbot für Bürger/innen»

Bürger und Bürgerinnen können mit Hilfe eines Chatbots gezielt Informationen zu verschiedenen Themen wie Öffnungszeiten und Standorten von Ämtern, Bauvorschriften, Parkregelungen, etc. abfragen. Als Benutzerschnittstelle dient typischerweise ein

[84] https://www.parlament.ch/de/suche
[85] https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-104110.html

eingeblendetes Fenster auf einer einschlägigen Webseite. Die vom Chatbot generierte Antwort basiert auf öffentlich zugänglichen Informationsquellen wie spezifischen Webseiten, Hilfeseiten, Diskussionsforen oder auch offiziellen Dokumenten (Formulare, Wegleitungen, Regularien etc.) und sollte als zusätzliche Erklärung auch explizit auf die jeweilige Quelle verweisen. Der Chatbot sollte weiter in der Lage sein, sich an den Gesprächsverlauf anzupassen und beispielsweise Referenzen auf zurückliegende Äusserungen interpretieren zu können. Weitere wünschenswerte Funktionen sind die automatische oder vom Bürger oder von der Bürgerin angefragte Weiterleitung zu einem menschlichen Gesprächspartner, wobei ebenfalls der bisherige Chatverlauf übergeben werden sollte und die Unterstützung der Mehrsprachigkeit bei Ein- und Ausgabe. Für Qualitätsprüfung und als Grundlage für die technische Weiterentwicklung sollten Gesprächsverläufe ausgewertet werden können.

Als technische Basis zur Umsetzung eines Chatbots dient ein LM, welches speziell auf die relevanten Datenquellen (Webseiten und Dokumente) angepasst wird, beispielsweise durch Einbindung einer klassischen Suchmaschine mit Hilfe von RAG[86].

UC2 «Office-Applikationen»

Mitarbeitende der öffentlichen Verwaltung verwenden im Arbeitsalltag Office-Applikationen wie beispielsweise eine Textverarbeitung, um Briefe oder Berichte zu erstellen. Ein LM kann dabei den Mitarbeitenden unterstützen, indem Textbausteine erstellt, Textpassagen zusammengefasst, in einen anderen Stil (z.B. einfache Sprache) umgeschrieben oder eine andere Sprache übersetzt werden [73]. Diese verschiedenen Funktionen werden bei Bedarf vom Mitarbeitenden aufgerufen, beispielsweise per Kontextmenü in Bezug auf eine zuvor markierte Textpassage. Die Daten werden dabei aus dem Dokument extrahiert, an das LM übergeben, und das Ergebnis wieder in das Dokument eingefügt. Für Qualitätsprüfung und als Grundlage für die technische Weiterentwicklung sollten ausgewertet werden können, wobei datenschutzrechtliche Vorgaben stets berücksichtigt und sensible Daten angemessen geschützt werden müssen.

UC3 «Fallbearbeitung mit Fachapplikation»

Mitarbeitende in öffentlichen Verwaltungen und anderen Organisationen verwenden häufig Fachapplikationen für die spezifische Fallbearbeitung, beispielsweise im Sozialwesen [88]. Die Mitarbeitende verwenden dabei neben strukturierten Daten zum Klienten (demographische Daten, Lebenssituation etc.) und Fall (Typisierung, Finanzen, zeitlicher Verlauf etc.) vor allem zahlreiche Texte und Dokumente (Gesprächsprotokolle, Aktennotizen, Briefe, Formulare etc.). Eine automatisch generierte Fallzusammenfassung soll für den Mitarbeitenden den aktuellen Status und die wichtigsten fallspezifischen Details aus diesen unterschiedlichen Datenquellen übersichtlich darstellen. Je nach Aufgabe sollte diese Zusammenfassung durch den Mitarbeitenden in Bezug auf Inhalt, Länge und Darstellung konfigurierbar sein. Zur Nachbearbeitung, Qualitätsprüfung und als Grundlage für die technische Weiterentwicklung sollte ein Mitarbeitender generierte Zusammenfassungen kommentieren oder verbessern können.

UC4 «Sprach-Erkennung und -Protokollierung»

In vielen Gesprächssituationen im Umfeld des öffentlichen Sektors sehen sich Mitarbeitende mit der Erstellung von Protokollen zur Dokumentation, Veröffentlichung oder weiteren Analyse konfrontiert, beispielsweise bei Beratungsgesprächen oder Gerichtsanhörungen. Durch Einsatz eines geeigneten Spracherkennungssystems, welches ein gegebenes Sprach-Audiosignal in Text transkribiert, sollen Mitarbeitende bei der Protokollerstellung unterstützt werden. Das System sollte die gängigsten Sprachen und Dialekte verarbeiten, mit fachspezifischem Vokabular umgehen und in Situationen mit mehreren Sprechenden zwischen diesen differenzieren können. Zur Nachbearbeitung, Qualitätsprüfung und als Grundlage für die technische Weiterentwicklung sollte ein Mitarbeitender generierte Protokolle kommentieren oder verbessern können.

[86] https://www.elastic.co/search-labs/tutorials/chatbot-tutorial/welcome

5 Lösungsansätze für den Einsatz von GenAI

In diesem Kapitel werden vier strategische Ansätze für den Einsatz von GenAI skizziert, wobei der Fokus auf der Auswahl zwischen öffentlichen und privaten Sprachmodellen liegt. Techniken wie das Fine-Tuning und Retrieval Augmented Generation werden in diesem Kontext kurz erklärt. Weiter werden Best Practices für die Sicherheit des Betriebs einer solchen Umgebung sowie des Prompt-Managements aufgezeigt.

5.1 Modellauswahl

Beim Betrieb von *Language Models* (LM) gibt es verschiedene Paradigmen, die angewandt werden können. Man verwendet *Public Models*, z.B. ChatGPT, mit deren proprietären Tools resp. Web Frontends interaktiv, oder indirekt über Schnittstellen (*Application Programming Interface,* API). Diese Art der Anwendung ist pragmatisch, aber auch limitiert. Nur einfache Anwendungsfälle, z.B. Formulieren, Zusammenfassen, Rechtschreibeprüfung, Übersetzen, etc. von nicht schützenswerten Daten (in Form von Dokumenten, Textblöcken, etc.) können so durchgeführt werden. Da diese *Foundation Models*[87] (Basismodelle) sehr generisch sind, muss allenfalls eine Anpassung an die eigene Domäne erfolgen. Aufgrund der rechenintensiven Anforderungen für eine Adaptation ist dies nur eingeschränkt möglich. Als eine Möglichkeit zur lokalen Anpassung bietet sich die Auswahl von vorgegebenen Prompts (Instruktionen) aus einer kuratierten Sammlung an; eine *Prompt Library*.

Im Gegensatz dazu stehen die *Private Models*: diese werden in einer kontrollierten und sicheren Umgebung installiert und betrieben, um die Integrität der Datenflüsse und der Datenspeicherung zu gewährleisten. Auch hier können die Modelle mittels Schnittstellen oder Tools angesprochen werden, gleich wie bei den Public Models. Dabei kann in der Regel auf *in-house* Ressourcen zurückgegriffen werden, oder ein spezialisierter, vertrauenswürdiger Hosting Dienst kann diesen Service direkt anbieten. Private Modelle haben weitere Charakteristika, die sie von den öffentlichen Modellen unterscheiden. Private Modelle können lokal trainiert werden; entweder in Form von *fully-trained* oder *pre-trained*. Der Aufwand dabei ist aber bei lokalen LM beträchtlich, sehr variabel, und es ist nicht garantiert, dass ein gutes, angepasstes Modell resultiert. Mit *fine-tuning* kann ein vortrainiertes Modell weiter an die Domäne angepasst werden. Dabei sind die wichtigsten Voraussetzungen ein qualitativ hochwertiger, aufgabenspezifischer Trainingsdatensatz, ausreichende Rechenressourcen und das nötige Fachwissen. Das vortrainierte Modell wird entweder mit allen oder einer Untermenge seiner Parameter entsprechend weiter angepasst. Ein typisches Verfahren ist die *low-rank adaptation*[88]. *Transfer Learning* ist dabei ein allgemeiner Begriff im Kontext des fine tunings von Modellen.

Als Alternative zu einem Training bietet sich *Retrieval Augmented Generation* (RAG)[89] an. RAG erweitert dabei die Fähigkeiten von LM, dynamisch Domänenwissen aus einer Dokumentendatenbank in den Abfrageprozess zu integrieren. Das LM generiert dabei eine Antwort, die sowohl auf seinem Modellwissen als auch auf den aktuell abgerufenen Informationen aus dem Retrieval basiert, was zu präziseren und besser verifizierbaren Ergebnissen führt; das potenzielle Halluzinieren wird dabei reduziert.

Eine Optimierungsmöglichkeit gibt es mit den *Small Language Models* (SLM), die einen optimierten und daher kleineren Parameterraum im Vergleich zu den grossen LM aufweisen. Deren Grösse ist dabei nicht genau definiert und variiert je nach Anbieter. Allen gemeinsam ist, dass die SLM von verschiedenen Anbietern zur lokalen Verwendung zur Verfügung gestellt werden. Diese Art der Modelle haben aus Ausgangsbasis immer ein Foundation Model und wurden kondensiert, damit sie in einem Infrastruktur Umfeld betrieben werden können, wo Ressourcen nur limitiert zur Verfügung stehen. Durch den vorgängigen Komprimierungsvorgang ist es nun aber wieder möglich, ein volles Training zur Prägung dieser

[87] Im Gegensatz zu den *Frontier Models*, die jeweils Gegenstand der aktuellen Forschung sind.
[88] https://en.wikipedia.org/wiki/Fine-tuning_(deep_learning)#Low-rank_adaptation
[89] https://de.wikipedia.org/wiki/Retrieval_Augmented_Generation

Sprachmodelle durchzuführen. Wie bei allen Trainingsvarianten spielt auch hier die Qualität der Trainingsdaten eine zentrale Rolle. Ein Nachteil von SLM, bedingt durch ihre interne Datenstruktur (Quantisierung), ist ihre Tendenz des Halluzinierens [150, 157].

Im Open-Source- bzw. Open-Weight-Umfeld [66] lokaler Sprachmodelle konkurrieren derzeit verschiedene Open Models miteinander:

- DeepSeek[90]: 70 Milliarden Parameter
- Dolly[91]: 12 Milliarden Parameter
- Falcon[92]: 40 Milliarden Parameter
- Gemma[93]: 4 – 27 Milliarden Parameter
- Llama[94]: 70 Milliarden Parameter
- Mixtral[95]: gemischter Parameterraum, 8 x 7 Milliarden
- Pharia[96]: 7 Milliarden Parameter
- Phi[97]: 14 Milliarden Parameter

Ein Vergleich der Leistung von Sprachmodellen kann in der *LM Arena*[98] interaktiv angefragt werden: man lässt die Modelle gegeneinander antreten, indem man Anfragen durchführt und die Antworten darauf in Bezug auf deren Qualität beurteilt.

Ein Augenmerk sollte auch auf die Herkunft (EU, China, USA) solcher Modelle und deren z.T. spezialisierter Anwendungsgebiete (Sprachkompetenz, Textgenerierung; Image-to-Text, Speech-to-Text STT, etc.) gelegt werden. Ein wichtiges Kriterium ist das *Kontextfenster* dieser Modelle. Dieses definiert dabei die Menge an Text (Tokens, wobei ein Wort ca. 1.5-2 Tokens entspricht), die es gleichzeitig verarbeiten kann und ist entscheidend für das Verständnis langer Abhängigkeiten und die Kohärenz der Ausgaben. Grössere Kontextfenster ermöglichen komplexere Aufgaben, erfordern aber auch mehr Rechenleistung und Speicher. Weiter stellt sich die Frage, ob die Modelle über eine dokumentierte API verfügen, damit nicht nur interaktiv, also direkt durch Anwendende, sondern auch via Schnittstellen, mit diesen Modellen interagiert werden kann. Der Betrieb der kleinen Modelle kann lokal erfolgen, sofern die notwendigen Hardwareressourcen vorhanden sind. Aktuelle GPU/NPU Inferenzsysteme, geeignet für einfache Anwendungsfälle, können mit diesen Modellen gut umgehen.

Für die An- und Einbindung in einen potenziellen Workflow eignen sich *OpenWeb UI*[99], *Ollama*[100] oder *LM Studio*[101]. Die Kombination OpenWeb UI/Ollama ist aktuell an der Berner Fachhochschule (BFH-TI) für Forschung und Lehre im Einsatz. Der Plattform-unabhängige Service Ollama läuft dabei auf einem Apple System und bietet verschiedene Open-Weight Modelle zur Inferenz an. Die Weboberfläche OpenWeb UI kann auf einem beliebigen System betrieben werden; das UI orientiert sich dabei an demjenigen von ChatGPT. Ebenfalls kann über OpenWeb UI via API auf die Modelle zugegriffen werden. Eine weitere Abstraktionsebene kann durch die Implementierung des *Model Context Protocol* (MPC)[102] Paradigmas eingefügt werden. Dieses verbindet standardisiert Modelle mit diversen internen und externen Datenquellen, was deren Einbettung weiter vereinfacht.

[90] https://www.deepseek.com/
[91] https://github.com/databrickslabs/dolly/
[92] https://huggingface.co/tiiuae/falcon-40b/
[93] https://ai.google.dev/gemma/
[94] https://ollama.com/library/llama3.3/
[95] https://mistral.ai/news/mixtral-of-experts/
[96] https://aleph-alpha.com/phariaai/
[97] https://azure.microsoft.com/en-us/products/phi/
[98] https://lmarena.ai/
[99] https://openwebui.com/
[100] https://ollama.com/
[101] https://lmstudio.ai/
[102] https://modelcontextprotocol.io/introduction/

5.2 Best Practices für IT-Sicherheit, Betrieb und Wartung

Unter Annahme, dass die Modelle bzw. darauf aufbauende Services selbst keine datenspezifischen Sicherheitsaspekte verletzen, bedarf es einer sicheren Umgebung (Rechenzentrum) mit entsprechender Zutrittskontrolle. Der Grad dieses Schutzaspekts kann naturgemäss variieren, abhängig von den Anforderungen an die verwendeten Daten, die verarbeitet werden. Unterliegen die Daten besonderem Schutz, dann muss neben deren physischer Unversehrtheit (Manipulation, Datensicherheit) auch auf Ebene der Applikation die Integrität und der Datenschutz eingehalten werden. Ein besonderes Augenmerk liegt auf dem Backup (Sicherung) der verwendeten Daten, die durch die Modelle verarbeitet werden. Die Sicherung dieser Daten muss unter Umständen verschlüsselt werden, während die Modelle, sofern nicht lokal trainiert, lediglich vor direkter Manipulation (z.B. Anpassen von Modellparametern, etc.) geschützt werden müssen.

LM sind wie klassische Client-Server-Systeme mit verschiedenen Risiken behaftet. Das Open Web Application Security Project (OWASP) veröffentlicht dazu jährlich eine Liste[103] der 10 kritischsten Sicherheits-Risiken zusammen mit Massnahmen, um diese zu minimieren. Mögliche Angriffsszenarien, die sich basierend auf der OWASP Liste ergeben, sind:

- *Model- / Data Poisoning* beginnt bereits beim Aufbau eines nachtrainierten Models, in dem gefälschte oder voreingenommene Daten in das Training einfliessen. Die daraus entstandenen *generates* können als weitere Angriffsvektoren dienen, oder das Vertrauen in ein Modell nachhaltig schädigen. Auch eine RAG-basierte Anfrage kann davon betroffen sein, da angeforderte, den Prompt ergänzende Dokumente evtl. kompromittiert sind.
- *Prompt Injection*, als aktive Einflussnahme, hat zum Ziel, das Modellverhalten durch wiederholtes Stellen einer böswilligen Fragenkette an inhärent nicht-öffentliche Informationen zu gelangen.
- *Sensitive Information Disclosure* erfolgt, wenn durch eine Prompt Injection oder unbeabsichtigt (passiv) eine Interaktion mit einem Modell erfolgt und dadurch inhärente nicht-öffentliche Daten preisgegeben werden.

Es ist festzuhalten, dass es keine umfassende Sicherheitslösung geben kann, welche die genannten Risiken abdeckt und Angriffsszenarien komplett verhindert. Ein ganzheitlicher Ansatz und stetes Anpassen und Testen der Sicherheitsmassnahmen verringern die Wahrscheinlichkeit eines nicht-konformen Verhaltens des Modells.

Zu den wichtigsten Massnahmen gehören:

- Den *Netzwerkverkehr überprüfen*, filtern und isolieren, um Anwendungen vor gefährdeten LM zu schützen bzw. kompromittierte LM zu entdecken.
- *Web Application Firewall*-verwaltete Regelsätze zur Blockade von LM-Angriffen auf Basis von SQL-Injection, XSS und anderen Web-basierten Angriffsformen (*attack vectors*) verwenden.
- *LM Sandboxing;* dabei werden aktive Aufrufe von Tools oder Anfragen an externe APIs während des Generierungsprozesses bewusst eingeschränkt; aber auch die eigene API wird vor unbefugten Zugriffen geschützt. Erreicht wird dies durch den Einsatz einer container runtime resp. sandbox der Modelle und deren API-Software.
- *Limitierung der LM-Tasks;* als abstraktes Konzept kann eine Limitierung beinhalten, dass ein LM-Agent keine eigenständigen Entscheidungen treffen darf. Konkret kann dieser Aspekt beinhalten, dass aufgrund von Regeln (*Policies*) gewisse Funktionen nur zeit- oder ortsabhängig durchgeführt werden dürfen.
- *RBAC, ABAC*; Rollen- und Attribut (Eigenschaften)-basierte Zugriffskontrolle auf Ebene der Modelle selbst oder auf Schnittstellen zu diesen Modellen.
- Prüfung von Benutzereingaben und den generierten Outputs; dies kann regelbasiert erfolgen, oder aber durch Datenbereinigung (*Sanitization*), die vor- oder nach der

[103] https://genai.owasp.org/llm-top-10/

Generierung auf verdächtige Muster angewendet wird (z.B. Vermögensangaben oder Gesundheitsdaten, etc.).
- Verwendung eines Zero Trust-Sicherheitsmodell, um die Angriffsfläche zu verkleinern.
- Remote-Browserisolierung (RBI)[104] einsetzen, um Nutzer vor Modellen mit eingeschleustem Schadcode zu schützen.

Es gibt von der Netzwerkebene bis zur LM-Anwendung eine breite Palette von Angriffsformen. Dazu gibt es auch entsprechende Gegenmassnahmen, die, je nach Anforderung, stark in den Abfrageprozess eingreifen können resp. müssen, damit sich die LM-Umgebung regelkonform verhält. Dieser Abschnitt stellt eine Zusammenfassung der Artikelserien in [14, 92] dar.

5.3 Aufbau und Betrieb

Eine lokale Infrastruktur für GenAI aufzubauen ist ein Vorhaben, das eine sorgfältige Planung in verschiedenen Bereichen nötig macht. Je nach Anforderung (Erstellen eines Basismodells, Nachtrainieren, Inferenz) unterscheiden sich Aufbau und Betrieb erheblich. Die Auswahl der richtigen Hardware für KI-Workloads, insbesondere für LMs, hängt stark von den spezifischen Bedürfnissen ab. Die verschiedenen Rechnerarchitekturen wie CPUs, GPUs, TPUs, FPGAs und ASICs haben Stärken und Schwächen, und eignen sich je nach Einsatzszenario für die jeweiligen Workloads unterschiedlich. Standard *CPUs* als Allrounder sind ideal für kleinere Modelle oder für das Prototyping, wo Flexibilität gefordert ist. Sie stehen in der Regel rasch zur Verfügung und sind günstig in der Anschaffung, jedoch aufgrund ihrer seriellen Arbeitsweise, und daher einer geringeren Parallelisierungsfähigkeit, für grosse LMs ineffizient. *GPUs* (ursprünglich für *Graphics Processing Units*) haben die Möglichkeit, massiv-parallele Operationen durchzuführen (*Single Instruction, Multiple Data*; SIMD) und sind ideal für das Training und die Inferenz von LMs, auch wenn sie teuer und wenig energieeffizient sind im Vergleich zu hochspezialisierten Chips. TPUs, FPGAs und ASICs sind Architekturen, die eine sehr gute Energieeffizienz und hohe Geschwindigkeit für spezifische KI-Workloads besitzen. Diese Chips können in der Regel jedoch nicht ohne weiteres in einem lokalen Umfeld, also *on-premise*, betrieben werden, da diese meistens durch Hyperscaler Service Providers in der Cloud angeboten oder nur durch spezifische Hersteller für Nischenanwendungen hergestellt werden.

Falls ein neues Basismodell das Ziel ist, ist die Wahl eines geeigneten Rechenzentrums mit adäquater Stromversorgung sowie Klimatisierung unerlässlich. Geeignete Hochleistungsrechner (HPC) mit GPUs von NVIDIA oder, Stand 2025 weniger geeignet, den Alternativen von AMD und Intel, bilden die Grundlage jeder solcher Rechnerinfrastruktur. Ergänzend dazu muss eine Netzwerk- und Datenspeicherlösung implementiert werden, die grosse Mengen an Trainingsdaten schnell und effizient den *compute nodes* zur Verfügung stellen kann. Distributed Storage Systeme bieten hierbei solche Möglichkeiten. Weiter ist Expertise gefordert, welche von der Netzwerkebene, der Serverhardware bis hin zur spezialisierten Software möglichst viel abdeckt. Beträchtliche Investitionen sind daher nötig, welche zudem wegen dem raschen Technologiewandel fortlaufend getätigt werden müssen. Die Anforderungen für das Nachtrainieren von Basismodellen sind ebenfalls hoch.

Ist das Ziel lediglich das Training und der Betrieb eines SLM oder der Betrieb eines Modells zur Inferenz, so sind die technischen Anforderungen und die Investitionen entsprechend tiefer. Zunächst einmal sind die Rechenanforderungen deutlich geringer, da keine neuen Modelle erstellt, sondern bestehende Modelle angepasst und zur Vorhersage verwendet werden. Dies führt zu einem geringeren Energieverbrauch und reduziert die Notwendigkeit für leistungsstarke und teure Hardware. Zudem sind die Datenanforderungen, zumindest bei der Inferenz, wesentlich geringer, da nur die Eingabedaten und die bereits trainierten Modelle benötigt werden. Auch die Wartung und Skalierbarkeit der Infrastruktur gestalten sich einfacher. Es können Standard-Server und -Netzwerkkomponenten verwendet werden, und die Systeme lassen sich leichter an zukünftige steigende Anforderungen anpassen. Dies unterstützt eine effektivere Kosten- und Risikosteuerung.

[104] https://www.cloudflare.com/learning/access-management/what-is-browser-isolation/

Die KI-Hardwareentwicklung von NVIDIA (Stand März 2025) zeigt eine Fokussierung auf unterschiedliche Anwendergruppen resp. deren spezifische Bedürfnisse. Die aktuellen *consumer-grade* GeForce RTX GPUs (32 GB VRAM) sind für lokale Inferenz und das Training von SLM geeignet. Parallel dazu hat NVIDIA die Angebotspalette für grosse KI-Anwendungen optimiert. Im *prosumer-grade* Bereich gibt es GPUs, die sich an Nutzende in einem Workstation-Umfeld richten; dies insbesondere mit den Modellen der NVIDIA RTX Pro GPUs (96 GB VRAM). Diese Hardware eignet sich speziell für das Training von SLM. Als Alternative gibt es die all-in-one Lösungen DGX Spark (früher Project DIGITS) und DGX Station. Die DGX Spark[105] ist dabei eine kompakte Hardware-Plattform, die sich für das Prototyping und das Testen von KI-Anwendungen eignet. Eine Alternative ist die Apple KI-Hardware mit den *Apple Silicon* Chips (CPU/GPU, NPU: *Neural Processing Unit*). Diese bietet ebenfalls Lösungen für maschinelles Lernen an. Die aktuelle Chip Generation kann dabei auch komplexe Sprachmodelle unterstützen, sowohl zur Inferenz wie auch zum Training.

Auf der Softwareseite hat NVIDIA das Open-Source Framework *Dynamo*[106] und das proprietäre *Blueprints*[107] Toolkit eingeführt. Dynamo wird im NVIDIA Jargon als „Betriebssystem für KI-Fabriken" beschrieben. Es ist darauf ausgelegt, die Inferenz generativer KI- und Reasoning-Modelle in verteilten Umgebungen zu optimieren, indem die GPU-Auslastung von compute nodes orchestriert wird. *Blueprints* dienen der Entwicklung von Workflows für ML-Applikationen. Ein ähnliches Tool wie Dynamo gibt es auch mit Support für die Apple Silicon-Architektur in Form der Open-Source Runtime *exo*[108]. Hier spielt es keine Rolle, welcher Art die Hardware ist, da ein heterogenes Umfeld (Apple CPUs/GPUs, NVIDIA GPUs, eingeschränkt auch für AMD GPUs) unterstützt wird. Ein von Apple entwickeltes ML-Framework für ihre CPU/GPU Architektur ist MLX[109].

Realisierbare Implementierungen basieren auf dem folgenden exemplarischen Szenario (UC2); dabei ist anzumerken, dass die anderen Anwendungsfälle (UC3, UC4) so oder ähnlich ebenfalls betrieben werden können. Dabei gilt es jedoch zu beachten, dass die Anforderungen, insbesondere in Bezug auf den Datenschutz und die Datensicherheit, jeweils variieren können; siehe dazu auch die Liste in Abschnitt 6.4.

Der Anwendungsfall UC2, im *Kontext der Inferenz*, kann mit folgender User Story beschrieben werden: «Als eine Person, die als Sachbearbeitende die Office-Suite unserer Abteilung nutzt, möchte ich die Funktionen verwenden, die von unserem lokalen Sprachmodell innerhalb von Anwendungen wie einer Textverarbeitung oder eines E-Mail-Programms für Aufgaben wie das Schreiben von Dokumenten und das Zusammenfassen von E-Mails unterstützt werden. Ich will so von den Applikationen profitieren, ohne dass sensible Informationen ausserhalb der lokalen sicheren Umgebung prozessiert werden. Dadurch wird der Datenschutz bei der Nutzung von KI gewährleistet».

Eine auf dieser User Story aufbauende Erweiterung, im Kontext eines *vortrainierten lokal angepassten* Modells, könnte lauten: «Ich möchte ein Modell verwenden, das speziell für die Anforderungen und das Vokabular meiner Domäne angepasst wurde, um Aufgaben wie das Verfassen von fachlich korrekten Dokumenten, das Extrahieren von relevanten Informationen aus Texten oder das Generieren von präzisen Zusammenfassungen sicher durchführen zu können. Dabei will ich sicherstellen, dass das Modell lokal und ohne externe Datenverarbeitung läuft, um die Datenschutz- und Compliance-Anforderungen meiner Domäne zu erfüllen.»

Ein alternativer Ansatz ist die Anwendung eines *RAG-basierten* Modells. Dabei werden der UC2 resp. dessen Erweiterung so miteinander verknüpft, dass ein lokales, nicht spezifisch vortrainiertes Modell bei einer Anfrage mit dem Kontext aus einem kuratierten Wissenskorpus angereichert wird. Dieses Dokumenten-basierte Vorgehen verbindet die Vorteile der lokalen Verarbeitung mit der Fähigkeit, domänenspezifische Informationen zu nutzen.

[105] https://www.nvidia.com/en-eu/products/workstations/dgx-spark/
[106] https://www.nvidia.com/en-us/ai/dynamo/
[107] https://build.nvidia.com/blueprints
[108] https://github.com/exo-explore/exo
[109] https://github.com/ml-explore/mlx

Durch die Integration von Dokumenten in den Kontext der Anfrage kann das Modell präzisere und relevantere Antworten generieren, ohne dass ein vollständiges vortrainiertes Modell für jede spezifische Domäne erforderlich ist. Natürlich ist es dabei wichtig, dass der Wissenskorpus relevante Information enthält, damit diese Art der Anwendung gewinnbringend eingesetzt werden kann.

Ein mögliches Setup für den RAG-Anwendungsfall kann wie folgt getestet werden. Im konkreten Fall ist ein Webserver als Gateway mit OpenWeb UI so konfiguriert, dass dieser entweder interaktiv den Zugriff auf eine Auswahl an installierten Modellen erlaubt, die Weboberfläche orientiert sich dabei an derjenigen von ChatGPT, oder als API *end point* dient. OpenWeb UI bringt bereits eine einfache Möglichkeit für eine RAG-Umgebung mit. API-Aufrufe werden mit dem Command Line Tool *curl* an den end point geschickt, und mit dem Command Line Tool *jq* wird die JSON-Antwort formatiert [Listing 1].

Wichtig ist dabei die fortlaufende Überprüfung der generierten Resultate z.B. durch User Surveys, damit die Schwelle für die Benutzung nicht-konformer externer KI-Tools möglichst hoch ist (Verhindern der Benutzung einer «Shadow AI»). Dies erreicht man nur, wenn die Modelle resp. die daraus resultierenden generierten Antworten richtig und relevant sind, und rasch zur Verfügung stehen.

```
curl -s -X POST https://gateway:3000/api/chat/completions \
-H "Authorization: Bearer <jwt>" \
-H "Content-Type: application/json" \
-d '{"model": "<model>", "messages":
     [{"role": "user", "content": "Berechne Quadratmeter der Seen Kt. Luzern"}],
     "files": [{"type": "collection", "id": "<workspace>"}]
    }' | jq
```

Listing 1: Anfrage an ein <model>, mit Support durch RAG mit der id <workspace> (<…> als generische Platzhalter).

5.4 Prompt Management

Klarheit und Präzision sind entscheidend bei der Formulierung von Prompts, um Mehrdeutigkeiten zu vermeiden und spezifische Ergebnisse zu erzielen. Ein *definierter* Kontext – inklusive Nennung der Zielgruppe bzw. Fachdomäne – sowie Beispiele verbessern die Modellantwort. *Einschränkungen* wie Länge oder Stil sollten genannt werden, um unerwünschte Ergebnisse auszuschliessen. *Strukturierte Prompts*, die Aufgaben in Schritte gliedern oder Vorlagen nutzen, halten die Ausgabe fokussiert. Der Artikel «Gespräche mit LLMs» [130] zeigt Paradigmen auf, die beim Prompting Anwendung finden. Eine iterative Optimierung durch Testen und Anpassen bewährter Prompts bleibt wichtig, um Qualität und Konsistenz zu steigern. Die Verwaltung gelingt am besten über ein zentrales Repository mit sprechenden Namen und Metadaten. Wichtig ist auch die Anpassung an die Eigenheiten des jeweiligen Modells sowie der Einsatz von Schutzmechanismen (siehe Abschnitt 6.4) zur Sicherung von Robustheit und Zuverlässigkeit. Es gibt verschiedene Template-Frameworks und Applikationen zur Umsetzung dieser Kriterien; die Produkte von Agentatech[110] (Berlin) bieten einen niederschwelligen Einstieg. Eine Alternative stellt das Open-Source-Framework der Entwicklergruppe Latitude[111] (Barcelona) dar. Beide sind on-premise einsetzbar. Ein Testframework für LMs stellt Confident[112] bereit.

[110] https://docs.agenta.ai/prompt-management/overview
[111] https://latitude.so/
[112] https://docs.confident-ai.com/

6 Datenschutzkonforme Implementierung

Eine datenschutzkonforme Implementierung von GenAI/LM zielt darauf ab, Datenschutzrisiken zu minimieren und regulatorische Vorgaben (z.B. DSG [136]/DSGVO [43], EU AI Act[113]) einzuhalten. Zudem soll sie Transparenz und Nachvollziehbarkeit bei der Datennutzung und Entscheidungsfindung gewährleisten. Ein weiteres zentrales Ziel ist die Sicherstellung der Datensicherheit über den gesamten Lebenszyklus der Modelle hinweg.

Beim Einsatz von GenAI/LLMs müssen die folgenden Phasen unterschieden werden, da jede Phase spezifische Risiken für die Grundrechte und -freiheiten der Betroffenen, insbesondere hinsichtlich der Rechtfertigung der Datenverarbeitung nach Art. 6 DSGVO [43] mit sich bringt [45]:

1. Erhebung von Trainingsdaten
2. Vorbereitende Datenverarbeitung
3. Training
4. Operation: Verarbeitung von Prompts und Outputs
5. (Optional) Training mit Prompts

Die entsprechend der fünf Phasen aufgezeigten Massnahmen sollen einen Überblick geben, müssen aber in jedem Fall an den Anwendungsfall und das Betriebskonzept angepasst werden. Eine detaillierte Auflistung potenzieller Angriffe und entsprechender Gegenmassnahmen findet sich in [32], die für den jeweiligen Einsatzfall genauer analysiert werden sollte. Diese müssen dann zusätzlich zu den in Abschnitt 5.2 aufgezeigten allgemeinen Best Practices zum Betrieb umgesetzt werden.

6.1 Phase 1: Erhebung von Trainingsdaten

Eine klare Definition des Anwendungsfalls bildet die Grundlage für die gezielte Auswahl und Klassifizierung von Trainingsdaten. Ziel sollte es sein, personenbezogene Daten von Anfang an zu minimieren (Privacy-by-Design) [111].

Trainingsdaten sind alle strukturierten oder unstrukturierten Daten, die zum initialen Trainieren und der Weiterentwicklung der Modelle verwendet werden. Trainingsdaten können auch Prompts enthalten. Das Training mit Prompts wird auf Grund seiner Besonderheiten separat in Abschnitt 6.5 behandelt.

- **P1.1 - Datenquellenbewertung**: Vor der Nutzung von Trainingsdaten muss geprüft werden, ob personenbezogene oder andere vertrauliche Daten enthalten sind.
    - Die Daten müssen anhand der **Datenschutzstufen klassifiziert** werden (z.B. öffentlich, vertraulich, personenbezogen, besonders schützenswert nach Art. 5 DSG [136]).
    - **Öffentliche Datenquellen**: In der Schweiz sind viele Dokumente der Verwaltung öffentlich zugänglich (z.B. Parlamentsprotokolle, Verordnungen, Bundesgerichtsentscheide). Diese können für das Training genutzt werden, sofern sie keine schützenswerten Informationen enthalten.
    - **Interne Datenquellen**: Falls nicht-öffentliche Verwaltungsdaten verwendet werden, ist vorab eine Datenschutz-Folgenabschätzung (DSFA)[114] erforderlich.
    - **Daten aus Drittquellen**: Bei der Nutzung externer Daten (z. B. Open Data, akademische Korpora) ist sicherzustellen, dass keine Lizenz- oder Datenschutzverstösse vorliegen.
- **P1.2 - Datensparsamkeit**: Nur die für den spezifischen Anwendungsfall erforderlichen Daten dürfen erhoben und verarbeitet werden (**Prinzip der Zweckbindung**).

[113] https://artificialintelligenceact.eu/de/
[114] https://www.edoeb.admin.ch/de/datenschutz-folgenabschaetzung

- **P1.3 - Rechtliche Absicherung**: Klärung der Rechtsgrundlagen für die Verarbeitung, insbesondere in Bezug auf Datenschutzgesetze wie die DSGVO [43], DSG [136] oder kantonale Datenschutzverordnungen.
- **P1.4 - Informationspflicht**[115]: Bei der Verwendung von personenbezogenen Daten ist es wichtig, die betroffenen Personen klar und verständlich über die Verwendung als Trainingsdaten[116] zu informieren (siehe auch Art 19 DSG [136]). Dies ist besonders relevant, wenn Nutzende persönliche Angaben bei der Verwendung der GenAI/LM-Tools machen und diese aufgezeichnet und ggf. als Trainingsmaterial wiederverwendet werden. Je nach Anwendungsfall muss eine explizite Zustimmung der Betroffenen eingeholt werden.

6.2 Phase 2: Vorbereitende Datenverarbeitung

In dieser Phase werden die in **Phase 1** ausgewählten Datenquellen analysiert, bereinigt und für das Training (Phase 3) aufbereitet. Das Ziel ist, **nur relevante, qualitativ hochwertige und datenschutzkonforme Daten** für das Modelltraining zu verwenden. Besondere Sorgfalt ist bei der Aufarbeitung von personenbezogenen Daten aufzuwenden.

- **P2.1 - Bereinigung und Filterung der Daten** (*Data sanitization*):
  - **Datenqualitätsprüfung**: Identifikation und Entfernung irrelevanter, unvollständiger oder fehlerhafter Daten, um eine höhere Qualität des Trainingsmodells sicherzustellen.
  - **Filterung sensibler Informationen**: Entfernung aller nicht für den Anwendungsfall benötigten Daten, insbesondere personenbezogener und geschützter Informationen.
  - **Bias-Analyse**: Prüfung auf gesellschaftliche und sprachliche Verzerrungen in den Daten. Falls Bias identifiziert wird, sind Gegenmassnahmen wie Nachjustierung der Datenverteilung oder Ergänzung ausgewogener Trainingsdaten zu prüfen.
- **P2.2 – Pseudonymisierung oder Anonymisierung der Daten**
  - **Pseudonymisierung oder Maskierung** zur Minimierung der Re-Identifizierbarkeit.
  - **Aggregationsmechanismen**, falls lediglich Gruppenmerkmale erforderlich sind (z.B. Durchschnittswerte statt individueller Angaben).
  - **Prüfung des Einsatzes von Anonymisierungstechniken**[117], wie k-Anonymität [146] oder Differential Privacy [38]. Allerdings bleibt immer ein **Restrisiko der Re-Identifizierung**, insbesondere aufgrund der nicht vorhersehbaren Interaktionen mit LLMs und deren generativer Natur.
- **P2.3 - Dokumentation und Nachvollziehbarkeit**
  - **Erstellung eines Transparenzberichts**: Alle Verarbeitungsschritte, insbesondere Filter- und Anonymisierungsmassnahmen, müssen dokumentiert werden, um Revisionssicherheit und Einhaltung gesetzlicher Vorgaben (DSG [136], DSGVO [43],…) sicherzustellen.
  - **Festlegung von Verantwortlichkeiten**: Klare Definition, wer für Datenqualität, Anonymisierung und Dokumentation verantwortlich ist.

6.3 Phase 3: Training

Während des Trainingsprozesses müssen technische, organisatorische und sicherheitsrelevante Massnahmen implementiert werden, um Datenschutzrisiken zu minimieren und Angriffe auf das Modell zu verhindern.

[115] Die Informationspflicht kann entfallen, wenn für die Datenbearbeitung eine gesetzliche Grundlage besteht [19]. Siehe auch Art 19 DSG [136].

[116] In der Regel müssen Betroffene bereits vor oder kurz nach der Datenerhebung über die Verwendung der Daten informiert werden. Dies muss berücksichtig werden, wenn in der Vergangenheit gesammelte Daten in das Training einbezogen werden sollen. Es sollte abgeklärt werden, ob eine zusätzliche Zustimmung der Betroffenen erforderlich ist.

[117] **Anonymisierung** ist kein einmaliger Vorgang, sondern ein kontinuierlicher Prozess, der regelmässig überprüft und bei der Integration neuer Daten angepasst werden muss. Da Wirksamkeit der Anonymisierung immer vom spezifischen Anwendungszweck der Daten abhängt, ist eine generelle und kontextunabhängige Anonymisierung von Trainingsdaten für LMs kaum möglich [47].

- **P3.1 - Wahl der Infrastruktur (On-Premise vs. Cloud)**: Bei der Verwendung von personenbezogenen Daten ohne Maskierung, Pseudonymisierung/Anonymisierung oder Aggregation sowie anderer sensibler Daten, wie z.B. Firmengeheimnisse, sollte das Training idealerweise auf einer lokalen Infrastruktur erfolgen, um volle Kontrolle über die Daten zu gewährleisten. Falls eine Cloud-Lösung genutzt wird, sind datenschutzkonforme Anbieter mit entsprechenden Sicherheitsrichtlinien zu wählen.
- **P3.2 - Wahl des Modells**: Beim **Pre-Training** (vortrainiertes Modell/Foundation Model) sollte nach Möglichkeit auf Open Source-Modelle mit dokumentierten Trainingsdaten gesetzt werden.
- **P3.3 - Kontrolle der Trainingsumgebung & Schutz vor Datenmanipulation**: Um Risiken wie Training Data Poisoning [5], Model Extraction [110] oder Adversarial Attacks [10] zu minimieren, müssen umfassende Sicherheitsmassnahmen ergriffen werden:
    - **Strikte Nutzung der in Phase 2 validierten Daten**: Keine nicht-geprüften oder externen Datenquellen dürfen während des Trainings verwendet werden.
    - **Sandboxing und Netzwerkkontrolle**: Es darf keine Verbindung zu externen APIs oder unbekannten Datenquellen während des Trainingsprozesses geben.
    - **MLSecOps** (Machine Learning Security Operations): Durch definierte Prozesse zur Bereitstellung und Wartung der Modelle sowie beim Training können eine konsistente und zuverlässige Leistung in Produktionsumgebungen gewährleistet und Angriffe frühzeitig entdeckt werden (siehe auch Abschnitt 2.3).
    - **Einsatz von Test-Sets**, um Angriffe und Manipulationen zu erkennen.
    - Prüfung auf **Model Inversion Attacks**[118] (Rückgewinnung von Trainingsdaten aus dem Modell).
- **P3.4 - Auditing und Dokumentation**: Der Trainingsprozess sollte umfassend dokumentiert werden, um Transparenz über verwendete Daten und angewandte Schutzmassnahmen sicherzustellen.

6.4 Phase 4: Operation - Verarbeitung von Prompts und Outputs

Nach dem Training muss sichergestellt werden, dass die Nutzung des Modells keine Datenschutzrisiken birgt. Insbesondere gilt es zu verhindern, dass sensible Informationen durch ungeeignete Prompts, unkontrollierte Modellantworten oder gezielte Angriffe offengelegt werden.

- **P4.1 - Prompt-Filterung & Eingabe-Validierung**: Durch technische Schutzmassnahmen wie Eingabe-Validierungen mit entsprechenden Warnhinweisen kann verhindert werden, dass Nutzende ungewollt oder absichtlich sensible oder personenbezogene Daten eingeben. Zusätzlich sollten Anfragen, die auf Model Inversion oder Prompt Injection abzielen, gesperrt werden. Automatisierte Angriffe können durch die Begrenzung der Anfragen pro User und Zeitraum erschwert werden.
- **P4.2 - Knowledge Sanitization**[119]: Um zu verhindern, dass das Modell – absichtlich oder unabsichtlich - sensible oder personenbezogene Daten ausgibt, kann das Modell durch Re-Training, Fine-Tuning mit entsprechendem Trainingsmaterial oder Gradient Unlearning[120] angepasst werden.
- **P4.3 - Kontrollierte Ausgabe**: Durch Analyse der generierten Texte können sensible Daten identifiziert und anschliessend maskiert oder unterdrückt werden.
- **P4.4 - Einschränkung externer Datenquellen**: Durch die Begrenzung des Zugriffs auf externe Datenquellen können kontrollierte und sichere Antworten gewährleistet werden.

[118] Bei **Model Inversion Attacks** versucht ein Angreifer oder eine Angreiferin, Informationen über die Trainingsdaten eines Modells zu rekonstruieren. Dies kann dazu führen, dass personenbezogene oder vertrauliche Daten aus einem trainierten Modell extrahiert werden.

[119] **Knowledge Sanitization** bezeichnet die gezielte Bereinigung von trainierten KI-Modellen, um das Risiko der ungewollten Preisgabe sensibler oder personenbezogener Informationen zu minimieren. Dabei werden nach dem Training sensible, fehlerhafte oder sicherheitskritische Informationen entfernt, sodass das Modell sie weder absichtlich noch unabsichtlich reproduzieren kann.

[120] **Gradient Unlearning** ist eine Technik zur gezielten Löschung von spezifischen Trainingsdaten. Es ermöglicht das Entfernen von unerwünschten oder schützenswerten Informationen, ohne das gesamte Modell neu trainieren zu müssen.

- **P4.5 - Zugriffskontrollen**: Bei GenAI-Anwendungen mit sensiblen Daten, sollten nur berechtige Personen das Modell verwenden dürfen. Ggf. sind auch in Abhängigkeit der Benutzerrolle differenzierte Antworten (z.B. detaillierte Infos für Poweruser vs. allgemeinere Antworten für normale Nutzende) vorzusehen.
- **P4.5 - Echtzeit-Monitoring**: Laufende Überwachung und Logging der Modellnutzung können helfen, unerwünschte oder unsichere Anfragen, aber auch Datenlecks, Missbrauch und Angriffe frühzeitig zu erkennen.

6.5 Phase 5: Training mit Prompts (optional)

Falls das Modell durch Interaktionen mit Nutzenden weiter trainiert oder angepasst wird (z.B. durch Reinforcement Learning from Human Feedback [79]) müssen zusätzlich zu den Massnahmen aus Phase 4 weitere Massnahmen ergriffen werden, um personenbezogene Daten zu schützen und unerwünschte Modellveränderungen zu verhindern.

- **P5.1 - Nutzertransparenz und Einwilligung**: Nutzende müssen klar und verständlich informiert werden, ob und wie ihre Eingaben für das Modelltraining genutzt werden. Eine ausdrückliche Zustimmung (Opt-in) ist erforderlich. Ggf. können Nutzende selbst entscheiden können, ob ihre Interaktionen für das Training genutzt werden dürfen**.**
- **P5.2 - Qualitätskontrolle der Prompts**: Um eine ungewollte Verzerrung oder Sicherheitslücken zu vermeiden, sollten Filter verwendet werden, um eine ungewollte Verzerrung oder Sicherheitslücken zu vermeiden, und regelmässig Stichproben geprüft werden.
- **P5.3 - Aufbereitung der Prompt-Daten** (*Data Sanitization*): Ziel ist es, zu verhindern, dass bei der Weiterentwicklung des Modells sensible Daten aus dem Prompt in das Modell gelangen.
    - **Filterung sensibler Informationen**: Entfernung aller personenbezogener und geschützter Informationen.
    - **Pseudonymisierung oder Maskierung** von personenbezogenen Daten.
- **P5.4 - Dokumentation der Trainingssätze**: Alle für das Training verwendeten Prompts und Outputs müssen dokumentiert werden, um Nachvollziehbarkeit und Transparenz sicherzustellen

6.6 Einordnung der Massnahmen

Die in diesem Kapitel beschriebenen Massnahmen für die fünf Phasen – von der Datenvorbereitung über das Training bis hin zur sicheren Nutzung und kontinuierlichen Optimierung – bieten eine umfassende Grundlage für den datenschutzkonformen Einsatz von GenAI und LMs. Allerdings ist nicht jede Massnahme in jedem Szenario gleich sinnvoll oder erforderlich.

Die konkrete Auswahl und Umsetzung hängen massgeblich ab von:

- dem spezifischen Anwendungsfall,
- dem Betriebsmodell (lokal vs. Cloud-Umgebung),
- den regulatorischen Anforderungen,
- den verfügbaren Ressourcen (technische Infrastruktur, Fachpersonal).

Ein **risikobasierter Ansatz** ist daher entscheidend: Je höher das Risiko für Datenschutzverletzungen oder missbräuchliche Nutzung, desto strenger müssen die Schutzmassnahmen ausfallen. Eine DSFA ist ein geeignetes Instrument, um strukturiert die datenschutzrechtlichen Risiken zu erfassen, zu verhandeln und zu mindern. Organisationen sollten damit gezielt eine individuelle Bewertung der Anwendungsfälle vornehmen und jene Massnahmen umsetzen, die einen echten Mehrwert für Sicherheit, Nachvollziehbarkeit und Datenschutz bieten, ohne unnötige Komplexität zu erzeugen.

Die durchgeführten Experteninterviews thematisierten ebenfalls das Spannungsfeld zwischen Innovation und Datenschutz.

6.7 Beispielhafte Anwendung auf die exemplarischen Anwendungsfälle

In der nachfolgenden Tabelle wird die Anwendbarkeit der beschriebenen Massnahmen aufgezeigt.

	UC1 «Chatbot für Bürger/innen»	**UC2 «Office-Applikationen»**[121]	**UC3 «Fallbear-beitung mit Fachapplikation»**	**UC4 «Sprach-Erkennung und -Protokollierung»**[122]
Phase 1 – Erhebung von Trainingsdaten				
P1.1 – Datenquellen-bewertung	Öffentlich zugängliche Informationen	Keine eigenen Trainingsdaten, ausser bei kuratierten Prompts oder RAG	Vorrangig interne Datenquellen, personenbezogene oder auch schützenwerte Daten	(optional) Vorrangig interne Datenquellen (z.B. vorhandene Transkripte und Protokolle) zum Fine Tuning des LMs zur Textgenerierung
P1.2 - Datensparsamkeit	Nicht relevant	Nur öffentliche oder unkritische Daten sollten verwendet werden	Nur die für die Fachapplikation relevanten Daten sollten berücksichtigt werden	Nur bedingt anwendbar
P1.3 - Rechtliche Absicherung	Auskunftserteilung mit expliziten Hinweis auf Quelle; Falschaussagen/Hallu-zinationen müssen abgesichert werden	Abklärung AGB der KI-Anbieter	bei Verarbeitung personenbezogener Daten gesetzliche Grundlage notwendig	gesetzliche Grundlage notwendig
P1.4 – Informations-pflicht	Bei Aufzeichnung und Verwendung von Prompts zu Trainingszwecken, Qualitätssicherung etc.	Nicht anwendbar	Information der betroffenen Personen (Transparenz & informationelle Selbstbestimmung)	Information der betroffenen Personen (Transparenz & informationelle Selbstbestimmung))
Phase 2 – Vorbereitende Datenverarbeitung				
P2.1 - Bereinigung und Filterung der Daten	Entfernung aller nicht benötigten Daten, Prüfung Datenqualität und Bias-Analyse	Entfernung aller nicht benötigten Daten	Entfernung aller nicht benötigten Daten, Bias-Analyse	Entfernung aller nicht benötigten Daten, (Bias-Analyse)[123]

[121] Bei der Auswahl der geeigneten Massnahmen für UC2 wurde von einer «Standard-Implementierung» ausgegangen, bei der der KI-Anbieter grundsätzlich auf alle Daten der Nutzenden zugreifen kann und diese auch zur Weiterentwicklung seiner Lösung verwendet. Bei der Verwendung von eigenen Modellen deckt sich sonst UC2 mit UC3.

[122] Bei UC4 wurde davon ausgegangen, dass das ASR mit einen LM ergänzt wird, das mit Hilfe von bereits vorhandenen Transkripten und Protokollen auf den entsprechenden Anwendungsfall, z.B. juristische Sprache, trainiert wird.

[123] Stimm- und sprachbezogene Merkmale werden für technische Zuverlässigkeit sowie Sprecher-Differenzierung bzw. Sprecher-Erkennung benötigt, enthalten aber potential immer einen Bias.

	UC1 «Chatbot für Bürger/innen»	UC2 «Office-Applikationen»[121]	UC3 «Fallbearbeitung mit Fachapplikation»	UC4 «Sprach-Erkennung und -Protokollierung»[122]
P2.2 – Pseudonymisierung oder Anonymisierung der Daten	Nicht relevant	Pseudonymisierung oder Maskierung von Personendaten	Nur notwendig, wenn Unbefugte Zugang zu den personenbezogenen Daten haben	Nur notwendig, wenn Unbefugte Zugang zu den personenbezogenen Daten haben
P2.3 – Dokumentation und Nachvollziehbarkeit	Verantwortlicher für Datenqualität	Muss durch Best Practices geregelt werden	Transparenzbericht, Verantwortliche für Datenquelle und Bereinigung/ Filterung	Transparenzbericht
Phase 3 - Training				
P3.1 - Wahl der Infrastruktur	LM mit Anpassungen auf die speziellen Datenquellen	LM	on-Premise oder bei datenschutzkonformen Anbietern	on-Premise oder bei datenschutzkonformen Anbietern
P3.2 - Wahl des Modells	Pre-Modell mit idealerweise dokumentierten Trainingsdaten	Pre-Modell mit idealerweise dokumentierten Trainingsdaten	Pre-Modell mit idealerweise dokumentierten Trainingsdaten	Pre-Modell mit idealerweise dokumentierten Trainingsdaten
P3.3 - Kontrolle der Trainingsumgebung	Notwendig	Nicht möglich	Notwendig	Notwendig
P3.4 - Auditing und Dokumentation	Dokumentation des Trainingsprozesse, speziell der Anpassungen	Gesprächsverläufe sollten zur Qualitätssicherung ausgewertet werden können	Dokumentation des Trainingsprozesse	Dokumentation des Trainingsprozesse
Phase 4 - Operation				
P4.1 - Prompt-Filterung & Eingabe-Validierung	Erforderlich	Warnhinweise	Teilweise erforderlich	Nicht erforderlich
P4.2 - Knowledge Sanitization	Nicht relevant	Ggf. sinnvoll, z.B. bei Bias	Ggf. sinnvoll, z.B. bei Bias	Notwendig, bei personenbezogenen oder sensiblen Daten
P4.3 - Kontrollierte Ausgabe	Nicht notwendig	Eher nicht notwendig	Teilweise erforderlich	Nicht erforderlich

	UC1 «Chatbot für Bürger/innen»	UC2 «Office-Applikationen»[121]	UC3 «Fallbearbeitung mit Fachapplikation»	UC4 «Sprach-Erkennung und -Protokollierung»[122]
P4.4 – Einschränkung externer Datenquellen	Ggf. sinnvoll	Eher nicht notwendig	Unbedingt notwendig	Unbedingt notwendig
P4.5 – Zugriffskontrollen	Nicht notwendig	Evtl. nur für bestimmte Mitarbeitergruppen zugänglich machen	Erforderlich	Erforderlich
P4.5 - Echtzeit-Monitoring	Sinnvoll	Sinnvoll, um Verwendung von Personendaten in Prompts aufzudecken	Erforderlich	Erforderlich
Phase 5: Training mit Prompts				
P5.1 – Nutzertransparenz und Einwilligung	Explizite Zustimmung der Nutzenden erforderlich	Explizite Zustimmung erforderlich	Explizite Zustimmung der Nutzenden erforderlich	Explizite Zustimmung der Nutzenden erforderlich
P5.2 – Qualitätskontrolle der Prompts	Ggf. sinnvoll	Notwendig, um Verwendung von personenbezogenen oder sensiblen Daten auszuschliessen	Regelmässige Kontrolle	Regelmässige Kontrolle
P5.3 - Aufbereitung der Prompt-Daten	Notwendig, um zu verhindern, dass personenbezogene oder sensible Daten ins Modell gelangen[124]	Notwendig, um zu verhindern, dass personenbezogene oder sensible Daten ins Modell gelangen	Ggf. sinnvoll	Ggf. sinnvoll
P5.4 – Dokumentation der Trainingssätze	erforderlich	erforderlich	erforderlich	erforderlich

Tabelle 1 - Beispielhafte Massnahmen zur datenschutzkonformen Umsetzung

[124] Benutzer können personenbezogene Daten eingeben (auch wenn diese nicht notwendig wären).

7 Handlungsempfehlungen

Unter Beachtung der übergeordneten strategischen Ziele zur Nutzung von KI in öffentlichen Institutionen wie der Verbesserung der angebotenen Dienstleistungen oder der effizienteren Gestaltung von Verwaltungsgeschäften bei gleichzeitiger Wahrung der digitalen Souveränität (siehe Abschnitt 4.1) und Einhaltung der rechtlichen und ethischen Rahmenbedingungen (siehe Abschnitt 4.2), sind Umsetzung und Betrieb eigener GenAI-Lösungen zumindest für kritische Anwendungsfelder notwendig. Als Grundlage hierfür schlagen wir in den nachfolgenden Abschnitten empfehlenswerte begleitende Massnahmen vor.

7.1 Daten- und KI-Kultur

Die Verfügbarkeit von Daten in der notwendigen Menge und Qualität ist eine zentrale Voraussetzung für die Entwicklung einer KI-Lösung. Dies ist nicht nur eine technische Frage in Bezug auf Datenmodellierung und -speicherung, Daten- und Qualitätsmanagement, sondern erfordert die Etablierung einer Datenkultur (siehe Abschnitt 4.1), bei der es zur gelebten Normalität wird, dass **Mitarbeitende** zielgerichtet und regelkonform mit Daten umgehen und diese effektiv und effizient einsetzen können – nicht nur, aber auch für KI-Lösungen. Entsprechend sollten die Mitarbeitenden auch frühzeitig in die Diskussion um eine Daten- und KI-Kultur mit einbezogen werden, sei es durch strukturierte Dokumentation (zu Datensätzen, technischer Infrastruktur, Verantwortlichkeiten, etc.), regelmässige Information (z.B. via Newsletter), Weiterbildung oder Workshops (z.B. zum Erfahrungsaustausch oder der Diskussion zu einem spezifischen Anwendungsfall), bis hin zur aktiven Mitarbeit bei Entwicklungsprojekten, beispielsweise bei der Definition von Anwendungsfällen, dem Erstellen von Trainings- und Testdaten und der Evaluierung des UI. Mitarbeitende der Verwaltung des Kantons Luzern tauschen sich beispielsweise in einer *Community of Practice* zum Thema KI aus. Eine solche systematische Auseinandersetzung mit der **aktuellen Entwicklung** ist angesichts des schnellen technischen Fortschritts, der neben besseren auch neue Anwendungsmöglichkeiten von KI erlaubt, dringend notwendig.

7.2 Weiterbildung und Rekrutierung von Mitarbeitenden

Sowohl Einsatz als auch Eigenentwicklung von KI-Lösungen erfordern, dass Mitarbeitende nicht nur in der **Anwendung** von KI-Funktionen versiert sind, sondern auch über potenzielle **Probleme** wie falsche oder diskriminierende KI-Ergebnisse, über das Risiko bei der Verwendung von Shadow AI und über die geltenden rechtlichen und ethischen Regeln Bescheid wissen und mit diesen in ihrem Arbeitsalltag, wie in den exemplarischen Anwendungsfällen skizziert, auch umgehen können. Eine Organisation sollte ihre Mitarbeitenden entsprechend informieren und weiterbilden. In einigen öffentlichen Verwaltungen wurden bereits öffentlich einsehbare **Leitlinien** zum Einsatz von GenAI herausgegeben, beispielsweise durch die Stadt San Francisco[125], die Staatskanzlei St. Gallen[126], den EU-Datenschutzbeauftragten (EU Data Protection Supervisor (EUDPS))[127] oder die Schweizerische Bundesverwaltung[128]. Diese thematisieren z.B., das KI-generierte Inhalte kritisch zu überprüfen sind und Mitarbeitende einem öffentlichen KI-Modell keine nicht-öffentlichen Daten übermitteln dürfen.

Für die teilweise oder vollständige Entwicklung sowie den Betrieb eigener KI-Lösungen wird weiterhin auch interne technische Expertise im KI-Umfeld benötigt. Neben der **Rekrutierung** von externen Spezialisten (Data Scientist, ML/KI Engineer, Data Architect, etc.) ist vor allem die **Weiterbildung** der vorhandenen IT-Mitarbeiter essenziell, um dem Fachkräftemangel entgegenzuwirken (siehe Abschnitt 2.3). Auch die bereits erwähnte KI-Strategie der

[125] https://media.api.sf.gov/documents/Generative_AI_Guidelines_-_CCSF.pdf
[126] https://www.berichte.sg.ch/geschaeftsbericht-der-regierung-2023/berichte/staatskanzlei/leitlinien-zur-verwendung-von-chatgpt.html
[127] https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/2024-06-03-first-edps-orientations-euis-using-generative-ai_en
[128] https://www.bk.admin.ch/bk/de/home/digitale-transformation-ikt-lenkung/ikt-vorgaben/strategien-teilstrategien/sb021-strategie-einsatz-von-ki-systemen-in-der-bundesverwaltung.html

Schweizerischen Bundesverwaltung definiert den Kompetenzaufbau der Mitarbeitenden als zentrales Handlungsfeld.

7.3 Gesamtbetrachtung der rechtlichen Rahmenbedingungen

Aus rechtlicher Sicht sind für den Einsatz von KI in der öffentlichen Verwaltung eine Reihe von rechtlichen Grundsätzen relevant, die neben dem in diesem Bericht besonders betrachteten Datenschutz insgesamt zu berücksichtigen sind (siehe Abschnitt 4.2): Notwendigkeit zur Legitimierung einer KI-Lösung für bestimmte Aufgaben, Recht auf vorgängige Äusserung und Mitwirkung bei Verwaltungsentscheiden durch Betroffene, Diskriminierungsverbot und Begründungspflicht. Aus diesen rechtlichen Anforderungen lassen sich teilweise auch direkt **technische Funktionen** ableiten, die eine KI-Lösung entsprechend umzusetzen hat, z.B. Privacy by Design, Bias-Analyse, Zugriffskontrollen, Maskierung von sensitiven Daten, XAI, etc. (siehe Abschnitt 4.2 und Kapitel 6).

Zu beachten ist auch, dass die meisten der geltenden Bestimmungen lange vor der weiten Verbreitung von KI verabschiedet wurden und es zu erwarten ist, dass diese in Zukunft entweder angepasst oder durch **KI-spezifische Regelungen** ergänzt werden. So auch im Fall des kürzlich beschlossenen EU AI Act (siehe Abschnitt 4.2), dessen weitreichende Implikationen und die daraus resultierenden Massnahmen zur Umsetzung zurzeit noch intensiv diskutiert werden.

7.4 Applikationen und Infrastruktur

Als Basis jeglicher KI-Strategie, und den daraus entstehenden Handlungen, bedarf es eines soliden technologischen Fundaments in Form von Applikationen und Infrastruktur, die in ausgewählten Anwendungsfeldern auch durch die Organisation selbst entwickelt und betrieben werden sollte[129]:

- **Verhindern einer Shadow AI:** Um dies umzusetzen, d.h. das unkontrollierte Anwenden von externen KI-Tools zu unterbinden, und zeitgleich die Effizienz kritischer Anwendungsfälle (z.B. UC2) sicherzustellen, empfehlen wir die Anschaffung resp. den Aufbau eigener Ressourcen in Bezug auf Daten und die Infrastruktur: es gibt aktuell leistungsfähige Hardware und die passenden Language Models (SLM), welche für viele dieser Anwendungsfälle, insbesondere auf Basis von RAG, geeignet sind. Allenfalls gilt es abzuklären, ob spezialisierte Service Provider mit Sitz in der Schweiz unterstützend mitwirken können; entweder im Bereich der Infrastruktur oder als externe Spezialkräfte. Das Trainieren von grossen Modellen ist eher keine Option; allerdings ist es grundsätzlich möglich, kleinere Modelle an die jeweilige Fachdomäne mittels *fine tuning* anzupassen.
- **Betriebskonzepte:** Die Sicherstellung der Datensicherheit und entsprechende Zugangs- und Zugriffskontrollen sind wesentliche Aspekte eines sicheren Betriebs der KI-Infrastruktur. Ein konstantes Monitoring ermöglicht es, die Leistungsfähigkeit der Infrastruktur und der Modelle zu überwachen und proaktiv auf Probleme einzugehen und zu reagieren, was gleichzeitig auch den Einsatz von Shadow AI verhindern oder verringern sollte.

7.5 Transparenz gegenüber der Öffentlichkeit und Politik

Eine transparente Gestaltung von GenAI-Systemen ist essenziell, um **Vertrauen** bei Nutzenden und Betroffenen aufzubauen. Dazu sollten folgende Massnahmen berücksichtigt werden:

- **Menschliche Kontrolle:** Vermeidung vollständiger Automatisierung, indem menschliche Kontrolle und Entscheidungsfindung gewährleistet werden. Bei Bedarf sollte immer eine Abwicklung der Geschäftsprozesse ohne den Einsatz von KI möglich sein.

[129] Dies wurde beispielsweise auch als zentrale Voraussetzung für die Umsetzung der KI-Strategie des Schweizerischen Bundes definiert, siehe https://www.bk.admin.ch/bk/de/home/digitale-transformation-ikt-lenkung/ikt-vorgaben/strategien-teilstrategien/sb021-strategie-einsatz-von-ki-systemen-in-der-bundesverwaltung.html

- **Klar kommunizierte Datennutzung:** Die Verwendung von Daten, insbesondere von personenbezogenen Daten, sollte verständlich und nachvollziehbar erklärt werden.
- **Berücksichtigung des *Privacy Calculus*** [65]**:** Durch gezieltes Framing und die Bereitstellung qualitativ hochwertiger Argumente sollte eine ausgewogene Wahrnehmung der Chancen und Risiken bei der Nutzung von GenAI und LM gefördert werden.

7.6 Risikomanagement

Für eine sichere und datenschutzkonforme Nutzung von GenAI und LM soll ein risikobasierter Ansatz verfolgt werden (siehe auch Abschnitt 6.6). Die Auswahl geeigneter Schutzmassnahmen sollte individuell per Anwendungsfall erfolgen und sich an folgenden Faktoren orientieren:

- spezifische Datenschutzanforderungen,
- Betriebsmodell,
- verfügbare Ressourcen (technische Infrastruktur, Fachpersonal)

Je höher das Risiko für Datenschutzverletzungen oder missbräuchliche Nutzung, desto strikter sollten die Schutzmassnahmen ausfallen. Dabei ist ein pragmatischer Ansatz entscheidend, der Innovation nicht unnötig behindert, sondern gezielt Sicherheit, Nachvollziehbarkeit und Datenschutz stärkt.

Ein geeignetes Mittel zum systematischen Risikomanagement ist die Datenschutz-Folgenabschätzung (DSFA), die einen strukturierter Prozess zur Identifikation, Bewertung und Minderung von Risiken für die Rechte und Freiheiten betroffener Personen bei der Bearbeitung von Personendaten beinhaltet. Eine DSFA sollte schon in den frühen Phasen eines Vorhabens in Betracht gezogen werden, besonders wenn sensible oder personenbezogene Informationen verarbeitet werden (siehe auch Merkblatt Datenschutz-Folgenabschätzung und Vorabkonsultation des Kantons Luzern[130]).

7.7 Bedarfsanalyse und Wirtschaftlichkeit

Angesichts der vielfältigen Anwendungsmöglichkeiten von GenAI-Lösungen im Umfeld öffentlicher Verwaltungen, beispielsweise bei Chatbots oder der Erstellung, Bearbeitung und Auswertung von Texten (siehe Kapitel 3 und Abschnitt 4.3), einerseits und begrenzter Ressourcen für die Umsetzung andererseits bedarf es auch einer Kosten-Nutzen-Analyse [75] zur **Priorisierung** der Anwendungsfälle und zur Abwägung zwischen den jeweiligen Umsetzungsalternativen. Ziel der Gegenüberstellung von Nutzen und Kosten ist somit auch zu vermeiden, dass KI-Lösungen realisiert werden, die aufgrund des typischen Technologie-Hype Cycles[131] momentan zwar populär sind, aber keinen wirklichen Mehrwert für den konkreten Anwendungsfall bieten.

Der **Nutzen** kann dabei qualitativ (z.B. neues Informationsangebot) oder quantitativ (z.B. eingesparte Arbeitszeit) sein und sich sowohl auf Seiten der Verwaltung/Mitarbeitenden oder der Betroffenen manifestieren [107].

Die Kostenanalyse sollte auf einer Bewertung der **Gesamtkosten** (eng. *Total Cost of Ownership (TCO)*) [156] beruhen, die alle relevanten Teilkosten für Anschaffung (z.B. Hardware und einmalige Lizenzen), Entwicklung (z.B. Personalkosten ML Engineer und Energiekosten Training ML-Modell) und Betrieb (z.B. nutzenbasierte Lizenzen und Support) der KI-Lösung beinhaltet. Die Berechnung des TCO für verschiedene Umsetzungsalternativen ist naturgemäss aufwändig und für etliche Teilkosten können häufig nur Kostenrahmen geschätzt werden. Als Hilfestellung

[130] https://datenschutz.lu.ch/-/media/Datenschutz/Dokumente/Merkblaetter/neuesKDSG/Merkblatt_Datenschutz_Folgenabschtzung_und_Vorabkonsultation_V10.pdf
[131] https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence

können Tools wie der «AI Total Cost of Ownership Calculator» von HuggingFace[132] eingesetzt werden.

7.8 Vernetzung

Neben einer organisationsinternen Vernetzung der Mitarbeiter, beispielsweise in einer Community of Practice (siehe Abschnitt 7.1) sollte es angesichts der Komplexität und des raschen Fortschritts von KI einerseits sowie den begrenzten Ressourcen von öffentlichen Verwaltungen andererseits auch eine enge Kooperation mit externen Partnern geben:

- **Kommerzielle Partner** wie Beratungen, Auftragsentwickler und Cloud-Anbieter können helfen, eine KI-Lösung zu entwerfen, zu entwickeln, oder zu betreiben, wobei es aufgrund der rechtlichen Rahmenbedingungen empfehlenswert ist, vorrangig mit lokalen Partnern zusammenzuarbeiten.
- **Forschungspartner** und unabhängige NGOs wie AlgorithmWatch[133] können bei der Konzeptionierung und Bewertung von Lösungsalternativen bzw. bei der fortlaufenden Qualitätssicherung und dem Risikomanagement unterstützen.
- Empfehlenswert ist weiter eine Kooperation mit anderen **öffentlichen Institutionen** auf kantonaler, nationaler und internationaler Ebene zum Erfahrungsaustausch bis hin zu gemeinsam durchgeführten Projekten. Beispielsweise unterstützt das vom Bundesamt für Statistik (BFS) betriebene Kompetenznetzwerk für künstliche Intelligenz (CNAI)[134] einen nationalen Erfahrungsaustauch. Und das ebenfalls vom BFS getragene Kompetenzzentrum für Datenwissenschaft (DSCC)[135] bietet Beratungsleistungen und Unterstützung bei der initialen Entwicklung einer KI-Lösung an.
- Zu prüfen ist aus Gründen der Wirtschaftlichkeit, Transparenz und dem damit verbundenen Erfahrungsaustausch auch eine Verwendung von bzw. aktive Mitarbeit in **Open Source**-basierten KI-Projekten. Beispielsweise hat die EU mit «OpenEuroLLM[136]» ein Projekt zur Erstellung von LM für alle EU-Sprachen (plus ausgewählte weitere) lanciert, wobei sowohl Daten, Code und die LM unter der Apache 2.0[137] Open Source-Lizenz veröffentlicht werden sollen.

[132] https://huggingface.co/blog/dhuynh95/ai-tco-calculator
[133] https://algorithmwatch.ch/de/
[134] https://cnai.swiss
[135] https://www.bfs.admin.ch/bfs/de/home/dscc/dscc.html
[136] https://interoperable-europe.ec.europa.eu/collection/open-source-observatory-osor/news/eurollm-pioneering-european-open-source-ai und https://openeurollm.eu
[137] https://www.apache.org/licenses/LICENSE-2.0

8 Zusammenfassung und Ausblick

Der technologische Fortschritt bei ML und GenAI hat in den letzten Jahren zahlreiche neue Anwendungen im Arbeitsalltag ermöglicht und LMs können zunehmend für die Bearbeitung von komplexen Aufgaben wie die automatisierte Beantwortung von Supportanfragen oder die Erstellung von fachlichen Texten verwendet werden. Gerade auch für öffentliche Verwaltungen ergibt sich somit ein grosses Potential zur Unterstützung der Mitarbeitenden im Arbeitsalltag, zur Optimierung der Verwaltungsprozesse und zur effizienten Interaktion mit Bürger/innen und anderen Institutionen.

Um Lösungen auf Basis von GenAI und LM für öffentliche Verwaltungen praktisch nutzbar zu machen, bedarf es neben der Adressierung technischer, organisatorischer und wirtschaftlicher Herausforderungen dieser Technologien der besonderen Berücksichtigung der übergeordneten Digitalisierungsstrategie (u.a. mit dem Ziel der Herstellung der digitalen Souveränität) sowie der geltenden rechtlichen Rahmenbedingungen (Legitimierung, Begründungspflicht, Diskriminierungsverbot und Datenschutz). In dieser Studie wurden daher die Alternativen zur technischen Umsetzung einer GenAI-Lösung inklusive der jeweils umzusetzenden Datenschutz-Massnahmen anhand exemplarischer Anwendungsfälle (Suchmaschinen und Chatbots, KI-Assistenten für Office- und Fach-Applikationen, Spracherkennung) eingehend analysiert. Weiterhin empfiehlt die vorliegende Studie eine Reihe von begleitende Massnahmen für die erfolgreiche Konzeptionierung, Realisierung und Betrieb einer GenAI-Lösung: die Etablierung einer Daten- und KI-Kultur, die Priorisierung der GenAI-Anwendungsfälle mit Hilfe einer Kosten-Nutzen-Analyse, den Aufbau zusätzlicher Infrastruktur und personeller Ressourcen durch Weiterbildung und Rekrutierung, vertrauensbildende Massnahmen und Transparenz gegenüber Mitarbeitenden und Betroffenen, Risikomanagement, und Zusammenarbeit mit externen Partnern.

Die durchgeführte Markt- und Literaturanalyse hat gezeigt, dass viele der gefundenen GenAI-Lösungen im Umfeld öffentlicher Verwaltungen entweder Forschungsprototypen sind und insbesondere rechtliche und ethische Aspekte ausklammern oder sich erst in einem frühen Entwicklungsstadium befinden; entsprechend sind erst vereinzelt eigene GenAI-Lösungen im produktiven Einsatz mit Schwerpunkt auf Chatbots zum Zugriff auf öffentlich verfügbare Daten. Von zentraler Bedeutung zum Aufbau von eigenen Erfahrungen ist daher auch die konkrete Umsetzung ausgewählter Anwendungsfälle, beispielsweise der hier skizzierten UC1-4, über den gesamten Lebenszyklus von Konzeption, Implementierung und insbesondere auch produktivem Einsatz und unter Berücksichtigung aller in dieser Studie identifizierter technischer, wirtschaftlicher, organisatorischer, rechtlicher und ethischer Rahmenbedingungen.

Angesichts des enormen Ressourceneinsatzes durch Industrie und Forschung zur technologischen Weiterentwicklung von KI sind auch in Zukunft fortlaufende Verbesserungen, neue Methoden und neue Anwendungsmöglichkeiten zu erwarten. Öffentliche Verwaltungen sollten diese entsprechend mitverfolgen und, soweit möglich, auch mitgestalten (beispielsweise durch Forschungskooperationen). Analoges gilt für die KI-spezifischen rechtlichen Rahmenbedingungen, die wie im Fall des EU AI Acts gerade erst definiert werden und deren praktische Implikationen noch auszulegen sind.

9 Tabellenverzeichnis

Tabelle 1 - Beispielhafte Massnahmen zur datenschutzkonformen Umsetzung 35

10 Glossar

AI (Artificial Intelligence) und **KI** (Künstliche Intelligenz)	KI beschäftigt sich mit der Automatisierung intelligenten Verhaltens, z.B. der Fähigkeit Informationen zu erfassen und zu verarbeiten, Probleme zu lösen, Ziele zu erreichen, oder Sprache zu verstehen und zu verwenden.
API (Application Programming Interface)	Eine Programmierschnittstelle definiert eine standardisierten Funktions- und Datenaustausch zwischen Software-Komponenten, z.B. in der Form einer Bibliothek oder eines Dienstes.
BERT (Bidirectional Encoder Representations from Transformers)	Ein LM auf Basis der Transformer-Architektur, das auf die Vorhersage von Worten, die zu einer gegebenen Wortsequenz passen, trainiert wurde.
Bias	Eine systematische Verzerrung in Daten oder Funktionen, die zu fehlerhaften, unfairen oder voreingenommenen Ergebnissen führen kann.
Chatbot	Eine textbasiertes Dialogsystem auf der Basis von natürlich-sprachigen Nachrichten (z.B. Fragen und Antworten), die zwischen einem menschlichen Nutzenden und dem Chatbot in (nahezu) Echtzeit ausgetauscht werden. Der Chatbot kann vollautomatisch auf Anfragen reagieren und Aktionen auslösen (z.B. einen Antrag ausfüllen und einreichen). Die technische Umsetzung erfolgt heutzutage meist auf Basis eines LM.
Datenschutz-Folgenabschätzung (DSFA)	Die DSFA ist ein strukturierter Prozess, mit dem datenschutz-rechtliche Risiken erfasst, bewertet und behandelt werden können.
DSGVO (Datenschutzgrundverordnung) und **DSG** (Datenschutzgesetz)	Die DSVGO bezeichnet die Verordnung der EU 2016/679 [43] zur Regelung der Verarbeitung personenbezogener Daten. In der Schweiz wurde mit dem revidierten DSG eine mit dem DSGVO kompatible Fassung erlassen.
Foundation Model	Ein auf einer grossen Datenmenge vortrainiertes ML-Modell, das durch eine nachgelagerte Anpassung auf ein spezifisches Problem angewendet werden kann.
GenAI (Generative AI)	Ein ML-Modell zur Erzeugung neuer Daten (Texte, Bilder, Videos etc.), die den ursprünglichen Trainingsdaten ähneln, und so kreative Anwendungen wie das Erstellen von Textdokumenten ermöglicht.
GPT (Generative Pre-Trained Transformer)	Ein LLM auf Basis der Transformer-Architektur, das auf die Vorhersage des nachfolgenden Worts zu einer gegebenen Wortsequenz trainiert wurde.
Halluzinieren	Halluzinieren ist ein typischer Fehler von LMs, bei dem das LM zwar einen syntaktisch korrekten Text erzeugt, dieser aber inhaltlich falsch und für den unkritischen Menschen dennoch überzeugend formuliert ist.

LM (Language Model), **LLM** (Large Language Model) und **SLM** (Small Language Model)	Ein LM ist ein ML-Modell, das auf Basis eines neuronalen Netzes anhand einer Menge von Referenztexten darauf trainiert wurde, natürliche Sprache zu verarbeiten und bestimmte Aufgaben wie die Klassifikation oder maschinelle Übersetzung von Dokumenten zu lösen; je nach Umfang der dafür benötigten Ressourcen in Bezug auf Datenmenge, Speicher- und Rechenkapazität sowie Komplexität spricht man entweder von LLM oder SLM, wobei es aber keine klar definierte Abgrenzung gibt.
ML (Maschinelles Lernen)	ML trainiert anhand von Lernalgorithmen Lösungen für Probleme auf der Basis von Beispieldaten (Trainingsdaten). Ein Lernalgorithmus bildet vorgegebene Beispieldaten auf ein mathematisches Modell so ab, dass dieses von den Beispieldaten auf neue Fälle verallgemeinert werden kann. Nach dem Training ist der gefundene Lösungsweg in einem ML-Modell gespeichert.
Model Inversion Attack	Bei Model Inversion Attacks versucht ein Angreifer oder eine Angreiferin, Informationen über die Trainingsdaten eines Modells zu rekonstruieren. Dies kann dazu führen, dass personenbezogene oder vertrauliche Daten aus einem trainierten Modell extrahiert werden.
NGO (Non-Governmental Organization)	Eine Nichtregierungsorganisation ist eine private Organisation, die gemeinwohlorientierte, beispielsweise soziale, gesellschaftliche oder umweltschützende, Ziele verfolgt und nicht gewinnorientiert arbeitet.
NLP (Natural Language Processing)	NLP (auch: Textanalyse) beschäftigt sich mit der Verarbeitung von natürlicher Sprache in der Form von Text- oder Sprachdaten.
Privacy by Design und **Privacy by Default**	Datenschutz durch Technikgestaltung (Privacy by Design) und datenschutzfreundliche Voreinstellungen (Privacy by Default) stellen sicher, dass betroffene Personen darauf vertrauen können, dass die grundsätzlichen Datenschutzanforderungen grundsätzlich und ohne weiteres Zutun durch die technische Umsetzung gewahrt sind.
Privacy Calculus	Der Privacy Calculus beschreibt das Entscheidungsmodell, bei dem Individuen bewusst die Vorteile und Risiken der Preisgabe persönlicher Daten abwägen. Dabei geben sie ihre Daten eher preis, wenn der erwartete Nutzen (z. B. personalisierte Dienste) die wahrgenommenen Risiken (z.B. Datenschutzverletzungen) überwiegt.
Prompt	Ein Prompt ist die natürlich-sprachige Anweisung an ein LM zur Generierung des erwünschten Ergebnisses.
RAG (Retrieval Augmented Generation)	RAG kombiniert ein allgemeines LM mit einer oder mehreren zusätzlichen spezialisierten Informationsquellen wie einer Datenbank oder einer Suchmaschine, deren relevante Informationen in den Prompt an das LM eingebettet wird, um die Ergebnisqualität des LM zu verbessern.
Trainingsdaten, **Testdaten**	Trainingsdaten sind die strukturierten oder unstrukturierten Daten, die zum Trainieren generativer KI-Modelle (GenAI) und grosser Sprachmodelle (LLMs) verwendet werden. Testdaten werden zur qualitativen (augenscheinlichen) und quantitativen (gemessener Qualitätsmetriken) Überprüfung (Evaluierung) des KI-Modells verwendet und beinhalten ggf. eine manuelle Bewertung.

UC (Use Case)	Ein Use Case (Anwendungsfall) beschreibt aus Sicht der Anwendenden, wie dieser mit Hilfe der Funktionen, die ein System zur Verfügung stellt, ein oder mehrere Ziele erreicht.
UI (User Interface)	Das UI ist die (meist graphische oder sprachliche) Benutzerschnittstelle, über die der menschliche Benutzer mit dem Computer interagiert.
Word Embedding	Ein Word Embedding bildet ein gegebenes Wort auf einen zahlenbasierten Vektor in einem mehrdimensionalen Raum ab, da sich Vektoren besser für die algorithmische Verarbeitung eignen als die menschenlesbare Zeichenkette des Worts. In den letzten Jahren wurden zahlreiche Verfahren zur Berechnung von Word Embeddings entwickelt, darunter die ML-basierten word2vec und BERT.
XAI (eXplainable AI)	XAI zielt darauf ab, das von einem ML-Modell erzeugte Ergebnis in einer für den menschlichen Nutzenden verständlichen Form darzustellen.

11 Literaturverzeichnis

[1] Mostafa Abbasi, Rahnuma Islam Nishat, Corey Bond, John Brandon Graham-Knight, Patricia Lasserre, Yves Lucet, and Homayoun Najjaran. 2024. A Review of AI and Machine Learning Contribution in Predictive Business Process Management (Process Enhancement and Process Improvement Approaches). https://doi.org/10.48550/ARXIV.2407.11043

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6, (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Jay Alammar and Maarten Grootendorst. 2024. *Hands-on large language models: language understanding and generation* (1st edition ed.). O'Reilly, Beijing Boston Farnham.

[4] Abdulaziz Aldoseri, Khalifa N. Al-Khalifa, and Abdel Magid Hamouda. 2023. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Applied Sciences* 13, 12 (June 2023), 7082. https://doi.org/10.3390/app13127082

[5] Alexander Wan, Eric Wallace, Sheng Shen, Dan Klein. 2023. Poisoning Language Models During Instruction Tuning. In *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, 2023. 35413–35425. https://arxiv.org/abs/2305.00944

[6] Ethem Alpaydin. 2020. *Introduction to machine learning*. MIT press.

[7] E. Ameisen. 2020. *Building Machine Learning Powered Applications: Going from Idea to Product*. O'Reilly. Retrieved from https://books.google.ch/books?id=-3mfxgEACAAJ

[8] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, May 2019. IEEE, Montreal, QC, Canada, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

[9] Armenia Androniceanu and Irina Georgescu. 2021. E-Government in European Countries, a Comparative Approach Using the Principal Components Analysis. *NISPAcee Journal of Public Administration and Policy* 14, 2 (December 2021), 65–86. https://doi.org/10.2478/nispa-2021-0015

[10] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. 2023. .

[11] Adebayo Yusuf Balogun, Olufunke Cynthia Metibemu, Abayomi Titilola Olutimehin, Adekunbi Justina Ajayi, Damilola Comfort Babarinde, and Oluwaseun Oladeji Olaniyi. 2025. The Ethical and Legal Implications of Shadow AI in Sensitive Industries: A Focus on Healthcare, Finance and Education. *SSRN Journal* (2025). https://doi.org/10.2139/ssrn.5137049

[12] Marco Barenkamp. 2023. Einflussnahme großer Sprachmodelle auf die moderne Arbeitswelt. *Informatik Spektrum* 46, 4 (August 2023), 185–188. https://doi.org/10.1007/s00287-023-01546-8

[13] Liam Barkley and Brink van der Merwe. 2024. Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models. https://doi.org/10.48550/ARXIV.2410.19385

[14] Marco Bertenghi. 2025. Angriffe auf große Sprachmodelle. Retrieved from https://www.heise.de/select/ix/2025/1/2424713350370301071

[15] Christopher M. Bishop. 2006. *Pattern recognition and machine learning*. Springer, New York.

[16] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. 2021. LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2021. IEEE, Nashville, TN, USA, 1102–1110. https://doi.org/10.1109/CVPRW53098.2021.00121

[17] Nadja Braun Binder. 2019. Vollautomatisierte Verwaltungsverfahren, vollautomatisiert erlassene Verwaltungsakte und elektronische Aktenführung. In *Braun Binder, Nadja (2019). Vollautomatisierte Verwaltungsverfahren, vollautomatisiert erlassene Verwaltungsakte und elektronische Aktenführung. In: Seckelmann, Margrit. Digitalisierte Verwaltung: Vernetztes E-Government (2. Auflage). Berlin: Erich Schmidt Verlag, 311-326.*, Margrit Seckelmann (ed.). Erich Schmidt Verlag, Berlin, 311–326. Retrieved April 21, 2023 from https://www.zora.uzh.ch/id/eprint/179185/

[18] Nadja Braun Binder. 2019. Künstliche Intelligenz und automatisierte Entscheidungen in der öffentlichen Verwaltung. *Schweizerische Juristen-Zeitung (SJZ)* 15 (2019), 467–476.

[19] Nadja Braun Binder, Matthias Spielkamp, Catherine Egli, Laurent Freiburghaus, Eliane Kunz, Nina Laukenmann, Michele Loi, Anna Mätzener, Liliane Obrecht, and Jessica Wulf. 2021. Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen. (2021). https://doi.org/10.5451/UNIBAS-EP83846

[20] Jonathan Bright, Florence E. Enock, Saba Esnaashari, John Francis, Youmna Hashem, and Deborah Morgan. 2024. Generative AI is already widespread in the public sector. https://doi.org/10.48550/ARXIV.2401.01291

[21] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/ARXIV.2005.14165

[22] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2023. *Generative AI at Work*. National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w31161

[23] Beatriz M. Cabrera, Luiz E. Luiz, and João P. Teixeira. 2025. The Artificial Intelligence Act: Insights regarding its application and implications. *Procedia Computer Science* 256, (2025), 230–237. https://doi.org/10.1016/j.procs.2025.02.116

[24] Stefano Ceri, Alessandro Bozzon, Marco Brambilla, Emanuele Della Valle, Piero Fraternali, and Silvia Quarteroni. 2013. *Web Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39314-3

[25] Dhivya Chandrasekaran and Vijay Mago. 2022. Evolution of Semantic Similarity—A Survey. *ACM Comput. Surv.* 54, 2 (March 2022), 1–37. https://doi.org/10.1145/3440755

[26] Tzuhao Chen and Mila Gasco-Hernandez. 2024. Uncovering the Results of AI Chatbot Use in the Public Sector: Evidence from US State Governments. *Public Performance & Management Review* (August 2024), 1–26. https://doi.org/10.1080/15309576.2024.2389864

[27] Xiang Chen, Chaoyang Gao, Chunyang Chen, Guangbei Zhang, and Yong Liu. 2025. An Empirical Study on Challenges for LLM Application Developers. *ACM Trans. Softw. Eng. Methodol.* (January 2025), 3715007. https://doi.org/10.1145/3715007

[28] Sunitha Cheriyan, Shaniba Ibrahim, Saju Mohanan, and Susan Treesa. 2018. Intelligent Sales Prediction Using Machine Learning Techniques. In *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, August 2018. IEEE, Southend, United Kingdom, 53–58. https://doi.org/10.1109/iCCECOME.2018.8659115

[29] Mark Considine, Michael Mcgann, Sarah Ball, and Phuc Nguyen. 2022. Can Robots Understand Welfare? Exploring Machine Bureaucracies in Welfare-to-Work. *J. Soc. Pol.* 51, 3 (July 2022), 519–534. https://doi.org/10.1017/S0047279422000174

[30] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21, 2022. ACM, Seoul Republic of Korea, 864–876. https://doi.org/10.1145/3531146.3533150

[31] Leandro DalleMule and Thomas H Davenport. 2017. What's your data strategy. *Harvard business review* 95, 3 (2017), 112–121.

[32] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* 57, 6 (June 2025), 1–39. https://doi.org/10.1145/3712001

[33] Fabiola Del Toro Osorio, Victoria Eugenia Ospina Becerra, and Elsa Estévez. 2023. Designing a Data Strategy for Organizations. In *Cloud Computing, Big Data & Emerging Topics*, Marcelo Naiouf, Enzo Rucci, Franco Chichizola and Laura De Giusti (eds.). Springer Nature Switzerland, Cham, 143–156. https://doi.org/10.1007/978-3-031-40942-4_11

[34] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2021. The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. https://doi.org/10.48550/ARXIV.2105.03354

[35] Kevin C. Desouza, Gregory S. Dawson, and Daniel Chenok. 2020. Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons* 63, 2 (March 2020), 205–213. https://doi.org/10.1016/j.bushor.2019.11.004

[36] Alejandro Díaz, Kayvaun Rowshankish, and Tamim Saleh. 2018. Why data culture matters. *McKinsey Quarterly* 3, 1 (2018), 36–53.

[37] Catherine D'Ignazio and Rahul Bhargava. 2015. Approaches to building big data literacy. In *Proceedings of the Bloomberg data for good exchange conference*, 2015. .

[38] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2008. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC 2008)*, 2008. 1–19. https://doi.org/10.1007/978-3-540-79228-4_1

[39] Abdussalam Elhanashi. 2025. *Deep Learning in Action: Image and Video Processing for Practical Use* (1st ed ed.). Elsevier, Chantilly.

[40] Zeynep Engin and Philip Treleaven. 2019. Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies. *The Computer Journal* 62, 3 (March 2019), 448–460. https://doi.org/10.1093/comjnl/bxy082

[41] Wolfgang Ertel. 2025. *Introduction to Artificial Intelligence*. Springer Fachmedien Wiesbaden, Wiesbaden. https://doi.org/10.1007/978-3-658-43102-0

[42] Beat Estermann, Marianne Fraefel, Alessia C. Neuroni, and Juergen Vogel. 2018. Conceptualizing a national data infrastructure for Switzerland. *IP* 23, 1 (February 2018), 43–65. https://doi.org/10.3233/IP-170033

[43] EU. 2016. Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung). Retrieved from https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32016R0679

[44] Virginia Eubanks. 2019. *Automating inequality: how high-tech tools profile, police, and punish the poor* (First Picador edition ed.). Picador St. Martin's Press, New York.

[45] European Data Protection Board. 2024. Report of the work undertaken by the ChatGPT Taskforce. (2024). Retrieved from https://www.edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-chatgpt-taskforce_en

[46] European Parliament. Directorate General for Parliamentary Research Services. 2020. *The impact of the general data protection regulation on artificial intelligence.* Publications Office, LU. Retrieved March 26, 2025 from https://data.europa.eu/doi/10.2861/293

[47] Fabian Engl. 2023. Verfügbarkeit von personenbezogenen Daten unter dem Aspekt der Anonymisierung. *Anwendungen und Konzepte der Wirtschaftinformatik* (2023). Retrieved from https://akwi.hswlu.ch/article/view/3827

[48] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. Generative AI. *Bus Inf Syst Eng* 66, 1 (February 2024), 111–126. https://doi.org/10.1007/s12599-023-00834-7

[49] Luciano Floridi and Mariarosaria Taddeo. 2016. What is data ethics? *Phil. Trans. R. Soc. A.* 374, 2083 (December 2016), 20160360. https://doi.org/10.1098/rsta.2016.0360

[50] Nils Freyer, Hendrik Kempt, and Lars Klöser. 2024. Easy-read and large language models: on the ethical dimensions of LLM-based text simplification. *Ethics Inf Technol* 26, 3 (September 2024), 50. https://doi.org/10.1007/s10676-024-09792-4

[51] Nicolad Garneau and Olivier Bolduc. 2024. The State of Commercial Automatic French Legal Speech Recognition Systems and their Impact on Court Reporters et al. https://doi.org/10.48550/ARXIV.2408.11940

[52] Stephen Gilbert. 2024. The EU passes the AI Act and its implications for digital medicine are unclear. *npj Digit. Med.* 7, 1 (May 2024), 135. https://doi.org/10.1038/s41746-024-01116-6

[53] Andreas Glaser. 2018. Einflüsse der Digitalisierung auf das schweizerische Verwaltungsrecht. *Schweizerische Juristen-Zeitung* 114, 8 (April 2018), 181–190.

[54] Jake Goldenfein. 2019. Algorithmic transparency and decision-making accountability: Thoughts for buying machine learning algorithms. *Jake Goldenfein,'Algorithmic Transparency and Decision-Making Accountability: Thoughts for buying machine learning algorithms' in Office of the Victorian Information Commissioner (ed), Closer to the Machine: Technical, Social, and Legal aspects of AI (2019)* (2019).

[55] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation." *AIMag* 38, 3 (October 2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[56] Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchán. 2023. A survey of Generative AI Applications. https://doi.org/10.48550/ARXIV.2306.02781

[57] Stephan Grimmelikhuijsen. 2023. Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making. *Public Administration Review* 83, 2 (March 2023), 241–262. https://doi.org/10.1111/puar.13483

[58] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Sci. Robot.* 4, 37 (December 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[59] Ulrich Häfelin, Georg Müller, and Felix Uhlmann. 2020. *Allgemeines Verwaltungsrecht* (8., überarbeitet Auflage ed.). Dike, Zürich St. Gallen.

[60] Dominik Halvoník and Jozef Kapusta. 2024. Large Language Models and Rule-Based Approaches in Domain-Specific Communication. *IEEE Access* 12, (2024), 107046–107058. https://doi.org/10.1109/ACCESS.2024.3436902

[61] Jiawei Han, Jiankang Lu, Ying Xu, Jin You, and Bingxin Wu. 2024. Intelligent Practices of Large Language Models in Digital Government Services. *IEEE Access* 12, (2024), 8633–8640. https://doi.org/10.1109/ACCESS.2024.3349969

[62] Lauren Harrington. 2023. Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Front. Commun.* 8, (July 2023), 1165233. https://doi.org/10.3389/fcomm.2023.1165233

[63] Moreen Heine, Anna-Katharina Dhungel, Tim Schrills, and Daniel Wessel. 2023. *Künstliche Intelligenz in öffentlichen Verwaltungen: Grundlagen, Chancen, Herausforderungen und Einsatzszenarien*. Springer Fachmedien Wiesbaden, Wiesbaden. https://doi.org/10.1007/978-3-658-40101-6

[64] Moreen Heine, Anna-Katharina Dhungel, Tim Schrills, and Daniel Wessel. 2023. *Künstliche Intelligenz in öffentlichen Verwaltungen: Grundlagen, Chancen, Herausforderungen und Einsatzszenarien*. Springer Fachmedien Wiesbaden, Wiesbaden. https://doi.org/10.1007/978-3-658-40101-6

[65] Hess, Thomas, Christian Matt, Verena Thürmel, and Mena Teebken. 2022. Zum Zusammenspiel zwischen Unternehmen und Verbrauchern in der Datenökonomie: Herausforderungen und neue Gestaltungsansätze. In *Die Zukunft von Privatheit und Selbstbestimmung*. Springer Vieweg, 93–124. Retrieved from https://doi.org/10.1007/978-3-658-35263-9_3

[66] Illugi Torfason Hjaltalin and Hallur Thor Sigurdarson. 2024. The strategic use of AI in the public sector: A public values analysis of national AI strategies. *Government Information Quarterly* 41, 1 (March 2024), 101914. https://doi.org/10.1016/j.giq.2024.101914

[67] Zijin Hong, Zheng Yuan, Qinggang Zhang, Hao Chen, Junnan Dong, Feiran Huang, and Xiao Huang. 2024. Next-Generation Database Interfaces: A Survey of LLM-based Text-to-SQL. https://doi.org/10.48550/ARXIV.2406.08426

[68] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2 (March 2025), 1–55. https://doi.org/10.1145/3703155

[69] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 03, 2021. ACM, Virtual Event Canada, 560–575. https://doi.org/10.1145/3442188.3445918

[70] Chip Huyen. 2025. *AI engineering: building applications with foundation models*. O'Reilly, Sebastopol,CA.

[71] Independent High Level Expert Group on Artificial Intelligence set up by the European Commission. 2019. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. Retrieved April 28, 2023 from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[72] Todor Ivanov and Valeri Penchev. 2024. AI Benchmarks and Datasets for LLM Evaluation. https://doi.org/10.48550/ARXIV.2412.01020

[73] Sonia Jaffe, Neha Parikh Shah, Jenna Butler, Alex Farach, Alexia Cambon, Brent Hecht, Michael Schwarz, and Jaime Teevan. 2024. *Generative AI in Real-World Workplaces*. Microsoft. Retrieved from https://www.microsoft.com/en-us/research/publication/generative-ai-in-real-world-workplaces/

[74] Qilu Jiao and Shunyao Zhang. 2021. A Brief Survey of Word Embedding and Its Recent Development. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, March 12, 2021. IEEE, Chongqing, China, 1697–1701. https://doi.org/10.1109/IAEAC50856.2021.9390956

[75] Per-Olov Johansson and Bengt Kriström. 2016. *Cost-benefit analysis for project appraisal*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781316392751

[76] Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed.). Retrieved from https://web.stanford.edu/~jurafsky/slp3/

[77] Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. 2024. Evaluating LLMs at Detecting Errors in LLM Responses. https://doi.org/10.48550/ARXIV.2404.03602

[78] Qinqing Kang and Xiong Ding. 2021. Urban management image classification approach based on deep learning. *AIS* 13, 5 (September 2021), 347–360. https://doi.org/10.3233/AIS-210609

[79] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A Survey of Reinforcement Learning from Human Feedback. https://doi.org/10.48550/ARXIV.2312.14925

[80] Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion* 109, (September 2024), 102422. https://doi.org/10.1016/j.inffus.2024.102422

[81] D Kiron, PK Prentice, and R Ferguson. 2012. The strategic and operational impact of big data. *MIT Sloan Management Review* 53, 2 (2012), 21–31.

[82] Tanja Klenk, Frank Nullmeier, and Göttrik Wewer (Eds.). 2025. *Handbuch Digitalisierung in Staat und Verwaltung*. Springer Fachmedien Wiesbaden, Wiesbaden. https://doi.org/10.1007/978-3-658-37373-3

[83] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, November 2024. Miami, Florida, United States, 1–46. Retrieved from https://hal.science/hal-04789354

[84] Alma Kolleck and Carsten Orwat. 2020. *Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick*. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB). https://doi.org/10.5445/IR/1000127166

[85] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. 2023. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access* 11, (2023), 31866–31879. https://doi.org/10.1109/ACCESS.2023.3262138

[86] Ulla Kruhse-Lehtonen and Dirk Hofmann. 2020. How to Define and Execute Your Data and AI Strategy. *Harvard Data Science Review* (July 2020). https://doi.org/10.1162/99608f92.a010feeb

[87] Miroslav Kubat. 2015. *An Introduction to Machine Learning*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-20010-1

[88] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I. Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3580882

[89] Mascha Kurpicz-Briki. 2023. *More than a Chatbot: Language Models Demystified*. Springer Nature Switzerland, Cham. https://doi.org/10.1007/978-3-031-37690-0

[90] Samuli Laato, Miika Tiainen, A.K.M. Najmul Islam, and Matti Mäntymäki. 2022. How to explain AI systems to end users: a systematic literature review and research agenda. *INTR* 32, 7 (December 2022), 1–31. https://doi.org/10.1108/INTR-08-2021-0600

[91] Anna Grøndahl Larsen and Asbjørn Følstad. 2024. The impact of chatbots on public service provision: A qualitative interview study with citizens and public service providers. *Government Information Quarterly* 41, 2 (June 2024), 101927. https://doi.org/10.1016/j.giq.2024.101927

[92] Patrick Levi. 2025. RAG-Systeme absichern. Retrieved from https://www.heise.de/select/ix/2025/1/2427011035369852053

[93] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, 2020. Curran Associates, Inc., 9459–9474. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[94] Nut Limsopatham. 2021. Effectively Leveraging BERT for Legal Document Classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, 210–216. https://doi.org/10.18653/v1/2021.nllp-1.22

[95] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2022. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* 54, 2 (March 2022), 1–36. https://doi.org/10.1145/3436755

[96] J. Loof, D. Falavigna, R. Schluter, D. Giuliani, R. Gretter, and H. Ney. 2010. Evaluation of automatic transcription systems for the judicial domain. In *2010 IEEE Spoken Language Technology Workshop*, December 2010. IEEE, Berkeley, CA, USA, 206–211. https://doi.org/10.1109/SLT.2010.5700852

[97] D. Lu and Q. Weng. 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* 28, 5 (March 2007), 823–870. https://doi.org/10.1080/01431160600746456

[98] Abdul Majeed and Sungchang Lee. 2021. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access* 9, (2021), 8512–8545. https://doi.org/10.1109/ACCESS.2020.3045700

[99] Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison Cottrell. 2020. Speech Recognition and Multi-Speaker Diarization of Long Conversations. https://doi.org/10.48550/ARXIV.2005.08072

[100] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. Artificial Intelligence Index Report 2024. https://doi.org/10.48550/ARXIV.2405.19522

[101] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (July 2022), 1–35. https://doi.org/10.1145/3457607

[102] Mohammad I. Merhi. 2023. An evaluation of the critical success factors impacting artificial intelligence implementation. *International Journal of Information Management* 69, (April 2023), 102545. https://doi.org/10.1016/j.ijinfomgt.2022.102545

[103] Slava Jankin Mikhaylov, Marc Esteve, and Averill Campion. 2018. Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration. *Phil. Trans. R. Soc. A.* 376, 2128 (September 2018), 20170357. https://doi.org/10.1098/rsta.2017.0357

[104] Aman Kumar Mishra, Amit Kumar Tyagi, Sathian Dananjayan, Anand Rajavat, Hitesh Rawat, and Anjali Rawat. 2024. Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration. In *Conversational Artificial Intelligence* (1st ed.), Romil Rawat, Rajesh Kumar Chakrawarti, Sanjaya Kumar Sarangi, Piyush Vyas, Mary Sowjanya Alamanda, Kotagiri Srividya and Krishnan Sakthidasan Sankaran (eds.). Wiley, 607–634. https://doi.org/10.1002/9781394200801.ch34

[105] Prafful Mishra. 2025. *A Guide to Implementing MLOps: From Data to Operations*. Springer Nature Switzerland, Cham. https://doi.org/10.1007/978-3-031-82010-6

[106] Lokke Moerel and Paul Timmers. 2021. Reflections on digital sovereignty. *EU cyber direct, research in focus series* (2021). Retrieved from https://ssrn.com/abstract=3772777

[107] Holger F. H. Mühlenkamp. 2010. Zur Kosten-Nutzen-Analyse von E-Government-Projekten. In *E-Government*, Bernd W. Wirtz (ed.). Gabler, Wiesbaden, 177–193. https://doi.org/10.1007/978-3-8349-6343-7_9

[108] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A Comprehensive Overview of Large Language Models. https://doi.org/10.48550/ARXIV.2307.06435

[109] Whitney Nelson, Min Kyung Lee, Eunsol Choi, and Victor Wang. Designing LLM-based support for homelessness caseworkers. In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*, .

[110] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel. 2020. Extracting Training Data from Large Language Models. In *Proceedings of the 30th USENIX Security Symposium*, 2020. . https://arxiv.org/abs/2012.07805

[111] Nima Oshnooei. 2022. Data Science: Privacy by Design-Strategien bei KI/ML-Systemen. Retrieved from https://www.dr-datenschutz.de/data-science-privacy-by-design-strategien-bei-ki-ml-systemen/

[112] Krishna Kumar Nirala, Nikhil Kumar Singh, and Vinay Shivshanker Purani. 2022. A survey on providing customer and public administration based services using AI: chatbot. *Multimed Tools Appl* 81, 16 (July 2022), 22215–22246. https://doi.org/10.1007/s11042-021-11458-y

[113] Farhad Nooralahzadeh, Yi Zhang, Ellery Smith, Sabine Maennel, Cyril Matthey-Doret, Raphaël de Fondville, and Kurt Stockinger. 2024. StatBot.Swiss: Bilingual Open Data Exploration in Natural Language. https://doi.org/10.48550/arXiv.2406.03170

[114] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Min & Knowl* 10, 3 (May 2020), e1356. https://doi.org/10.1002/widm.1356

[115] Mohammad Nuruzzaman and Omar Khadeer Hussain. 2018. A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, October 2018. IEEE, Xi'an, 54–61. https://doi.org/10.1109/ICEBE.2018.00019

[116] Takeshi Okadome. 2025. *Essentials of Generative AI*. Springer Nature Singapore, Singapore. https://doi.org/10.1007/978-981-96-0029-8

[117] Ole Olesen-Bagneux. 2023. *The Enterprise Data Catalog*. O'Reilly Media, Inc.

[118] Ville Paananen, Jonas Oppenlaender, and Aku Visuri. 2024. Using text-to-image generation for architectural design ideation. *International Journal of Architectural Computing* 22, 3 (September 2024), 458–474. https://doi.org/10.1177/14780771231222783

[119] Peter Parycek, Verena Schmid, and Anna-Sophie Novak. 2023. Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions. *J Knowl Econ* 15, 2 (June 2023), 8390–8415. https://doi.org/10.1007/s13132-023-01433-3

[120] Jayesh Patel. 2019. Bridging Data Silos Using Big Data Integration. *IJDMS* 11, 3 (June 2019), 01–06. https://doi.org/10.5121/ijdms.2019.11301

[121] Jayesh Patel. 2019. Bridging Data Silos Using Big Data Integration. *IJDMS* 11, 3 (June 2019), 01–06. https://doi.org/10.5121/ijdms.2019.11301

[122] Werner Pfeil. 2020. „Leinen los und Fahrt voraus!" – Evolutionäre Algorithmen, Künstliche Intelligenz und Legal Tech ändern das Recht, die Rechtsordnung und den Zugang zum Recht. *InTeR* (2020), 17–25.

[123] David Plotkin. 2014. *Data stewardship: an actionable guide to effective data management and data governance*. Elsevier/Morgan Kaufman, Amsterdam ; Boston.

[124] Julia Pohle and Thorsten Thiel. 2020. Digital sovereignty. *Internet Policy Review* 9, 4 (December 2020). https://doi.org/10.14763/2020.4.1532

[125] Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. 2024. End-to-End Speech Recognition: A Survey. *IEEE/ACM Trans. Audio Speech Lang. Process.* 32, (2024), 325–351. https://doi.org/10.1109/TASLP.2023.3328283

[126] Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (Almost) Dead. https://doi.org/10.48550/ARXIV.2309.09558

[127] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, March 07, 2017. ACM, Oslo Norway, 117–126. https://doi.org/10.1145/3020165.3020183

[128] Meike Ramon, Alexandre Barbey, and Sylvain Métille. 2024. Face Recognition Technology in Swiss Law Enforcement. *AJP / AJA : Aktuelle Juristische Praxis / Pratique Juridique Actuelle* 2024, 02 (February 2024), 128–143.

[129] David Rechsteiner. 2018. Der Algorithmus verfügt. *Jusletter* (November 2018), 1–16.

[130] Manuel Riesen and Peter von Niederhäusern. 2025. Gespräche mit LLMs. Retrieved from https://www.societybyte.swiss/2025/01/10/conversations-with-llms-patterns-and-paradigms-for-better-prompting/

[131] Richard H R Roberts, Stephen R Ali, Thomas D Dobbs, and Iain S Whitaker. 2024. Can Large Language Models Generate Outpatient Clinic Letters at First Consultation That Incorporate Complication Profiles From UK and USA Aesthetic Plastic Surgery Associations? *Aesthetic Surgery Journal Open Forum* 6, (January 2024), ojad109. https://doi.org/10.1093/asjof/ojad109

[132] Susanne Rönnecke. 2024. Die europäische KI-Verordnung. Ein Weg hin zu diskriminierungsfreien Algorithmen? *Journal Netzwerk Frauen- und Geschlechterforschung NRW* 2024, (December 2024), 65. https://doi.org/10.17185/DUEPUBLICO/82756

[133] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[134] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. https://doi.org/10.48550/ARXIV.2402.07927

[135] SBFI. 2020. Leitlinien «Künstliche Intelligenz» für den Bund. Retrieved from https://www.sbfi.admin.ch/sbfi/de/home/bfi-politik/bfi-2021-2024/transversale-themen/digitalisierung-bfi/kuenstliche-intelligenz.html

[136] Schweizerische Eidgenossenschaft. 2023. Revidiertes Datenschutzgesetz der Schweiz. Retrieved from https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-90134.html

[137] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, January 29, 2019. ACM, Atlanta GA USA, 59–68. https://doi.org/10.1145/3287560.3287598

[138] Friso Selten, Marcel Robeer, and Stephan Grimmelikhuijsen. 2023. 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review* 83, 2 (March 2023), 263–278. https://doi.org/10.1111/puar.13602

[139] Sajani Senadheera, Tan Yigitcanlar, Kevin C. Desouza, Karen Mossberger, Juan Corchado, Rashid Mehmood, Rita Yi Man Li, and Pauline Hope Cheong. 2024. Understanding Chatbot Adoption in Local Governments: A Review and Framework. *Journal of Urban Technology* (February 2024), 1–35. https://doi.org/10.1080/10630732.2023.2297665

[140] Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access* 13, (2025), 18253–18276. https://doi.org/10.1109/ACCESS.2025.3533217

[141] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (November 1999), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

[142] Vinay Sridhar, Sriram Subramanian, Dulcardo Arteaga, Swaminathan Sundararaman, Drew Roselli, and Nisha Talagala. 2018. Model Governance: Reducing the Anarchy of Production ML. In *Proceedings of the 2018 USENIX Conference on Usenix Annual Technical Conference* (*USENIX ATC '18*), 2018. USENIX Association, USA, 351–357.

[143] Jess Stratton. 2024. *Copilot for Microsoft 365: Harness the Power of Generative AI in the Microsoft Apps You Use Every Day*. Apress, Berkeley, CA. https://doi.org/10.1007/979-8-8688-0447-2

[144] Tao Sun, Ming Shan, Xing Rong, and Xudong Yang. 2022. Estimating the spatial distribution of solar photovoltaic power generation potential on different types of rural rooftops using a deep learning network applied to satellite images. *Applied Energy* 315, (June 2022), 119025. https://doi.org/10.1016/j.apenergy.2022.119025

[145] Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. The Use of Generative Search Engines for Knowledge Work and Complex Tasks. https://doi.org/10.48550/ARXIV.2404.04268

[146] Latanya Sweeney. 2002. k-Anonymity: A model for protecting privacy. 10, 5 (2002), 557–570. https://doi.org/10.1142/S0218488502001648

[147] Lukas Teufelberger, Xintong Liu, Zhipeng Li, Max Möbus, and Christian Holz. 2024. Demonstrating LLM-for-X: Application-agnostic Integration of Large Language Models to Support Writing Workflows. In *The 37th Annual ACM Symposium on User Interface Software and Technology*, October 13, 2024. ACM, Pittsburgh PA USA, 1–3. https://doi.org/10.1145/3672539.3686757

[148] Daniela Thurnherr. 2021. Automatisierte Verwaltungsverfahren auf Bundesebene (Schweiz). In *Auswirkungen der Digitalisierung auf die Erlassung und Zuordnung behördlicher Entscheidungen*, Nadja Braun Binder, Peter Bußjäger and Mathias Eller (eds.). New academic press.

[149] Roberto Togneri and Daniel Pullella. 2011. An Overview of Speaker Identification: Accuracy and Robustness Issues. *IEEE Circuits Syst. Mag.* 11, 2 (2011), 23–61. https://doi.org/10.1109/MCAS.2011.941079

[150] Lukas Tuggener, Pascal Sager, Yassine Taoudi-Benchekroun, Benjamin F. Grewe, and Thilo Stadelmann. 2024. So you want your private LLM at home? A survey and benchmark of methods for efficient GPTs. In *2024 11th IEEE Swiss Conference on Data Science (SDS)*, May 30, 2024. IEEE, Zurich, Switzerland, 205–212. https://doi.org/10.1109/SDS60720.2024.00036

[151] United Nations Department of Economic and Social Affairs. 2024. *United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development - With the addendum on Artificial Intelligence*. United Nations.

[152] Colin Van Noordt and Gianluca Misuraca. 2022. Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Government Information Quarterly* 39, 3 (July 2022), 101714. https://doi.org/10.1016/j.giq.2022.101714

[153] Ian Wallis. 2021. *Data strategy: from definition to execution*. BCS Learning and Development, Swindon, UK.

[154] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. https://doi.org/10.48550/ARXIV.2308.13387

[155] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. https://doi.org/10.48550/ARXIV.2206.07682

[156] Martin Wild, Sascha Herges, and Justus Liebig University Giessen. 2000. *Total Cost of Ownership (TCO) : Ein Überblick*. Universitätsbibliothek Gießen. https://doi.org/10.22029/JLUPUB-2098

[157] Christian Winkler. 2025. Small Language Models: Grosse Sprachmodelle werden klein. Retrieved from https://www.heise.de/hintergrund/Uebersicht-ueber-kleine-KI-Modelle-die-sich-fuer-den-lokalen-Betrieb-eignen-10306979.html

[158] Johannes Wirth and René Peinl. 2024. ASR Bundestag: A Large-Scale Political Debate Dataset in German. In *Intelligent Systems and Applications*, Kohei Arai (ed.). Springer Nature Switzerland, Cham, 190–202. https://doi.org/10.1007/978-3-031-47724-9_13

[159] Bernd W. Wirtz, Jan C. Weyerer, and Carolin Geyer. 2019. Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration* 42, 7 (May 2019), 596–615. https://doi.org/10.1080/01900692.2018.1498103

[160] Heinrich Amadeus Wolff, Stefan Brink, and Marion Albers. 2022. *Datenschutzrecht: DS-GVO, BDSG, Grundlagen, bereichsspezifischer Datenschutz: Kommentar* (2. Auflage ed.). C.H. Beck, München.

[161] Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When Search Engine Services Meet Large Language Models: Visions and Challenges. *IEEE Trans. Serv. Comput.* 17, 6 (November 2024), 4558–4577. https://doi.org/10.1109/TSC.2024.3451185

[162] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans. Knowl. Discov. Data* 18, 6 (July 2024), 1–32. https://doi.org/10.1145/3649506

[163] Tan Yigitcanlar, Rita Yi Man Li, Tommi Inkinen, and Alexander Paz. 2022. Public Perceptions on Application Areas and Adoption Challenges of AI in Urban Services. *Emerg Sci J* 6, 6 (September 2022), 1199–1236. https://doi.org/10.28991/ESJ-2022-06-06-01

[164] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric Artificial Intelligence: A Survey. *ACM Comput. Surv.* 57, 5 (May 2025), 1–42. https://doi.org/10.1145/3711118

[165] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. 2023. Text-to-image Diffusion Models in Generative AI: A Survey. https://doi.org/10.48550/ARXIV.2303.07909

[166] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. https://doi.org/10.48550/ARXIV.2210.03493

[167] Nan-ning Zheng, Zi-yi Liu, Peng-ju Ren, Yong-qiang Ma, Shi-tao Chen, Si-yu Yu, Jian-ru Xue, Ba-dong Chen, and Fei-yue Wang. 2017. Hybrid-augmented intelligence: collaboration and cognition. *Frontiers Inf Technol Electronic Eng* 18, 2 (February 2017), 153–179. https://doi.org/10.1631/FITEE.1700053