Künstliche Intelligenz (KI) im Bereich Geodaten: Large Language Models (LLM) für die semantische Suche und Abfragen entwickeln

Aktion 5-25-2

(Version 1.0, 24.11.2025)

Projektleitung | *Direction du projet*	
Projektkoordination (PROK SGS) *Coordination de projet (PROK SGS)*	Pasquale Di Donato
Team	Raphaëlle Arnaud, Christine Najar, Pasquale Di Donato
Vertragssumme inkl. MWST *Montant contractuel avec TVA*	-
Vertragsende | *Fin du contrat*	31.12.2025

Management Summary

Die Massnahme 5-25-2 des Aktionsplans Geoinformation Schweiz 2025 untersucht den Einsatz von grossen Sprachmodellen (LLMs) für die semantische Suche und Abfragen im Bereich Geodaten. Ziel ist es, die Potenziale und Grenzen dieser Technologien systematisch zu erfassen und eine Entscheidungsgrundlage für die strategischen Gremien zu schaffen. Im Rahmen der Arbeiten wurden mehrere Prototypen entwickelt, die den Zugang zu Geodaten über natürliche Sprache erleichtern und sowohl Fachleuten als auch Laien neue Möglichkeiten eröffnen.

Die Analyse zeigt, dass LLMs erhebliche Chancen für die Verbesserung der Auffindbarkeit und Nutzung von Geodaten bieten, gleichzeitig aber auch technische, ethische, ökologische und rechtliche Risiken bergen. Zu den zentralen Herausforderungen zählen Halluzinationen, mangelnde Reproduzierbarkeit, der Black-Box-Effekt sowie Sicherheitsrisiken wie Prompt Injection oder Datenlecks. Ergänzend wurden Fragen zu Bias, schädlichen Inhalten und Umweltbelastungen durch Energie- und Wasserverbrauch untersucht.

Als Minderungsmassnahmen wurden unter anderem der Einsatz kleinerer Modelle, verbessertes Prompt Design, Explainable AI sowie die Kopplung mit vertrauenswürdigen Datenquellen über RAG identifiziert. Parallel dazu müssen regulatorische Entwicklungen in der Schweiz berücksichtigt werden sowie das nationale LLM «Apertus» und die Standards eCH-0272.

Die Arbeiten wurden in enger Abstimmung mit Fachleuten aus Verwaltung, Forschung und Industrie durchgeführt, um Synergien zu nutzen und die Prototypen praxisnah zu testen. Erste Ergebnisse zeigen, dass LLMs bei räumlichen Fragestellungen unterstützen können, jedoch spezifische Use Cases für die weitere Validierung erforderlich sind.

Für die nächsten Schritte wird die Umsetzung und Validierung der Minderungsmassnahmen empfohlen, ebenso die Fortführung des Use Cases «Nutzung von Geodaten» mit konkreten Beispielen. Zudem soll die Umsetzbarkeit der semantischen Suche in ersten Umsetzungsschritten überprüft werden.

Das Projekt liefert damit eine fundierte Grundlage, um über die Integration von LLMs in die Nationale Geodaten-Infrastruktur (NGDI) ab 2026 zu entscheiden. Es zeigt klar auf, dass Chancen und Risiken sorgfältig abgewogen werden müssen, um eine verantwortungsvolle und nachhaltige Nutzung sicherzustellen.

Inhaltsverzeichnis

1 Einleitung

1.1 Hintergrund

Derzeit ermöglichen Geodaten-Infrastrukturen hauptsächlich den Zugriff auf Geodaten, die entweder als Luftbild, Pixelkarte oder strukturierte Vektorgeodaten gespeichert sind. Nutzeranfragen erfolgen dabei typischerweise über die Eingabe von Schlüsselwörtern in ein Suchfeld, und die gelieferten Antworten bestehen aus einer strukturierten Liste verschiedener Elemente, die das gesuchte Schlüsselwort oder die gesuchten Schlüsselwörter enthalte. Um das Gewünschte zu finden, muss der Nutzer über Branchenkenntnisse und Fachwissen verfügen, seine Anfrage richtig formulieren können, die Strukturierung der Geodaten kennen und wissen, in welcher Form die Geodaten verfügbar sind.

Das Aufkommen von grossen Sprachmodellen (LLMs) und generativer künstlicher Intelligenz (KI) hat jedoch neue Möglichkeiten eröffnet, um Fragen mit räumlicher Dimension durch eine Frage-Antwort-Interaktion in natürlicher Sprache zu unterstützen.

Die Einführung von ChatGPT im November 2022 war ein Meilenstein für den Einsatz von LLMs und generativer KI in verschiedenen Branchen und brachte sowohl grosse Vorteile als auch neue Herausforderungen mit sich.

Die Geoinformationsbranche hat sofort Interesse an diesen neuen Technologien gezeigt und in den letzten Jahren wurden mehrere Demo-Anwendungen und wissenschaftliche Artikel veröffentlicht. Die Demo-Anwendungen sind im Wesentlichen zweierlei: Einige richten sich an ein Laienpublikum, um den Zugang zu Geoinformationen zu vereinfachen; andere sind eher für Experten gedacht, um die Interaktion mit Datenbanken und Datenverarbeitungswerkzeugen zu vereinfachen. Die meisten wissenschaftlichen Artikel konzentrieren sich auf die Analyse der geografischen «Kompetenzen» von LLMs.

Das Thema wurde 2024 im Rahmen einer spezifischen Aktion des Aktionsplans der Strategie Geoinformation Schweiz (SGS) behandelt. Es wurden zwei separate Projekte gestartet, eines mit Schwerpunkt auf der Nutzung von LLM und generativer KI zur Verbesserung der Auffindbarkeit von Geodaten, das andere mit Schwerpunkt auf der Verbesserung der Zugänglichkeit von Geodaten. Es wurde ein Stand der Technik bei LLM und Geodaten, deren Potenziale und Grenzen, erstellt.

Beide Projekte untersuchten zwei Hauptfragen: (i) Wie können zuverlässige Geodatenquellen eingebunden werden, um die Genauigkeit, Zuverlässigkeit und Relevanz der Antworten sicherzustellen? (ii) Wie kann man diese Tools mit räumlichen Denkfähigkeiten ausstatten?

Der Ansatz war, LLM-Agenten zu entwickeln, die die Interaktion zwischen LLM und anderen Ressourcen wie zuverlässigen Datenquellen (z.B. über Retrieval Augmented Generation - RAG) und Berechnungstools (Tool Calling) koordinieren.

1.2 Ziele

Zum Zweck einer Entscheidungsfindung sieht der Aktionsplan 2025 der SGS eine zusätzliche Massnahme vor. Einerseits sollen die oben genannten technischen Probleme besser identifiziert, analysiert und gegebenenfalls gelöst werden, andererseits sollen die ethischen,

ökologischen und rechtlichen Fragen im Zusammenhang mit dem Einsatz von LLM vertieft werden.

Die strategischen Gremien müssen entscheiden können, wie die LLMs ab 2026 in den Rahmen der NGDI integriert werden sollen. In diesem Zusammenhang wird dieses Dokument erstellt, das als Entscheidungsgrundlage dient.

Ziel dieses Dokuments ist es, alle Informationen zusammenzufassen, die in Gesprächen mit Fachleuten, aus wissenschaftlichen Artikeln und anderen Dokumenten sowie aus den Ergebnissen der verschiedenen Prototypen gewonnen wurden. Es zielt darauf ab, die Grenzen, Einschränkungen und Risiken aus ethischer, ökologischer, rechtlicher, technischer und sicherheitstechnischer Sicht zu identifizieren und zu klassifizieren sowie mögliche Minderungsmassnahmen zu ermitteln. Auf dieser Grundlage werden verschiedene Varianten für die nächsten Jahre vorgeschlagen.

1.3 Vorgehen

Im Rahmen der Massnahme 5-25-2 des Aktionsplans 2025 sind zwei Ziele definiert:

- Die Weiterführung der Ergebnisse und Erkenntnisse aus den 2024 durchgeführten Arbeiten.
- Die Ausarbeitung eines Dokuments, das eine Entscheidungsgrundlage für die strategischen Gremien darstellt, um über den künftigen Umgang mit LLM im Rahmen der NGDI (ab 2026) entscheiden zu können.

Die Weiterführung der Ergebnisse und Erkenntnisse aus den 2024 durchgeführten Arbeiten zielt insbesondere darauf ab:

- den Zugang zu den zwei funktionsfähigen Prototypen bis zum 31.12.2025 sicherzustellen;
- die bekannte Fehler zu dokumentieren und zu beheben;
- die Erklärung unerwarteter Verhaltensweisen und Lösungsvorschläge;
- der Einbezug möglicher laufender Entwicklungen in den Prototyp;
- Testen der Prototypen durch ein Publikum aus Fachleuten und Personen, die nicht im geografischen Bereich tätig sind;
- die Erstellung einer Zusammenfassung der Ergebnisse.

Ein dritter Prototyp wurde ebenfalls im Jahr 2025 gestartet, um Geodaten mit LLMs zugänglich zu machen.

Es war möglich, sich über Projekte in anderen Ämtern oder Organisationen auszutauschen und mögliche Synergien zu finden.

Mit mehreren Schweizer Fachleuten fand ein Austausch auf dem Gebiet der KI statt: CNAI (BFS: Bertrand Loison, Christine Choirat, Raphaël de Fondeville), der ETHZ (Imanol Schlag) und armasuisse (Michael Rüegsegger).

Die GKG und die KGK wurden ebenfalls regelmässig über den Fortschritt dieses Projekts informiert.

2 Einsatz von Sprachmodellen: Fragestellungen

Dieses Kapitel wurde auf der Grundlage von Literaturrecherchen und Gesprächen mit Experten geschrieben.

Das Auftauchen von grossen Sprachmodellen (LLMs) hat in den letzten Jahren in vielen Bereichen für einen radikalen Wandel gesorgt, von fortschrittlichen Entscheidungshilfesystemen bis hin zur automatisierten Erstellung von Inhalten. Diese komplexen Modelle haben mithilfe von Deep-Learning-Techniken riesige Datensätze verarbeitet und menschenähnliche Texte erstellt.

Grosse Sprachmodelle wie GPT-4 von OpenAI sind ein echter Sprung in der KI-Technologie und zeigen, wie gut sie Texte schreiben und verstehen können, die fast wie von Menschen gemacht sind. Diese Modelle können spannende Texte verfassen, Fragen beantworten, Sprachen übersetzen, in verschiedenen Themenbereichen unterrichten und sogar Computerprogramme schreiben. Je tiefer wir in das Zeitalter der KI vordringen, desto deutlicher wird die Leistungsfähigkeit dieser Modelle. Sie können Informationen demokratisieren, indem sie jedem mit Internetzugang Zugang zu Wissen und Lernwerkzeugen ermöglichen, und haben das Potenzial, Branchen von der Gesundheitsversorgung bis zum Bildungswesen zu revolutionieren.

Trotz des grossen Potenzials scheint der Einsatz dieser Technologien in der Praxis derzeit aber nicht so einfach zu sein.

Obwohl LLMs mittlerweile beeindruckende Leistungen zum Beispiel bei medizinischen Aufgaben zeigen, gab es bei Versuchen, LLMs in klinische Umgebungen zu integrieren, um Menschen dabei zu helfen, bessere Entscheidungen zu treffen, einige Probleme.

Wenn man von Standardtests zu natürlichen Dialogen mit Menschen wechselt, zeigen selbst die besten KI-Modelle einen deutlichen Rückgang bei der Genauigkeit der Diagnosen, und das klinische Wissen in LLMs lässt sich nicht in erfolgreiche Interaktionen mit Menschen umsetzen: Eine aktuelle Studie[1] hat gezeigt, dass KI-Modelle wie GPT-4o bei medizinischen Benchmarks wie MedQA[2] besser abschneiden als Menschen und in 95 % der Fälle von vorgegebenen Szenarien die richtigen medizinischen Diagnosen stellen. Wenn Menschen das LLM benutzt haben, ist die Genauigkeit aber auf etwa 34 % der Fälle gesunken. Die Studie hat gezeigt, dass das Design der Benutzerinteraktion immer noch ein Problem ist.

Ausserdem ist die Leistung dieser Technologien mit einer hohen Komplexität verbunden. Das schnelle Wachstum von LLMs hat gezeigt, dass es dringend nötig ist, ihre ethischen Auswirkungen, versteckten *Bias*, Auswirkungen auf die Umwelt und andere Probleme genauer zu anschauen[3].

2.1 Ethische Fragen

Sprachmodelle sind KI-Systeme, die Deep-Learning-Techniken nutzen, um natürliche Sprache zu verstehen und zu generieren. Sie sind mit neuronalen Netzen aufgebaut, die aus mehreren Schichten miteinander verbundener Verarbeitungseinheiten bestehen und die

[1] Andrew M. Bean et al. 2025. Clinical knowledge in LLMs does not translate to human interactions - https://arxiv.org/abs/2504.18919v1
[2] Di Jin et al. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams - https://arxiv.org/abs/2009.13081
[3] Für einen Überblick über einige dieser Themen siehe den Bericht « Limites et contraintes del LLM » - https://backend.geoinformation.ch/fileservice/sdweb-docs-prod-geoinformatch-files/files/2025/04/30/fe70e90f-6104-43dd-ab91-f73c0f0abf64.pdf

Informationsverarbeitungsmechanismen des menschlichen Gehirns nachahmen. LLMs lernen in einer Vorbereitungsphase aus riesigen Mengen von Textdaten Sprachmuster, indem sie das nächste Wort in einem Satz vorhersagen. In dieser Phase eignen sich die Modelle Sprachkenntnisse, Grammatik, Syntax und Semantik an. Anschliessend können LLMs mithilfe kleinerer Datensätze für bestimmte Aufgaben fein abgestimmt werden, um ihre Leistung für bestimmte Anwendungen zu verbessern.

Da der Bedarf an umfangreichen Trainingsdaten sehr hoch ist, werden die meisten LLMs angeblich mit «dem gesamten Internet» trainiert. Diese Datenmenge kann zwar die Fähigkeiten des Modells verbessern und ist der Grund, warum es menschliche Sprache so gut versteht und reproduziert, aber sie enthält auch Daten, die man lieber nicht in den Ergebnissen des Modells sehen möchte. Dazu gehören unter anderem Sprache oder Inhalte, die schädlich oder diskriminierend sind, insbesondere gegenüber marginalisierten oder geschützten Gruppen. Um dem entgegenzuwirken, ist eine sorgfältige Auswahl der Trainingsdaten (und deren Feinabstimmung) entscheidend.

2.1.1 Schädliche Inhalte in LLMs

Die Ursachen für toxische und schädliche Inhalte in LLMs lassen sich auf die Daten zurückführen, die für ihr Training verwendet wurden.

Um schädliche Inhalte aus einer LLM-Ausgabe zu entfernen (oder zumindest einzuschränken), musst man (i) dem Modell Beispiele für solche Inhalte geben; (ii) das Modell anhand der Beispiele lernen lassen diese Art von Inhalten in den Trainingsdaten zu erkennen und dann rauszufiltern.

Um Beispiele für toxische Inhalte zu finden, werden menschliche Annotatoren eingesetzt, um Datenausschnitte zu kennzeichnen. Leider greifen KI-Unternehmen für diese Aufgabe oft auf versteckte menschliche Arbeitskraft im globalen Süden zurück, die oft schädlich und ausbeuterisch sein kann. Diese Arbeiter bleiben am Rande der Gesellschaft, obwohl ihre Arbeit zu Milliardenindustrien beiträgt.

GPT-3, z.B., hatte schon beeindruckende Leistungen gezeigt, aber es war schwer zu verkaufen, weil es auch dazu neigte, toxische Inhalte zu generieren. OpenAI hat dann Zehntausende von Textausschnitten an eine Outsourcing-Firma in Kenia geschickt. Ein Grossteil dieser Texte scheint aus dem Dark Web zu stammen. Einige davon beschrieben detailliert Situationen wie sexuellen Missbrauch, Mord, Selbstmord und Selbstverletzung. Die Arbeiter wurden mit etwa 2 Dollar pro Stunde oder weniger bezahlt. Einige der von der Times befragten Arbeiter sagten, dass sie durch die Arbeit psychisch traumatisiert seien.[4]

Das ist nicht der einzige Fall. Zum Beispiel nutzt das finnische Unternehmen Metroc Gefängnisarbeit, um sein Modell zu trainieren.[5]

[4] Billy Perrigo. 2023. OpenAI Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic - https://time.com/6247678/openai-chatgpt-kenya-workers/
[5] Morgan Meaker. 2023. These prisoners are training AI. https://www.wired.com/story/prisoners-training-ai-finland/

Das Risiko schädlicher Inhalte kann limitiert werden, indem man sich nicht auf die internen Infos des LLM verlässt, sondern das Modell eher als Text-Prozessor/Generator nutzt und es mit einer vertrauenswürdigen externen Wissensdatenbank (z. B. über RAG) ergänzt.

Dieser Ansatz kann das Problem zwar etwas abmildern, ist aber keine komplette Lösung. Das LLM könnte immer noch Infos aus der vertrauenswürdigen Wissensdatenbank mit schädlichen Mustern aus seinen eigenen Trainingsdaten mischen.

2.1.2 Bias in LLMs

Bias in LLMs ist ein echtes ethisches Problem, das die Fairness und Gerechtigkeit von KI-Anwendungen beeinträchtigen kann. Bias bedeutet, dass bestimmte Gruppen, Ansichten oder Eigenschaften systematisch bevorzugt oder benachteiligt werden.

Die Ursachen für Bias in LLMs lassen sich auf die Daten zurückführen, die für das Training verwendet werden. Voreingenommenheit in den Trainingsdaten, die oft gesellschaftliche Vorurteile widerspiegeln, kann vom Modell gelernt und anschliessend weitergegeben werden.

Auch beim Design von Algorithmen können Bias auftreten. Programmierfehler, zum Beispiel wenn ein KI-Entwickler Faktoren im Entscheidungsprozess unfair gewichtet, können unbemerkt ins System gelangen. Gewichtungen sind oft eine Technik, um Bias zu vermeiden. Allerdings können sie Annahmen seitens der Entwickler erfordern, was zu Ungenauigkeiten und Verzerrungen führen kann. Entwickler können auch subjektive Regeln in den Algorithmus einbauen, die auf ihren eigenen bewussten oder unbewussten Bias basieren.

Bias in LLMs kann sich auf verschiedene Arten zeigen[6], die unterschiedliche Auswirkungen auf die Ergebnisse und Entscheidungen haben, die diese Modelle liefern. Hier sind ein paar gängige Arten von Bias:

- **Rassistische und ethnische Bias:** Rassistische und ethnische Bias treten auf, wenn LLMs Ergebnisse liefern, die bestimmte Ethnien unfair benachteiligen. Im Finanzbereich, z.B., können KI-Modelle, die für Kreditvergaben genutzt werden, Vorurteile aus historischen Daten übernehmen, was zu diskriminierenden Praktiken wie *Redlining* oder unfairen Kreditbedingungen führen kann.
- **Geschlechtsspezifische Bias:** Geschlechtsspezifische Bias können zu einer ungleichen Darstellung und Behandlung der Geschlechter im generierten Text führen. Zum Beispiel, obwohl Frauen einen beträchtlichen Anteil der medizinischen Fachkräfte ausmachen, verwenden LLMs häufig männliche Pronomen («er») oder Namen, wenn sie aufgefordert werden, Sätze wie «Ein Arzt betrat das Krankenhaus. __ untersichte den Patienten sorgfältig.» zu vervollständigen.
- **Kulturelle Bias:** Kulturelle Bias können zu Missverständnissen oder falschen Darstellungen verschiedener kultureller Kontexte führen. Zum Beispiel, LLMs, die hauptsächlich mit westlich geprägten Datenquellen trainiert wurden, werden Ergebnisse liefern, die westliche kulturelle Normen und Werte bevorzugen.

[6] Für einen Überblick siehe den Bericht « Limites et contraintes del LLM » - https://backend.geoinformation.ch/fileservice/sdweb-docs-prod-geoinformatch-files/files/2025/04/30/fe70e90f-6104-43dd-ab91-f73c0f0abf64.pdf

- **Geografische Bias:** Das ist der Fall, wenn die Ergebnisse von LLMs Leute aufgrund ihres Standorts bevorzugen oder benachteiligen.
- **Politische Bias:** LLMs können bestimmte politische Ideologien bevorzuge, was die Neutralität der Information beeinträchtigen kann.

Die Bias eines LLM können durch den Einsatz einer Wissensdatenbank ausserhalb des Modells selbst in gewisser Weise begrenzt werden. Hier gelten die gleichen Überlegungen wie im vorherigen Kapitel in Bezug auf schädliche Inhalte.

2.2 Umweltfragen

Die schnelle Entwicklung und der Einsatz von generativen KI-Modellen haben Auswirkungen auf die Umwelt[7]. Die Umweltbelastung durch LLMs und generative KI ist ziemlich gross, vor allem wegen des hohen Energie- und Wasserverbrauchs, der für ihr Training und ihren Betrieb nötig ist.

Die Rechenleistung, die zum Trainieren generativer KI-Modelle mit oft Milliarden von Parametern erforderlich ist, kann einen enormen Stromverbrauch verursachen. Darüber hinaus erfordert die Bereitstellung dieser Modelle in realen Anwendungen sowie ihre Feinabstimmung einen hohen Energieverbrauch.

Neben dem Strombedarf wird auch viel Wasser benötigt, um die Hardware zu kühlen, die für das Training, den Einsatz und die Feinabstimmung generativer KI-Modelle verwendet wird.

Die Entwicklung generativer KI hat auch die Nachfrage nach leistungsstarker Computerhardware angeheizt, was weitere indirekte Auswirkungen auf die Umwelt mit sich bringt.

Hier ein paar Beispiele[8]:

- Eine 1000 Wörter lange E-Mail mit ChatGPT zu schreiben, verbraucht etwa 500 ml Wasser.
- 10 bis 50 Suchanfragen verbrauchen etwa 2 Liter Wasser.
- Das Training eines Modells wie GPT-3 kann 5,4 Millionen Liter Wasser verbrauchen.
- Der weltweite Wasserbedarf für KI wird im Jahr 2027 voraussichtlich zwischen 4,2 und 6,6 Milliarden Kubikmeter betragen und damit den jährlichen Wasserverbrauch von 4 bis 6 Ländern von der Grösse Dänemarks übersteigen.
- Eine Suchanfrage braucht 2,9 Wh, was 6- bis 10-mal so viel Strom ist wie eine normale Google-Suche.
- Eine 1000 Wörter lange E-Mail mit ChatGPT zu schreiben, verbraucht etwa 140 Wh Strom, was 7 vollständigen Ladevorgängen eine iPhone Pro Max entspricht.
- Das Training von GTP-4 hat über 50 GWh Strom verbraucht, was ungefähr 0,02 % des jährlichen Stromverbrauchs des Bundesstaates Kalifornien und dem 50-fachen des Stromverbrauchs für das Training von GPT-3 entspricht.

[7] Sasha Luccioni et al. 2024. The Environmental Impact of AI - Primer. https://huggingface.co/blog/sasha/ai-environment-primer
[8] Save the AI - https://savethe.ai/

- Zum Beispiel haben Rechenzentren in den USA im Jahr 2023 etwa 4,4 % des Stroms des Landes verbraucht, und das könnte bis 2028 auf bis zu 12 % des US-Stroms steigen.

Es gilt daher, die Risiken und negativen direkten und indirekten Umweltwirkungen des Einsatzes generativer KI im Blick zu behalten und diese zu minimieren. Kleine Sprachmodelle (SLMs) sind kompakter und effizienter als die grossen Modelle. Deshalb brauchen SLMs weniger Speicherplatz und Rechenleistung: Mit SMLs kann man den Energieverbrauch echt runterfahren.

2.3 Technische Fragen

Trotz ihrer beeindruckenden Fähigkeiten haben LLMs auch einige technische Einschränkungen[9]. Zu den kritischsten gehören Halluzinationen, mangelnde Reproduzierbarkeit der Ergebnisse und die Black-Box-Natur der internen Prozesse. Jede dieser Herausforderungen kann das Vertrauen der Nutzer und die Zuverlässigkeit der Modelle beeinträchtigen

2.3.1 Halluzinationen

Bei Halluzinationen handelt es sich um Ereignisse, bei denen ein LLM eine Ausgabe erzeugt, die zwar kohärent und grammatikalisch korrekt, aber sachlich falsch oder unsinnig ist. Halluzinationen werden oft als technischer Fehler angesehen, der durch die Einstellung bestimmter Parameter des Modells (z.B. die sogenannte «Temperatur»[10]) begrenzt werden kann, aber in Wirklichkeit sind sie ein wesentliches Merkmal von LLMs. LLMs erfassen statistische Sprachmuster, ohne wirkliches semantisches Verständnis oder ein eingebautes System zur Überprüfung von Fakten: Sie neigen irgendwas zu finden und sind nicht dafür gemacht, die Wahrheit im eigentlichen Sinne zu sagen. Das Problem der Halluzinationen ist sehr präsent, wenn LLMs als Wissensdatenbank genutzt werden: Es tritt in geringem Umfang auf, wenn das LLM mit einer externen, hochwertigen Wissensdatenbank gekoppelt ist. Hier gilt das Gleiche wie bei schädlichen Inhalten und Bias.

2.3.2 Reproduzierbarkeit

LLMs können unterschiedliche Ergebnisse für dieselbe Eingabe liefern. Diese Zufälligkeit kann für Kreativität nützlich sein, aber bei wissenschaftlichen, rechtlichen oder auditrelevanten Anwendungen, bei denen konsistente Ergebnisse wichtig sind, kann das ein Problem sein. Um das zu mildern, kann man deterministische Decodierungsmethoden – wie zum Beispiel die *Greedy Decoding* – oder kontrollierte Sampling-Techniken nutzen[11] , um die Variabilität zu verringern und besser vorhersagbare Ergebnisse zu bekommen.

2.3.3 Black-Box-Effekt

Black-Box-Effekt: Der «Black-Box-Effekt» bei LLMs meint, dass es echt schwierig ist, zu verstehen, wie diese Modelle zu bestimmten Vorhersagen, Entscheidungen oder Antworten kommen. Diese Opazität kommt von der Architektur von LLMs, die oft Milliarden von Parametern und Schichten nichtlinearer Transformationen haben, was ihre interne Funktionsweise selbst für

[9] Für einen Überblick siehe den Bericht « Limites et contraintes del LLM » - https://backend.geoinformation.ch/fileservice/sdweb-docs-prod-geoinformatch-files/files/2025/04/30/fe70e90f-6104-43dd-ab91-f73c0f0abf64.pdf
[10] Joshua Noble. 2024. What is LLM Temperature - https://www.ibm.com/think/topics/llm-temperature
[11] IBM. 2025. Foundation model parameters: decoding and stopping criteria - https://www.ibm.com/docs/en/watsonx/saas?topic=prompts-model-parameters-prompting

ihre Entwickler schwer zu analysieren oder zu erklären macht. Die mangelnde Transparenz der Entscheidungsprozesse von LLMs wirft mehrere Fragen auf:

- Mangelnde Erklärbarkeit: Anders als bei traditionellen Softwaresystemen, wo die Logik explizit programmiert ist und der Datenfluss leicht nachvollziehbar ist, funktionieren LLMs auf der Grundlage statistischer Korrelationen und gelernter Muster in riesigen Datensätzen. Daher ist es nicht einfach zu bestimmen, warum eine bestimmte Ausgabe erzeugt wurde. Das ist besonders problematisch in Bereichen, in denen es entscheidend ist, die Gründe für Entscheidungen zu verstehen, wie zum Beispiel bei juristischen Argumentationen, medizinischen Diagnosen oder Kreditgenehmigungen.
- Verantwortlichkeit: In Bereichen, wo Entscheidungen echt wichtig sind (z. B. im Gesundheitswesen oder im Finanzwesen), kann es Probleme geben, wenn man nicht erklären kann, warum ein Modell eine bestimmte Entscheidung getroffen hat.
- Vertrauen: Nutzer, Entscheidungsträger und Aufsichtsbehörden sind oft zurückhaltend gegenüber Systemen, deren Funktionsweise sie nicht verstehen.

Erklärbare KI (*Explainable AI, XAI*)[12], also eine Reihe von Prozessen und Methoden, mit denen Menschen die Ergebnisse und Outputs von KI-Algorithmen verstehen und ihnen vertrauen können, kann den «Black-Box-Effekt» von LLMs abmildern. Die «*Chain of Thought*» (CoT)[13] ist zum Beispiel eine Methode, die vor allem in LLMs verwendet wird, wo das Modell dazu gebracht oder trainiert wird, bei der Lösung komplizierter Probleme Zwischenschritte der Argumentation zu generieren.

2.4 Sicherheitsfragen

LLMs bringen erhebliche Sicherheitsrisiken mit sich. Diese reichen von Prompt-Injektionen und Datenlecks bis hin zur böswilligen Nutzung durch Dritte. Bei der Integration dieser Technologien in reale Anwendungen ist es wichtig, ihre Schwachstellen zu kennen und sie systematisch zu sichern.

Drei der dringendsten Sicherheitsrisiken im Zusammenhang mit LLMs sind Prompt Injection, Datenlecks und Automatisierungsrisiken.

2.4.1 Prompt Injection

Bei der Prompt-Injection manipuliert jemand die Eingabe in ein LLM, um die beabsichtigten Instruktionen oder das Verhalten zu ändern. Das ist ähnlich wie bei der Code-Injection in der traditionellen Software-Sicherheit. Ein böswilliger Nutzer kann schädliche Befehle oder Eingabeaufforderungen in eine Eingabezeichenfolge einfügen, sodass das Modell sensible Infos preisgibt, Sicherheitsvorkehrungen umgeht oder ungewollte Aktionen ausführt. Hier sind ein paar Massnahmen, um das Problem zu mildern:

- Eingabesäuberung: Benutzeraufforderungen bereinigen, um verdächtige Muster oder Anweisungen zu entfernen.

[12] IBM. 2023. What is Explainable AI (XAI)? - https://www.ibm.com/think/topics/explainable-ai
[13] IBM. What is a chain of thought (CoT) prompting? - https://www.ibm.com/think/topics/chain-of-thoughts

- Prompt-Design: Benutzeraufforderungen klar von Systemaufforderungen trennen.
- Rollenbasierte Zugriffskontrolle: Unterschiedliche Zugriffsebenen auf Modellfunktionen je nach Benutzerberechtigung implementieren. Sicherstellen, dass Benutzeraufforderungen die Systemaufforderungen nicht überschreiben können.

2.4.2 Datenlecks

Datenlecks passieren, wenn ein LLM sensible oder geschützte Infos preisgibt. Das kann durch das Training mit vertraulichen Daten, schlechte Zugriffskontrollen oder Nutzeraufforderungen passieren, die zu unbeabsichtigten Offenlegungen führen. Datenlecks sind die unbeabsichtigte Offenlegung vertraulicher oder sensibler Infos durch die Ausgaben eines LLM. Das kann an mehreren Stellen im Modelllebenszyklus passieren und ernsthafte Risiken für den Datenschutz und die Compliance mit sich bringen. Die Lösung für dieses Problem ist im Grunde eine Frage der Daten-Governance: Sensible und vertrauliche Daten müssen aus den Trainingsdaten entfernt werden.

2.4.3 Automatisierungsrisiken

Automatisierungsrisiken entstehen, wenn Modelle ohne ausreichende menschliche Kontrolle handeln, was zu kritischen Fehlern, Missbrauch oder Ausnutzung führen kann. Ein LLM-gestützter E-Mail-Assistent könnte zum Beispiel selbstständig falsche oder sensible Infos an den falschen Empfänger schicken oder bei Missbrauch bösartigen Code erzeugen. Hier sind ein paar Massnahmen, um das Problem zu mildern:

- *Human-in-the-Loop*: Bevor das Modell sensible Aufgaben oder Ausgaben ausführen kann, muss eine Überprüfung durch einen Menschen erfolgen.
- Ausgabevalidierung: Es sollten regelbasierte Systeme oder sekundäre Modelle verwendet werden, um die Ausgaben des LLM zu validieren, bevor sie angewendet werden.
- Systemdesign: Aufgaben nach ihrem Risikograd kategorisieren und LLMs auf Aufgaben mit geringem Risiko beschränken. Ausführungsberechtigungen basierend auf Kritikalität der Aufgabe festlegen, z.B. «*Read-only*»-Modus für risikoreiche Aufgaben.

2.5 Rechtliche Fragen

Da LLMs mit riesigen Datenmengen aus dem Internet trainiert werden, besteht die Gefahr, dass sie persönliche Infos und auch urheberrechtlich geschützte Infos enthalten. Es besteht das Risiko, dass diese Modelle sensible Infos preisgeben und/oder gegen geistige Eigentumsrechte verstossen.

- Urheberrechtsverletzungen: Die Trainingsdaten für LLMs können urheberrechtlich geschütztes Material wie Bücher, Artikel, Quellcode usw. enthalten. Wenn diese Modelle Inhalte generieren, die urheberrechtlich geschützten Werken stark ähneln oder diese replizieren, stellt sich die Frage nach möglichen Urheberrechtsverletzungen.
- Datenschutz und Persönlichkeitsrechte: LLMs können versehentlich persönliche Daten speichern und wiedergeben, insbesondere wenn solche Daten ohne Zustimmung in den Trainingsdatensätzen enthalten waren. Dies kann gegen Datenschutzgesetze verstossen.

Im Grunde geht es hier nochmals um Daten-Governance: Persönliche Infos und urheberrechtlich geschütztes Material müssen aus den Trainingsdaten entfernt werden.

3 Governance für eine ethische und verantwortungsvolle KI

Es gibt zahlreiche ethische Bedenken und potenzielle Probleme im Zusammenhang mit Grundmodelle (*Foundation Models*) wie LLMs. Obwohl viele dieser Bedenken generell für Künstliche Intelligenz gelten, werden sie angesichts der Macht und der weiten Verfügbarkeit von Grundmodellen besonders dringlich.

Die Entwicklung von LLMs – und von Grundmodellen im weiteren Sinne – die ethisch, unvoreingenommen und gesellschaftlich vorteilhaft sind, erfordert ein starkes Engagement für eine verantwortungsvolle KI-Entwicklung.

Im Kern bedeutet dies, ethische Überlegungen in den gesamten Lebenszyklus eines KI-Systems einzubetten – von der Datenerhebung und dem Modelltraining bis hin zur Bereitstellung und der Überwachung nach der Veröffentlichung.

Ethische KI und verantwortungsvolle KI verfolgen parallele Wege, befassen sich aber mit unterschiedlichen Aspekten.

Ethische KI bezieht sich auf die moralischen und ethischen Aspekte der Entwicklung und Nutzung von KI. Sie konzentriert sich darauf, KI mit menschlichen Werten in Einklang zu bringen, betont die moralischen Implikationen von KI-Entscheidungen und setzt sich für eine KI ein, die Gerechtigkeit und Grundrechte respektiert und Schaden vermeidet. Ethische KI befasst sich mit Fragen wie: (i) Ist das KI-System fair und unvoreingenommen? (ii) Respektiert es die Würde und Freiheit des Menschen? (iii) Ist es in seinen Entscheidungen transparent? Im Mittelpunkt steht die Frage, was richtig oder falsch, gerecht oder ungerecht und fair oder unfair ist, wenn es um die Auswirkungen von KI auf den Einzelnen und die Gesellschaft geht.

Verantwortungsvolle KI ist dagegen eher praktisch und handlungsorientiert. Sie konzentriert sich auf die praktische Umsetzung, Governance und Compliance-Mechanismen, die dafür sorgen, dass KI sicher, fair und gesetzeskonform eingesetzt wird. Sie schliesst die Lücke zwischen ethischer Theorie und umsetzbarer Praxis, indem sie ethische Grundsätze in Geschäftsprozesse und technische Systeme einbaut.

Verantwortungsvolle KI ist die Entwicklung und Nutzung von KI-Systemen, die der Gesellschaft nützen und gleichzeitig das Risiko negativer Folgen minimieren. Es geht darum, KI-Technologien zu entwickeln, die nicht nur unsere Fähigkeiten verbessern, sondern auch ethische Bedenken berücksichtigen – insbesondere in Bezug auf Voreingenommenheit, Transparenz und Datenschutz. Dazu gehört auch, Probleme wie den Missbrauch personenbezogener Daten, voreingenommene Algorithmen und das Potenzial von KI, bestehende Ungleichheiten zu perpetuieren oder zu verschärfen, anzugehen. Das Ziel ist es, vertrauenswürdige KI-Systeme zu entwickeln, die gleichzeitig zuverlässig, fair und mit menschlichen Werten im Einklang stehen.

Verantwortungsvolle KI befasst sich mit Fragen wie: (i) Wer ist verantwortlich, wenn die KI Schaden verursacht? (ii) Sind die Systeme sicher, robust und geschützt? (iii) Wurden angemessene Tests, Dokumentationen und menschliche Kontrollen durchgeführt?

Hier eine Liste der wichtigsten Grundsätze für verantwortungsvolle KI:

- Transparenz – KI-Systeme sollten offenlegen, wie sie funktionieren, auf welcher Datenbasis sie basieren und welche Prozesse ihre Entscheidungen beeinflussen. Diese Nachvollziehbarkeit schafft Vertrauen und erleichtert die Kontrolle.
- Erklärbarkeit – Die Ergebnisse und Entscheidungen von KI müssen für Menschen verständlich und interpretierbar sein. Anwender und Betroffene sollen nachvollziehen können, warum ein bestimmtes Ergebnis zustande kam.
- Fairness – KI sollte so gestaltet sein, dass sie keine Gruppen benachteiligt. Ziel ist die Gleichbehandlung aller Menschen – unabhängig von Alter, Geschlecht, Herkunft oder anderen Merkmalen. Verzerrungen in Daten und Algorithmen müssen erkannt und reduziert werden.
- Verantwortlichkeit und Governance – Es braucht klare Zuständigkeiten für alle Phasen im Lebenszyklus von KI – von der Entwicklung bis zum Einsatz. Nur so kann sichergestellt werden, dass Entscheidungen nachvollzogen und Verantwortliche benannt werden können.
- Robustheit und Sicherheit – KI-Systeme müssen stabil, widerstandsfähig und geschützt gegenüber Störungen, Fehlern oder Angriffen sein. Sie sollten auch unter schwierigen Bedingungen zuverlässig funktionieren.
- Datenschutz und Datenethik – Der Schutz personenbezogener Daten steht im Mittelpunkt. Der Umgang mit Daten muss verantwortungsbewusst, gesetzeskonform und transparent erfolgen – vom Sammeln bis zur Nutzung.
- Inklusivität – KI sollte die Vielfalt der Gesellschaft widerspiegeln und auf die Bedürfnisse unterschiedlichster Nutzergruppen eingehen. Der Zugang und Nutzen von KI-Technologie dürfen nicht auf privilegierte Gruppen beschränkt bleiben.
- Rechtliche Konformität – Der Einsatz von KI muss im Einklang mit geltenden Gesetzen und branchenspezifischen Vorgaben stehen. Rechtssicherheit ist die Grundlage für einen verantwortungsvollen Umgang mit der Technologie.

Ethische und verantwortungsvolle KI sind eine Frage der Governance und brauchen einen regulatorischen Rahmen, aber die Umsetzung ist nicht so einfach.

Die Stadt Rotterdam hat zum Beispiel ein KI-System für die Verwaltung von Kinderbetreuungsleistungen eingesetzt: Leider wurden Tausende von Familien fälschlicherweise des Betrugs beschuldigt. Der vom System verwendete Algorithmus war voreingenommen und stufte Sozialhilfeempfänger anhand ihrer Kleidung und ihrer Niederländisch Kenntnisse ein und zielte gezielt auf alleinerziehende Migrantinnen ab.

Aus dieser Erfahrung hat die Stadt Amsterdam gelernt und ein risikoreiches Experiment unter der Leitung von verantwortungsvolle AI gestartet. Die Stadt unternahm den Versuch, die Bias des

Modells gegenüber nicht-niederländischen Staatsangehörigen und Personen mit nicht-westlicher Staatsbürgerschaft zu korrigieren. Andere demografische Verzerrungen wurden jedoch nicht berücksichtigt. Die Trainingsdaten des Modells wurden neu gewichtet – ein Verfahren, das sein primäres Ziel erreichte: Ein Grossteil der Benachteiligung von Bewerber mit Migrationshintergrund verschwand.

Nach dem scheinbaren Erfolg der Neugewichtung wurde das System in einem Pilotprojekt eingesetzt. Die Ergebnisse dieses Piloten zeigten jedoch, dass neue Bias im Modell aufgetreten waren. Tatsächlich tendierte das neu gewichtete Modell nun häufig dazu, genau die gegenteiligen Gruppen falsch zu markieren als das ursprüngliche Modell. So hatten Frauen eine um 22 Prozent höhere Wahrscheinlichkeit, fälschlicherweise markiert zu werden als Männer. Ebenso wurden Antragsteller mit niederländischer Staatsangehörigkeit häufiger falsch markiert als solche ohne niederländische Staatsangehörigkeit. Das Projekt wurde eingestellt[14] [15].

[14] Lighthouse Reports. 2025. The Limits of Ethical AI - https://www.lighthousereports.com/investigation/the-limits-of-ethical-ai/
[15] Lighthouse Reports. 2025. How we investigated Amsterdam's attempt to build a 'fair' fraud detection model - https://www.lighthousereports.com/methodology/amsterdam-fairness/

4 KI-bezogene Aktivitäten und Regulierungsrahmen für KI in der Schweiz

4.1 Auf dem Weg zu einer Schweizer KI-Gesetzgebung

Der Einsatz von KI bringt neue Herausforderungen für die Gesellschaft und das Recht mit sich. Deshalb wurden auf internationaler Ebene Regelungen wie die KI-Konvention des Europarats oder das KI-Gesetz der EU verabschiedet[16].

Die **KI-Konvention des Europarats**[17] ist das erste rechtsverbindliche internationale Abkommen über KI und steht vollkommen im Einklang mit dem KI-Gesetz der EU, der weltweit ersten umfassenden KI-Regelung. Die Konvention sieht einen gemeinsamen Ansatz vor, um sicherzustellen, dass KI-Systeme mit den Menschenrechten, der Demokratie und der Rechtsstaatlichkeit vereinbar sind und gleichzeitig Innovation und Vertrauen ermöglichen. Es enthält eine Reihe von Prinzipien aus dem KI-Gesetz der EU, wie einen risikobasierten Ansatz, Transparenz entlang der Verpflichtungen von KI-Systemen und KI-generierten Inhalten, detaillierte Dokumentationspflichten für als risikoreich eingestufte KI-Systeme und Risikomanagementpflichten mit der Möglichkeit, Verbote für KI-Systeme einzuführen, die eine eindeutige Bedrohung für die Grundrechte darstellen.

Während die KI-Konvention darauf abzielt, dass KI-Systeme mit den Menschenrechten, der Demokratie und der Rechtsstaatlichkeit im Einklang stehen, ist das **KI-Gesetz**[18] eine rechtsverbindliche Verordnung, die sich auf den EU-Binnenmarkt konzentriert und darauf abzielt, sichere und vertrauenswürdige KI-Systeme zu gewährleisten, indem deren Entwicklung, Einsatz und Nutzung innerhalb der EU reguliert werden.

Das KI-Gesetz verfolgt einen risikobasierten Ansatz. So legt er eine Klassifizierung von KI-Systemen nach dem Grad ihres Risikos für Grundrechte, Gesundheit und Sicherheit fest. Entsprechend dieser Einstufung sieht das KI-Gesetz verschiedene Verpflichtungen vor[19]:

- Inakzeptables Risiko: Alle KI-Systeme, die eine klare Gefahr für die Sicherheit, die Lebensgrundlage und die Rechte von Menschen darstellen, sind verboten.
- Hohes Risiko: Dazu gehören KI-Systeme, die ernsthafte Risiken für Gesundheit, Sicherheit oder Grundrechte mit sich bringen können und als risikoreich eingestuft werden. Hochrisiko-KI-Systeme unterliegen strengen Verpflichtungen (z.B. Risikobewertung, Risikominderung, Datenqualität, hohe Robustheit, Cybersicherheit und Genauigkeit).
- Begrenztes Risiko: Dies bezieht sich auf die Risiken, die mit der Notwendigkeit der Transparenz in Bezug auf den Einsatz von KI verbunden sind. Menschen sollten wissen, dass sie mit einer Maschine interagieren, damit sie eine informierte Entscheidung treffen können, und KI-generierte Inhalte müssen erkennbar sein.

[16] Bundesamt für Justiz. 2025. Künstliche Intelligenz - https://www.bj.admin.ch/bj/de/home/staat/gesetzgebung/kuenstliche-intelligenz.html
[17] Council of Europe. 2024. Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law - https://rm.coe.int/1680afae3c
[18] Verordnung (EU) 2024/1689 - https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32024R1689
[19] Siehe für die Details https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

- Minimales oder kein Risiko: Für KI-Systeme, die kaum oder gar keine Auswirkungen auf Rechte oder Sicherheit haben, gibt es im EU-KI-Gesetz keine verbindlichen Vorgaben. Anbieter können freiwillig Verhaltenskodizes und/oder ethische Leitlinien befolgen.

In der Schweiz gibt es noch keine umfassende Gesetzgebung speziell für KI. Deshalb hat der Bundesrat am 22. November 2023 das UVEK (BAKOM) und das EDA (Abteilung Europa) damit beauftragt eine Auslegeordnung zu möglichen Regulierungsansätzen für KI zu erarbeiten[20], die mit der KI-Konvention des Europarats und der KI-Verordnung des KI-Gesetzes der EU kompatibel sind.

Die Auslegeordnung umfasst juristische, wirtschaftliche und europapolitische Analysen und erforderte daher eine enge Zusammenarbeit aller Departemente. Die übergreifenden Ziele der Auslegeordnung sind: (i) die Stärkung des Innovationsstandorts Schweiz; (ii) die Wahrung des Grundrechtsschutzes inklusive der Wirtschaftsfreiheit; (iii) die Stärkung des Vertrauens der Bevölkerung in KI. Um diese Ziele zu erfüllen, schlägt die Auslegeordnung drei mögliche Regulierungsansätze vor, welche die Ziele in unterschiedlichem Masse erfüllen würden:

- Fortführung der themen- und sektorspezifischen Regulierungsaktivitäten: Der Regulierungsbedarf würde weiterhin thematisch und sektoriell identifiziert und umgesetzt. Wenn nötig, wird das auch Querschnittsthemen wie den Datenschutz betreffen. Der Bundesrat würde keine umfassende Schweizer Regulierung für KI-Anwendungen anstreben.
- Ratifikation der KI-Konvention des Europarats
- Ratifikation der KI-Konvention des Europarats und Umsetzung in Anlehnung an das KI-Gesetz der EU: In Anlehnung an das KI-Gesetz der EU wäre mit der Einführung eines risikobasierten Ansatzes für den Umgang mit Produkten mit KI-Bestandteilen dieselben Voraussetzungen für Schweizer Unternehmen wie in der EU gegeben.

Auf der Grundlage der Auslegeordnung hat der Bundesrat am 12. Februar 2025 entschieden, sich an folgenden Eckwerten zu orientieren[21]:

- Die KI-Konvention des Europarats wird ins Schweizer Recht übernommen.
- Wo Gesetzesanpassungen nötig sind, sollen diese möglichst sektorbezogen ausfallen. Eine allgemeine, sektorübergreifende Regulierung beschränkt sich auf zentrale, grundrechtsrelevante Bereiche, wie beispielsweise den Datenschutz.
- Neben der Gesetzgebung werden auch rechtlich nicht verbindliche Massnahmen zur Umsetzung der Konvention erarbeitet.

Der Bundesrat hat darüber hinaus das weitere Vorgehen definiert: Das EJPD erarbeitet gemeinsam mit dem UVEK und dem EDA bis Ende 2026 eine Vernehmlassungsvorlage. Diese dient der Umsetzung der KI-Konvention des Europarats und legt die erforderlichen gesetzlichen

[20] Bundesamt für Kommunikation. 2025. Auslegeordnung zur Regulierung von künstlicher Intelligenz - https://www.bj.admin.ch/dam/bj/de/data/staat/gesetzgebung/kuenstliche-intelligenz/ber-bakom-ki.pdf.download.pdf/ber-bakom-ki-d.pdf
[21] Bundesamt für Kommunikation. 2025. KI-Regulierung: Bundesrat will Konvention des Europarats ratifizieren - https://www.news.admin.ch/de/nsb?id=104110

Massnahmen fest – insbesondere in den Bereichen Transparenz, Datenschutz, Nichtdiskriminierung und Aufsicht.

Ausserdem will der Bundesrat die Koordination von KI in der Bundesverwaltung stärken und hat das Eidgenössische Departement des Innern EDI sowie die Bundeskanzlei BK beauftragt, bis Ende 2025 einen Vorschlag zur Weiterentwicklung der Koordination zu KI zu erarbeiten. Aktuell sind die Zuständigkeiten wie folgt:

- **CNAI:** Kompetenzzentrum und Netzwerk für Künstliche Intelligenz. Es fungiert als zentrale Ansprechpartnerin für das Thema KI und dient als Anlaufstelle für verwaltungsinterne Fragestellungen.
- **Bundesamt für Justiz (BJ):** Erarbeitet bis Ende 2026 die Vernehmlassungsvorlage zur KI-Regulierung.
- **Bundesamt für Kommunikation (BAKOM):** Entwickelt ergänzende, nicht rechtsverbindliche Massnahmen im Bereich KI.
- **Direktion für Völkerrecht (DV):** Verantwortlich für Fragen der internationalen KI-Regulierung sowie für die Koordination und Zusammenarbeit mit dem BJ und dem BAKOM zur Umsetzung internationaler Vorgaben ins nationale Recht.
- **Bereich DTI der Bundeskanzlei (BK):** Erarbeitet Strategien, Richtlinien und konkrete Umsetzungsmassnahmen für den Einsatz von KI-Systemen innerhalb der Bundesverwaltung.

4.2 Das schweizerische LLM «Apertus»

«Apertus» ist der Name des LLM – entwickelt von Forschern der EPFL, der ETH Zürich und anderen Schweizer Unis zusammen mit Ingenieuren vom CSCS (*Centro Svizzero di Calcolo Scientifico*)[22].

Das Modell wurde auf dem Supercomputer «Alps» des CSCS in Lugano trainiert und ist seit Anfang September 2025 verfügbar. Hier ein paar besondere Merkmale des Modells:

- Modellgrösse: Apertus gibt's in zwei Grössen – mit 8 Milliarden (8B) und 70 Milliarden (70B) Parametern.
- Offenheit: Quellcode und Gewichte sind frei zugänglich. Auch die Trainingsdaten sind transparent dokumentiert und reproduzierbar. Das Modell steht unter der Apache-2.0-Lizenz. Architektur, Trainingsmethoden und Nutzungsrichtlinien sind in einer ausführlichen Dokumentation beschrieben und ermöglichen eine transparente Wiederverwendung sowie Weiterentwicklung.
- Vielsprachigkeit: Das Basismodell wurde mit einem großen Datensatz in über 1500 Sprachen trainiert (60 % Englisch, 40 % andere Sprachen), ergänzt durch Daten aus Programmiercode und Mathematik. Dank dieser breiten Grundlage besitzt es eine hohe internationale Anwendbarkeit.

[22] Ein Sprachmodell im Dienste der Gesellschaft - https://ethz.ch/de/news-und-veranstaltungen/eth-news/news/2025/07/ein-sprachmodell-im-dienste-der-gesellschaft.html

- Rechtliche und regulatorische Grundlagen: Bei der Entwicklung wurden die schweizerischen Datenschutzgesetze, das Urheberrecht sowie die Transparenzanforderungen der KI-Verordnung der EU berücksichtigt.

Damit trägt das Modell dazu bei, einige der in Kapitel 1 dieses Dokuments dargestellten Herausforderungen zu adressieren.

4.3 eCH-0272

Der Standard eCH-0272[23] wurde im April 2025 vom Verein eCH verabschiedet und bietet Orientierungsrahmen für den Einsatz von KI-Systemen in der Verwaltung.

Bei eCH-0272 handelt es sich um einen konzeptionellen Standard, der Anforderungen an die Transparenz, Erklärbarkeit und Risiken der KI-Systeme definiert. Der Standard gibt vor, wie Risiken von KI-Systemen bewertet, dokumentiert und kontrolliert werden sollen. Er umfasst 27 Anforderungen, die den gesamten Lebenszyklus eines KI-Systems abdecken – von der Entwicklung über den Einsatz bis hin zur Wartung und zum Rückbau: es wird zusätzlich gezeigt an welchen Grundlagen sich die Anforderungen orientieren.

eCH-0272 bezieht sich auf vortrainierte und lokal betriebene KI-Systeme. Cloud-basierte und kontinuierlich trainierte *GenAI*-Systeme, die z. B. auf Plattformen wie OpenAI laufen, fallen nicht darunter, weil sie zusätzliche Datenschutz- und Sicherheitsanforderungen haben.

Dieser Standard ist definitiv ein Hilfsmittel für den Einsatz von KI-gestützten Anwendungen in der öffentlichen Verwaltung.

4.4 KI-gestützte Anwendungen auf Bundesebene

Zum Zeitpunkt der Erstellung dieses Dokuments gibt es in der Schweiz eine bekannte KI-gestützte Anwendung auf Bundesebene:

- RoBIT: RoBIT ist der IT-Support-Chatbot des BIT (Bundesamt für Informatik und Telekommunikation) und hilft den Mitarbeitern der Schweizer Bundesverwaltung, einfache Antworten rund um die IT und Support zu Fachanwendungen zu finden. RoBIT verwendet als Large Language Modell GPT- 3.5 Turbo. Dieses LLM wird über Microsoft Azure mit Serverstandort in der Schweiz bezogen. Es wird erklärt, dass keine Daten ins Ausland übertragen werden und Microsoft verspricht, dass die Eingaben, die RoBIT erhält und die Antworten, die RoBIT generiert, nicht mit OpenAI oder Dritten geteilt werden. Kann dieses Versprechen garantieren, dass wirklich keine Daten an die Server von OpenAI oder einer anderen amerikanischen Organisation gehen? Bei einer öffentlichen Anhörung im französischen Senat am 10. Juni 2025 räumte Microsoft unter Eid ein, dass US-Behörden Zugriff auf EU-Daten erhalten könnten.[24] Das könnte schon für die Daten aus der Schweiz gelten.[25]

[23] Verein eCH. 2025. eCH-0272 – Transparenz, Erklärbarkeit und Risiken von KI-Systemen - https://www.ech.ch/de/ech/ech-0272/1.0.0
[24] Commande publique : audition de Microsoft - https://videos.senat.fr/video.5460497_6847c70b82594.commande-publique--audition-de-microsoft
[25] https://www.nzz.ch/wirtschaft/was-wenn-trump-den-stecker-zieht-der-us-praesident-hat-den-durchgriff-auf-daten-von-schweizer-firmen-ld.1893786

Es gibt auch eine ganze Reihe von Pilotprojekten. Eine Übersicht bietet die Datenbank des CNAI, die Projekte der Bundesverwaltung auflistet, die mit KI zu tun haben oder auf KI-Technologien basieren.[26]

5 LLMs und Geodaten-Infrastrukturen

Das Aufkommen von LLMs hat neue Möglichkeiten eröffnet, um Fragen mit räumlicher Dimension durch eine Frage-Antwort-Interaktion in natürlicher Sprache zu beantworten.

Die Einführung von ChatGPT im November 2022 war ein Meilenstein für den Einsatz von LLMs in verschiedenen Branchen und brachte sowohl grosse Vorteile als auch neue Herausforderungen mit sich.

Die Geoinformationsbranche hat sofort Interesse an diesen neuen Technologien gezeigt und In den letzten Jahren wurden mehrere Demo-Anwendungen und wissenschaftliche Artikel veröffentlicht. Die Demo-Anwendungen sind im Wesentlichen zweierlei: Einige richten sich an ein Laienpublikum, um den Zugang zu Geoinformationen zu vereinfachen; andere sind eher für Experten gedacht, um die Interaktion mit Datenbanken und Datenverarbeitungswerkzeugen zu vereinfachen. Die meisten wissenschaftlichen Artikel konzentrieren sich auf die Analyse der geografischen «Kompetenzen» von LLMs[27]. Die Ergebnisse sind bisher nicht so begeisternd, da diese Tools nur wenig Ahnung von Geografie und den räumlichen Beziehungen zwischen geografischen Einheiten haben[28].

Trotz ihrer derzeitigen Einschränkungen sind LLMs schon interessante Hilfsmittel. Die Verbindung von LLMs mit Geodaten-Infrastrukturen (GDI) kann die Art und Weise, wie Leute mit räumlichen Daten umgehen, sie analysieren und daraus Erkenntnisse gewinnen, ziemlich verändern. LLMs können den Zugang zu Geodaten demokratisieren, indem sie intuitive Schnittstellen bieten, für die man kein spezielles Wissen braucht.

5.1 LLMs und Geospatial

AI-Basismodelle haben sich bisher hauptsächlich auf das Verstehen und Erzeugen von Text (LLM), Audio (*Audio & Speech Foundation Models*) und Video/Bildern (*Vision Foundation Models*)[29] konzentriert, während geographischen Einheiten und räumliche Beziehungen vernachlässigt wurden. ChatGPT kann zum Beispiel einen Ortsnamen richtig erkennen, aber es gibt ziemlich ungenaue Koordinaten und falsche Antworten auf Fragen zu topologischen Beziehungen[30].

Vision-FMs und ihre georäumlichen Varianten sind im Geodatenbereich weit verbreitet und haben zur ersten Generation von Geo-FMs geführt, die mit grossen Mengen von Satellitenbildern

[26] CNAI. Projektdatenbank - https://cnai.swiss/dienstleistungen/projektdatenbank/
[27] Für einen Überblick siehe den Bericht « LLM et Géodonnées» - https://backend.geoinformation.ch/fileservice/sdweb-docs-prod-geoinformatch-files/files/2025/04/30/6f8c5a3f-5c99-4f54-9964-2af711dc3c14.pdf
[28] Alan Turing Institute. 2024. Geospatial AI for Land Use - https://www.gov.uk/government/publications/geospatial-ai-for-land-use-by-the-alan-turing-institute
[29] Die verschiedenen Arten von Eingabe- und Ausgabedaten, die ein Modell verarbeiten kann, werden als „Modalitäten" bezeichnet. Es gibt Basismodelle, die zwei oder mehr Modalitäten kombinieren können: Diese werden als multimodale Basismodelle bezeichnet. Beispiele für solche Modelle sind GPT-4o, Gemini und DALL·E 3.
[30] Gengchen Mai et al. 2024. On the opportunities and challenges of foundation models for GeoAI - https://dl.acm.org/doi/10.1145/3653070

vorab trainiert wurden. Beispiele für solche Modelle sind Prithvi[31] (von der NASA und IBM veröffentlicht) und das kürzlich veröffentlichte AlphaEarth Foundations[32] (Google DeepMind).

Andererseits kann keines der bestehenden Modelle Vektordaten wie Punkte, Linien und Polygone verarbeiten[33]. Die Entwicklung eines generischen Geo-FMs bleibt aufgrund der multimodalen Natur von Geodaten, die Geoinformationen wie Text-, Raster- und Vektordaten umfassen, eine Herausforderung. Jede Modalität weist spezielle Strukturen auf, die eine eigene Darstellung erfordern. Diese multimodale Natur der Geoinformationen verhindert eine direkte Anwendung bestehender LLMs in GeoAI-Aktivitäten.

Ein möglicher Ansatz zur Lösung dieser Probleme ist die Entwicklung von LLM-Agenten, die die Interaktion zwischen dem LLM und anderen Ressourcen koordinieren, wie z. B. externen Wissensdatenbanken (um die LLM-Ausgaben in autoritativen Geodaten und Metadaten zu verankern) und Rechenwerkzeugen (um das räumliche Denken des LLM zu verbessern).

5.2 LLMs und die Nationale Geodaten-Infrastruktur

Die Verbindung von LLMs mit den Komponenten der Nationalen Geodaten-Infrastruktur (NGDI) kann die Art und Weise verändern, wie Leute mit räumlichen Daten umgehen. LLMs können den Zugang zu Geodaten demokratisieren, indem sie intuitive Schnittstellen bieten, für die man kein spezielles Wissen braucht. Die Möglichkeit, über natürliche Sprache mit diesen Tools zu interagieren, kann die Erfahrungen von Nicht-Experten verbessern und ihnen ermöglichen, von der riesigen Menge an zuverlässigen verfügbaren Geodaten zu profitieren.

Hier sind zwei wichtige Anwendungsbereiche, in denen diese Synergie die Benutzerfreundlichkeit für Leute ohne Fachwissen verbessern kann:

- **Semantische Suche in natürlicher Sprache**: Damit können Leute relevante Infos zu einem bestimmten Thema in natürlicher Sprache und ohne Fachwissen finden (z. B. «Wo finde ich Infos zum Solarpotenzial meines Hauses?»).
- **Abfragen in natürlicher Sprache**: Damit können Leute Geodaten abfragen und mit ihnen interagieren in natürlicher Sprache und ohne Fachwissen finden (z. B. «Zeig mir die Grundstücke rund um mein Haus»).

Inwieweit können LLMs genutzt werden, um solche Fragen anhand von Infos aus Geodaten-Infrastrukturen zu beantworten? Im Wesentlichen geht es um zwei Arten von Fragestellungen:

- **Wie kann man LLMs dazu bringen, mit Daten aus der NGDI zu arbeiten**? LLMs kennen nur die Informationen, mit denen sie trainiert wurden. Eine Feinabstimmung (*Fine-Tuning*) ist im Moment keine Option. Geodaten werden oft über Rest-APIs und in Vektorform bereitgestellt. Mit keinem der aktuellen Basismodelle kann man im Moment eine Feinabstimmung mit Vektordaten machen.
- **Wie kann man LLMs dazu bringen, raumbezogene Fragen anhand von Daten aus der NGDI zu beantworten**? LLMs haben wenig Verständnis für Geografie und die

[31] https://www.earthdata.nasa.gov/news/nasa-ibm-openly-release-geospatial-ai-foundation-model-nasa-earth-observation-data
[32] https://deepmind.google/discover/blog/alphaearth-foundations-helps-map-our-planet-in-unprecedented-detail/
[33] Gengchen Mai. 2024. Geo-Foundation Models - https://www.researchgate.net/publication/377981578_Geo-Foundation_Models

räumlichen Beziehungen (Topologie) zwischen geografischen Entitäten. Die Generierung von Antworten auf raumbezogene Fragen erfordert eine (oft komplexe) räumliche Analyse der Daten. LLMs, die solche Aufgaben lösen sollen, müssen in der Lage sein, externe Tools zu nutzen.

Um diese Fragen zu beantworten, sind folgende Schritte nötig:

- Die LLM mit Funktionen für die Datensuche (semantisch) und den Datenzugriff ausstatten.
- Die LLM mit Funktionen für die (räumliche) Datenanalyse ausstatten.
- Die LLM über verfügbare Metadaten/Daten und Funktionen für die Datensuche/den Datenzugriff und die Datenanalyse «informieren».

Die sogenannte agentenbasierte RAG-Architektur[34] ist ein echter Fortschritt bei KI-gesteuerten Anwendungen und kann hier eine Lösung bieten.

Bei traditionelle RAG (*Retrieval-Augmented Generation*) wird die Ausgabe eines LLMs optimiert, sodass es auf eine Wissensbasis ausserhalb seiner Trainingsdaten verweist, bevor eine Antwort generiert wird. Das kann verwendet werden, um LLM-Ausgaben in zuverlässigen georäumlichen Metadaten/Daten zu verankern.

Agentenbasierte RAG bringt die Vorteile von traditionellem RAG mit den Fähigkeiten von KI-Agenten zusammen, Arbeitsabläufe zu planen, Tools zu integrieren und auszuführen.

Dieser Ansatz wird gerade in ein paar *Proof-Of-Concepts* getestet, die als Teil der Aktionen SGS-4-24-02 (Aktionsplan 2024, SGS) und SGS-5-25-2 (Aktionsplan 2025, SGS) entwickelt wurden.

Wenn wir diese Anwendungen im Rahmen des KI-Gesetzes der EU betrachten wollen, sollten sie als begrenztes Risiko eingestuft werden: Da diese Anwendungen von einer Bundesbehörde bereitgestellt werden, sind Transparenzpflichten sinnvoll[35]. Das heisst, wenn Leute KI-Systeme benutzen, sollten sie wissen, dass sie mit einer Maschine interagieren, damit sie eine gute Entscheidung treffen können: Das soll das Vertrauen aufrechterhalten.

Diese Anwendungen sind ausserdem nicht besonders riskant, weil:

- offene, nicht sensible und nicht personenbezogene Daten verarbeitet werden.
- der Zugriff auf die Komponenten der NGDI nur im Lesemodus möglich ist.
- es sich um Informationsanwendungen handelt. Es werden keine Entscheidungen oder Prozesse automatisch vom System gestartet.

Die Analyse der Testergebnisse der Prototypen zeigt folgende Punkte:

- Prototyp 1 - Geodaten mit LLMs auffindbar machen
  - Potenzial für die Auffindbarkeit von Datensätzen.
  - Mehrsprachige und hybride Suche wird von den Nutzern geschätzt.
  - Wichtige Themen: geografische Filterung, Komplexität der Agenten, Vertrauen.

[34] https://www.ibm.com/de-de/think/topics/agentic-rag
[35] Das EMBAG, z. B., verpflichtet Behörden auf Bundesebene zur transparenten Erfüllung ihrer Aufgaben mittels elektronischer Mittel.

- Notwendige Verbesserungen in Bezug auf Benutzerfreundlichkeit und Transparenz.

- Prototyp 2a - Geodaten mit LLMs zugänglich machen
  - Vielversprechend, aber fragil: grosses Potenzial, aber Probleme mit Zuverlässigkeit und Skalierbarkeit.
  - Ideal für einfache Aufgaben (kleine Gebiete, einfache Abfragen).
  - Die Qualität der Daten ist entscheidend (verwirrende Datensatznamen und inkonsistente Ergebnisse behindern eine breitere Akzeptanz).
  - Benutzer benötigen Unterstützung (bessere Verwaltung der natürlichen Sprache, Anleitungen und Datensatzkataloge sind unerlässlich).
  - Verbesserung der Stabilität, Leistung und Benutzerfreundlichkeit, um einen höheren Mehrwert zu generieren.
- Prototyp 2b - Geodaten mit LLMs zugänglich machen
  - Es gelten mehr oder weniger die gleichen Überlegungen wie für den Prototyp 2a.

Allgemeine Schlussfolgerungen:

- Die semantische Suche (über verschiedene Ressourcen wie Geometadaten, Geodaten, Geodatenmodelle usw.) scheint ein Anwendungsfall zu sein, in den man investieren kann und der im Rahmen von NGDI-Komponenten getestet werden kann.
- Die Implementierung eines Systems, das unabhängig von bestimmten Anwendungsfällen und/oder Themen ist, kann ziemlich kompliziert sein, da man nicht vorhersagen kann, welche Fragen die Nutzer stellen werden und welche Analysefunktionen man braucht, um diese Fragen zu beantworten. In diesem Fall ist es wichtig, klar zu sagen, was das System kann und wo seine Grenzen sind.
- Die Implementierung eines themenspezifischen Systems sollte weniger kompliziert sein, da man einen begrenzten Satz an notwendigen Daten festlegen und die möglichen Fragen der Nutzer vorhersehen kann.

5.3 Von Prototypen zur Produktion

Wenn man von der Prototyp-Phase zur Implementierung in einer Produktionsumgebung übergehen will, muss man diese Anwendungen im Zusammenhang mit den in Kapitel 2 dieses Dokuments aufgezeigten Fragen betrachten und gleichzeitig versuchen, mögliche Minderungsmassnahmen zu konkretisieren.

Die folgende Tabelle zeigt die verschiedenen Fragestellungen, die entsprechenden Minderungsmassnahmen und eventuelle Aktionen.

Tabelle 1: Fragestellungen, Massnahmen und Aktionen

Fragestellung		Massnahmen	Aktionen
Ethische Fragen	Schädliche Inhalt in LLMs	«Apertus» und nötige Guardrails implementieren	Testfälle festlegen, um zu überprüfen, ob das Modell anfällig für das Problem ist.
	Bias in LLMs	«Apertus» und nötige Guardrails implementieren	Testfälle festlegen, um zu überprüfen, ob das Modell anfällig für das Problem ist (alle Bias in Kapitel 1.1.2).
Umweltfragen		«Apertus» 8B implementieren	
Technische Fragen	Halluzinationen	• RAG implementieren • Klares und starkes System-Prompt konfigurieren (e.g., «Antworte nur mit dem angegebenen Kontext. Wenn du nicht sicher bist, sag einfach 'Ich weiss es nicht'») • Die «Temperatur» des Modells niedriger einstellen, damit die Ausgabe deterministischer und weniger fantasievoll wird	Testfälle festlegen, um zu überprüfen, ob die Massnahmen das Problem der Halluzinationen mildern.
	Reproduzierbarkeit	Deterministische Decodierungsmethoden oder kontrollierte Sampling-Techniken implementieren	Testfälle festlegen, um zu überprüfen, ob die Massnahme die Variabilität verringert.
	Black-Box-Effekt	Erklärbare KI implementieren (e.g., «*Chain of-Thought*»)	
Sicherheitsfragen	Prompt Injection	Starkes Prompt-Design	Testfälle festlegen, um zu überprüfen, dass Benutzer-Prompts die System-Prompts nicht überschreiben können.
	Datenlecks	«Apertus» und nötige Guardrails implementieren	
	Automatisierungsrisiken	Starkes System-Design	Agenten haben nur Leserechte.
Rechtliche Fragen		«Apertus implementieren	

Wichtige Leitplanken für die Arbeit mit der Minderungsmatrix:

- **Die Digitale Souveränität ist sehr wichtig und muss gewährleistet sein**:

  Das bedeutet, dass die eingesetzten Systeme und Anwendungen jederzeit unter der Kontrolle der verantwortlichen Institutionen bleiben müssen. Datenhaltung, Modelltraining und operative Nutzung sollen möglichst auf nationaler oder europäischer Infrastruktur erfolgen, um Abhängigkeiten von externen, nicht kontrollierbaren Akteuren zu vermeiden.

  Praktische Konsequenzen:

  - Einsatz von Schweizer oder europäischen LLMs, die den rechtlichen und ethischen Rahmenbedingungen entsprechen.
  - Sicherstellung, dass sensible Daten nicht in proprietäre Black-Box-Systeme abfliessen.
  - Transparente Dokumentation der Herkunft, Nutzung und Weiterentwicklung der Modelle.

- **Verantwortlichkeit und Governance:**
  - Klare Zuständigkeiten für die Umsetzung und Kontrolle von Minderungsmassnahmen.
  - Ein Governance-Rahmen, der sicherstellt, dass ethische, rechtliche und technische Fragen nicht isoliert, sondern integriert betrachtet werden.

- **Kontinuierliche Weiterentwicklung:**
  - Die Matrix ist ein lebendes Dokument: sie muss regelmässig überprüft, angepasst und erweitert werden.
  - Neue Risiken (z. B. durch technologische Trends) werden systematisch aufgenommen.

6 Variantenvorschläge: LLMs in der NGDI

Als Entscheidungshilfe für die nächsten Schritte im Hinblick auf LLM und NGDI werden hier einige Varianten vorgeschlagen. Die Varianten basieren auf Aufgabenpaketen. Die identifizierten Aufgaben sind in der folgenden Tabelle aufgelistet.

Tabelle 2: Aufgaben

A1	Erfahrungsaustausch	Austausch
A2	Beobachtung der Entwicklungen der gesetzlichen und rechtlichen Grundlagen	Gesetz/Recht
A3	Weiterführung der Arbeit an Massnahmen zur Minderung der ethischen Fragestellungen	Ethik

A4	Weiterführung der Arbeit an Massnahmen zur Minderung der ökologischen Fragestellungen	Umwelt
A5	Weiterführung der Arbeit an Massnahmen zur Minderung der technischen Fragestellungen	Technik
A6	Weiterführung der Arbeit an Massnahmen zur Minderung der sicherheitsrelevanten Fragestellungen	Sicherheit
A7	Prototyp 1 (Geodaten mit LLMs auffindbar machen) weiterführen	Prototype 1
A8	Prototyp 2a (Geodaten mit LLMs zugänglich machen) weiterführen	Prototype 2a
A9	Prototyp 2b (Geodaten mit LLMs zugänglich machen) weiterführen	Prototype 2b
A10	Nutzertests vertiefen	Nutzertests
A11	Implementierung der semantische Suche in swissgeo prüfen	swissgeo
A12	Themenspezifischer Prototyp «Geodaten mit LLMs zugänglich machen» implementieren (anhand Anwendungsfall)	Themenspezifischer Prototyp
A13	Semantische Suche in geocat.ch implementieren	geocat.ch

Auf der Grundlage dieser Aufgaben werden die folgenden Arbeitspakete definiert.

Tabelle 3: Arbeitspakete

P0	A1+A2	Beobachtung der rechtlichen Situation und Austausch weiterführen mit diversen Stakeholder (Bund, Fachleute).
P1	A3+A4+A5+A6+A7+A10	Überprüfung der Minderungsmassnahmen nur im Prototyp 1
P2	A3+A4+A5+A6+A8+A10	Überprüfung der Minderungsmassnahmen nur im Prototyp 2a
P3	A3+A4+A5+A6+A9+A10	Überprüfung der Minderungsmassnahmen nur im Prototyp 2b
P4	A3+A4+A5+A6+A7+A9+A10	Überprüfung der Minderungsmassnahmen in den Prototypen 1, 2a und/oder 2b

P5	A3+A4+A5+A6+A13+A10	Semantische Suche in geocat.ch, inkl. die Minderungsmassnahmen
P6	A11+A10	Semantische Suche in swissgeo (Prototyp) (ohne Minderungsmassnahmen)
P7	A12+A10	Themenspezifischer Prototyp zu «Geodatenmit LLMs zugänglich machen» (ohne Minderungsmassnahmen)

Auf der Grundlage dieser Aufgabenpakete werden die folgenden Varianten vorgeschlagen.

Tabelle 4: Varianten

V0	Keine spezifische Aktion
V1	Weiterführung Prototypen (1, 2a und/oder 2b) (Aufgabenpakete P0+P4)
V2	Weiterführung Prototyp «Nutzen von Geodaten» (2a und/oder 2b) und Umsetzung semantische Suche (Erfahrung von Prototyp 1) in geocat.ch (Aufgabenpakete P0+P3+P5+P7)
V3	Weiterführung Prototyp «Nutzen von Geodaten» (2a und/oder 2b) und Umsetzung semantische Suche (Erfahrung von Prototyp 1) in geocat.ch sowie Prototyp in swissgeo. (Aufgabenpakete P0+P3+P5+P6+P7)

Bemerkungen: Varianten sind abhängig von Planungshorizont und Budget. Etappenweises Vorgehen. Langfristiges Ziel ist V3. Ggf. muss man (je nach Budget und Ressourcensituation) mit V1 anfangen oder mit V2 weiterfahren.

7 Fazit

Sie enthalten konkrete Vorschläge für die nächsten Jahre in Form eines Projektinitialisierungsauftrags.

7.1 Ergebnisse aus der Diskussion GKG und KGK Workshop

Die vorliegenden Arbeiten des Jahres 2025 bildeten eine Entscheidungsgrundlage zur strategischen Bedeutung von LLMs für Geodaten-Infrastrukturen. Im Rahmen des KGK-Workshops vom 11.9.2025 und des GKG-Workshops vom 31.101.2025 wurden die Ergebnisse der Massnahme 5-25-2 präsentiert und diskutiert.

Dabei sind folgende Rückmeldungen eingegangen:

- Generell:
    - Die Initiative wird als wichtig und sinnvoll erachtet.
    - Arbeitspakete und Vorgehen sind korrekt und schlüssig aufgebaut.
    - Es besteht der Wunsch, Synergien mit anderen Bundesämtern zu nutzen (z. B. Einsatz von RAG für Bundesratsprotokolle).
    - Ein breites Publikum soll angesprochen werden, gleichzeitig soll die Nutzerinteraktion systematisch ausgewertet werden.
    - Es ist wichtig sich auf die Funktionen der NGDI zu konzentrieren, welche man mit LLM alimentieren möchte. Dafür sind beiden gewählten Anwendungsfälle («Geodaten suchen» und «Geodaten nutzbar machen») geeignet.
    - Varianten für das weitere Vorgehen wurden diskutiert (gestaffeltes Vorgehen, Variante 1, Kombination von Variante 1 und 3).
- Minderungsmatrix:
    - Die Minderungsmatrix enthält alle relevanten Punkte, welche im Umgang mit generativen KI zu beachten sind. Jedoch muss diese Minderungsmatrix flexibel gehalten werden und weiterentwickelt werden, da die Technologie sich rasant verändert und z.B. in den Wissenschaften viele Arbeiten in diesem Bereich laufen.
    - Das Implementieren von «Apertus» zur Risikominimierung wird nicht überall als geeignet betrachtet, da Apertus (gemäss eigenen Autoren) nicht als finales Produkt herausgegeben wurde, sondern in einer frühen Modellphase (nach dem Fine-Tuning)
    - Erwartungshaltung an den Output von LLMs muss präzisiert werden: Welche Fragen sollen beantwortet werden, welche Ziele verfolgt werden, wie soll kommuniziert werden.
    - Fokus auf geodatenbezogene Fragen, um die Relevanz zu sichern.

- Qualität der Metadaten ist entscheidend – widersprüchliche oder verstreute Informationen mindern die Leistungsfähigkeit der KI.
- Einheitliche Bereitstellung von Geobasisdaten wäre ein starker Anreiz und Voraussetzung für bessere Ergebnisse.

- Nächsten Schritte:
  - Aus den Diskussionen ist klar hervorgegangen, dass man die Arbeiten weiterführen muss und eine Weiterentwicklung der NGDI ohne künstliche Intelligenz keine Option ist.
  - Der Fokus kurzfristig: die Prototypen weiterführen und die entsprechenden Minderungsmassnahmen dazu definieren;
  - Mittelfristig: Live-Version auf **swissgeo** vorbereiten.
  - Arbeitspakete abstimmen und personell zuordnen:
    - P4: Überprüfung der Minderungsmassnahmen in Prototypen 1 und 2b (13 Personen)
    - P5: Semantische Suche in geocat.ch inkl. Minderungsmassnahmen (2 Personen)
    - P6: Semantische Suche in SWISSGEO (Prototyp, ohne Minderungsmassnahmen) (10 Personen)
    - P7: Themenspezifischer Prototyp (ohne Minderungsmassnahmen) (8 Personen)

      Varianten prüfen und konsolidieren (gestaffeltes Vorgehen, Variante 1, Kombination von Variante 1 und 3).

7.2 Generelle Fazit

Die bisherigen Arbeiten haben den Stand der Technik zu LLMs im Bereich Geodaten aufgearbeitet und zentrale ethische, technische, rechtliche und ökologische Fragestellungen bezüglich Arbeit mit generativen KI identifiziert. Es wurden mehrere Prototypen entwickelt, die den Zugang zu Geodaten über semantische Suche und natürliche Sprache erleichtern und sowohl Fachleuten als auch Laien zugänglich gemacht wurden. Bekannte Fehler und Risiken wie Halluzinationen, Bias oder Sicherheitsprobleme wurden dokumentiert und erste Minderungsmassnahmen wie RAG-Ansätze oder Explainable AI geprüft. Zudem erfolgte ein breiter Austausch mit Fachleuten und Institutionen, um Synergien zu nutzen und die Grundlage für eine strategische Entscheidung zur Integration von LLMs in die NGDI ab 2026 zu schaffen.

Für die Arbeiten in den folgenden Jahren sollen stehen folgende Schwerpunkte im Vordergrund:

- Vertiefung der Massnahmen zur Risikominimierung
  - Testen von Apertus anhand dedizierter Testfälle (z. B. Apertus small)
  - Weiterentwicklung des Prompt Designs

    - Einsatz von Methoden zur erklärbaren KI (Explainable AI)
- Vertiefung und Implementierung von LLM und Geodaten anhand der beiden angefangenen Anwendungsfällen
    - Nutzung und Analyse komplexer räumlicher Daten
    - Entwicklung eines spezifischen Use Cases für komplexe Spatial Analysis

Lieferobjekte für das Jahr 2026 sollten folgende sein:

- Implementierung und Validierung der Minderungsmassnahmen;
- Fortsetzung des Anwendungsfalls «Nutzung von Geodaten» (evtl. zusätzlich anhand eines konkreten Anwendungsfalls);
- Die Umsetzbarkeit der semantischen Suche in geocat.ch und in der Testumgebung von SWISSGEO überprüfen.