Konzept für die KI-Grundversorgung

Dieses Konzept soll dabei unterstützen eine nachhaltige Lösung für die Bereitstellung von generativen KI-Diensten für die Bildung mit Fokus auf Hochschulen zu erreichen um Teilhabe Aller an KI-Services sicherzustellen. Bei der Entwicklung des Konzepts wurden insbesondere die Erfahrungen durch den Betrieb des KI-Servicezentrums KISSKI berücksichtigt. Unser Ziel ist der Aufbau eines kompletten KI-Ökosystems, um Nutzende zu einer effizienten Nutzung von KI-Services zu befähigen und auch über die eigenen Angebote hinaus mit anderen Anbietern zusammen zu arbeiten. Dabei sollen verschiedene KI-Services auch vernetzbar sein um die KI-Landschaft in Deutschland insgesamt zu stärken und gleichzeitig den Nutzenden mehr Expertise, Selbstbestimmung und Auswahl zu bieten. Aufgrund der benötigten Fachkenntnis und Investitionsvolumina können wir nur als länderübergreifende und kooperierende Community langfristig die benötigten Dienstleistungen bereitstellen, welche einzelne Akteure nicht finanzieren können. Das Dokument soll als Diskussionsgrundlage dienen und im Rahmen von Workshops weiterentwickelt werden.

Inhaltsverzeichnis

1. Status Quo der KI-Versorgung

Die Versorgungslandschaft mit KI-Lösungen an Deutschen Hochschulen ist derzeit – wo überhaupt vorhanden - sehr heterogen. So gibt es Entwicklungen von Brower-basierten Schnittstellen, die externe Modelle für die Angehörigen bereitstellen wie etwa „HAWKI[1]“ oder „LibreChat[2]“, Instituts-interne

[1] https://hawk-digital-environments.github.io/HAWKI2-Documentation/
[2] https://www.librechat.ai/

Lösungen wie Helmholtz‘ „Blabladior[3]“ und einige wenige Lösungen, die ganze Bundesländer versorgen wollen wie KI:Connect[4] für Nordrhein-Westfalen. Auf Bundesebene gibt es die vom BMFTR geförderten KI-Servicezentren (KISSKI[5], WestAI[6], HessianAI[7], KISZ-BB[8]), deren Auftrag die allgemeine KI-Unterstützung von Nutzenden in ganz Deutschland ist.

Bei vielen Lösungen besteht das Grundproblem darin, dass nicht skalierbar geplant wird, sondern von Anfang an die Verwendung durch einen begrenzten Nutzendenkreis vorgesehen ist. Dabei werden auch Hardwareressourcen und Personal nur für die geplante, begrenzte Lösung finanziert, zusätzliche Mittel um die eigene Lösung doch größer zu skalieren sind schwer einzuwerben. Dabei kommt es zwischen verschiedenen Anbietern oft zu Parallelentwicklungen, welche zur Verschwendung von Ressourcen führen und Personal binden, welches sonst an der Weiterentwicklung der KI-Technologien im Ganzen arbeiten könnten. Weiterhin besteht dadurch beim Endnutzenden ein Überangebot vieler funktional ähnlicher Lösungen, denen bekannte, kommerzielle Anbieter gegenüberstehen, die dadurch als bekannter/sicherer wahrgenommen werden. Eine bundesweite Koordination der KI-Entwicklungen oder eine Verständigung auf wenige, technisch geeignete Versorger, fehlt - wäre aber aus Effizienz- und Kostengründen angeraten.

2. Definition KI-Grundversorgung

Mit KI-Grundversorgung bezeichnen wir den Zugang zu generativen KI-Diensten, die regelmäßig und interaktiv zur Unterstützung von Endnutzern genutzt werden und dort typische Use-Cases abdecken. Diese umfassen beispielsweise Zusammenfassen oder Generieren von Texten, Literaturrecherche, Unterstützung bei der Erstellung von Lehr- und Lernhilfen aber auch Unterstützung im Studium. Typischerweise sind damit Inferenzanfragen mit Large-Language-Modellen (LLMs) gemeint. Diese Anfragen werden von den End-Nutzenden direkt im Webbrowser oder mittels Werkzeugen indirekt übermittelt und Nutzendenanfragen müssen dabei meist zeitnah beantwortet werden (typischerweise innerhalb von Sekunden).

Grundversorgung bedeutet nicht einfach nur das bloße Vorhandensein von einem oder mehreren Werkzeugen, sondern auch die rechtssichere Bereitstellung und elementare Unterstützung und Training der Anwender mit diesen Werkzeugen. Aufbauend auf einer Grundversorgung ist ein Community-basiertes Ökosystem angedockt, dass Nutzenden ermöglicht KI effektiv und verantwortungsvoll für viele weitere Anwendungsszenarien bequem zu nutzen.

[3] https://helmholtz.cloud/services/?serviceID=d7d5c597-a2f6-4bd1-b71e-4d6499d98570
[4] https://kiconnect.pages.rwth-aachen.de/pages/
[5] https://kisski.gwdg.de/
[6] https://westai.de/
[7] https://hessian.ai/
[8] https://hpi.de/ki-servicezentrum/

Nicht verstehen wir mit der Grundversorgung individuelle Bedürfnisse von Forschenden an KI-Training und Zeit-unkritische Inferenz von komplexen Modellen. Diese sind nicht nur schwierig zu quantifizieren, sondern hängen stark vom Nutzungsszenario ab. Diese können mit Rechenzentren bspw. im NHR hochqualitativ abgewickelt werden. Ein grundlegender Unterschied des Nutzungsszenarios Forschung ggü. Grundversorgung ist hierbei die Expertise der Nutzenden (mittel/hoch vs. gering) und die Zeitkritikalität der Anfragen (Stunden vs. interaktiv), welche sich unmittelbar auf die Verarbeitung auswirkt (Scheduling im Batch-Modus vs. hochverfügbarer Live-Service).

Ebenfalls können auch nicht die Bedürfnisse ALLER Nutzenden erfüllt werden, aber 95% der Use-Cases für generative KI-Dienste sollen - qualitativ zufriedenstellend - erfüllt werden.

3. Anforderungen

Eine langfristig nachhaltige Lösung für die KI-Grundversorgung muss unseren Ansichten nach Regulierungskonformität, Unabhängigkeit, Entscheidungsfreiheit, Anpassungsfähigkeit, Teilhabe, Qualität und eine langfristige Perspektive für Deutschland bereitstellen.

Unabhängigkeit bedeutet die Kontrolle über Daten und Gedanken in einem eigenen Ökosystem. Bereits jetzt sehen wir, wie LLMs genutzt werden um Meinungen in Sozialen Medien zu beeinflussen. Auch Antworten von ChatBots sind durch die Auswahl der Trainingsdaten ideologisch geprägt und können Nutzende in die Irre führen. Durch gezielte Beeinflussung der Antworten eines LLMs kann Einfluss auf die Werte und Normen von Gesellschaften genommen werden. Diese Einflüsse, bspw. auf die öffentliche Meinung in aktuellen politischen Geschehnissen oder vor Wahlen sollte durch Nutzung vertrauenswürdiger Modelle verhindert werden.

Regulierungskonformität bedeutet die Übereinstimmung mit den hohen Regularien und Standards in Deutschland, insbesondere dem Datenschutz und der Datensouveränität. Auch im Hochschulkontext gibt es viele Use-Cases mit sensiblen Daten. Im Fall kommerzieller „gratis“ Anbieter stellen die eingegebenen Daten jedoch den Preis für den Service dar und werden vom Anbieter ausgewertet oder für Training neuer Modelle nachgenutzt.

Entscheidungsfreiheit muss durch eine Governance-Struktur geregelt werden, bei der die Anbietenden und Nutzenden entscheiden können, wie ein Produkt aussieht und welche Funktionalitäten es bietet. In proprietären Ökosystemen sind die Nutzenden 100% auf den Hersteller angewiesen und haben oft keinen Einfluss auf Design- und technische Entscheidungen.

Anpassungsfähigkeit ist notwendig für die Anbindung an andere Software und Nutzung für ein breites Spektrum an Use-Cases. Im universitären Kontext ist hier beispielsweise die Anbindung an Lehrsoftware (z.B. Moodle, StudIP) zu nennen. Aber auch generell müssen Anpassungen möglich sein und offene Schnittstellen hierzu verfügbar. Hier werden oft proprietäre Schnittstellen bspw. für Vektor-Datenbanken

genutzt um ein Ökosystem zu kontrollieren. Was vermieden werden muss, ist dass alternative Anbieter ausgeschlossen werden, da sie mangels Zugriff auf Schnittstellen wenig Funktionalitäten bieten können.

Teilhabe bedeutet, das allen Menschen offenen Zugang zu den Funktionalitäten eines Services gewährt werden soll. Gerade im universitären Kontext muss allen Studierenden Zugang zu den gleichen Funktionalitäten geboten werden. Kommerzielle Anbieter operieren oft im „Freemium"-Model, bei denen Basisdienste und eine begrenzte Anzahl Zugriffe „gratis" angeboten werden, eine unbegrenzte Nutzung und Zugriff auf alle Funktionalitäten jedoch nur zahlenden Nutzenden zur Verfügung stehen. Insbesondere bei API-Nutzung sind Funktionen oft aufpreispflichtig.

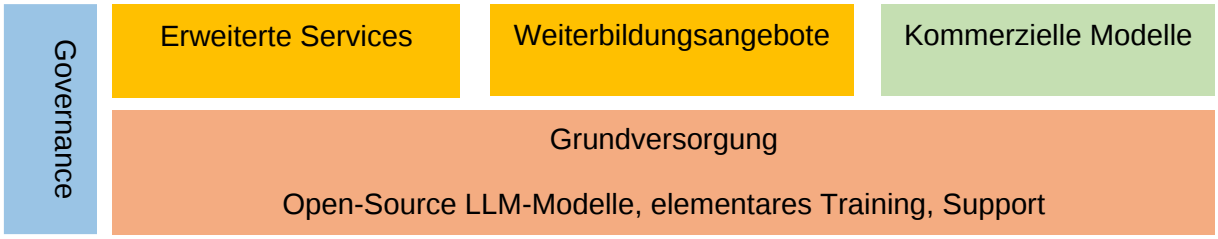
Qualität muss für jeden Service und jeden unterstützten Use-Case gleich und gut sein - andernfalls werden die Nutzenden die Services der Grundversorgung nicht nutzen und sich wieder hin zu kommerziellen Anbietern orientieren. Auch muss der Großteil der durch Hochschulen benötigten Use-Cases durch die Grundversorgung abgedeckt werden, da sonst das gleiche Problem auftritt. Wir streben an mindestens 95% der Use Cases für Hochschulen durch die Grundversorgung und das darauf aufbauende Ökosystem abzudecken.

Langfristige Perspektive: Eine Lösung sollte langfristig gedacht werden und eine Perspektive für den Standort bieten. Deutschland benötigt eine nicht von kommerziellen Interessen getriebene KI-Lösung, die Grundversorgung für Hochschulen und Bildung bietet sowie weitere Entwicklungen und Beiträge zu einem offenen Ökosystem leisten kann. Die langfristige Perspektive der KI-Grundversorgung in Deutschland ist nicht mit gewinnorientierten Firmen umsetzbar.

4. KI-Ökosystem

Für unsere Vision eines KI-Ökosystems habe wir als zentralen Baustein die **Grundversorgung** mit ihren Teilbereichen Open-Source LLM-Modelle, Training und Support identifiziert. Dazu kommen kommerzielle Modelle, die bspw. über APIs in Services eingebunden werden. Dabei ist es wichtig, dass die Nutzenden entweder informierte Entscheidungen treffen können zwischen Modellen mit verschiedenen Datenschutzstandards oder Institutionen entscheiden können für ihre Mitarbeitenden nur sichere Services verfügbar zu machen. Ein großes Feld sind **erweiterte Services** wie die Nutzung sicherer KI-Services in Angeboten von Dritten. Dies ist zB. über Nutzung marktdominierender Schnittstellen wie der OpenAI API möglich. So kann Chat AI bereits in bestehende Hochschulwerkzeuge wie Moodle und StudIP integriert werden. Statt dem Ausnutzen solcher Markteffekte ist auch eine Verständigung vieler Dienstleister auf gemeinsame Standards möglich, was eine kollektive Weiterentwicklung des KI Ökosystems befeuern würde. Weiterhin benötigt es Weiterbildungsangebote, die über ein Basistraining der Nutzenden hinausgehen. Dieses **Basistraining** umfasst die Befähigung der Nutzenden zu verstehen wie KI funktioniert, das richtige Werkzeug für ihre Zwecke auszuwählen und das Werkzeug einzusetzen um ihre Aufgaben vollständig umzusetzen. Im Gegenzug sehen wir weitere **Weiterbildungsangebote** im KI-Ökosystem im Bereich der Anwendungsentwicklung, Bereitstellung, oder Forschung. Dabei sollen diese Weiterbildungsangebote von allen Teilhabenden im KI-Ökosystem kommen, bspw. Anbietern von KI-

Services, kommerziellen Anbietern von KI, Cloudlösungen und Hardware, Anwendern (welche ihre Use-Cases und brancheninternen Anforderungen vorstellen) und auch Einzelpersonen, die Lösungsansätze vorstellen. Ein weiterer fundamentaler Grundstein des KI-Ökosystems ist **Governance**. Um ein kritisches Momentum im Ökosystem zu erreichen, sollten die Nutzenden in die Entwicklung von Services und das Angebot von Schulungen und weiteren Services eingebunden sein. Dies kann minimal durch die Interaktion über eine Community und die Umsetzung von Nutzendenanfragen geschehen. Ein wesentlich interaktiverer Ansatz ist jedoch auch Entwicklungen der Nutzenden zu integrieren und aktiv im Ökosystem zu bewerben und die Roadmap der weiteren Entwicklungen am Feedback und den Wünschen der Nutzenden auszurichten. Dieser Ansatz sichert nicht nur die aktive Mithilfe der Nutzenden und die Bindung an die Services, die sie aktiv mitgestaltet haben, sondern sichert auch das Vorhandensein einer Nutzendenbasis für neue Services.



Die hochschullokale Bereitstellung von KI-Dienstleistungen ist an den meisten Standorten sowohl mit Herausforderungen im regulatorischen Bereich wie Datenschutz und KI-Verordnung als auch mit technischen Herausforderungen (Verfügbarkeit, adaptive Skalierung, Zertifizierung) konfrontiert. Im Ökosystem ist daher über Partner eine **vertikale Diversifizierung nach Aufgaben** ökonomisch deutlich sinnvoller als die Replikation ähnlicher Funktion. Die Nutzenden sind durch externe Dienste eine hohe Qualität und Funktionalität gewohnt, eine Aufteilung nach Funktion gestattet es den Anbietern sich zu spezialisieren und eine hohe Qualität einzelner Services zu ermöglichen.

Ergebnisse aus Forschung und Entwicklung aller Partner und die daraus entwickelten Komponenten sollen regelmäßig in das Ökosystem integriert werden – dies wurde in KISSKI schon erfolgreich mit vielfältigen Komponenten durchgeführt, die von Mitgliedern der Community entwickelt wurden. Die Software übergebener Services müssen dann jedoch auch langfristig gewartet werden. Auch wurden Mitglieder der Community dabei unterstützt auf Grundlage der KISSKI Services eigene Services zu entwickeln und erfolgreich anzubieten.

5. Finanzierungskonzepte

Es besteht Bedarf für die Finanzierung der Hardware, aber auch für den langfristigen Betrieb der KI-Grundversorgung, hauptsächlich Energie- und Personalkosten. Für die Ermittlung des Bedarfs können wir uns auf die Nutzungsstatistiken von KISSKI stützen (Q1 2024 – Q2 2025). Wir verzeichnen dabei im Mittel

1 Anfrage / Tag und Nutzendem, perspektivisch werden mit der steigenden Nutzung bis zu 10 Anfragen / Tag und Nutzendem generiert. Für die zeitnahe Abarbeitung von LLM-Anfragen benötigen wir aktuell pro 50.000 Nutzenden

- 4 kleine, leistungsschwache Inferenzmodelle, bspw. 8B auf 1 GPU
- 2 großes, leistungsstarkes Inferenzmodell, bspw. 80B auf 8 GPUs

Dabei verzeichnen wir an Wochenenden prozentual weniger Anfragen, aber Studierende zeigen auch hier hohe Zugriffszahlen. Die nötigen Investitionen für beide Knoten zusammen betragen 250.000 €, zusätzlich werden als Betriebskosten bei aktuellem Niveau der Energiekosten ca. 7.000 € pro Jahr und Knoten nötig.

Um den Betrieb der Systeme (Administration), Wartung, Bearbeitung von Supportanfragen und Trainingsbereitstellung zu leisten, wird im Mischbetrieb pro 500.000 Nutzenden eine Person benötigt. Die relevante Weiterentwicklung der Services muss über Forschungs- und Entwicklungsprojekte erfolgen, nur die grundsätzliche Pflege/Wartung der Software und minimale Anpassungen sind enthalten (bspw. neue Modelle bereitstellen). Kosten dafür sind ca. 100.000 € pro Jahr. Eine kritische Anzahl an Personal (4 Personen) ist notwendig um alle Aufgaben grundlegend erfüllen zu können.

Hochrechnung für 3.6 Mio Nutzende aus Hochschulen (Studierende, Mitarbeitende, Administration) und 6 Jahre Betrieb mit einer Hardwaregeneration können wie folgt überschlagen werden:

3.6M Nutzende	**Bedarf**	**Kosten pro Jahr**	**Kosten für 6 Jahre**
Investition	72 - 4x GPU Knoten 72 - 16x GPU Knoten	—- (3 Mio €)	18 Mio €
Betriebskosten	~7000 € pro Knoten	1 Mio €	6 Mio €
Personalkosten	8 Personen	800.000 €	4.8 Mio €
Gesamt		4.8 Mio €	28.8 Mio €

Möglichkeiten zur Finanzierung einer KI-Grundversorgung sind a) ein individuelles Angebot an und Nutzung durch die Hochschulen, b) ein Service-Angebot über den DFN-Verein, c) Landeslösungen und d) nationale Maßnahmen.

Stand heute ist über die Academic Cloud für a) bereits ein Kostenmodell in Q1/2025 erstellt und auf den ersten CIO Workshops vorgestellt worden, b) das kostenlose KISSKI-Angebot bereits über den DFN verfügbar, die Erweiterung über das Kostenmodell wird vorbereitet.

c) Politisch sind individuelle Landeslösungen vermutlich attraktiv, allerdings bietet dies viele Nachteile für Nutzenden und würde in deutlich teureren Lösungen münden. Stattdessen ist die vertikale Funktionsaufteilung über die Grundversorgung hinaus (s.o. Ökosystem) von Partnern aus unterschiedlichen Länder deutlich attraktiver. So kann eine funktional kompetitive Lösung mit hoher Attraktivität für Nutzende erstellt werden.

d) Ein optimales Finanzierungskonzept für nationale Maßnahmen wäre die Bund-Länder-Finanzierung. Hierbei bietet sich der Königsteiner Schlüssel als Verteilungsmechanismus an, welcher auf dem Bevölkerungsanteil basiert. Beispielsweise würde Niedersachsen mit 9.4% einen Anteil von 451.000 € beitragen - bei ungefähr 225.000 Studierenden, Hochschulmitarbeitenden und Administration beträgt der Beitrag 2 € pro Jahr und Person.

Weitere Komponenten des Ökosystems (s.o.) wären on Top zur Grundfinanzierung zu finanzieren. Es ist für die Community erstrebenswert eine starkes Ökosystem mit vielen externen Beitragenden und finanzierten Partnern zu sichern, damit dieses langfristig bestehen kann.

Dieses Modell könnte ebenfalls für Schulen transferiert werden, die aktuelle Umsetzung für niedersächsische Schulen die bereits mit dem European Digital Innovation Hub for AI and Cybersecurity (DAISEC) gestartet sind wäre als Beispiel zu nennen.

6. Governance

Mit der Finanzierung der KI-Grundversorgung über die Länder könnten sowohl Betrieb und Wartung der Dienste als auch sehr begrenzt die Weiterentwicklung im Rahmung von Wartungsarbeiten. Dabei sollen Länder und Universitäten nicht nur als Geldgeber/Nutzende auftreten, sondern aktiv in Entscheidungsprozesse eingebunden werden. Dies geschieht bei KISSKI aktuell bereits über die “Chat AI Community” der Nutzenden und muss noch weiter ausgebaut werden. In dieser können Modell- und Servicewünsche direkt an die Entwickler weitergeben, sich über Vorträge aus der Community weitergebildet werden und Probleme im Kollektiv gelöst. Über das erweiterte Ökosystem könnten sich alle Beteiligten so absprechen und dies gemeinsam vorantreiben.

In der angedachten Open Governance Struktur werden Geldgeber und Nutzende in der regelmäßig tagenden Nutzendenvertretung über die Ausrichtung der Services und Nutzung der Ressourcen mitbestimmen und Lösungen für den dauerhaften Betrieb einbringen.

Appendix - Hintergründe

7. Das KI-Servicezentrum KISSKI

Die KI-Servicezentren sind seit 2022 vom BMFTR als deutschlandweite KI-Servicedienstleister für Wissenschaft, Kleine und Mittelständige Unternehmen (KMUs) und StartUps sowie öffentlicher Einrichtungen gefördert. Die GWDG ist Partner im Konsortium des KI-Servicezentrums KISSKI. Dabei stellt die GWDG als eines von zwei Rechenzentren des Konsortiums Hardware, Entwicklungs-/ Forschungsleistungen und Services im Bereich Computing und Datensicherheit bei. Dabei hat KISSKI seit 2023 Hardware und Services vorbereitet und seit Anfang 2024 diesen den Nutzenden zur Verfügung gestellt. Die Kernkompetenz von KISSKI ist die Bereitstellung sicherer und hochverfügbarer Dienstleistungen für die kritischen Infrastrukturen Medizin und Energiewirtschaft. Dabei ist KISSKI aber branchenoffen und überträgt seine Erfahrungen mit sicheren Entwicklungen auch auf Bereiche wie generative KI-Services, Personendatenverarbeitung und skalierbare Infrastrukturlösungen. KISSKI hat zu knapp 1000 unterschiedliche Anwendungen unterstützt, 3500 Support Tickets beantwortet, und hat knapp 180 Projekten aus Wissenschaft und Industrie Rechenressourcen für KI-Entwicklungen zur Verfügung gestellt.

Lösungen die in KISSKI entwickelt werden, können von der GWDG in Produktivdienste überführt und im Markt angeboten werden. Stand August 2025 versorgt die GWDG über KISSKI bereits 400 Institutionen (ca. 700.000 Nutzende) mit KI-Inferenz-Diensten. Durch die Erfahrungen aus der Servicebereitstellung für eine exponentiell wachsende Nutzendengruppe lassen sich notwendige Infrastrukturen für die flächendeckende Versorgung der deutschen Universitäten abschätzen und durch die skalierbare Natur der KI-Services auch mit vergleichsweise wenig Aufwand umsetzen. Typische Nutzungsszenarien hochverfügbarer KI-Services erfordern dabei hohen Datenschutz, hohe Verfügbarkeit und eine adaptive Skalierung basierend auf den Anfragen der Nutzenden - die Kernkompetenzen von KISSKI. Dabei zeigt sich im Betrieb hochverfügbarer KI-Services ein deutlicher Mehraufwand im Vergleich zu traditionellen, in universitären Rechenzentren gehosteten HPC Anwendungen, bei denen die Nutzenden lediglich Rechenressourcen und einen sehr begrenzten Softwarestack zur Verfügung gestellt bekommen, den Großteil der Entwicklungsleistung jedoch selbst unternehmen. Beim Betrieb eines KI-Services liegt die Wartung, Weiterentwicklung und der Betrieb notwendiger periphärer Dienste wie bspw. Identitätsmanagement auf Seiten des Betreibers und muss durch die Nutzenden oder eine Förderung finanziert werden.

KISSKI ist als Servicezentrum bis Ende 2027 durch das BMFTR finanziert. Dabei wird lediglich bis Ende 2025 voll finanziert, für 2026 und 2027 werden die Betriebskosten des Clusters nur teilweise kompensiert, so dass ab Juni 2026 eine Überführung in einen langfristigen, kostenpflichtigen Servicebetrieb durchgeführt werden muss. Dies entspricht ebenfalls einer Auflage des DLR/BMFTR für die Phase 2, dass ein Servicebetrieb für die Inferenz die Forschungsaufgaben der KISSKI-Nutzenden nicht kannibalisiert. Dies wird aktuell durch Scheduling der HPC Batch-Jobs und Scale-to-0 der Inferenzmodelle in der Nacht

zusammen mit der Verlagerung von Anwendungen ohne Zeitkritikalität (bspw. Dokumentenvorbereitung) in die Nacht ermöglicht. Auch mit aktuellem Wachstumsrate der Nutzenden ist der bestehende Inferenz-Betrieb auch durch zusätzliche, eigens beschaffte Hardware bis Mitte 2026 gesichert und die Echtzeitbereitstellung der GenKI Services stehen den Kund:innen tagsüber zur Verfügung. Mit der sukzessiven Einbindung weiterer Nutzendengruppen in der Größenordnung der Hochschulen eines Bundeslandes können wir, Verfügbarkeit von General Purpose GPUs (GPGPUs) für Neubeschaffungen vorausgesetzt, die hochverfügbaren Inferenzservices weiterhin mit guten Antwortzeiten betreiben. Dennoch, ab 01.07.2026 muss das Angebot für die Dienstnutzung der generativen KI-Dienste für den Produktivbetrieb kostenpflichtig werden.

Derzeit besteht das LLM Portfolio in KISSKIs Chat AI Service aus lokal gehosteten, für alle Nutzenden frei verfügbare, Open Source Modellen und in Europa gehosteten OpenAI Modellen im Lizenzmodel. Die OpenAI Modelle werden für einige Nutzungsszenarien von Kund:innen in der Verwaltung und den Social Sciences bevorzugt genutzt. Es besteht auch eine Nachfrage Modelle anderer kommerzieller Anbieter (bspw. Gemini von Google, Claude von Anthropic) mit höheren Datenschutzanforderungen nutzen zu können und auch bspw. spezifische Modelle für die Verarbeitung von Rechts- und medizinischen Fachtexten zur Verfügung zu stellen. In diesen Fällen ist die Möglichkeit der Eingabe sensibler Kund:innen- oder Patient:innendaten ein Hindernis bei der Nutzung kommerzieller, nicht-europäischer Anbieter. Gleichzeitig stellen medizinische, KI-gestützte Assistenzprogramme, bspw. zur Transkription oder automatischen Erstellung von Patientenbriefen oder Berichten einen leicht zu erschließenden Wachstumsmarkt für Anbieter, die die Datensicherheit garantieren können.

Es besteht auch die Nachfrage nach KI-Diensten, deren Entwicklung das aktuell zur Verfügung stehende Finanzvolumen deutlich überschreiten. Dabei steht zum einen die Nutzbarkeit im Fokus, da besonders Nutzende unterer Expertisestufen vor allem das User Interface verschiedener KI-Services vergleichen und trotz datenschutzrechtlicher Bedenken zu großen kommerziellen Anbietern tendieren. Hier bieten z.B. Assistenzprogramme, die Nutzendeneingaben vor-verarbeiten oder ein breiteres Spektrum an Eingabetypen verarbeiten können, eine deutlich komfortablere Nutzendenerfahrung, benötigen aber einen ähnlich großen Entwicklungsaufwand wie der Kern-KI-Service selbst. Zum anderen verlangen neue LLM-Generationen nach wesentlich größeren, leistungsfähigeren Rechenknoten, welche gerade erst in der Breite verfügbar werden. Daher sind kontinuierliche Investitionen nötig, um entsprechende Hardware für diese neuen Modellgenerationen vorzuhalten. Kontinuierliche Investitionen sind bei öffentlich geförderten Projekten jedoch oft nicht vorgesehen, weshalb die Hardware am Ende eines Förderzeitraums typischerweise veraltet ist.

8. Angebote bei KISSKI

KISSKI bietet allen deutschen Nutzenden aus Wissenschaft, Industrie und öffentlicher Hand niedrigschwelligen Zugang zu seinen Services. Das Portfolio umfasst dabei neben generellen Beratungsleistungen und solche spezifisch für die Branchen Medizin und Energie, die Bereitstellung von

Rechenressourcen für Prototypisierung und Produktivbetrieb, Schulungen und Consultings, Daten- und Modellkataloge, sowie Zugriff auf verschiedene generative KI-Dienste über Webbrowser und einer API-Schnittstelle (Chat AI[9], RAG[10], Voice AI[11], Image AI[12], Protein AI[13], CoCo AI[14]). Diese API-Schnittstelle wurde nach dem Vorbild der OpenAI Schnittstelle programmiert, die als einer der Marktführer von verschiedensten Anbietern unterstützt wird, weshalb KISSKIs GenKI Services inhärent auch mit mit einer langen Liste an Drittanbietern kompatibel sind. Dabei war die Bereitstellung von GenKI Diensten nicht explizit im ursprünglichen Projektantrag von 2021 aufgeführt. Der für die interne Nutzung entwickelte Chat AI Service wurden jedoch aufgrund der beständig großen Nachfrage der Nutzenden schnell veröffentlicht und kontinuierlich durch Forschung- und Entwicklungsarbeit mit weiteren Services und Funktionalitäten ausgebaut.

Dabei besteht ein Großteil des angebotenen Portfolios aus lokal gehosteten Diensten unter hohen Datenschutzanforderungen (ISO27001, Datenschutzklasse C; derzeit Audit für ISO27701, C5-Testat, Datenschutzklasse D), welche im Gegensatz zu kommerziellen Anbietern keine Nutzendendaten serverseitig speichern oder in irgendeiner Weise nachnutzen. Auch die externen, als Lizenzmodel angebotenen Services, werden in Europa gehostet und unterliegen damit strengeren Datenschutzanforderungen als die üblichen kommerziellen Services, welche in oft in den USA oder China gehostet werden und wesentlich weniger reguliert werden. Dies ist neben der Anwendung in kritischen Infrastrukturen besonders in der Forschung wichtig, wo Zwischenstufen in der Veröffentlichung neuer Publikationen oder Patente bei kommerziellen Anbietern in Datenleaks oder Trainingsdaten auftauchen und von anderen Parteien genutzt werden könnten. Aber auch in der alltäglichen Anwendung sind Anwendungsfälle, die höhere Schutzstufen erfordern, oft verbreiteter als von den Nutzenden angenommen. Schon die Verarbeitung von Personendaten in Listen oder Eingabe von urheberrechtlich Geschützten Textpassagen kann rechtliche Konsequenzen haben.

Die den GenKI Diensten von KISSKI zugrunde liegende Architektur der Bereitstellung über die Inferenzplattform selbst ist vollständig im Göttinger High-Performance Computing (HPC) und der Private Cloud angesiedelt. Dabei steht Private Cloud für eine interne Cloud Lösung, bei der die Designentscheidungen hinsichtlich Architektur, Sicherheit, und Flexibilität ganz in unserer Hand liegen. Im Gegensatz zur Nutzung von Cloud Services bei externen Anbietern können wir so die Sicherheit und Skalierbarkeit unserer Services garantieren.

Bei der Implementierung der Architektur werden die Bedürfnisse der Nutzenden auch langfristig berücksichtigt:

[9] https://chat-ai.academiccloud.de/
[10] https://docs.hpc.gwdg.de/services/arcana/index.html
[11] https://voice-ai.academiccloud.de/
[12] https://image-ai.academiccloud.de/
[13] https://protein-ai.academiccloud.de/
[14] https://docs.hpc.gwdg.de/services/coco/index.html

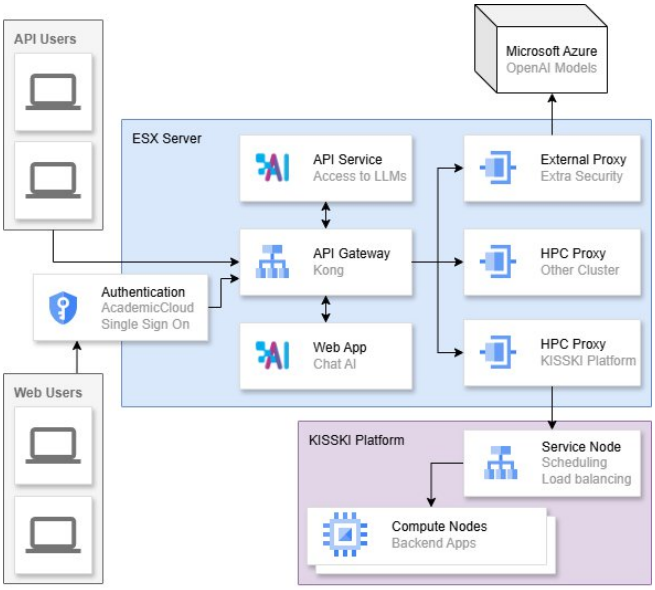


Figure 1: Aufbau der SAIA Plattform

Skalierbarkeit: Die GWDG hat mit den in KISSKI entwickelten Chat AI Services und der SAIA Plattform den Grundstein für eine skalierbare Versorgungsstruktur gelegt, die sich adaptiv an die erforderlichen Bedürfnisse der Nutzenden anpasst (Scale-out und Scale-to-0 basierend auf den beobachteten Nutzeranfragen und den Antwortzeiten).

Regulatorische Konformität: Die GWDG stellt ihre Services konform bereit, erfüllt regulatorische Auflagen wie KI-VO und befähigt Nutzende diese ebenfalls zu erfüllen. Es ist essentiell das die Dienste professionell betrieben werden und alle organisatorischen Richtlinien und Prozesse erfüllt werden.

Datenschutz: Die GWDG speichert lediglich für die Abrechnung notwendige Daten der Nutzenden. Prompts und Dateien werden in der Kommunikation verarbeitet. Hoher Datenschutz wird durch Zertifizierungen wie ISO27001 und ISO9000 ermöglicht. Die GWDG durchläuft gerade ein C5 Testat und ISO27701. Weiterhin wird die Schutzstufe der verarbeitbaren Daten bis Q4 2025 von C auf D angehoben um die Verarbeitung von personenbezogener Daten und medizinischer Daten für alle Anwendungsszenarien zu ermöglichen.

Training: Ein elementares Portfolio an Kursen für die sichere Nutzung der Dienste ist notwendig, dies umfasst sowohl Selbstlernkurse (über frei zugängliche Videokurse auf bspw. YouTube) und als Onlinekurse unter Anleitung von Trainern, welche einfach über die GWDG Academy[15] für die Nutzenden buchbar sind. Auch eigens für Kund:innen angepasste On-Site Kurse sind auf Anfrage möglich. Neben der Befähigung der Nutzenden ist es auch wichtig die lokale Multiplikatoren an den Hochschulen zu befähigen lokales Training und Support anzubieten - dies kann mittels dem “Train-the-Trainers”-Konzept durchgeführt werden - verfügbares Trainingsmaterial steht unter Open educational ressources (OER) zur Verfügung.

[15] https://academy.gwdg.de/index.xhtml

Support: Wir stellen den Nutzenden Support über ein Ticketsystem, verschiedene offen zugängliche Räume im Matrix-Chat und über die Nutzendencommunity selbst zur Verfügung. Dabei soll der Großteil der an den Hochschulen anfallende Supportanfragen über die Community und die lokalen Multiplikatoren abgewickelt werden. Beantwortung kritischer Tickets muss jedoch weiterhin über das Entwicklerteam erfolgen, wofür langfristig Mittel aus der Grundversorgung benötigt werden.

Offenes Ökosystem: KISSKI nutzt für die Umsetzung soweit möglich Offene Software um Vendor Lock-In zu vermeiden sowie offene Schnittstellen wie die OpenAI API. Die nahtlose Integration in ein größeres Ökosystem muss ermöglicht werden und ist von Anfang an mitgedacht. Dies hat außerdem den Vorteil, dass die Services mit diverser, auf diese Schnittstellen angepasste, Software direkt nutzbar ist. Wir begründeten mit der Chat AI Community Ende 2024 eine Governance-Struktur und bauen diese seitdem sukzessive aus.

9. Unsere Erfahrungen der Bedarfe aus KISSKI

Über die Nutzendencommunity, aber auch in direkten Kundengesprächen und in Gesprächskreisen, zu denen KISSKI explizit als Serviceanbieter geladen ist, haben wir Einblick in die Kund:innenwünsche und notwendigen Entwicklungen in der deutschen KI-Landschaft, insbesondere an den Hochschulen aber auch aus Industrie und Behörden. Durch die Erfahrung mit der Serviceerstellung und –bereitstellung sowie den Prognosen, die wir aus der Zusammensetzung der Nutzendenschaft ziehen, können wir weiterhin Vorraussagen über die zukünftigen Entwicklungen der KI-Nutzung treffen, beispielsweise die Zahl der zu erwartenden Nutzenden, auf welche Arten von Institutionen diese sich verteilen und auf welche Services diese vermutlich hauptsächlich zugreifen werden wollen. Aus diesen Vorhersagen haben wir für KISSKI einen Fahrplan entwickelt eine optimale Versorgung sicherzustellen, den wir versuchen auch in die entsprechenden Gremien zu tragen.

Insbesondere haben wir dieses Jahr im Mai in Gesprächen in CIO-Kreisen eine Grundversorgung definiert, welche die unmittelbaren Bedürfnisse der Hochschulen an GenKI-Dienstleistungen deckt und ausreichend Entwicklungspotential bietet. Diese haben wir auch in politische Gesprächen wie bspw. mit dem bayrischen Kultusministerium und dem niedersächsischen Schulverband weitergetragen. Dabei zeigen sich klare Unterschiede in der Servicenutzung zwischen Universitäten und KMUs/StartUps, welche auch bei der Entwicklung und Bereitstellung von Services beachtet werden muss. Während letztere überwiegend Beratungsleistungen und Rechenressourcen für Prototypisierungen anfragen, liegt bei Universitäten der Fokus klar auf der Nutzung von generativer KI, Chat AI und Retrieval Augmented Generation (RAG) Nutzungen und insbesondere auf der Einbindung via API (1000+ API Kunden) in bereits vorhandene Anwendungen wie StudIP oder Moodle. Dabei ist auch immer wieder ein Spannungsfeld zwischen bereits stattfindender KI-Nutzung der Einzelpersonen und einer angestrebten Implementierung institutionsübergreifender Lösungen ein Thema, Nutzende sollten von der privaten Nutzung kommerzieller Anbieter zu den von den Institutionen befürworteten sichereren Services wechseln. Dabei ist sind die Problematiken Urheberrecht und Datenschutz den Entscheidern in den Institutionen oft

bewusst, Nutzende vernachlässigen diese aber oft zugunsten bspw. komfortabler Servicenutzung oder zusätzlicher Funktionalitäten.

Oft ist auch die Reaktionsfähigkeit ein Thema, mit der die Universitäten auf den Bedarf an KI-Services reagieren. So haben wir oft bereits eine signifikante Anzahl Nutzende aus einer Institution, bevor die betreffende Verwaltung mit Anfragen auf uns zukommt oder Richtlinien zur Nutzung von KI-Diensten am Arbeitsplatz verabschiedet. Daher geht KISSKI pro-aktiv z.B. für den Abschluss von Auftragsdatenverarbeitungsverträgen (AV-Verträge) auf Institutionen zu, sobald wir hinreichend Nutzende aus einer Institution verzeichnen. Es wurden bereits über 100 AV-Verträge mit verschiedenen Institutionen abgeschlossen. Aus diesem Kontakt ergeben sich oft weitere Bedarfe bei den Institutionen, bspw. der Abschluss von Lizenzverträgen für die Nutzung externer OpenAI Modelle oder Consulting zur Implementierung interner RAG Lösungen, etwa ChatBots für Studierende, die Antworten der Modelle etwa aus Prüfungsordnungen und weiteren relevanten Dokumenten speisen.

Die Bedarfe unterscheiden sich auch an Universitäten zwischen den Nutzendengruppen Studierenden, Lehrende und Verwaltungsangehörigen erheblich. Während bei Studierenden der Fokus auf Literaturrecherche, Publikationszusammenfassung und Schreibunterstützung liegt, nutzen Dozierende KI vor allem zur Generierung von Lernmaterialien und Entwicklung von Services wie intelligenter Studierendenberatung und Lernplattformen. In der Verwaltung der Hochschulen sind vor allem KI-Assistenz bei Routineaufgaben benötigt. Darüber hinaus ist eine Versorgung mit weiterführenden KI-Services wie intelligente Lernplattformen und virtuelle Tutorien, Unterstützung und Beratung von Studierenden, Unterstützung bei Prüfungen/Hausarbeiten, sowie automatisierte Personal- und Finanzverwaltung bereits größtenteils mit universitären Partnern pilotiert und sollen ebenfalls von KISSKI bereitgestellt werden.

Während viele dieser Use-Cases auf den ersten Blick leicht zu implementieren scheinen oder mit kommerziellen Lösungen bereits umsetzbar sind, benötigen rechtssicher einsetzbare Lösungen hohen Datenschutz, Integration in bestehende Software wie Moodle oder StudIP sowie eine hohe Anpassungsfähigkeit an die Wünsche der einzelnen Institutionen, optimalerweise eine Möglichkeit für die Institutionen selbst die Nutzeroberflächen der Services zu verwalten. Wir setzen derzeit schon Lösungen für individuelle ChatBots und Lernunterstützungen mit verschiedenen Universitäten um.

Dabei sollte nicht unterschätzt werden wie alltäglich mittlerweile die Nutzung von KI-Services im Alltag der Lehrenden und Lernenden ist. Allerdings werden oft - in der Abwesenheit zentral angebotener, sicherer Dienste - persönliche Accounts bei Anbietern wie OpenAI und Google genutzt. Diese sind auf dem besten Weg eine Monopolstellung zu erreichen und auch aktuell gewöhnen sich Nutzer bereits – ähnlich wie bei Microsoft Office – an eine Nutzeroberfläche, bis Alternativen als inhärent weniger performant wahrgenommen werden und ein Wechsel nicht mehr in Betracht gezogen wird. Bei kommerziellen Anbietern besteht, gerade bei gratis Angeboten, große Bedenken hinsichtlich des Datenschutzes, des Urheberrechts und der Nachnutzung von Daten. So werden Kund:innendaten zum Training neuer Modelle genutzt und Eingaben der Nutzenden können später aus Modellen extrahiert

werden. Auch Datenleaks, bspw. bei DeepSeek haben in den vergangenen Jahren immer wieder Kundendaten exponiert, aber unter verschiedenen internationalen Gesetzgebungen ist auch das direkte Auslesen von Kund:innendaten durch die Firmen erlaubt.

Langfristig bestehen auch Bedenken bzgl. Vendor-Lock-In, mangelnde Anpassbarkeit, Verlust an Kompetenz und langfristig Planbarkeit von Finanzbudget. Dies kann durch nicht-optionale Integration von KI-Services in Betriebssysteme geschehen, wie es bei neueren iOS- und Windows-Systemen schon geschehen ist. Aber auch Schnittstellen, die nicht allgemein nutzbar sind, binden Kund:innen langfristig an Services. Zur Gewinnmaximierung können Anbieter zukünftig auch über die Nachnutzung der Daten hinausgehen. Freemium-Modelle, bei denen eine unbegrenzte Nutzung oder spezielle Funktionalitäten nur gegen eine Zahlung verfügbar sind, sind bereits üblich. Eine künftige Einbindung von Werbung – auch in Antworten der Modelle - ist denkbar. Durch globale Entwicklungen besteht zudem die Möglichkeit eines Versorgungsausfalls. Amerikanische Exportbeschränkungen haben Europa in der jüngsten Vergangenheit bereits betroffen, diese könnten in der Zukunft als politisches Druckmittel eingesetzt werden, sollte Deutschland nicht eine eigene Versorgung an Rechenressourcen, Modellen und Kompetenzen aufbauen.

Zusammengefasst sind kommerziell angebotene internationale KI-Services für viele Aufgaben im universitären Alltag nicht unmittelbar, oder nicht langfristig geeignet. In Deutschland gehostete, sichere KI-Services, welche nicht nur Einzelnutzenden zur Verfügung stehen, sondern auf Ebene von Institutionen/Ländern befürwortet und den Universitätsangehörigen zur Verfügung gestellt werden, können diese Lücke schließen. Weiterhin werden auf die angebotenen KI-Services zugeschnittene grundlegende Schulungsangebote benötigt, die als Video- oder Gruppenkurs bereitgestellt werden können und den Einsatz der Werkzeuge demonstriert. Weitere Schulungen werden benötigt um Nutzende über die rechtlichen Hintergründe, aber auch die IT-Mitarbeiter über die Möglichkeiten der Einbindung in bestehende Dienste zu informieren.