KI-Entwicklungen gemeinsam voranbringen: Erkenntnisse aus den Fokus-Sessions des Symposiums „Zugänge zu generativer KI schaffen – Lösungen zur technischen Bereitstellung an Hochschulen“

Claudia de Witt, Caroline Berger-Konen, Michael Hanses, Jonas Leschke, Stefan Göllner

Am 01. Juli 2024 trafen sich an der FernUniversität Hagen annähernd 100 Interessierte aus der gesamten Bundesrepublik auf dem Symposium „Zugänge zu generativer KI schaffen – Lösungen zur technischen Bereitstellung an Hochschulen zur Integration von Künstlicher Intelligenz (KI)“. Am Vormittag wurden im Rahmen von Fachvorträgen sowohl kommerzielle als auch Open-Source-Lösungen zur Bereitstellung generativer KI an Hochschulen vorgestellt. Die gezeigten Präsentationen können hier im Programm heruntergeladen werden. Dieser Blogbeitrag geht im Folgenden auf die drei Fokus-Sessions am Nachmittag zu folgenden Fragestellungen ein:

- **Fokus-Session Technologie:** Welche technologischen Entwicklungen gibt es und wie lassen sie sich umsetzen? Welche Vor- und Nachteile bieten jeweils Open-Source-basierte und kommerzielle Lösungen? Welche Entwicklungen sind noch zu erwarten?
- **Fokus-Session Hochschule:** Wo steht die einzelne Hochschule heute und welche nächsten Schritte werden in Erwägung gezogen? Welche technologischen Bedarfe bestehen im Bereich von Studium und Lehre, von Forschung und der Verwaltung? Welche Ansätze sind denkbar, um generative KI datenschutzkonform, rechtlich unbedenklich und nachhaltig umzusetzen?
- **Fokus-Session Bundesland:** Wo stehen die Bundesländer hinsichtlich der technologischen Infrastrukturen für generative KI gegenwärtig? Wie arbeiten Landes- und Hochschulpolitik für die Bereitstellung generativer KI zusammen? Welche Rolle spielt die hochschulübergreifende Zusammenarbeit? Welche nächsten Schritte sollten in Erwägung gezogen werden?

1. Fokus-Session Technologie

Michael Hanses (CATALPA/FernUni) moderierte die Fokussession „Technologie“. Hierbei tauschte sich die Community insbesondere zu den Themen „Infrastruktur“, „Software“ und „Prozesse“ aus. Diesbezüglich sind im Folgenden die Hauptaspekte bzw. -fragestellungen zusammengetragen.

Thema 1: **Infrastruktur**

Leitfragen: *Wieso werden leistungsstarke Grafikkarten (GPUs) für die technische Bereitstellung von Large Language Models (LLMs) an Hochschulen benötigt? Welche weiteren technischen Anforderungen bestehen im Hinblick auf die benötigte Infrastruktur?*

Als Basis für die Textgenerierung werden Vektordaten genutzt. Grafikkarten können Vektordaten optimal berechnen und eignen sich daher hervorragend für die Verarbeitung in großen Sprachmodellen (LLMs). Große Modelle bestehen teilweise aus bis zu Milliarden Parametern, wodurch die Verhältnisse der Vektoren zueinander äußerst komplex werden. Allerdings ist oft nicht die Rechenleistung, sondern die Speichergröße (RAM) insb. der Grafikspeicher, der limitierende Faktor. Auch die parallele Verarbeitung wird durch GPUs besonders unterstützt. Insbesondere durch Frameworks für maschinelles Lernen wie TensorFlow, PyTorch, CUDA und cuDNN.

Welche Infrastruktur konkret benötigt wird, hängt primär von der Fragestellung ab, was genau lokal zur Verfügung gestellt werden soll. Ist es nur das Sprachmodell, oder auch die begleitenden Applikationen wie Chat-Interface, API oder sonstige Eigenentwicklungen? Dabei gibt es verschiedene Ansätze, wie beispielsweise das User-Interface lokal bereitzustellen und eine kommerzielle API zu nutzen oder das gesamte LLM-Stack lokal bereitzustellen. Jede Hochschule hat dabei äußerst diverse Anforderungen, insbesondere bei den Faktoren Zeit, Skalierbarkeit und dem jeweiligen Use Cases, sodass eine pauschale Aussage hier nicht getroffen werden kann.

Ein Lösungsansatz bei Ressourcenmangel ist das "Scale to Zero"-Modell: Eine Anfrage wird erhalten, verarbeitet und das Modell danach heruntergefahren. Dies ermöglicht eine beliebige Nutzung der Hardware und eine bedarfsweise Nutzung spezifischer Modelle auf Anfrage.

Welche Erfahrungen bestehen mit mehreren kleineren GPUs innerhalb eines Servers?

Die Teilnehmer berichten, dass mehrere kleine, günstige GPUs teurere Modelle durchaus schlagen können. Durch Virtualisierung wie Kubernetes kann weiterhin für jeden Anwendungsfall eine Ressourcenzuordnung erfolgen, z. B. welche Grafikkarte, wie viel RAM und wie lange diese genutzt werden sollen. Sodass sich die verschiedenen Anwendungen nicht gegenseitig die Ressourcen blockieren.

Ein Leistungsproblem entsteht oft, wenn Daten zwischen Festplatte und Grafikkarte oder Arbeitsspeicher hin- und her geladen werden müssen. Dies kann dazu führen, dass der Prozess verlangsamt wird. Eine Lösung können spezialisierte Chips sein, auf denen alle Prozesse im Kern laufen, wie z. B. Apples System-on-a-Chip. Ein Teilnehmer berichtete von guten Performance-Erfahrungen mit Ollama auf einem Mac Studio. Ein Fazit der Teilnehmer lautet:

Kleinere Anwendungsfälle benötigen auch nur kleinere Systeme und nicht für jeden LLM-Anwendungsfall ist zwingend eine große Investition notwendig.

Inwiefern wird große Hardware benötigt?

Je nach Bedarf sind eventuell Serversysteme nicht notwendig, die Skalierbarkeit muss je nach Anwendungsfall ermittelt werden. Nicht jede Universität kann große Hardware anschaffen. Der Fokus sollte vielmehr auf die gewünschten Angebote gerichtet werden: Welche Angebote benötigen welche Lösung? Bei größeren Bedarfen bieten sich Kooperationen an, die eine Ressourcenteilung zwischen Institutionen ermöglichen. Bei kleineren Bedarfen können auch lokale Lösungen auf kleineren Geräten ausreichend sein. Trotz geringen initialen Bedarfs an Rechenleistung müssen die Systeme skalierbar gestaltet werden, um bei steigenden Anfragen mithalten zu können.

Warum wurden lokale Lösungen gegenüber Cloud-basierten Services gewählt?

Es besteht der Wunsch, die Abhängigkeit von kommerziellen Systemen anderer Anbieter zu vermeiden. API-Schnittstellen ermöglichen austauschbare Systeme (lokale Modelle vs. Abruf aus Cloud-Lösungen, Open Source vs. kommerziell). Es bestehen teilweise besondere Anforderungen an die Systeme durch institutionelle Voraussetzungen, wie z. B. besonders schützenswerte Datensätze. Lokal bedeutet jedoch nicht immer gleich sicher, da z. B. geklärt werden muss, wer lokal Zugriff auf welche Räume hat, sowie die Notwendigkeit von Patches, Updates und Netzwerksicherheit. Eine nachhaltige Vorhaltung der Daten und Modelle ist notwendig, sofern eine curriculare Einbindung gegeben ist.

Werden sich AI-MPUs (Mikroprozessoren) durchsetzen?

Ein Vorteil dieser Prozessoren ist die optimierte Matrixmultiplikation, was sie kosteneffizienter machen könnte. Aktuell gibt es jedoch noch wenig Bewegung auf dem Markt in dieser Hinsicht.

Zusammenfassend ist festzuhalten, dass bei der Wahl der „Infrastruktur“ folgende Fragestellungen zu beachten sind:

- ➔ Was ist der konkrete Anwendungsfall?
- ➔ Ist eine Skalierung kurzfristig zu erwarten oder kann zunächst mit einem kleineren System gestartet werden?
- ➔ LLMs bilden nur einen Schritt im gesamten Prozess ab, die Anwendungsfälle erfordern meist verschiedene Systeme. Nicht alle dieser Prozesse müssen physisch auf dem LLM-Server verortet sein.
- ➔ Aktuell werden oft kleinere Lösungen fokussiert. Das Bereitstellen von Systemen zum Experimentieren hat Priorität, eine Skalierung kann ggf. später vorgenommen werden.

Thema 2: Software

Leitfragen: *Welche Software wird zur Bereitstellung von LLMs genutzt? Gibt es Erfahrungen mit Betriebssystemen, Treibern, Frameworks oder Virtualisierung? Kann Software für besondere Einsatzgebiete wie Image-Generation, Music-Generation o.Ä. empfohlen werden?*

Aus dem Plenum gab es die Frage nach Erfahrungen mit Systemen, die gute Musik oder Tonfolgen generieren können. Allerdings gab es hier im Plenum kaum Erfahrungen zu. Es gab einen Austausch über die aktuellen Rechtsstreitigkeiten bezüglich Trainingsdaten und Urheberrecht, insbesondere bei Open Source Projekten.

Der Erfahrungsaustausch zu Chatbots in Moodle ergab, dass es die Möglichkeit gibt, einen Chatbot als Plug-In in das Moodle-Frontend zu integrieren. Im Marktplatz sind bereits einige Lösungen vorhanden, und einige Universitäten arbeiten an eigenen Entwicklungen.

Zu den Erfahrungen mit Retrieval-Augmented Generation (RAG) wurde festgestellt, dass mehrere Open Source-Varianten verfügbar sind. Der Retriever-Algorithmus kann jedoch qualitativ starke Unterschiede verursachen, was Optimierungsbedarf mit sich bringt. Es fehlt an Vergleichen zwischen den verschiedenen RAG-Implementierungen. Wichtig ist, dass die Benutzeroberfläche (UI) und das Backend getrennt werden. Eine bedarfsgerechte Zusammensetzung und ein gutes Schnittstellendesign sind notwendig. Aktuell funktionieren RAGs für viele große Sprachmodelle (LLMs) nicht sehr gut, weshalb modulare Lösungen benötigt werden. Zudem fehlen noch gute Evaluationstools für eine RAG-Bewertung.

Unterschiede zwischen Dokumentformaten für RAGs stellen ein weiteres Problem dar. Nicht maschinenlesbare Dokumentformate verursachen Schwierigkeiten. Während .txt-Dateien höhere Leistungen ermöglichen, geht dabei oft wichtige Information verloren, wie z. B. Bildinformationen oder bestimmte Tabellenformate.

Um mit der Dynamik in Prozessen, wie beispielsweise Verwaltungsprozessen, umzugehen, wurde der Einsatz von Webcrawlern als ein Lösungsansatz genannt. Zudem ist die Verschlagwortung von Verwaltungsvorgängen sowie das Fein-Tuning in sehr spezialisierten Bereichen, wie z. B. Domänenwissen, hilfreich.

Innerhalb der Community besteht der Wunsch nach einer standardisierten API. Die OpenAI-API wird häufig für Open Source-Ansätze übernommen oder zumindest werden OpenAI-kompatible Lösungen genutzt. Ollama ist derzeit noch nicht stabil veröffentlicht, aber es ist wahrscheinlich nur eine Frage der Zeit, bis dies der Fall ist.

Lokale Installationen erfordern regelmäßige Wartungen. Es besteht Interesse an einem Austauschforum zwischen Institutionen, um Implementierungsprozesse dauerhaft zu unterstützen. Derzeit wird überlegt, ein neues Austauschformat in Form eines Open Think Tanks am KI-ExpertLab Hochschullehre anzubieten.

Thema 3: Prozesse & Allgemeines

Leitfragen: *Welche Prozesse sind notwendig, um ein LLM in die Hochschule zu integrieren? Welche sonstigen Faktoren begünstigen oder verhindern einen schnellen Roll-Out? Platz für sonstige Themen rund um LLMs.*

Transparenz über die Ressourcennutzung ist ein wichtiger Aspekt. Dies kann beispielsweise durch eine Hardware-Überwachung erreicht werden. Innerhalb der Institution sollte die Nutzung dieser Ressourcen klar kommuniziert werden. Viele kleine Server können diese Prozesse zwar erschweren, jedoch haben kleinere Serverlösungen einen geringeren bzw. bedarfsgerechten Verbrauch. Eine Lösung für das Monitoring von Hardware-Ressourcen ist beispielsweise CheckMK.

Im Hinblick auf den Datenschutz haben Studierende Bedenken. Darüber hinaus gibt es Fragen zur Wahlfreiheit und zu weiteren offenen Herausforderungen. Ein Erfahrungsaustausch zu diesen Punkten wäre hilfreich, um die Anliegen der Studierenden besser zu verstehen und anzugehen.

Die Bereitstellung von Ressourcen sollte bedarfsorientiert erfolgen. Es wird erwartet, dass sich anbieterorientierte Lösungen und Retrieval-Augmented Generations (RAGs) durchsetzen werden. Die Entwicklung eigener Lösungen ist nach ersten Einschätzungen eher nicht skalierbar und umsetzbar.

Innerhalb der Fokussession wurden viele Fragestellungen intensiv diskutiert und dieser Blog-Beitrag gibt nur einen kleinen Ausschnitt aus den vielfältigen Diskussionen wieder. Die Teilnehmer der Fokussession hätten den Austausch sicherlich noch bis in die späten Abendstunden fortführen können. Ein Follow-Up auf der Learning-Aid wird angeregt.

Zwischenzeitlich ist auch das Paper „FernUni LLM Experimental Infrastructure (FLEXI) - Enabling Experimentation and Innovation in Higher Education Through Access to Open Large Language Modelsder“ (Zesch et al., 2024) als Preprint veröffentlicht worden und kann hier abgerufen werden: https://arxiv.org/abs/2407.13013 .

2. Fokus-Session Hochschule

Stefan Göllner (Stifterverband/KI-Campus) moderierte die Fokussession „Hochschule“. In dieser tauschte sich die Community im Rahmen eines World-Cafés zu den folgenden drei Fragestellungen aus:

Wo steht die einzelne Hochschule und welche nächsten Schritte werden in Erwägung gezogen?

Insgesamt wurde deutlich, dass die Hochschulen bereits zahlreiche Aktivitäten unternehmen, um den Einsatz von LLMs zu ermöglichen. Dies schließt konkrete Projekte, Workshops und Veranstaltungen zum Einsatz von generativen KI-Tools ein, aber auch die Einrichtung von Experimentierumgebungen, die den Studierenden zumeist kostenfrei zur Verfügung gestellt werden. Begleitend findet eine Betrachtung didaktischer Konzepte unter den Vorzeichen der

LLM-Nutzung statt. Ein Weg, studentische Projekte und Prüfungsformate grundsätzlich zu überdenken. Deutlich wurde auch: Die anwesenden Hochschulen stehen zum Teil an sehr unterschiedlichen Punkten. So reicht das Spektrum von fertigen und skalierungsfähigen Eigenentwicklungen, die bereits im Einsatz sind, über die Adaption von Best-Practices (z. B. HAWKI) bis hin zu Akteuren, die noch ganz am Anfang standen und die Veranstaltung als Anlass für eine erste Positionierung im Thema nutzen wollten.

Nächste Schritte sollten aus Sicht der Teilnehmenden konkrete Fort- und Weiterbildungen, Workshops für Lehrende oder Veranstaltungen zur Reflexion und Diskussion sein, die den Lehrenden einen Einstieg in das Thema „on the job“ ermöglichen. Praxisbeispiele und konkrete Anwendungssituationen sollten dabei im Vordergrund stehen. Für besonders wichtig halten die Teilnehmenden außerdem einen verstärkten Austausch unter den Lehrenden. Ihnen sollten konkrete Hilfestellungen zur Erprobung und Einbindung von KI-Technologien gegeben werden und entsprechende Austauschformate angeboten werden. Wichtig erscheint, Lehrende an dieser Stelle des „Ausprobierens“ abzuholen, bei der Nutzung der Tools, jedoch auch aus didaktischer, rechtlicher und datenschutzrechtlicher Sicht. Übergreifend bestand zudem der Wunsch, KI-Landesentwicklungen nutzen bzw. sich an diesen orientieren zu können.

Welche technologischen Bedarfe bestehen im Bereich von Studium und Lehre, von Forschung und der Verwaltung?

Bei dieser Fragestellung standen für die Teilnehmenden Rahmenbedingungen und Zugangsvoraussetzungen für die Nutzung von LLMs durch Studierende und Lehrende an erster Stelle. Insbesondere die Kosten sind aufgrund der vielfältigen Angebote und Anbieter und eines unübersichtlichen Marktgeschehens aktuell schwer abschätzbar. Falsche Entscheidungen können schnell hohe Kosten verursachen und stellen damit insbesondere für kleinere Hochschulen ein Risiko dar. Relevant für technologische Entscheidungen sind aber auch Datenschutzfragen, die oft ungeklärt sind und auch von den Anbietern nicht einheitlich gelöst werden. Die Teilnehmenden sehen deshalb vielfältige Weiterbildungsbedarfe, gerade im Hinblick auf technische Aspekte und Wirkungsweise der Systeme. Aber auch Schnittstellen, APIs, User Interfaces und Standardisierung sind Themenbereiche, in denen Weiterbildungsbedarf besteht. Darüber hinaus fehlen konkrete und gut dokumentierte Übersichten laufender Projekte, die Lehrenden über ihre Organisation hinaus Orientierung bieten können und Vernetzung begünstigen. Gewünscht wird deshalb eine Übersicht, aus der ersichtlich wird, welche Hochschulen KI-Systeme in welcher Form einsetzen.

In Bezug auf den AI-ACT wird das Erfordernis gesehen, Ausarbeitungen und Konkretisierungen für das Geschehen an den Hochschulen im Blick zu behalten. Nicht zuletzt die Frage nach der noch ungenauen Einstufung von KI-Tools als »kritische Infrastruktur« birgt Fragen, die letztlich aktuelle Initiativen und Überlegungen an den Hochschulen obsolet machen könnten und grundlegende Fragen zum »richtigen« Engagement aufwerfen.

Welche Rahmenbedingungen werden benötigt?

Der Fokus lag hier auf dem Wunsch nach kontinuierlichen, konkreten Austauschformaten im Hinblick auf Positionierung, Rahmenbedingungen und Organisation von Verwaltung, Lehre und Forschung. Diese werden innerhalb der Hochschulen benötigt, aber auch überregional und institutionsübergreifend. Diesbezüglich wurden zum einen der KI-Zugang als Standortfaktor, eine datenschutzkonforme Implementierung und Nutzung, aber auch über Leitlinien für einen erlaubten Einsatz diskutiert. Darüber hinaus binden Rahmenbedingungen stets finanzielle und auch personelle Ressourcen. Dezidierte Stellen, wie die, eines / einer KI-Beauftragten könnten hier zielführender sein. Zur Frage, wie solche Stellen fruchtbar ausgestaltet werden könnten, wurde erneut der institutionsübergreifende Austausch als wertvoll erachtet.

3. Fokus-Session Bundesland „Designing Infrastructures for GenAI“

Jonas Leschke (KI:edu.nrw/Ruhr-Universität Bochum) moderierte die Fokussession „Bundesland“, an der Teilnehmende aus Hamburg, Baden-Württemberg, Nordrhein-Westfalen, Brandenburg, Niedersachen, Bayern und Rheinland-Pfalz teilnahmen. Die Community tauschte sich zu folgenden vier Fragestellungen aus:

1. *Initiative – Welche landesweiten und übergreifenden Initiativen zur Bereitstellung generativer KI gibt es?*
2. *Kooperation – Welche Rolle spielt die hochschulübergreifende Zusammenarbeit?*
3. *Politik – Wie arbeiten Landes- und Hochschulpolitik für die Bereitstellung generativer KI zusammen?*
4. *Ausblick – Welche nächsten Schritte sollten in Erwägung gezogen werden?*

Initiative

Derzeit bestehen zahlreiche Initiativen aus den am Workshop teilnehmenden Bundesländern und darüber hinaus. Diese wurden im Rahmen der Kleingruppendiskussionen ohne Anspruch auf Vollständigkeit gesammelt und sind im Folgenden nach Bundesländern aufgelistet:

Baden-Württemberg

- Dach Dialogprozess: Zusammenarbeit zwischen Landes- und Hochschulpolitik
- Jupiter Hub
- BWgpt

Niedersachsen

- GWDG – Initiative
- Stack-Channel
- HAWKI / UH.HH.GPT & Uni Gießen / Slack

NRW

- KI:edu.nrw
- Open Source KI-NRW
- KI:connect.nrw
- KI-Campus-Hub NRW

Brandenburg

- Landesstrategie KI Brandenburg
- Austausch-Vernetzung: ZDT, eZB
- E-Learning-Verbund Brandenburg

Rheinland-Pfalz

- KI-Allianz RLP (Lehre und Studium, Hochschule zu Wirtschaft)
- Virtueller Campus RLP
- Rechenzentrumallianz

Thüringen

- AG KI: eTeach-Netzwerk
- Strategierat
- HS-ITZ (zentraler IT-Bereich)

Kooperation und Politik

Aus den Bereichen „Kooperation“ und „Politik“ sind bereits an einigen Hochschulen Piloten gestartet, die einen Zugang zu generativer KI über die HAWKI-Schnittstellenlösungen anbieten. Dennoch werden Anschubfinanzierungen für weitere Entwicklungen und Projekte benötigt, ebenso wie die Förderung für Lehrprojekte. Anhand von Konsortien sollte ein weiterer intensiver Austausch fortbestehen und (neue) Netzwerke gebildet werden, mit dem Ziel, bestehende Allianzen zu KI zu stärken.

Weitere Fragestellungen, die im Hinblick auf die Bereiche diskutiert wurden, lauteten wie folgt:

1. *Wo werden Abstimmungen vorgenommen? Bisher finden sie h*auptsächlich auf Landes- und Bundesebene: KMK, HRK, Bundesministerium für Bildung und Forschung etc. statt.
2. *Sind wir uns bewusst, WAS wir wirklich tun wollen bzw. welches Ziel wir verfolgen? Inwiefern ist unser Vorhaben an einem Verständnis von Hochschulbildung angedockt?*
3. *Inwiefern soll sich das Feld (generativer KI auf Länderebene) weiterentwickeln? Was ist wünschenswert?*

Ausblick

Abschließend ist festzustellen, dass die Kultusministerkonferenz (KMK), Hochschulallianzen und das Bundesministerium für Bildung und Forschung (BMBF) als Stakeholder zu betrachten sind, die in den Dialog zur Bereitstellung generativer KI an Hochschulen einbezogen werden müssen. Als bedeutsam betonten die Teilnehmenden aller Fokus-Sessions die Moderation auf Länderebene und einen fortschreitenden Dialog auf Bundesebene. Zudem bleibt allgemein zu erwarten, dass sowohl Lernende als auch Lehrende neben Open-Source-Modellen zusätzlich weiterhin kommerzielle Anwendungen nutzen (möchten).

Gerne möchten wir alle Interessierten zu einem Follow-Up auf der Learning AID 2024 am 2. und 3. September 2024 an der Ruhr-Universität Bochum einladen. Am zweiten Veranstaltungstag soll die Diskussion zur technischen Bereitstellung von generativer KI an Hochschulen fortgeführt, vertieft und konkretisiert werden. Moderiert wird die Session von Prof. Claudia de Witt und PD Dr. Malte Persike. Weitere Informationen erhalten Sie hier.