

**FAIRD: Faire und
transparente KI
für Forschung
und Lehre**

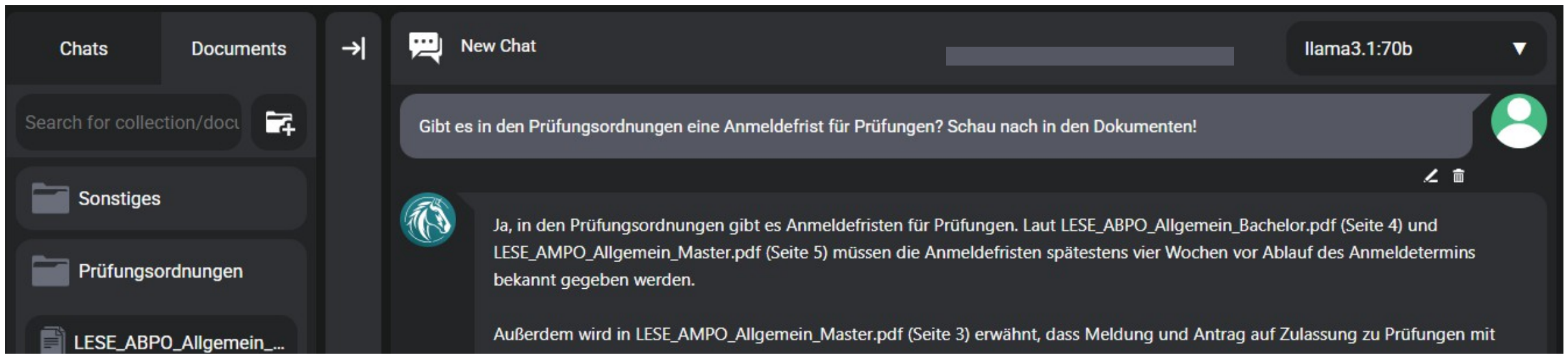


Hochschule
Kaiserslautern
University of
Applied Sciences



Was ist FAIRD?

- On-Premises Open Weights LLM mit RAG
- System zur LLM-basierten Auskunft auf Basis hinterlegter Dokumente
- Modell-agnostisch
- Selbst gehostetes LLM (ZW)



The screenshot shows a chat interface with a dark theme. On the left, there is a sidebar with 'Chats' and 'Documents' tabs. Under 'Documents', there is a search bar and a list of folders: 'Sonstiges', 'Prüfungsordnungen', and 'LESE_ABPO_Allgemein...'. The main chat area has a 'New Chat' button and a model selector set to 'llama3.1:70b'. A user message asks: 'Gibt es in den Prüfungsordnungen eine Anmeldefrist für Prüfungen? Schau nach in den Dokumenten!'. The LLM response, with a horse icon, states: 'Ja, in den Prüfungsordnungen gibt es Anmeldefristen für Prüfungen. Laut LESE_ABPO_Allgemein_Bachelor.pdf (Seite 4) und LESE_AMPO_Allgemein_Master.pdf (Seite 5) müssen die Anmeldefristen spätestens vier Wochen vor Ablauf des Anmeldetermins bekannt gegeben werden. Außerdem wird in LESE_AMPO_Allgemein_Master.pdf (Seite 3) erwähnt, dass Meldung und Antrag auf Zulassung zu Prüfungen mit'.

FAIRD-Projekt

- Fair AI R&D
- Fachbereichsübergreifende Initiative:
 - Prof. Dr. Jan Conrad (IMST)
 - Prof. Klaus Knopper (BW)
 - Prof. Dr. Eugen Staab (BW)
 - Prof. Dr. Dieter Wallach (IMST)
 - ...
- Unterstützt durch den Senatsausschuss Qualität und Lehre (SQL)
- Ziele in Strategiedokument festgehalten



HCI^{2B}
GROUP

Ziele

- Digitale Souveränität
- Finanzielle Unabhängigkeit
- Datenhoheit
- Echtsystem für Abschlussarbeiten
- Open-Source-System zur Verwendung an anderen Hochschulen



Chats

Dokumente



Neuer Chat

unsloth/Qwen3-235B

+ Neuer Chat



Bachelorarb...



Bachelorarb...

Gibt es eine Frist für die Anmeldung der Bachelorarbeit im Studiengang Wirtschaftsinformatik? Fasse dich kurz!



Ja, im Studiengang Wirtschaftsinformatik muss das Thema der Bachelorarbeit spätestens zu Beginn des Semesters, nach dem alle Prüfungen gemäß der Fachprüfungsordnung erbracht wurden, erhalten werden. Wird diese Frist um zwei Semester versäumt, gilt die Bachelorarbeit als erstmals nicht bestanden (Quelle: LESE_ABPO_Allgemein_Bachelor.pdf, Seite 12).

Prüfungsordnungen



Dokumentensuche

Schreibe eine Nachricht...



Was genau sind eigentlich CNNs?



CNN steht für Convolutional Neural Networks. Es handelt sich dabei um eine Form des Machine Learnings, die bei der Untersuchung und Identifikation visueller Daten Verwendung findet. Strukturell setzen sich CNNs aus verschiedenen Schichten zusammen, darunter Convolutional-Layer, Pooling-Layer und Fully-Connected-Layer. (Quelle: Erkennungsmethoden zur Identifizierung KI-generierter Bilder_VersionNr2.pdf, Seite 2)



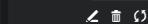
Wie genau funktionieren diese unterschiedlichen Layer?



Die Convolutional-Layer dienen dazu, bestimmte Merkmale eines Bildes zu identifizieren. Dies geschieht durch eine Faltung, bei der ein Filter über das Bild wandert und bestimmte Muster erkennt. Die Ergebnisse werden dann in einer Feature-Map zusammengefasst.

Die Pooling-Layer dienen dazu, die räumliche Auflösung der Feature-Maps zu reduzieren und somit die Anzahl der Parameter im Netzwerk zu verringern. Dies wird erreicht, indem das Bild in kleinere Bestandteile aufgeteilt und dann die wichtigsten Informationen extrahiert werden.

Die Fully-Connected-Layer dienen dazu, die Merkmale, die durch die Convolutional-Layer und Pooling-Layer identifiziert wurden, zu klassifizieren. Hierbei werden die Merkmale miteinander verbunden, um eine Vorhersage treffen zu können. (Quelle: Erkennungsmethoden zur Identifizierung KI-generierter Bilder_VersionNr2.pdf, Seite 2)



KI-Themen ▾

 Dokumentensuche

Schreibe eine Nachricht...



I. EINLEITUNG

Durch die rasante Entwicklung von bildgenerierenden KI-Modellen, wie unter anderem Dall-E von OpenAI oder Images von Google, wird es zunehmend schwieriger, echte Bilder von KI-generierten Bildern zu unterscheiden. Was auf der einen Seite eine bahnbrechende technische Errungenschaft ist, birgt auf der anderen Seite Gefahren, die es in dieser Form zuvor noch nie gegeben hat. So können frei zugängliche Image-Generatoren als mächtige Tools für böswillige Akteure dienen, die diese Technologie unter anderem zur Verbreitung von Fehlinformationen, Betrug oder Identitätsdiebstahl nutzen können. Mit dem Vormarsch der KIs und deren ständiger Weiterentwicklung wird auch die Notwendigkeit von Tools und Methoden zur Erkennung generierter Inhalte immer wichtiger. Viele dieser Tools und Methoden nutzen selbst KIs, um diese Inhalte zu erkennen und zu klassifizieren, verfolgen dabei jedoch verschiedene Herangehensweisen und Ansätze

II. CIFAKE

Jordan J. Bird und Ahmad Lotfi von der Nottingham Trent University veröffentlichten am 19. Januar 2024 einen wissenschaftlichen Artikel [1]. Dort beschreiben sie eine Studie, indem sie verschiedene tiefe neuronale Netzwerke mit Hilfe eines umfassenden eigens erstellten Datensatzes namens "CIFAKE" darauf trainieren, KI-generierte Bilder zu erkennen nach dem Prinzip "fighting fire with fire".

Horse -> herd, in a field, in the forest, outdoors, running, wildlife photography
Ship -> at sea, boat, cargo, cruise, on the water, river, sailboat, tug
Truck -> 18-wheeler, car transport, fire, garbage, heavy goods, lorry, mining, tanker, tow

1. Prompts zur Generierung (J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," in IEEE Access, vol. 12, pp. 15642-15650, 2024, doi: 10.1109/ACCESS.2024.3356122)

Vor allem war es ihnen wichtig, dass es nachvollziehbar ist, warum und durch welche Bildstellen die KI entscheidet. Dazu nutzten sie Gradient Class Activation Mapping, kurz Grad CAM, um die Aktivität der KIs durch eine Heatmap darzustellen.

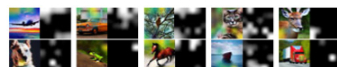
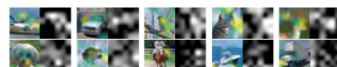
Die Forscher erreichten eine durchschnittliche Erkennungswahrscheinlichkeit von 91,79% und einen Bestwert von 93,55% bei einem Netzwerk bestehend aus 64 Neuronen.

Beim Auswerten der Heatmaps kamen sie zu einigen interessanten Schlüssen. Viele der durch KI-Generierung auftretenden Artefakten sind leicht zu erkennen, da Details fehlerhaft sind. So wurden unter anderem Flugzeuge ohne Cockpit oder anatomische Fehler bei den Tieren generiert. Bei LKWs liegt das Highlight der Heatmap auf den Kühlergrillen, da die KIs Schwierigkeiten haben, diese korrekt darzustellen.

Bei Flugzeugen, Pferden, Vögeln und Schiffen ist die Heatmap nahezu kalt. Es gibt kein bestimmtes Feature, das

sich als ausschlaggebend herausstellt. Die genutzte KI zur Generierung der falschen Bilder namens CompVis SD scheint laut Aussage der Forscher sehr gut darin zu sein, diese Klassen zu generieren.

Nach Abschluss dieser Studie stellten Jordan J. Bird und Ahmad Lotfi das CIFAKE Dataset öffentlich, sodass weitere Studien und Forschungsarbeiten diese Bilderkollektion zum Trainieren von KIs nutzen können.



2. Gradient Class Activation Maps verschiedener Klassen (J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," in IEEE Access, vol. 12, pp. 15642-15650, 2024, doi:10.1109/ACCESS.2024.3356122)

III. CONVOLUTIONAL NEURAL NETWORKS UND VISION TRANSFORMER

Eine Methode, in der das CIFAKE Dataset zum Einsatz kommt, ist die Erkennungsmethode mit sogenannten Convolutional Neural Networks (CNN) und Vision Transformers (ViT). Bei einem CNN handelt es sich um eine Form des Machine Learnings, die bei der Untersuchung und Identifikation visueller Daten Verwendung findet.

Strukturell setzen sich CNNs aus verschiedenen Schichten zusammen, darunter ein Input-Layer, ein Output-Layer und eine beliebige Anzahl an Filterschichten, sogenannten Convolutional Layers, mit denen bestimmte Merkmale eines Bildes identifiziert werden können. Der Filterprozess wird auch als Faltung bezeichnet. Alle relevanten Merkmale in einer zusätzlichen Schicht, dem Pooling-Layer, zusammengefasst.

Anders als CNNs untersuchen Vision Transformer geschieht bei einem Vision Transformer die Klassifizierung eines Bilders durch die Untersuchung einzelner Bildbestandteile und deren Beziehung zueinander. Dabei

Model	Accuracy	Precision	Recall
CNN Setup-1	96.31%	96.13%	96.50%
CNN Setup-2	95.94%	95.14%	96.83%
CNN Setup-3	95.82%	96.45%	95.15%
Bird & Lotfi [16]	92.98%	94.80%	90.90%
ViT S-16	87.09%	85.98%	88.63%

3. Leistungsvergleich von verschiedenen Modellen (M. Z. Hossain, F. Uz Zaman and M. R. Islam, "Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights", 2023 26th International Conference on Computer and Information Technology (ICCI), Cox's Bazar, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ICCI6459.2023.10440990)

In einer Studie von 2023 untersuchten Md. Zahid Hossain und Md. Rakibul Islam von der Ahsanullah University of Science and Technology, sowie Farhad Uz Zaman von der Southeast University verschiedene Variationen von CNN-Modellen in Kombination mit dem CIFAKE Dataset [2] und verglichen die Leistung dieser Modelle mit denen eines Vision Transformer Modells.

Die CNN-Modelle wurden dabei mit echten und gefälschten Bildern, jeweils 40.000 Stück, trainiert und mit jeweils 10.000 Bildern von beiden Kategorien validiert und getestet. Bei dem verwendeten Vision Transformer Modell mit der Bezeichnung ViT-S16, handelt es sich um ein vortrainiertes Modell, welches mit den ImageNet-21 und ImageNet1 Datensätzen trainiert wurde.

Dabei zeigten sich CNNs in ihrer allgemeinen Performance als recht zuverlässig, wobei das CNN-Modell mit den meisten Convolutional-Layers eine höhere Genauigkeit aufweist, auch wenn der Unterschied zu den anderen verwendeten Modellen verhältnismäßig gering ausfällt. Der signifikante Unterschied fällt bei der Leistung des Vision Transformer Modells auf, welche deutlich hinter den Ergebnissen von allen drei CNN-Modellen liegt.

IV. DIRE-MODEL

Generative Adversarial Networks, kurz "GANs", dienen als Trainingsmethode von generativen KIs. GANs setzen sich hauptsächlich aus zwei verschiedenen neuronalen Netzwerken zusammen, dem Generator und dem Diskriminator. Der Generator hat die Aufgabe, Daten zu generieren, die dann vom Diskriminator ausgewertet werden. Dabei muss der Diskriminator die generierten Daten des

Nutzungs- szenarien

(A) Allgemeine Verwendung

RAG zur Suche in eigenen Dokumenten.

(B) Vorlesungs-Coach

Interaktive KI-Unterstützung für Studierende zu Vorlesungsinhalten – Fragen stellen, Inhalte vertiefen

(C) Student-Services

Informationen zu Prüfungsordnungen, Prüfungen und anderen Verwaltungsthemen erfragen

(D) Auskunft-Chatbot

Auskunft für (internationale) Studieninteressierte zum Studium an einer bestimmten Hochschule/Studiengang.

(E) Verwaltungs-RAG

Mitarbeiter in der Verwaltung einer Hochschule können Fragen zu Leitlinien/Vorgängen/... stellen.

(E) Forschungsunterstützung

Integration mit FDM-Systemen

(G) Wahl-O-Mat für Studiengänge

KI-gestützte Studiengangsberatung: Nutzer beantworten Fragen, passende Studiengänge werden vorgeschlagen





Nutzungs- szenarien



Was ist Retrieval-Augmented Generation (RAG)?

Erweiterte LLM-Textgenerierung mit Informationsabruf

Extrahiert spezifische Daten aus verschiedenen Quellen (Datenbanken, Dateien, Live-Daten, Firmenwissen, Web Suche)

Ergänzt im Kontext die extrahierten Daten für genauere Antworten

Vorteile von RAG

Erweiterbarkeit mit
externem Wissen

Flexibilität und
Anpassungsfähigkeit

Geringere
Datenanforderungen

Kosteneffizienz

Erweiterte
Wissensabdeckung

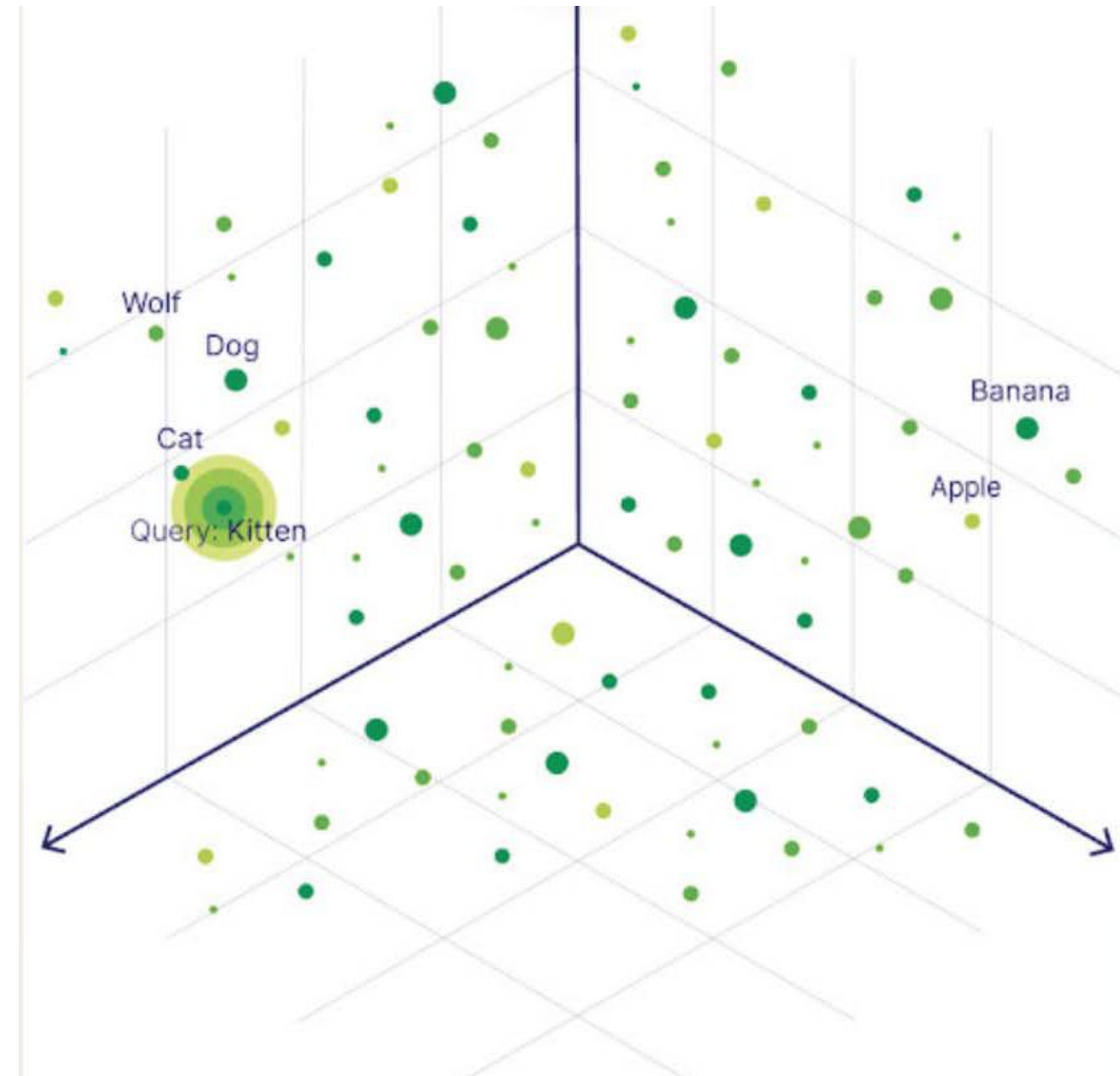
Verbesserte
Konsistenz und
Genauigkeit

Vielseitigkeit

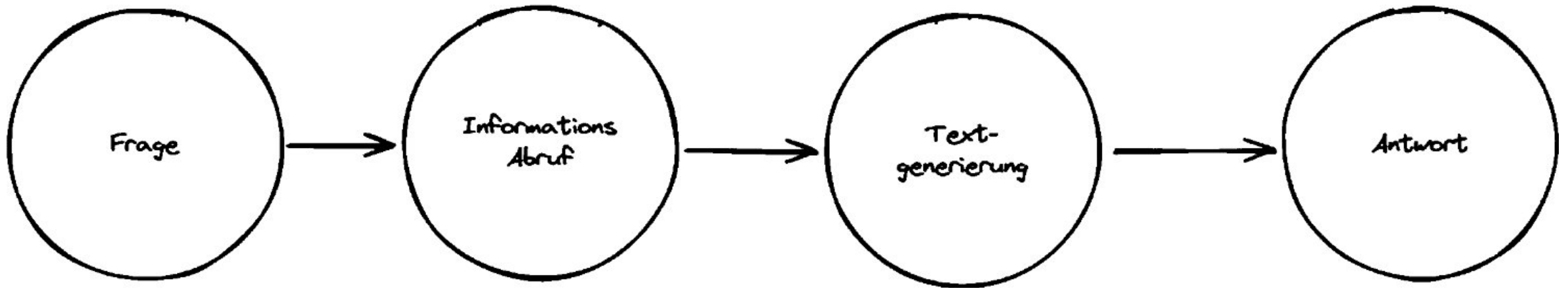
Quellenangaben

Vector Store Datenbanken

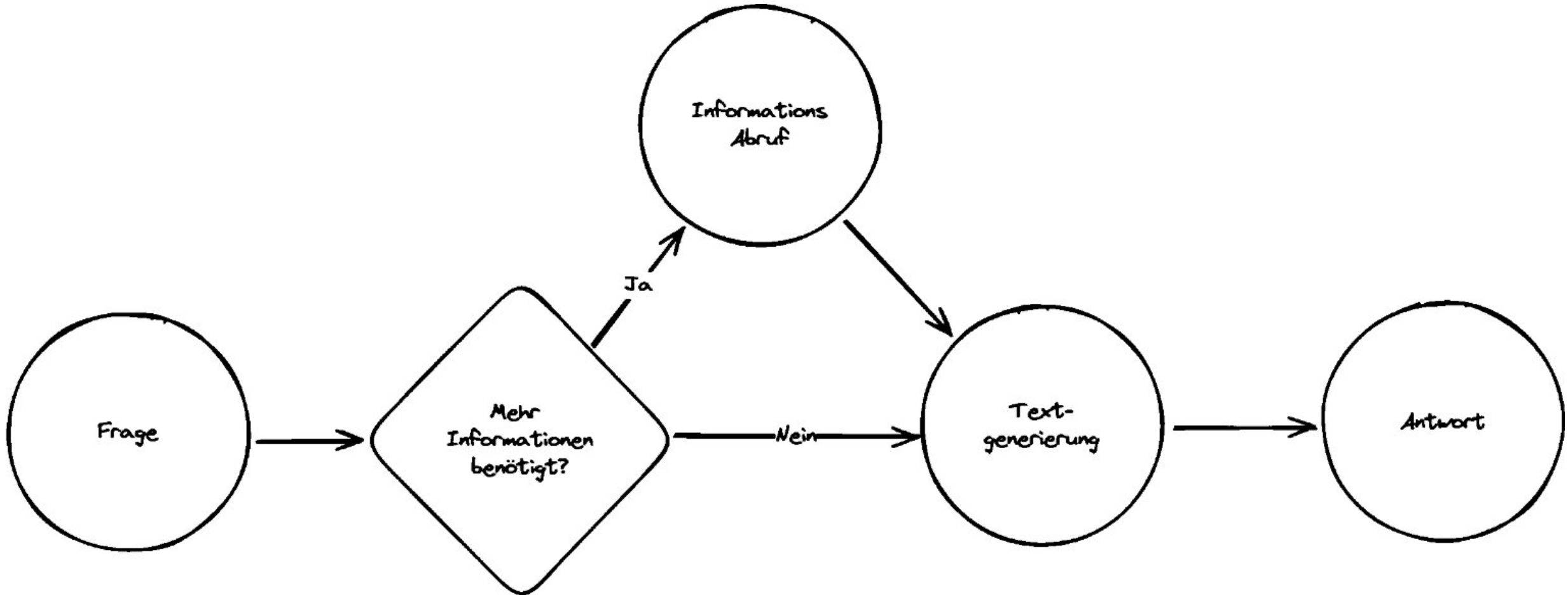
- Datenpunkte als Vektoren
- Semantische Klassifizierung
- Sehr schnelle Semantische Suche
- Skalierbar für große Datenmengen



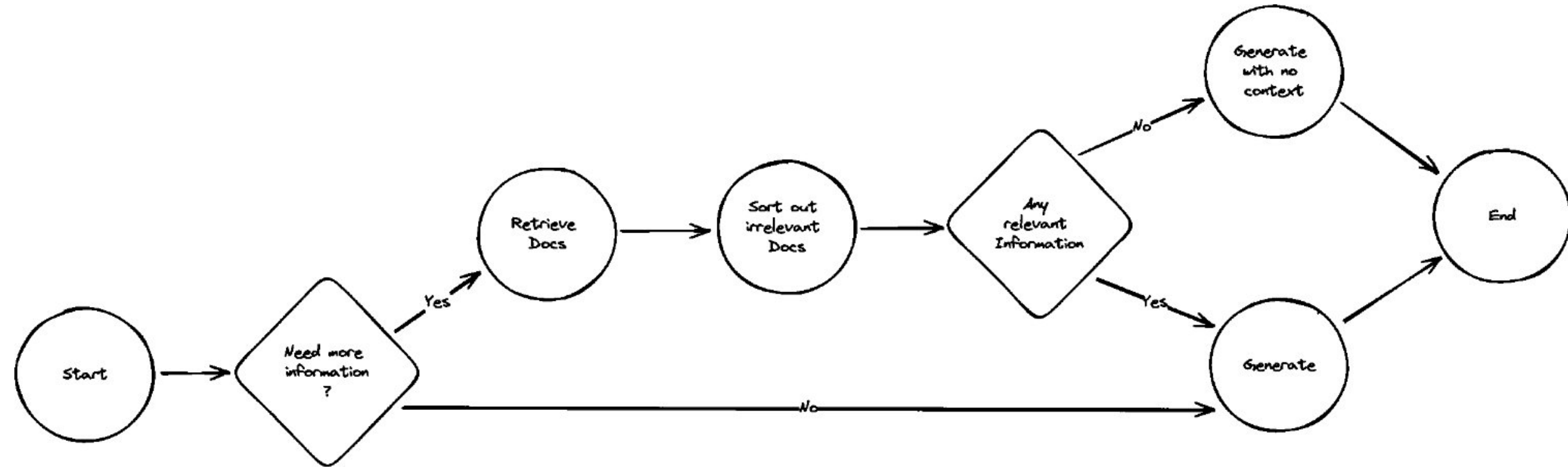
Einfaches RAG



Self Reflective RAG

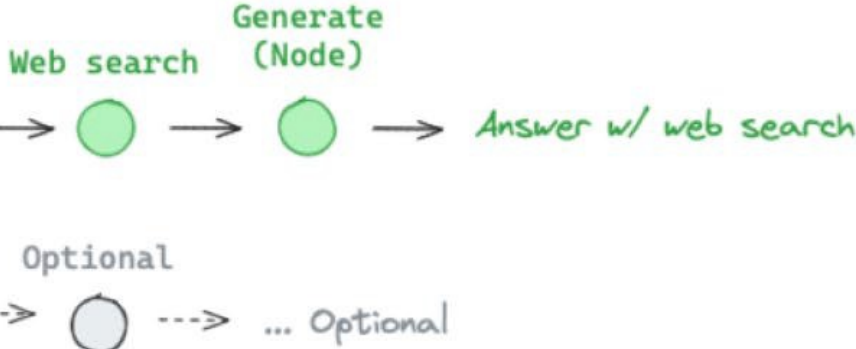
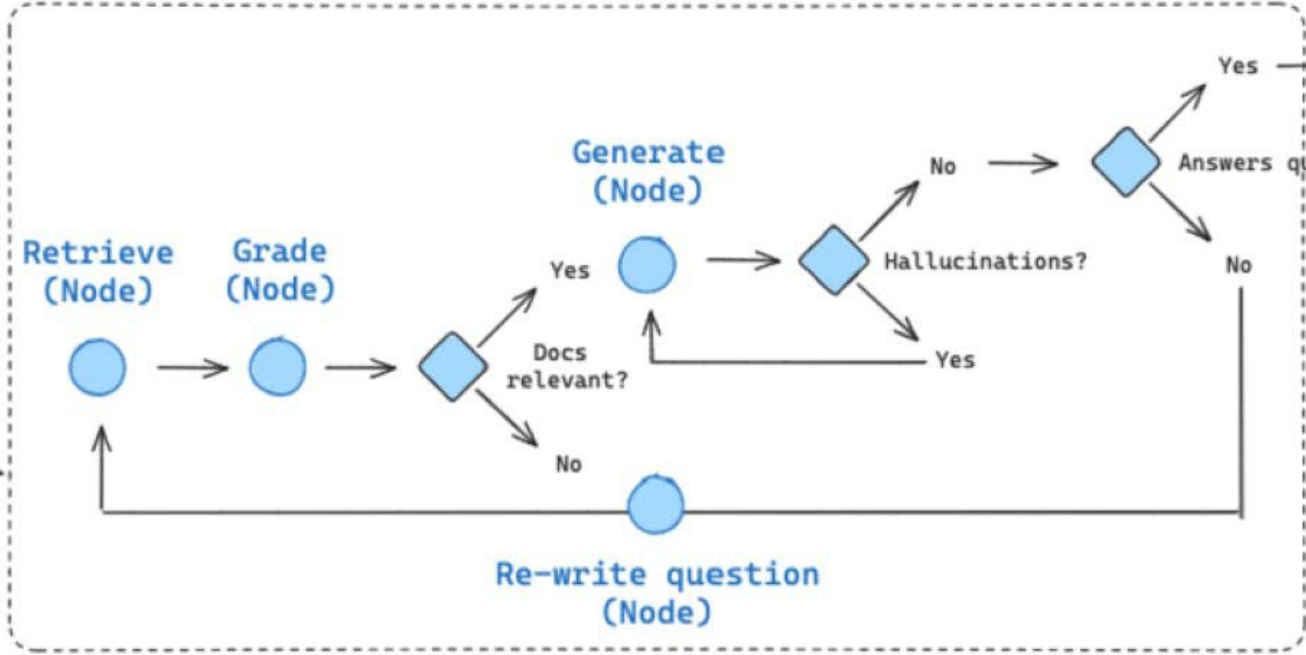
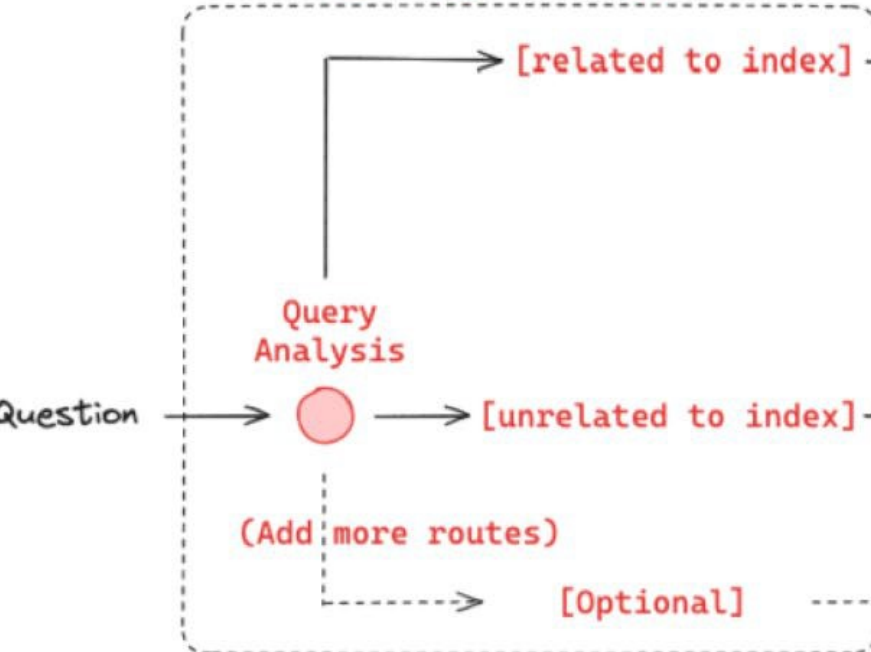


Self Reflective RAG mit Irrelevanz Check

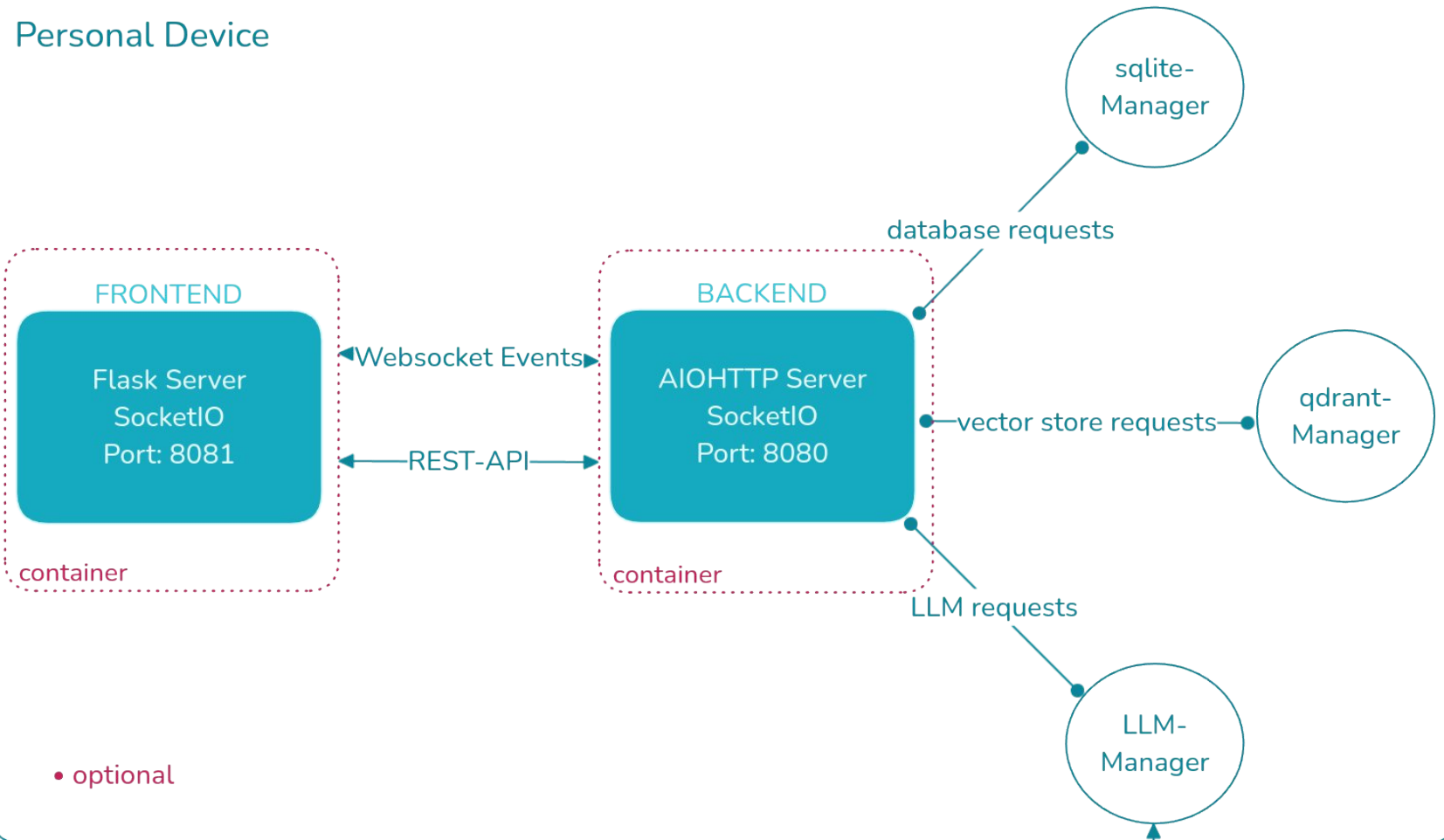


RAG + self-reflection

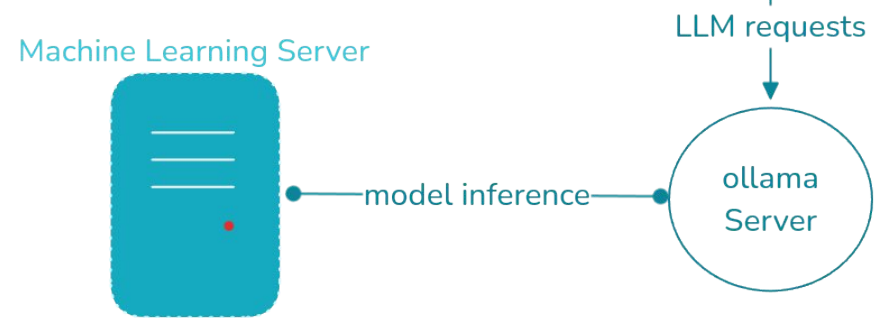
Query Analysis



Personal Device



HS-KL Intranet



- **Frontend:** Flask & Socket.IO (Port 8081)
- **Backend:** AIOHTTP & Socket.IO (Port 8080)
- **Kommunikation** via REST-API und WebSockets
- **Datenbanken:** SQLite (relational), Qdrant (Vektor-DB)
- **LLM-Verarbeitung** über externen Ollama-Server (HS-KL Intranet)
- **Containerisierung** der Dienste (Frontend/Backend) optional

Text to Speech (TTS) Speech to Text (STT)



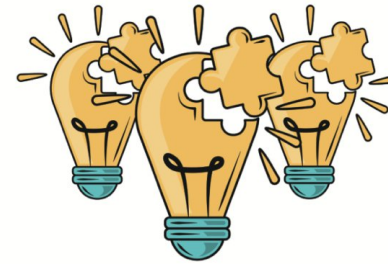
TTS STT „Joni“
Microtec Academy
Jonas Wallach

Kennst du schon die 6-3-5 Methode?

6 Teilnehmer



3 Ideen



5 Durchläufe



Jetzt testen!



offene digitalisierungsallianz pfalz



Hochschule Kaiserslautern
University of Applied Sciences

HCI^{2B} GROUP

Sophia Hertzler
Dieter Wallach

Nächste Schritte

- Open-Source-Setzung
- Entwicklung der Nutzungsszenarien:
 - Entwicklung Auskunftsbots und Einbindung in Webseite
 - Entwicklung Vorlesungs-Coach
- Teamverstärkung





FAIRD: Faire und transparente KI für Forschung und Lehre

