Anlagen zur

2. Konsultation des Deutschland-Stack

Anlage 1 Allgemeine DATABUND-Empfehlungen:

- Der DATABUND setzt sich für einen Digital-Stack der öffentlichen Verwaltung ein, der nicht als Technologieliste, sondern als strategische Infrastrukturentscheidung verstanden wird. Ziel ist eine souveräne, skalierbare und innovationsfähige Digitale Öffentliche Infrastruktur, die Staat, Wirtschaft und Gesellschaft dauerhaft verbindet. Daher sollte ein verbindlicher, marktfähiger und governance-robuster Digital-Stack für die öffentliche Verwaltung mit einer konsistenten und verbindliche Architektur entstehen.
- Governance ist dabei kein nachgelagerter Verwaltungsakt, sondern muss frühzeitig, klar und durchsetzbar etabliert werden. Ohne verbindliche Regeln entstehen Parallelstrukturen, Abhängigkeiten und ineffiziente Insellösungen. Governance ist die Voraussetzung für Skalierung, Sicherheit und Vertrauen.
- Die bestehende Tech-Stack-Landkarte ist ein wichtiger Ausgangspunkt, muss jedoch systematisch weiterentwickelt werden hin zu einer architekturgetriebenen Gesamtlogik, die Zuständigkeiten, Abhängigkeiten und Schnittstellen eindeutig beschreibt, die schlussendlich eine belastbare Architektur ermöglicht.
- Dabei muss der Stack nicht nur erklären was gebraucht wird, sondern auch wie es zusammenspielt. Kriterien müssen daher operabel sein. Formulierte Prinzipien und Kriterien entfalten nur Wirkung, wenn sie operationalisierbar, überprüfbar und entscheidungsrelevant sind, etwa für Beschaffung, Zulassung, Betrieb und Weiterentwicklung. Was nicht messbar, prüfbar und anwendbar ist, bleibt folgenlos.
- Bei der Planung und Umsetzung des Deutschland-Stack / Tech-Stack sollte eine Fokussierung auf Protokolle, Datenformate und Schnittstellen im Mittelpunkt stehen. Technologie ist vergänglich, daher sollte die verpflichtende Nutzung bestimmter Technologien nicht festgelegt werden. Das Ziel ist entscheidend, die passenden Lösungen sollte dann die deutsche Wirtschaft im Wettbewerb bereitstellen.
- Es ist wichtig, die richtigen Standards und dabei vor allem auf etablierte Standards zu setzen. Beispielsweise bei der Registermodernisierung / NOOTS soll ein neuer Protokollstandard für eine flächendeckende Kommunikationsinfrastruktur der Verwaltung eingeführt werden. Es gibt allerdings mit OSCI/XTA2 bereits heute ein funktionierendes System mit einem zuverlässigen gesicherten Betrieb, insbesondere unter den Aspekten der Funktionalität, der Zuverlässigkeit und der IT-Sicherheit,

welches in fast allen Verwaltungen in irgendeiner Form eingesetzt wird und das von vielen Fachverfahren technisch unterstützt wird. Eine Neuentwicklung verzögert die Umsetzung unnötig und ist aus unserer Sicht daher überflüssig. Auch im Geografie-Bereich (GIS) sind Datenformate schon bereits definiert, aber der Bund fängt hier wieder ganz von vorne an. Es sollte auf etablierte Standards, Datenformate und Schnittstellen gesetzt werden. Offenbar fehlt vielerorts eine gründliche Bestandsaufnahme. Problem für Software-Hersteller sind oft auch konkurrierende Standards. Kommunen haben Schwierigkeiten zu entscheiden, was nun angebunden und entwickelt werden muss. Wenn mehrere Schnittstellen und Datenformate angebunden werden müssen, wird ein Produkt entsprechend teurer und ist für Kommunen nicht mehr finanzierbar.

- Die Verwendung von Open Source Komponenten ist bereits heute in vielen Produkten Stand der Technik. Trotzdem darf bei der Diskussion über digitale Souveränität und Transparenz der Komponenten nicht vergessen werden, dass ein höheres Vertrauen in quelloffene Software nicht per se vorhanden sein kann. Software benötigt konstante (Weiter-)entwicklung und Wartung. Open Source kann nur bei einer ausreichend großen und kompetenten Community den Ansprüchen genügen, die bei kommerziellen Angeboten vorausgesetzt werden. Wir betonen daher die Notwendigkeit kommerzieller Lösungen für die speziellen Aufgaben der Behördendigitalisierung, die keine ausreichende Community bietet. Berücksichtigt werden muss auch, dass es für Open Source nie eine private Vorfinanzierung geben wird, da Open Source im Ergebnis als Wirtschaftsgut wertlos ist. Der Staat muss somit alle Kosten solcher Softwarelösungen sofort und allein tragen, auch die einer kontinuierlichen Weiterentwicklung und Pflege bis zum Ende des Lebenszyklusses.
- API-First-Ansatz: Der IT-Planungsratsbeschluss 2024-55 (API First) sollte konsequent umgesetzt werden. Software-Hersteller werden bereits verpflichtet, Daten über APIs zugänglich zu machen. Leider halten sich Behörden oftmals selber nicht an diese Vorgabe und bieten ihre Daten lediglich auf ihren Plattformen zum Download an. Beispielsweise fehlen auch bei der Nutzung von SAML oftmals Upload-Möglichkeiten für Meta-Daten. Hier müsste es auch eine Verpflichtung zur Bereitstellung über APIs geben.
- Gestartete Projekte konsequent umsetzen um Vertrauen zu schaffen: Ein bereits laufendes Projekt ist z. B. die BundID, für die laut Tagesspiegel Background bis zu 234 Mio. Euro veranschlagt werden. Bezüglich des Ziels mehr Vertrauen in den Staat zu schaffen und dabei auch verantwortungsvoll mit Steuergelder umzugehen, ist dies für Bürger:innen schwer nachvollziehbar, zumal die Usability die Erwartungen der Bürger/innen weit verfehlt. Bis zur nächsten Bundestagswahl sollten daher präsentable

funktionierende Ergebnisse verfügbar sein, in denen große Schmerzpunkte gelöst wurden. Es wäre daher wichtig, bereits Vorhandenes mit einer guten Usability in die Fläche bringen und dies auch (wie bei der Umstellung von 4 auf 5-stellige Postleitzahlen) mit einer Marketing-Kampagne zu verknüpfen, um dies den Bürger:innen näher zu bringen.

- Es bedarf daher einer umfassenden Bestandsanalyse und mehr brown- als greenfield-Denken bei Planung der Digitalisierung. Anstatt einem Top-Down-Ansatz, in der ein großes System für alle Herausforderungen die Lösung bieten soll, ist ein Bottom-up-Ansatz besser, in dem einzelne Probleme und Anforderungen definiert werden und die Wirtschaft entsprechende Lösungen entwickelt. Gute Lösungen werden sich dann auch selbst für andere Bereiche durchsetzen. Besehende Projekte z. B. DVC sollten auf neue Planungen abgestimmt werden.
- Anbindung an Testumgebungen ist immer ein Hindernis. Software zu testen in fremder Umgebung und dies im eigenen System nachzustellen erhöht den Aufwand. Verfügbare Testumgebungen wären daher wichtig. Dieser Bereich ist sehr vielfältig und hier könnte viel Zeit eingespart werden. Bei z. B. OSCI geht das nach Jahren der Etablierung gut, bei neuen Standards fängt man immer von vorne an. Überall wo auf Vorhandenes aufgesetzt werden kann geht es schneller voran.
- Technologielisten können Abhängigkeiten und Innovationshemmnisse erzeugen. Der Digital-Stack muss daher funktionale und architektonische Anforderungen definieren, aber nicht konkrete Produkte vorschreiben. Der Staat muss daher den Bedarf für die Architektur beschreiben, der Markt liefert die Lösungen. Die durch den Deutschland-Stack / Tech-Stack angebotenen Technologien dürfen also nicht als verbindlich gesetzt sein, sondern lediglich als Empfehlung dienen. Verbindlich sollten Festlegungen auf Datenformate, Protokolle, Schnittstellen und einige wenige zentrale Querschnittsdienste getroffen werden. Die Umsetzung sollte grundsätzlich durch die Wirtschaft erfolgen und Technologie-offen und -Technik-neutral sein.
- Der Stack muss funktionale Basiskomponenten einer Digitalen Öffentlichen Infrastruktur enthalten u.a. für Digitale Identitäten, Datenräume, Interoperabilität, Sicherheit und Betrieb. Diese Basiskomponenten sind Teil der öffentlichen Daseinsvorsorge im digitalen Raum. Digitale Infrastruktur ist kritische Infrastruktur und das auch jenseits von Netzen und Rechenzentren.
- Für cloudbasierte IT-Leistungen braucht es ein dynamisches Beschaffungssystem, das Innovation, Wettbewerb und Geschwindigkeit ermöglicht und gleichzeitig Rechtssicherheit und Transparenz wahrt. Dynamische Beschaffung anstatt statischer Vergabe ist daher ein wichtige Kriterium.

- Digitale Souveränität sollte durch die Nutzung von in Deutschland entwickelten Lösungen sichergestellt werden, mit einer Förderung des Deutschen Mittelstandes. Dabei ist auch eine Überwachung der Lieferketten notwendig, um Risiken zu minimieren.
- Projektsteuerung nachhaltig aufstellen: In den Ministerien wird regelmäßig auf Expertise von externen Berater:innen zurückgegriffen, die projektbasiert und daher auf Zeit ihr oft nicht ausreichendes Wissen und ihre Fähigkeiten einbringen. Um über Jahre eine stringente Steuerung und auch persönliche Expertise zu halten, müssen Projektteams mit eigenem Personal im BMDS etabliert werden. Das Vorgehen in Ministerien ist oftmals nicht nachhaltig, wenn externe Berater nach einem Projekt (Jahr) weiterziehen. Zudem gibt es immer noch verschiedene Ansprechpartner in den Ministerien, Parallel-Strukturen mit z. B. dem BMI sollten schnell aufgelöst werden.
- Zuständigkeit prüfen: Der BUND muss prüfen, ob er für bestimmte Bereiche überhaupt die Entscheidungskompetenz besitzt, Entscheidungen bis in die Kommune rechtssicher durchzusetzen zu können. Zudem sollten generell die verschiedenen Perspektiven der Kommunen und Fachverfahrenshersteller berücksichtigt werden.
- Planbarkeit schaffen: Ständige und willkürliche Änderung der Rahmenbedingungen sind Kostentreiber in der kommunalen Verwaltung, da diese in der Regel NICHT durch die Wartungsverträge der Softwarehersteller mit den Kommunen abgedeckt sind. In der Regel bilden diese die Umsetzung von Änderungen an fachlichen Gesetzen und Verordnungen ab. Die Kommunen müssen dann am Ende die Mehraufwände zahlen für Entscheidungen, an denen sie genauso wenig wie die Softwarehersteller beteiligt waren und die sie genauso wenig vertreten können. Ein Problem ist, dass Ausschreibungen und auch Gesetze und Verordnungen zum Teil auf einen bestimmten Anbieter ausgelegt sind. Andere Hersteller halten sich dann zurück und es gibt keinen Wettbewerb um Innovationen und niedrige Kosten, zum Nachteil der Kommunen und der Souveränität.
- Ein leistungsfähiger Digital-Stack entsteht nicht im Verwaltungsvakuum. Der strukturierte, kontinuierliche und institutionalisierte Austausch mit der Wirtschaft ist zwingend erforderlich, um Innovationszyklen, Marktfähigkeit und Skalierung sicherzustellen. Kooperation ist ein wichtiger Teil der Innovationssicherung auch für die Verwaltung.
- Der Stack muss so gestaltet sein, dass neue Marktlösungen schnell, sicher und standardisiert integriert werden können und dass ohne langwierige Sonderverfahren oder Neuarchitekturen. Ein guter Stack ist daher kein starres System, sondern ein offenes Ökosystem.

- Wir fordern als DATABUND eine Transparenz, sowie frühzeitige Einbindung und Beteiligung hinsichtlich der Planungen in unserem Kompetenzbereich und hinsichtlich der zu erwartenden Kosten gegenüber den Kommunen und Landkreisen. Es wird auch eine klare Roadmap benötigt, mit der Software-Hersteller planen können.

Anlage 2 Bewertung und Empfehlungen zum Themenbereich Künstliche Intelligenz im Deutschland-Stack

1. Ausgangslage und Einordnung

Der DATABUND begrüßt die Zielsetzung des Deutschland-Stacks, eine gemeinsame Grundlage für interoperable, föderal anschlussfähige und nachhaltig betreibbare digitale Infrastrukturen zu schaffen. Die Einordnung von Künstlicher Intelligenz als Enabler – insbesondere zur Automatisierung von Aufgaben, Abläufen und Regelwerken – adressiert einen realen Bedarf in der Verwaltungspraxis.

Die derzeitige Ausgestaltung des KI-Bereichs kombiniert …

a) konzeptionelle Aussagen zu agentischer und generativer KI,

b) die Benennung ausgewählter Protokolle/Standards (u. a. MCP, ANP, A2A, AG-UI) sowie

c) eine Technologie-Landkarte, in der Protokolle, Frameworks und Tools gemeinsam abgebildet

werden.

Aus DATABUND-Sicht ist dies ein sinnvoller Startpunkt. Damit die strategischen Ziele (Interoperabilität, Nachnutzung, digitale Souveränität, Vergabefähigkeit und Betriebssicherheit) im föderalen Betrieb zuverlässig erreicht werden, sollte der KI-Bereich jedoch um prüfbare Konformitätsanforderungen ergänzt werden, die unabhängig von einzelnen Tools gelten.

2. Positive Aspekte

- Protokollorientierung bei agentischen Systemen: Die Benennung von Protokollen für Kontextanlieferung sowie Agent-Agent- und Mensch-Agent-Interaktion ist ein wichtiger Schritt, um KI-Komponenten modular und austauschbar zu halten.
- Klar formulierte Nutzungsanforderungen an generative KI: Modellwahl, Einbindung fachspezifischer Quellen (RAG), Prompt-Nachnutzung sowie Nachvollziehbarkeit/Compliance adressieren zentrale Erfolgsfaktoren und mindern Lock-in-Risiken.
- Benennung offener Lücken: Der Hinweis auf künftig erforderliche Standards für Agenten-Überwachung und Qualitätssicherung ist fachlich korrekt und sollte zeitnah in konkrete, testbare Anforderungen überführt werden.

3. Nachschärfungen

3.1 Ebenentrennung: Standards vs. Konformitätsprofile vs. Referenzimplementierungen

Die Technologie-Landkarte führt Protokolle/Standards und Software-Implementierungen (Frameworks/Tools) nebeneinander. Für die praktische Nutzung im föderalen Betrieb ist jedoch eine zusätzliche, Tool-agnostische Ebene nötig: Konformitätsprofile beschreiben prüfbare Mindestanforderungen und Nachweise dafür, wann eine Implementierung als „Stack-konform“ gilt (z. B. Logging/Tracing, Rollenmodell, Lizenznachweise).

Aus Architektur- und Beschaffungssicht ist die aktuelle Darstellung problematisch, weil unklar bleibt,

- welche Schnittstellenregeln verbindlich festgelegt werden (Standards/Protokolle),
- welche Konformitätsanforderungen für Stack-konformen Betrieb gelten (Konformitätsprofile inkl. Test- und Nachweiskriterien; je nach Reifegrad empfohlen/verpflichtend) und
- welche Tools/Frameworks lediglich als mögliche Implementierungen bzw. Orientierung aufgeführt werden, ohne daraus eine Festlegung auf bestimmte Produkte abzuleiten (Toolübersicht; optional ergänzt um Referenzimplementierungen).

Ohne eine klare Trennung zwischen Standards/Protokollen, Konformitätsprofilen und Tools/Frameworks entsteht in der Praxis eine implizite Präferenzbildung über die Nennung in der Toolübersicht. In Beschaffungs- und Betriebsentscheidungen wird „im Stack gelistet“ typischerweise als Risiko- und Legitimitätsvorteil interpretiert; Kompetenzaufbau, Integrationen und Support fokussieren sich entsprechend auf die gelisteten Werkzeuge. Dadurch wird Konformität faktisch über Toolauswahl statt über Schnittstellen- und Nachweiskriterien definiert. Die parallele Listung mehrerer Tools schafft zwar nominelle Wahlfreiheit; echte Austauschbarkeit im föderalen Betrieb entsteht jedoch erst, wenn gemeinsame Schnittstellen- und Konformitätsprofile (inkl. Prüfkriterien) definiert sind. So wird ein Wechsel nicht zur projekthaften Migration, sondern zur beherrschbaren Betriebsentscheidung.

3.2 Agentische KI: Konformitätsprofil „Betriebs- und Nachweisfähigkeit“ (Control-Plane als Profil)

Agentische Systeme sind nur dann sicher, skalierbar und revisionsfähig betreibbar, wenn sie über eine belastbare Betriebs- und Nachweisfähigkeit verfügen. Diese sollte im Deutschland-Stack nicht als einzelnes Werkzeug („Control Plane“) verstanden werden, sondern als Tool-agnostisches KI-Konformitätsprofil, das jede Implementierung – unabhängig von Architektur- oder Tool-Wahl – erfüllen muss.

Dabei ist zu berücksichtigen, dass insbesondere komplexe Multi-Agent-Systeme mit asynchronen Aufrufen, parallelen Workflows und verteilter Tool-Nutzung technisch anspruchsvoll umzusetzen sind. Eine vollständige, durchgängige End-to-End-Rekonstruktion aller Agenteninteraktionen ist nicht in allen Szenarien unmittelbar realistisch oder erforderlich. Um sowohl Umsetzbarkeit als auch Governance-Ziele zu gewährleisten, sollte das Konformitätsprofil daher risikobasiert und gestuft ausgestaltet werden (z. B. Baseline/Advanced bzw. Draft/Recommended/Required).

Als prüfbare Mindestanforderungen (Baseline) sollten gelten:

- Referenzierbarkeit von Abläufen durch Correlation-/Trace-IDs auf Anfrage- bzw. Transaktionsebene,
- Event-Logging von Tool-Calls (Agent → Tool → Datenquelle → Ergebnis), inkl. Fehler- und Abbruchpfaden,
- Policy-Durchsetzung/Guardrails auf Rollen-, Tool- und Datenquellen-Ebene (z. B. Allow/Deny, Scope-Begrenzungen, Freigabe schreibender oder externer Aktionen),
- Datenabflussbegrenzung und -nachweis als systemische Aufgabe: Datenminimierung, getrennte Wissens- und Kontextbereiche, restriktive Tool- und Aktionsfreigaben sowie revisionsfähige Protokollierung relevanter Ausgaben,
- Änderungs- und Rollback-Fähigkeit (Versionierung von Agenten, Prompts/Templates, Tool-Schemas),

- mandantenfähige Logging- und Retention-Mechanismen als Grundlage für Audit und Revision.
- Qualitätssicherungs-Nachweise bei Änderungen („Regression-Light“): deterministischer Change-Impact-Check bei Änderungen an Modellen, Prompts/Templates, Tool-Schemas oder Policies; Prüfung definierter Invarianten (z. B. Schema-/Formatkonformität, Quellenpflichten, Rollen-/Berechtigungsregeln, Tool-Call-Beschränkungen sowie harte Fail-Conditions).

„Regression-Light“ ist dabei ausdrücklich als Verfahrensstandard zu verstehen und nicht als Wahrheits- oder semantischer Qualitätsstandard. Aufgrund stochastischer Eigenschaften großer Sprachmodelle sowie unterschiedlicher Trainings- und Gewichtungsregime sind modellübergreifende Ergebnisgleichheit oder semantische Regressionsprognosen nicht realistisch. Ziel von Regression-Light ist daher die regel- und Governance-konforme Stabilität (Invarianten) bei Änderungen, nicht die Vorhersage inhaltlicher Antwortgleichheit (vgl. Abschnitt 8.1).

Es ist fachlich anerkannt, dass eine vollständige semantische Verhinderung von Datenabfluss auf Ebene großer Sprachmodelle derzeit nicht zuverlässig lösbar ist, da sensible Informationen paraphrasiert, abstrahiert oder kombiniert werden können. Datenabflusskontrolle ist daher im Deutschland-Stack nicht als absolute Inhaltsverhinderung, sondern als mehrschichtiges Begrenzungs-, Kontroll- und Nachweiskonzept zu verstehen, das technische, organisatorische und verfahrensbezogene Maßnahmen kombiniert.

Erweiterte Anforderungen (Advanced) sind für definierte Risikoklassen bzw. produktive Multi-Agent-Szenarien verbindlich vorzusehen, insbesondere wenn personenbezogene Daten verarbeitet, rechtsrelevante Entscheidungen vorbereitet oder schreibende Aktionen ausgelöst werden. Dazu zählen u. a.:

- verteiltes Tracing (Spans) über Agenten, Tools und Datenquellen hinweg,
- rekonstruierbare Workflow-Strukturen (z. B. DAGs) für parallele und verzweigte Abläufe,
- gestufte Freigabe- bzw. Human-in-the-Loop-Mechaniken für kritische Tool-Calls und sensible Ausgaben,
- laufendes Monitoring inkl. Canary-/Shadow-Rollout und definierter Rollback-Optionen.

Ohne ein derart gestuftes Konformitätsprofil besteht die Gefahr, dass agentische Systeme entweder nicht ausgerollt oder informell betrieben werden. Beides unterläuft Governance-, Sicherheits- und Auditziele und verhindert eine föderal skalierbare Nutzung im Deutschland-Stack.

3.3 „Nutzenden-Compliance/Nachvollziehbarkeit“ als Konformitätskriterium

Die Forderung nach Nachvollziehbarkeit und Compliance ist richtig, bleibt aber ohne Mindestanforderungen schwer prüfbar. Der DATABUND empfiehlt, dies als Konformitätskriterium auszugestalten, mindestens mit:

- Provenienz-Nachweisen (Quellen, Lizenzen, Retrieval-Belege),
- Versionierung (Modell, Prompt/Template, Wissensbasis, Tools),
- Logging/Retention revisionssicher und mandantenfähig,
- Risikoklassen & Pflichtkontrollen (Human-in-the-Loop je nach Verfahren/Schwere).

4. Open Source, Lizenzmodelle und Deutschland-Cloud-Betrieb

Ein erheblicher Teil der KI-nahen Softwarebausteine ist Open-Source-Software und unterliegt Lizenzpflichten. Für den Betrieb in einer Deutschland-Cloud ist das ein Betriebs- und Governance-Faktor mit Auswirkungen auf Nachnutzung, Vergabefähigkeit und Exit-Fähigkeit.

4.1 Aktueller Stand: kein unmittelbarer Copyleft-Zwang aus den Top-Level-KI-Bausteinen

Viele der in der KI-Landkarte benannten Kernprojekte sind permissiv lizenziert (z. B. MIT/Apache/BSD). Dadurch ergibt sich typischerweise kein unmittelbarer AGPL/GPL/LGPL-Zwang allein durch diese Top-Level-Auswahl.

4.2 Praktische Risiken: entstehen in Abhängigkeiten und Artefakten

Relevante Risiken entstehen erfahrungsgemäß über transitive Dependencies, Plugins/Connectoren/Add-ons, Distribution von Artefakten (Container-Images, SDKs, Edge-Komponenten) sowie Netzwerkbereitstellung je nach Lizenzmodell.

4.3 Modelle und Daten: „Open Weights“ ist nicht gleich Open Source

Für den KI-Betrieb sind Lizenzen von Modellen, Embeddings und Datenquellen entscheidend. Diese unterliegen häufig restriktiven Nutzungsbedingungen, die Modellwahl, Fine-Tuning, Weitergabe und föderale Nachnutzung begrenzen können. Diese Ebene sollte im Stack als Katalogpflicht sichtbar werden.

4.4 Empfehlung: „License & Supply-Chain“ als Konformitätsprofil

Der DATABUND empfiehlt, für KI-Bausteine ein Konformitätsprofil „License & Supply-Chain“ vorzusehen (Tool-agnostisch), mindestens mit:

- SBOM pro Release (inkl. transitiver Dependencies),
- Lizenzinventar und automatisierte Prüfung als CI/CD-Gate,
- automatisierte Attribution/NOTICE,
- Supply-Chain-Security (Artefaktsignierung, Update-/Patch-Prozesse),
- Lizenzprofile für Modelle und Daten (Nutzung, Weitergabe, Fine-Tuning, Einschränkungen).

5. Offenes System statt abgeschlossenes Biotop: Aufnahme- und Lebenszyklus als Stack-Mechanik

Gerade im KI-Bereich entwickeln sich Technologien und Standards in kurzen Zyklen. Eine starre, geschlossene Festlegung würde das Risiko erhöhen, dass der Stack in der Praxis umgangen wird oder notwendige Sicherheits-/Innovationsupdates zu langsam aufgenommen werden. Der DATABUND empfiehlt daher, die gesteuerte Weiterentwicklung als Stack-Mechanik auszuformen: Aufnahme, Reifegrad und Abkündigung werden über Profil- und Konformitätskriterien gesteuert.

5.1 Vorschlag: „Reifegradmodell“ statt „Prozessbeschreibung“

Anstelle einer prozesslastigen Darstellung bietet sich ein Reifegradmodell an, das unmittelbar in die Stack-Logik passt:

- Draft: erste Einordnung und Schnittstellenbeschreibung
- Recommended: erfüllt Konformitätsprofil(e) + Prüfkriterien
- Required: für produktive Nutzung in definierten Szenarien verpflichtend, inkl. Nachweisartefakten

Die dazugehörigen „Gates“ werden damit zu Konformitätsnachweisen (Artefakte), nicht zu externen Prozessvorgaben.

5.2 Lifecycle-Regeln (als Katalogfelder und Release-Kriterien)

- Versionierung + definierte Supportzeiträume,
- Deprecation-Pfad (Ankündigung → Parallelbetrieb → Abschaltung),
- Security-Fast-Track (kritische Lücken: schnelle Updates/Hotfix, Rollback),
- transparente Change-Logs/Roadmap.

6. Prompt-Wiederverwendung: als „Use-Case-Artefakt“ statt Prompt-Textsammlung

Die Forderung nach Prompt-Nachnutzung ist sinnvoll, muss jedoch so operationalisiert werden, dass sie modell- und plattformübergreifend praktikabel bleibt. Eine 1:1-Portabilität von Prompts zwischen unterschiedlichen LLMs ist nicht verlässlich (Instruction-Following, Tool-Calling, Kontextfenster, Tokenisierung, Safety/Policy unterscheiden sich).

Der DATABUND empfiehlt daher, Prompt-Nachnutzung im Stack als Use-Case-Artefakt zu definieren, das aus drei Teilen besteht:

- Use-Case-Definition (modellneutral): Zweck/Scope, Risiko-/Verfahrensklasse, Ein/Ausgabeanforderungen, Qualitätskriterien (Quellenpflicht, Normstellen, Format).
- Template (halb-portabel): standardisierte Struktur mit Platzhaltern für Kontext/RAG.
- Adapter (modell-/plattformbezogen): Tool-Schemas, Parameterprofile, Rendering.

Als Katalog-/Profilfelder sollten mindestens vorgesehen werden: Versionierung + Ownership/Freigabe, Klassifikation (keine sensiblen Beispiele), Evaluationsset, Kompatibilitätslabel („getestet mit …“), Deprecation-Regeln. Damit wird Prompt-Nachnutzung robust, auditierbar und föderal nachnutzbar.

7. Modellwechsel und Modellversionen: Konformitätskriterium „Change & Regression“

Die angestrebte Modellwahl und Austauschbarkeit im Deutschland-Stack ist zu begrüßen. Gleichzeitig ist im Verwaltungskontext zu berücksichtigen, dass Änderungen an eingesetzten Modellen oder Modellversionen unterschiedliche Risikoprofile aufweisen und daher differenziert zu behandeln sind. Nicht die Versionsbezeichnung, sondern die Art und Tiefe der Änderung ist maßgeblich für die erforderlichen Prüf- und Freigabemechanismen.

Dabei ist zwischen folgenden Änderungsarten zu unterscheiden:

- Wechsel der Modellgeneration oder Modellfamilie (z. B. GPT-4 → GPT-5), auch beim selben Anbieter: Eine solche Änderung stellt regelmäßig eine wesentliche Änderung dar, da sich Fähigkeiten, Fehlermuster, Sicherheitsmechanismen, Interaktionsverhalten sowie Randbedingungen der Modellnutzung signifikant unterscheiden können.

- Modellwechsel zwischen Anbietern oder Modellfamilien (z. B. Wechsel von GPT-5 zu einem anderen LLM): Auch dies ist unabhängig vom Einsatzzweck als wesentliche Änderung zu behandeln.
- Modellversionsupdate innerhalb derselben Modellgeneration (z. B. GPT-5 → GPT-5.1): Solche Updates sind als inkrementelle Änderungen einzuordnen, sofern keine funktionalen, sicherheits- oder governance-relevanten Änderungen deklariert sind.

Für wesentliche Änderungen (Modellgeneration oder Modellfamilie) empfiehlt der DATABUND, das Konformitätskriterium „Change & Regression“ verbindlich anzuwenden, mindestens mit:

- deterministischen Prüfungen (Schema-/Formatkonformität, Quellenpflicht, harte Fail-Conditions, Rollen- und Berechtigungsregeln),
- stratifizierten Stichproben für kritische Use Cases, verstanden als gezielte Auswahl definierter Fallklassen (z. B. Standard-, Grenz-, Konflikt-, Sonder- und risikobehaftete Fälle), um relevante Fehl- und Abweichungsmuster systematisch abzudecken; nicht als statistische Repräsentativitätsaussage,
- gestuftem Rollout (z. B. Shadow- oder Canary-Betrieb) inkl. Monitoring und klar definierter Rollback-Optionen.

Für inkrementelle Modellversionsupdates innerhalb derselben Generation sind demgegenüber vereinfachte Verfahren ausreichend, insbesondere:

- ein Regression-Light-Verfahrenscheck (Prüfung definierter Invarianten wie Format, Pflichtbestandteile, Rollen-/Policy-Regeln und verbotene Aktionsklassen),
- verstärktes Monitoring im Betrieb sowie Rollback-Fähigkeit.

Aufgrund der stochastischen Eigenschaften großer Sprachmodelle und unterschiedlicher Trainings- und Gewichtungsregime sind modellübergreifende Ergebnisgleichheit oder semantische Regressions-Prognosen nicht realistisch. Das Konformitätskriterium „Change & Regression“ ist daher als Verfahrensstandard zu verstehen, nicht als Wahrheits- oder Qualitätsstandard. Ziel ist die beherrschbare Stabilität, Nachweisbarkeit und Governance-Konformität bei Änderungen, nicht die Vorhersage identischer Antwortinhalte.

Ergänzend können für ausgewählte, besonders kritische Use Cases oder zur vertieften Ursachenanalyse DoE-ähnliche Vorgehensweisen (Design of Experiments) eingesetzt werden. Diese dienen der systematischen Untersuchung von Einflussfaktoren (z. B. Prompt-Varianten, Parameter, Tool-Konfigurationen) und sind ausdrücklich nicht als verpflichtender Bestandteil, sondern als optionale Erweiterung zu verstehen.

8. Stabiler Vergleichsanker gegen Modell- und Versionschaos: Konformitäts-Benchmark als Stack-Referenz

Angesichts der hohen Dynamik bei Modellen, Versionen und Betriebsvarianten besteht das Risiko, dass Evaluations- und Regressionstests langfristig zu einem beweglichen Ziel werden. Der Deutschland-Stack sollte daher einen Konformitäts-Benchmark vorsehen, der nicht als Wahrheits- oder Zielstandard fungiert, sondern als Verfahrens- und Vergleichsanker, um Drift, Regression und grobe Abweichungen über Zeit und Modellversionen hinweg nachvollziehbar zu machen.

8.1 Konformitäts-Benchmark als Verfahrensstandard (nicht als Wahrheitsstandard)

Etablierte Evaluations- und Benchmark-Testsets (z. B. MMLU, HellaSwag oder vergleichbare Verfahren) eignen sich im Kontext des Deutschland-Stacks nur eingeschränkt zur Bewertung fachlicher oder rechtlicher Korrektheit. Sie operieren überwiegend auf einer allgemeinkognitiven Ebene, sind zumeist englischsprachig geprägt und bilden die normativ strukturierte Fach- und Verwaltungssprache der deutschen öffentlichen Verwaltung (z. B. Recht, Qualitätssicherung, Behördenverfahren) nur unzureichend ab.

Gleichzeitig ist bei der Definition zentraler Benchmarks zu berücksichtigen, dass normativ gesetzte Zielmetriken das Risiko systematischer Fehlsteuerung bergen. Entsprechend dem Goodhart's Law würde ein staatlich definierter, verbindlicher Konformitäts-Benchmark mit hoher Wahrscheinlichkeit dazu führen, dass Modelle gezielt auf das Erfüllen dieser Tests optimiert werden. Dies würde die Aussagekraft der Benchmarks untergraben und Scheinkonformität begünstigen, ohne robuste Aussagen über Governance-Konformität oder Praxistauglichkeit im Verwaltungskontext zu liefern.

Vor diesem Hintergrund sollte der Konformitäts-Benchmark im Deutschland-Stack nicht als Qualitäts- oder Wahrheitsstandard, sondern ausdrücklich als Verfahrensstandard zur Drift- und Veränderungserkennung ausgestaltet werden. Ziel ist es, relevante Veränderungen von Modellen, Modellversionen und Systemverhalten über Zeit hinweg sichtbar und nachvollziehbar zu machen, ohne ein normatives Optimierungsziel vorzugeben oder fachliche Wahrheit zu definieren.

In dieser Rolle leisten externe, etablierte Testsets einen sinnvollen Beitrag als grobe, domänenunabhängige Vergleichsindikatoren. Sie können:

- Verhaltensänderungen über Modellversionen hinweg anzeigen,
- bei Einsatz von Kalibrierankern helfen, Modell-Drift von Mess- oder Bewertungsdrift zu unterscheiden,
- sowie eine baseline-artige Vergleichbarkeit herstellen, ohne fachliche Angemessenheit zu behaupten.

Die eigentliche Qualitätssicherung für fachliche, rechtliche oder verfahrensrelevante Anwendungsfälle kann und sollte hingegen nicht zentral benchmarkbasiert erfolgen. Angesichts der hohen Heterogenität der Verwaltungs-Use-Cases ist sie risikobasiert und use-case-spezifisch lokal umzusetzen, insbesondere durch:

- definierte fachliche Invarianten (z. B. Pflichtbestandteile, Quellen- und Normstellenangaben),
- regelbasierte Prüfungen und Regression-Light-Verfahren,
- stratifizierte Stichproben typischer und kritischer Fallklassen,
- sowie organisatorische Kontrollmechanismen (z. B. Human-in-the-Loop bei sensiblen Verfahren).

Der Konformitäts-Benchmark dient damit als frühzeitiger Indikator für relevante Veränderungen, nicht als Ziel- oder Qualitätsmaßstab. Er unterstützt Governance, Betrieb und Revision, ohne Innovationsanreize zu verzerren oder föderale Handlungsspielräume einzuschränken.

8.2 Kalibrieranker durch gepinnte Referenzmodelle

Zur Sicherstellung der Vergleichbarkeit über Zeit und Modellversionen hinweg können gepinnte Referenzmodelle als Kalibrieranker eingesetzt werden. Diese dienen ausschließlich als konstanter Vergleichspunkt innerhalb der Evaluationspipeline, um Veränderungen der Messergebnisse einordnen zu können (z. B. Unterscheidung zwischen Modell-Drift und Test- bzw. Bewertungsdrift).

Gepinnte Referenzmodelle sind nicht als Ziel-, Qualitäts- oder Referenzmodelle im normativen Sinne zu verstehen. Sie begründen weder eine qualitative Vorrangstellung noch eine Beschaffungs- oder Einsatzempfehlung und dürfen nicht als implizite Vorgabe für Modellwahl oder Architektur interpretiert werden. Ihre Funktion beschränkt sich auf die methodische Kalibrierung des Vergleichsverfahrens.

Der Einsatz solcher Kalibrieranker ist optional und sollte transparent dokumentiert werden (z. B. Auswahlkriterien, Pinning-Zeitraum, Aktualisierungsregeln), um Nachvollziehbarkeit und Revisionsfähigkeit sicherzustellen.

8.3 Governance und Weiterentwicklung

Der Konformitäts-Benchmark sollte einer klaren Governance unterliegen, insbesondere:

- Versionierung der eingesetzten Tests und Bewertungsverfahren,
- transparente Change-Logs bei Anpassungen,
- regelmäßige Überprüfung der Eignung externer Testsets,
- klare Abgrenzung zwischen Baseline-Vergleich (Benchmark) und use-case-spezifischen Prüfungen, die risikobasiert lokal erfolgen können.

Damit wird der Konformitäts-Benchmark zu einem stabilen, nachvollziehbaren Vergleichsinstrument, ohne Innovationsanreize zu verzerren oder föderale Handlungsspielräume einzuschränken.

9. Beantwortung der Fragen des Bundesministeriums

9.1 Gibt es pragmatische Umsetzungsbedingungen, die ergänzt werden müssen?

Ja. Empfohlen werden insbesondere:

- Konformitätsprofil „Betriebs- und Nachweisfähigkeit“ (Control-Plane-Anforderungen),
- Responsible-AI-Mindestanforderungen (Provenienz, Versionierung, Logging, Risiko-klassen),
- Konformitätsprofil „License & Supply-Chain“ (SBOM/Lizenzinventar/Notices),
- Lizenzprofile für Modelle und Daten als Katalogpflicht,
- Reifegradmodell (Draft/Recommended/Required) inkl. Lifecycle-Regeln,
- Use-Case-Artefakte für Prompt-Nachnutzung,
- Konformitätskriterium „Change & Regression“ für Modellwechsel,
- Konformitäts-Benchmark als stabiler Vergleichsanker.

9.2 Sind die priorisierten Technologiefelder mit relevanten Technologien und Standards unterlegt?

Teilweise. Positiv ist die Benennung erster Protokolle/Standards für agentische KI. Für eine belastbare Unterlegung sollten jedoch Konformitätsprofile, Prüfkriterien und

Nachweisartefakte ergänzt werden. Zudem ist die Trennung zwischen Standards und Tools/Frameworks zu schärfen.

9.3 Passt die Ausrichtung der Standards und Technologien zu den strategischen Zielen?

Im Zielbild ja, in der Operationalisierung noch nicht vollständig. Interoperabilität und Souveränität werden robust erreicht, wenn Standards als Schnittstellen plus Konformitätsprofile beschrieben sind, Nachvollziehbarkeit als prüfbares Kriterium gilt, Modell-/Datenlizenzen transparent geregelt sind und der Stack über Reifegrad- und Lifecycle-Mechaniken kontrolliert weiterentwickelt wird. Der Konformitäts-Bechmark stärkt langfristig Vergleichbarkeit und Auditierbarkeit.