



KMU KI-PROGRAMM 2025 BERICHT

Von Dokumenten zu Antworten

Retrieval-Augmented Generation (RAG)
für Zürcher KMU nutzbar machen.



Kanton Zürich
Volkswirtschaftsdirektion
Amt für Wirtschaft



Swiss Data Science Center



Inhaltsverzeichnis

Inhaltsverzeichnis	2
Projektinformationen	3
Zusammenfassung	4
Geschäftliche Herausforderungen	5
In Dokumenten gefangenes Wissen	5
Sensible Daten und Anforderungen an die Datensouveränität.....	6
Wichtigste Ergebnisse	6
Ansatz	7
Warum RAG?	7
Warum KI?	7
Warum zusammenarbeiten?.....	7
Der gemeinsame Prototyp.....	8
Ergebnisse	8
Operative Einblicke: Bereits aktive Unternehmen	8
Geschäftsnutzen	10
Technische Ergebnisse	10
Was andere Zürcher KMU lernen können	11
Die Grundlage.....	11
Wichtigste Erkenntnisse	12
A. Anwendungsfall und Testdatensatz-Design	13
B. Entwicklungsstruktur: Baseline und vier Feature-Module.....	13
C. Technologie-Stack	14
D. Zentrale Design-Prinzipien	15
Open-Source-Prototyp	16
Weiterführende Informationen & Links	17



Projektinformationen

Projektbeteiligte

- Paulina Körner (Data Scientist), Ivan-Daniel Sievering (Senior Data Scientist), Thibaut Loiseau (Machine Learning Engineer) und Anna Fournier (Principal Data Scientist), Swiss Data Science Center (SDSC). Mehr Infos: datascience.ch
- Janine Plüss (CEO), Cistra. Mehr Infos: www.cistra.ch
- Marco Meola (CEO), DATABIOMIX. Mehr Infos: www.databiomix.com
- Adelene Lai (AI Integration Lead), Econetta AG. Mehr Infos: www.econetta.com
- Gosia Kubat (Managerin und Miteigentümerin) und Alma Thalmann (Managerin und Miteigentümerin), LanguageMasters GmbH. Mehr Infos: www.languagemasters.ch
- Roger Leimer (Leiter IT), Microdul AG. Mehr Infos: www.microdul.com
- Wolfgang Loerli (CEO), SOORT AG. Mehr Infos: soort.com
- Detty Berta (Verwaltungsratspräsidentin), StratoServ Sciences AG. Mehr Infos: www.stratoserv.ch
- Marco Induti (Deputy CTO), Swiss Safety Center AG. Mehr Infos: www.safetycenter.ch
- Patrik Vonlanthen (Geschäftsführer), Vonlanthen INSIGHT. Mehr Infos: www.vonlanthen.tv
- Markus Müller (Co-Leiter Standortförderung) und Raphael von Thiessen (Programmleiter KI-Standort), Kanton Zürich. Mehr Infos: zh.ch/standort
- Chantal Stäubli (CEO) und Anna Zakharova (Head Innovationsberatung), Ahead Zurich. Mehr Infos: ahead-zh.ch

Zeitraum

- September 2025 - März 2026: Zusammenarbeit im Roundtable zur Entwicklung von KI-Prototypen
- April - Mai 2026: Übergang von KI-Prototypen zu Pilotprojekten

Zusammenfassung

Neun Unternehmen aus den Bereichen Fertigung, Beratung, Health Tech, Umweltdienstleistungen, Sicherheitsprüfung und Handel nahmen am Retrieval-Augmented-Generation-Modul (RAG) des KI-Programms für KMU des Kantons Zürich teil. Ihre Branchen waren unterschiedlich, ihre Herausforderung dieselbe: Wertvolles Wissen war bereits in den Unternehmen vorhanden, in Dokumenten, Handbüchern, Berichten und Gutachten, aber zu schwer zu finden, zu langsam abrufbar und zu weit über verschiedene Quellen verteilt, um es zu verlässlichen Antworten zusammenzuführen.

In über zwei Wochen gemeinsamer Entwicklungszeit mit dem Swiss Data Science Center (SDSC) baute die Kohorte einen gemeinsamen, modularen RAG-Prototypen, verankert in einem realistischen Compliance-Szenario und erweiterbar für den jeweiligen eigenen Anwendungsfall jedes Unternehmens.

Übernahme des Prototypen: Innerhalb weniger Wochen nach Programmende gingen mehrere Unternehmen vom Prototypen in den produktiven Einsatz über.



“Die wertvollste Erkenntnis war die Bestätigung, dass KI ohne die richtigen Daten nichts ist. Wir sind mit einem konkreten Anwendungsfall gestartet, mussten aber schnell feststellen, dass die eigentliche Herausforderung nicht die Technologie selbst war. Fragen rund um Datenverfügbarkeit, Zugänglichkeit, Verantwortlichkeit und Vertraulichkeit tauchten rasch auf. Wir haben zudem gelernt, dass Ansätze wie RAG zwar bereits etabliert sind, ihre erfolgreiche Umsetzung innerhalb einer Organisation jedoch weit mehr erfordert als das Anbinden eines KI-Modells an eine Datenbank.”

— Marco Induti, Deputy CTO, Swiss Safety Center AG (mit KI übersetzt)

Die zentrale Erkenntnis über die gesamte Kohorte hinweg: Die Datenqualität, nicht das KI-Modell, war der erste Engpass, dem jedes Unternehmen begegnete. Die gesamte Pipeline zu verstehen, statt sie als Blackbox zu behandeln, ist das, was es einem Unternehmen ermöglicht, sie auch nach Programmende weiter zu verbessern. Die Zusammenarbeit über verschiedene Branchen und KI-Reifegrade hinweg führte in zwei Wochen zu fundierteren Ergebnissen, als ein einzelnes Unternehmen im gleichen Zeitraum allein erreicht hätte.

Die entwickelte RAG-Pipeline ist Open Source verfügbar und kann hier abgerufen werden:

<https://github.com/SwissDataScienceCenter/sme-kt-zh-collaboration-rag>

Geschäftliche Herausforderungen

Die neun teilnehmenden Unternehmen reichten von solchen, die am Beginn ihrer KI-Reise standen, bis hin zu solchen, die maschinelles Lernen bereits produktiv einsetzten, doch sie kamen mit demselben strukturellen Problem, nicht mit einer Technologie-Anfrage. In jedem Fall war Wissen, das bereits im Unternehmen vorhanden war, entweder zu langsam zugänglich, zu verstreut, um es zusammenzuführen, oder zu sensibel, um es externen Diensten anzuvertrauen.

- **Wissen in Dokumenten gefangen:** Über Jahre gesammelte Berichte, Handbücher und Gutachten lagen elektronisch vor, erforderten aber die Zeit von Fachleuten, um sie zu finden und zu interpretieren.
- **Datensouveränität:** Mehrere Unternehmen verarbeiten vertrauliche oder regulierte Daten und können daher gängige Cloud-KI-Werkzeuge nicht einsetzen.
- **Verlässlicher Informationsabruf:** Die aus den Dokumenten generierten Antworten müssen auf ihre Quelle zurückführbar sein, damit eine Überprüfung möglich ist und das Risiko sinkt, auf Basis unvollständiger oder fehlinterpretierter Informationen zu handeln.

Die konkreten Fachgebiete unterschieden sich, das strukturelle Muster nicht. Drei Unternehmen zeigen, wie sich das in der Praxis auswirkte.

In Dokumenten gefangenes Wissen

Die **Microdul AG** ist spezialisiert auf miniaturisierte Elektronik für medizinische Geräte und industrielle Anwendungen und operiert in einem stark regulierten Umfeld. Ingenieurinnen und Ingenieure greifen laufend auf eine grosse Basis an Design-Dokumenten, Validierungsberichten, Produktionsanweisungen und Bauteil-spezifikationen zurück. Das richtige Dokument zuverlässig zu finden und zu erkennen, wenn Informationen in einem Dokument einem anderen widersprechen, war sowohl zeitaufwendig als auch ein Compliance-Risiko.

„Um unsere reibungslose Produktion sicherzustellen, leisten unsere Ingenieurinnen und Ingenieure ein beträchtliches Volumen an anspruchsvoller und oftmals wiederkehrender Konstruktions- und Produktionsanweisungsarbeit. Dabei entsteht eine grosse Basis an wissensintensiven Dokumenten.“

— Daniel Thommen, Leiter Module Engineering, Microdul AG (mit KI übersetzt)

Und ergänzend dazu:

„[...] Die erste Herausforderung besteht darin, die richtigen Informationen in dieser grossen Informationsbasis zu finden und abzurufen. Die zweite ist die automatisierte Analyse und Erkennung von Inkonsistenzen, die dabei helfen würde, wiederholte Arbeit zu reduzieren und ein hohes Qualitätsniveau durchzusetzen.“

— Roger Leimer, Leiter IT, Microdul AG (mit KI übersetzt)

Die **Econetta AG**, eine unabhängige Umweltberatungsfirma, stand vor einer anderen Variante desselben Problems. Projektteams brauchten Zugang zu regulatorischem Wissen, Projekthistorie und technischen Analysen, die über isolierte, überwiegend unstrukturierte Dateien verteilt waren. Der Zugriff auf dieses Wissen kostete Zeit, die eigentlich der Arbeit am Kundenprojekt gehören sollte, und die Fähigkeit, bestimmte Fragen zu beantworten, hing zu stark von einzelnen Teammitgliedern ab, die dieses Wissen informell besaßen.

Die **Swiss Safety Center AG**, eine akkreditierte Prüf- und Zertifizierungsstelle, sammelt im täglichen Betrieb grosse Mengen an Prüfdaten. Über die reine Informationsbeschaffung hinaus sah das Unternehmen einen potenziellen Mehrwert darin, Muster und Zusammenhänge zwischen Fällen zu erkennen, die eine manuelle Durchsicht übersehen könnte, etwa Risikosignale, die zu sichereren Ergebnissen beitragen könnten. Es fehlte jedoch die Infrastruktur, um diese systematisch sichtbar zu machen.



Sensible Daten und Anforderungen an die Datensouveränität

Mehrere Unternehmen verarbeiteten Daten, die ihre eigene Infrastruktur nicht verlassen dürfen. **Vonlanthen INSIGHT**, eine Schweizer Boutique-Beratung, arbeitet mit vertraulichen Kundendokumenten. Die Anforderung lautete: ein datenschutzkonformes RAG-System, das vollständig auf lokaler Hardware läuft, ohne dass Daten an externe Anbieter übertragen werden, und das dennoch Antworten liefert, die zuverlässig genug für den professionellen Einsatz beim Kunden sind.

“[...] Ich brauchte einen Weg, strukturierte Erkenntnisse aus grossen, heterogenen Dokumentenbeständen zu gewinnen – ohne proprietäre Daten an Cloud-Anbieter weiterzugeben. Die Herausforderung war, ein datenschutzkonformes RAG-System zu bauen, das vollständig auf lokaler Infrastruktur läuft, reale Dokumentenkomplexität bewältigt und Antworten liefert, die in professionellen Kundenprojekten verlässlich einsetzbar sind.”

— Patrik Vonlanthen, Vonlanthen INSIGHT

DATABIOMIX, ein Health-Tech-Unternehmen, das eine KI-Infrastruktur für die Darmmikrobiom-Analyse entwickelt, stand vor einer Herausforderung in der Grössenordnung: Tausende heterogener wissenschaftlicher Publikationen mussten zu evidenzbasierten Ernährungsempfehlungen mit Quellenangaben zusammengeführt werden, denen Klinikerinnen, Kliniker und Ernährungsfachleute in ihrer Praxis vertrauen und die sie überprüfen können.

Die übrigen Unternehmen der Kohorte, **Cisra**, **LanguageMasters GmbH**, **SOORT AG** und **StratoServ Sciences AG**, standen vor ähnlichen Herausforderungen bezüglich Wissenszugriff und Vertrauen. Ihre konkreten Anwendungsfälle werden hier jedoch nicht im Detail beschrieben.

Wichtigste Ergebnisse

9 Unternehmen bauten einen gemeinsamen RAG-Prototypen, der Open-Source frei verfügbar ist für jedes Schweizer KMU.	5 Unternehmen setzen den Prototyp innerhalb weniger Wochen nach dem Programm aktiv ein oder erweitern ihn.	2 produktive Kundenlösungen wurden von Vonlanthen INSIGHT auf Basis des Prototyps lanciert.
---	---	--

Ansatz

Warum RAG?

RAG ist ein Ansatz für ein konkretes Ziel: bestehendes Wissen, belegt mit Quellenangaben, über natürliche Sprache zugänglich zu machen. Anders als ein allgemeiner KI-Assistent, der auf vortrainiertes Wissen zurückgreift, ruft ein RAG-System Informationen aus den eigenen Dokumenten des Unternehmens ab und stützt Antworten auf das, was dort tatsächlich steht.

Das machte RAG aus drei Gründen zur richtigen Wahl für diese Unternehmen:

- Das Wissen war bereits im Unternehmen vorhanden. Ziel war der Zugriff, nicht die Generierung neuer Inhalte.
- Quellenangaben sind entscheidend. In der Fertigungs-Compliance, der Sicherheitsprüfung, der Beratung und im Gesundheitswesen ist eine Antwort ohne überprüfbare Quelle nicht handlungsfähig und birgt professionelle Risiken.
- Ein ehrliches „Ich weiss es nicht“ ist sicherer als eine selbstbewusste falsche Antwort. Ein System, das Antworten erfindet, ist gefährlicher als eines, das die Grenzen seiner Evidenz anerkennt. Die RAG-Architektur kann gezielt so gestaltet werden, dass sie die zugrundeliegenden Quellen angibt und fehlende Evidenz erkennbar macht.

Warum KI?

Der Umfang und die Heterogenität der Dokumente hatten in diesen Unternehmen bereits das übertroffen, was sich manuell praktikabel verwalten lässt. Klassische Stichwortsuche liefert Dokumente, keine Antworten. KI-gestützte Informationsbeschaffung liefert fundierte, mit Quellen belegte Antworten auf Fragen in natürlicher Sprache und verkürzt so den Weg von einer Frage zu einer Entscheidung. Für Unternehmen mit tausenden Dokumenten ist das qualitativ etwas anderes als alles bisher Verfügbare.



“Aufschlussreich war zu sehen, wie KMU aus völlig unterschiedlichen Branchen ihre KI-Herausforderungen formulierten. Die grundlegenden Muster – Datenqualität, Change Management und vor allem: Vertrauen – sind universell und gänzlich unabhängig von der Branche.”

— Patrik Vonlanthen, Vonlanthen INSIGHT

Warum zusammenarbeiten?

Neun Unternehmen arbeiteten an einer gemeinsamen Codebasis, statt jedes für sich zu entwickeln. Moderne Entwicklungswerkzeuge ermöglichten eine effiziente Zusammenarbeit, sodass die Stärken jedes Unternehmens über das gesamte Programm hinweg in eine solide gemeinsame Lösung einfließen konnten.

Unternehmen in früheren Phasen der KI-Adoption sahen aus erster Hand, was produktionsreife KI tatsächlich verlangt: nicht nur einen funktionierenden Prototypen, sondern Daten-Governance, Evaluations-Infrastruktur und einen Plan für den laufenden Unterhalt. Im Austausch mit ihnen mussten erfahrenere Unternehmen Design-Entscheidungen artikulieren, die sie zuvor als selbstverständlich vorausgesetzt hatten.

Der gemeinsame Prototyp stiess ausserdem auf Fehlerfälle, die kein einzelnes Unternehmen allein hätte aufdecken können. Ein System, das für regulatorische Compliance-Dokumente entworfen wurde, verhielt sich sehr unterschiedlich, wenn es gegen wissenschaftliche Literatur oder Lieferanten-Selbstdeklarationen getestet wurde, und diese Unterschiede führten zu Design-Verbesserungen, von denen alle profitierten.

Der gemeinsame Prototyp

Der Prototyp war in einem konkreten Nachhaltigkeits-Compliance-Szenario für ein fiktives Verpackungsunternehmen verankert, das Fragen zu Lieferantenaussagen, Umweltzertifizierungen und Beschaffungsrichtlinien beantwortete. Damit erhielt jede Design-Entscheidung einen testbaren, realistischen Kontext. Die Architektur deckte die vollständige RAG-Pipeline ab: Dokumentaufnahme und -indexierung, Retrieval, Antwortgenerierung sowie eine funktionierende Chat-Oberfläche mit Quellenangaben. Der fiktive Kontext mit realen Dokumenten ermöglichte eine reibungslose Zusammenarbeit und ein gemeinsames Zielbild unter den Teilnehmenden, während jedes Unternehmen dennoch seinen eigenen Blickwinkel verfolgen konnte.

Die Entwicklung war um eine Baseline-Pipeline und vier weiterführende Feature-Module strukturiert: Evaluation, strukturierte Outputs, fortgeschrittenes Retrieval und agentische Reasoning-Ansätze. Damit konnten neun Unternehmen mit unterschiedlichen Prioritäten gleichzeitig zur selben Architektur beitragen und von ihr profitieren. Der gesamte Code ist Open Source, wird vom SDSC öffentlich auf [GitHub gepflegt und steht jedem Schweizer KMU zur freien Nutzung und Anpassung zur Verfügung](#).

Ergebnisse

Operative Einblicke: Bereits aktive Unternehmen

Das deutlichste Signal aus dem Programm war das Tempo, mit dem die Unternehmen vom Prototypen zur Anwendung übergingen. Mehrere begannen bereits innerhalb weniger Wochen nach den gemeinsamen Sitzungen, ohne auf zusätzliche Unterstützung zu warten.

Vonlanthen INSIGHT kam am weitesten. Innerhalb weniger Monate wurde aus dem SDSC-Prototyp TRAG, ein Open-Source-RAG-Framework für systematische Parametertests und den produktiven Einsatz. Zwei Kundenprojekte laufen inzwischen darauf: eines für den dialogorientierten Zugriff auf persönliche Dokumentarchive und eines für die On-Premise-Dokumentindexierung für den öffentlichen Sektor und KMU-Kunden, wo die lokale Verarbeitung vollständige Datensouveränität gewährleistet.



“Der SDSC-Prototyp aus dem Programm ist zu TRAG geworden (github.com/voninsight/TRAG). Ein RAG-Framework für systematisches Parameter-Testing und Produktivbetrieb. Zwei meiner aktuellen Projekte bauen darauf auf: ONE, ein experimentelles Wissens-Cockpit für den konversationellen Zugang zum eigenen Dokumentenuniversum; und Archivio, ein On-Premise-System zur Indexierung und Abfrage von Dokumentenbeständen – wahlweise mit lokalen Modellen für volle Datensouveränität oder falls zulässig auch mit Cloud-Anbindung.”

— Patrik Vonlanthen, Vonlanthen INSIGHT

Die **Microdul AG** hat das Programm-Repository unmittelbar nach Ende der Sitzungen geforkt. Zwei Mitarbeitende des Unternehmens erweiterten es um produktionsorientierte Funktionen, Unterstützung für zusätzliche Modell-Backends und ein einfaches Rollenmanagement. Der Rollout in Kombination mit einem internen Sprachmodell ist inzwischen im Gang.

„Nachdem die App bereit ist, läuft nun die eigentliche Implementierung. Der Betrieb eines internen LLM in Kombination mit dem RAG-System eröffnet weiteres Innovationspotenzial [...] Über standardisierte APIs lässt sich das LLM entweder eigenständig über eine Chat-Oberfläche nutzen oder in andere digitale Prozesse einbinden.“

— Roger Leimer, Leiter IT, Microdul AG (mit KI übersetzt)

DATABIOMIX integriert die RAG-Schicht direkt in seine Webanwendung, sodass Ernährungsfachleute Darmmikrobiom-Profile in evidenzbasierte, mit Quellen belegte Ernährungsempfehlungen übersetzen können.

Die **Swiss Safety Center AG** erkannte während des Programms, dass eine erfolgreiche Umsetzung interne Verantwortlichkeit voraussetzt. Das Unternehmen stellte eine Data Scientist ein und entwirft nun eine vollständige Retrieval-Pipeline für die Integration mit seiner internen Prüf-Wissensbasis.



„Das Programm hat uns geholfen zu verstehen, was technisch möglich und was erforderlich ist, damit es funktioniert. Als nächstes übertragen wir die Erkenntnisse in unsere eigene Umgebung. Derzeit stehen wir vor typischen Herausforderungen bezüglich Integration, Governance und begrenzten KI-Ressourcen. Unser Ziel ist, einen funktionsfähigen Agenten einzusetzen, der mit unserer internen Wissensdatenbank verbunden ist. Darauf aufbauend wollen wir die von uns entworfene vollständige Retrieval-Pipeline implementieren und sie stabil, messbar und wartbar machen, bevor wir sie mit echten Nutzern und Anwendungsfällen validieren.“

— Marco Induti, Deputy CTO, Swiss Safety Center AG
(mit KI übersetzt)

Die **Econetta AG** arbeitet an mehreren RAG-Anwendungsfällen und kombiniert interne Memos und regulatorische Empfehlungen mit europäischen Gesetzgebungen. Das Unternehmen baut Pipelines auf, um externe öffentliche Daten aufzunehmen und regulatorische Entwicklungen zu verfolgen.

„Dieses Programm hat uns einen greifbaren Ausgangspunkt gegeben, um KI als Lösung einzubinden, sowohl inhaltlich als auch praktisch. Inhaltlich gaben uns die Entmystifizierungs-Sitzungen ein solides Verständnis von RAG auf angewandter Ebene, praktisch konnten wir dank der MIT-lizenzierten RAG-Codebasis, die das SDSC bereitgestellt hat, sofort mit dem Prototyping starten. Verglichen mit anderen offen verfügbaren RAG-Lösungen, die oft eine ‚Blackbox‘ sind, ist die SDSC-Lösung offen, modular und wird mit Lernmaterialien geliefert, die uns helfen, Fehlerfälle nachzuvollziehen.“

— Adelene Lai, Econetta AG

Geschäftsnutzen

Über die gesamte Kohorte hinweg lieferte das Programm geschäftlichen Nutzen in drei konkreten Ausprägungen:

- **Beschleunigter Kompetenzaufbau:** In nur zwei Wochen Prototyping hatten die Unternehmen einen validierten Prototypen verfügbar und das grundlegende Verständnis erlangt, ihn zu erweitern – Arbeit, für die in anderen Fällen Monate an Ausprobieren nötig gewesen wären.
- **Kostenfreie Infrastruktur:** Der Open-Source-Prototyp kann von jedem Schweizer KMU ohne Lizenzkosten eingesetzt werden und senkt damit die finanzielle Einstiegshürde für die KI-Adoption unmittelbar.
- **Schnellere interne Akzeptanz:** Ein funktionierender Prototyp und das Verständnis, wie er entstanden ist, gab den Unternehmen den Nachweis und das Vokabular, um intern die Zustimmung für weitere KI-Investitionen zu sichern.

Die Unternehmen, die das Programm abgeschlossen haben, verfügen über alles, was für einen echten geschäftlichen Nutzen nötig ist: einen funktionierenden Prototypen, den sie selbst gebaut haben und von innen heraus verstehen. Ausserdem sind die organisatorischen Weichenstellungen nun bereits erfolgt, mit neuen Einstellungen, neuen Rollen und aktiven Umsetzungen. Die Frage ist nicht mehr, ob KI in ihrem Kontext funktionieren kann, sondern wie schnell die nun aufgebauten Systeme sich in gewonnene Zeit, verbesserte Entscheidungen und Wissen umsetzen lassen, das zuvor in Dokumenten verschlossen war und nun tatsächlich zugänglich wird.



„Das eigentliche Ergebnis dieses Programms ist nicht der Prototyp allein, sondern das, was die Unternehmen daneben aufgebaut haben: klarere KI-Roadmaps, Budgets für KI-Initiativen und Personen, die in KI-Rollen eingestellt wurden. Mehrere Unternehmen haben bereits weitere Entwicklungen bestätigt, und deshalb ist die Dynamik nicht mit dem Abschluss der ersten Runde des KI-Programms für KMU zum Stillstand gekommen.“

— Anna Fournier, Swiss Data Science Center

Technische Ergebnisse

Drei Design-Erkenntnisse erwiesen sich in allen Anwendungsfällen als durchgängig wichtig. Die vollständigen technischen Details finden sich im Technischen Anhang.

- Die Retrieval-Qualität wird nicht von einer einzelnen Entscheidung bestimmt. Wie Dokumente aufgeteilt werden, wie sie indexiert werden und wie Abfragen formuliert werden, jeder Schritt trägt zur endgültigen Antwortqualität bei. Kein einzelner Schritt dominiert, stattdessen wirken alle Schritte kumulativ.
- Ohne eine Möglichkeit, die Qualität zu messen, ist schwer zu erkennen, welche Änderungen helfen und welche nicht. Einen Evaluationsschritt früh einzubauen, bevor weitere Funktionen aufgeschichtet werden, macht jede Verbesserung nachvollziehbar und verhindert, dass sich Probleme unbemerkt anhäufen.
- Der Workshop testete semantische Suche, Stichwortsuche und hybride Kombinationen. Hybrides Retrieval ist so ausgelegt, dass es sowohl mit unterschiedlichem Vokabular als auch mit exakten Suchbegriffen umgehen kann, was es über verschiedene Abfragetypen hinweg konsistenter macht, wenngleich die Ergebnisse je nach Anwendungsfall variieren.

Was andere Zürcher KMU lernen können

Die Grundlage

Ein KI-System ist nur so nützlich wie die Informationen, auf die es zurückgreift. Dokumente, die gescannt statt digital vorliegen, uneinheitlich strukturiert, veraltet oder in inkompatiblen Formaten gespeichert sind, erfordern deutlich mehr Vorbereitungsarbeit, bevor sie ein KI-System zuverlässig nutzen kann. Die Qualität Ihrer Daten bestimmt, wie viel Aufwand zwischen Ihnen und einem funktionierenden System steht. Das ist kein Grund zu zögern. Es ist ein Grund, eine Bewertung der Datenqualität als expliziten ersten Schritt jedes KI-Projekts vorzusehen, noch vor jeder Anbieterauswahl und jeder technischen Arbeit.

Die organisatorische Bereitschaft ist ebenso wichtig: Ein funktionierender Prototyp wird in dem Moment zur Belastung, in dem niemand die Verantwortung für die Dokumente, die Zugriffsrechte oder die Freigabe der Ergebnisse übernimmt. Die Entscheidung der Swiss Safety Center AG, vor dem produktiven Einsatz einen Data Scientist einzustellen, spiegelt genau das wider: Eine technische Lösung braucht einen technischen Verantwortlichen.

Die erste Umsetzung ist ein Ausgangspunkt. Die Retrieval-Qualität verbessert sich mit dem Bereinigen der Dokumente, dem Verfeinern der Prompts und dem Bearbeiten von Grenzfällen. Planen Sie Iteration, nicht Installation. Unternehmen, die KI als eine sich entwickelnde Fähigkeit behandelten, kamen durchweg weiter als solche, die nach einer einmaligen Lösung suchten. Auch das Feld selbst bestärkt dies: Modelle, Werkzeuge und Retrieval-Techniken entwickeln sich schnell weiter, und eine Organisation, die engagiert bleibt, ist besser positioniert, um von diesen Fortschritten zu profitieren, als eine, die das Projekt als abgeschlossen betrachtet.

Ist RAG relevant für Ihr Unternehmen? Beginnen Sie mit diesen Fragen.

Zu Ihrem Wissen und Ihren Dokumenten:

- ⇒ Wo verbringen Ihre Mitarbeitenden die meiste Zeit mit der Suche nach Informationen?
- ⇒ Welche Entscheidungen erfordern das Zusammenführen mehrerer Dokumente, bevor gehandelt werden kann?
- ⇒ Hängt kritisches Wissen an einzelnen Personen, und was geschieht, wenn diese nicht verfügbar sind oder das Unternehmen verlassen?

Zu Datenschutz und Daten:

- ⇒ Welche Daten können Sie mit einem Cloud-KI-Anbieter teilen, und welche müssen On-Premise bleiben?
- ⇒ Gibt es regulatorische oder vertragliche Einschränkungen, wie Ihre Dokumente verarbeitet werden dürfen?
- ⇒ Wer besitzt die Dokumente und hat die Befugnis, einer KI Zugriff darauf zu gewähren?

Zur Bereitschaft:

- ⇒ Wie einheitlich, aktuell und gut strukturiert sind Ihre internen Dokumente?
- ⇒ Wer in Ihrer Organisation würde ein KI-Wissenssystem verantworten und unterhalten?
- ⇒ Können Sie definieren, wie eine gute Antwort aussieht, bevor Sie das System einsetzen?



„Was in dieser Kohorte heraussticht, ist, dass neun Unternehmen aus unterschiedlichen Branchen und in unterschiedlichen Stadien der KI-Reife kollaborativ eine Lösung entwickelt haben, die umfassender ist als alles, was ein einzelnes KMU allein hätte schaffen können.“

— Raphael von Thiessen, Kanton Zürich

Wichtigste Erkenntnisse

- **Beginnen Sie mit dem Geschäftsproblem, nicht mit der Technologie.** Jedes Unternehmen, das sinnvolle Fortschritte gemacht hat, begann damit, eine konkrete operative Schwierigkeit zu benennen, nicht eine Technologie-Ambition. „Wo verlieren wir Zeit bei der Dokumentensuche?“ und „Wo kosten uns fehlende Informationen am meisten?“ sind produktivere Fragen als „Wie können wir KI einsetzen?“ Die Wahl der Technologie folgt aus der Antwort, nicht umgekehrt.
- **Definieren Sie Erfolgskriterien, bevor Sie beginnen.** Das Problem zu benennen ist nicht dasselbe wie zu wissen, ob man es gelöst hat. Die Unternehmen, die am schnellsten in die Pilotphase kamen, hatten ein klares Bild davon, mit welchen Dokumenten das System arbeiten musste. Diese Klarheit mit einer messbaren Ausgangsbasis zu verbinden, macht die Argumentation für weitere Investitionen deutlich stärker. Diese Ausgangsbasis kann einfach sein: Wie lange brauchen Mitarbeitende heute, um eine typische Anfrage zu beantworten, oder wie häufig treten Fehler auf? Vor der Umsetzung erfasst, gibt Ihnen eine solche Zahl später etwas Konkretes, um Vergleiche anzustellen. Mehrere Unternehmen sagten, dass sie diesen Messansatz in einem zukünftigen Projekt sogar noch früher aufsetzen würden.
- **Die Datenaufbereitung ist der Ort, an dem der eigentliche Aufwand liegt und an dem er sich auszahlt.** Über die gesamte Kohorte hinweg brauchte die Arbeit des Bereinigens, Strukturierens und Vereinheitlichens von Dokumenten mehr Zeit als erwartet und prägte unmittelbar, was das System leisten konnte. Der Ertrag: Eine gut strukturierte Dokumentenbasis mit einem geradlinigen Retrieval-Ansatz übertraf durchweg ein raffiniertes Modell, das mit schlecht aufbereiteten Daten arbeitete.
- **Spezifisch für RAG: Die Pipeline ist das Produkt.** Ein RAG-System zu bauen bedeutet, eine Folge zusammenhängender Entscheidungen zu treffen, wie Dokumente aufgenommen und in Segmente aufgeteilt werden, mit welchem Verfahren sie in eine durchsuchbare Form gebracht werden, wie Abfragen formuliert werden, wie Ergebnisse gerankt werden und wie die endgültige Antwort erzeugt wird. Jede Entscheidung beeinflusst die Qualität der nächsten. Unternehmen, die die gesamte Pipeline verstanden, konnten Fehlerfälle diagnostizieren und methodisch verbessern, jene, die das System als Blackbox behandelten, konnten das nicht.
- **In professionellen Kontexten ist ein fundiertes „Ich weiss es nicht“ wertvoller als eine selbstbewusste falsche Antwort.** Für Unternehmen in regulierten Branchen, in der Compliance-Arbeit oder in kundennahen Rollen ist eine KI, die die Antwort verweigert, wenn die Evidenz nicht ausreicht, ein Merkmal, kein Mangel. Der gemeinsame Prototyp wurde ausdrücklich so gestaltet, dass er verifizierte Fakten von unbestätigten Aussagen unterscheidet und die Antwort verweigert, wenn die Evidenz keine Schlussfolgerung stützt. Das beseitigt das Risiko von Halluzinationen nicht vollständig, senkt es aber deutlich und macht die Unsicherheit des Systems sichtbar statt verborgen. In sensiblen Kontexten ist diese Transparenz genauso wichtig wie die Antwort selbst.

„[...] Ein RAG ist nicht ein Modell, sondern eine vollständige Pipeline, und die eigentliche Arbeit liegt in den Schritten dazwischen. Vom Roh-Korpus in den produktiven Einsatz zu kommen bedeutet, unzählige Entscheidungen zu treffen und zu justieren: Segmentierungsstrategie, Embedding-Modell, Retrieval und Neuordnung, Prompt-Design und die Parameter, die auf jeder Stufe gesetzt werden. Zu sehen, wie sich jede Entscheidung in der Qualität des Endergebnisses aufsummiert, war für mich der grösste Perspektivwechsel beim Bau eines solchen Systems.“

— Marco Meola, Mitgründer und CEO, DATABIOMIX



Technischer Anhang

Dieser Anhang richtet sich an Leserinnen und Leser mit Interesse an der technischen Umsetzung. Er fasst die Design-Entscheidungen und die Architektur des gemeinsamen Prototyps zusammen. Die vollständige Codebasis, einschliesslich Notebooks und Evaluations-Datensätze, sowie eine ausführliche README zu Installation, Feature-Modulen und Repository-Struktur ist verfügbar unter:

<http://github.com/SwissDataScienceCenter/sme-kt-zh-collaboration-rag>

A. Anwendungsfall und Testdatensatz-Design

Der Prototyp wurde rund um ein Nachhaltigkeits-Compliance-Szenario für ein fiktives Verpackungsunternehmen (PrimePack AG) aufgebaut, gezielt entworfen, um das System gegen jene Fehlerfälle zu testen, die naive Chatbots zeigen. Das Dokumentenkorpus (~40 Dokumente) umfasste von Dritten verifizierte Environmental Product Declarations (EPDs), Lieferanten-Datenblätter mit selbstdeklarierten Nachhaltigkeitsaussagen, Referenzstandards (GHG Protocol, CSRD-Leitlinien, ISO-Standards) sowie bewusst fehlerhafte, künstlich erstellte Dokumente mit fehlenden Daten, unbestätigten Aussagen und zeitlichen Widersprüchen.

Die fehlerhaften Dokumente wurden aufgenommen, um zu prüfen, ob das System die Qualität der Evidenz unterscheiden und die Grenzen seines Wissens kommunizieren kann, getestet wurden Halluzinationen (Antworten zu Produkten ausserhalb des Portfolios), Selbstüberschätzung (Umgang mit unbestätigten Aussagen), Fabrikation (Verhalten bei fehlenden Daten) und stille Fehler (Auflösung widersprüchlicher Quellen). Ein Ground-Truth-Evaluationsdatensatz wurde für fünf Kategorien vorbereitet: Portfolio-Scope, Aussagenverifikation, fehlende Daten, Quellenkonflikt und Richtlinien-Compliance.

B. Entwicklungsstruktur: Baseline und vier Feature-Module

Baseline RAG Pipeline

Die sechs Kernstufen: Ingest (Rohdokumente laden und in ein einheitliches Format bringen), Segment (Dokumente in abrufbare Segmente aufteilen), Embed (in Vektordarstellungen umwandeln), Store (in einer Vektordatenbank indexieren), Retrieve (relevante Segmente zu einer Abfrage finden) und Generate (eine fundierte Antwort erzeugen). Zwei Dokumentenparser wurden getestet und verglichen; Segmentierungsstrategien wurden für unterschiedliche Dokumentstrukturen und Inhaltstypen evaluiert.

Modul 1: Evaluation und multimodales Retrieval

Das RAGAS-Evaluationsframework wurde integriert und liefert vier Qualitätsmetriken: Faithfulness (widerspricht die Antwort ihren Quellen?), Answer Relevancy, Context Precision und Context Recall. Ein synthetischer Frage-Antwort-Datensatz wurde aus dem Korpus generiert, um Evaluation im grösseren Massstab zu ermöglichen. Multimodales Retrieval wurde ebenfalls ergänzt: Bilder wurden eingebettet und in einer separaten Vektorsammlung gespeichert, was den Abruf von visuellen Belegen wie Produktbildern, Etiketten und Zertifizierungsstempeln zusammen mit Text ermöglicht.

Modul 2: Strukturierte Outputs

Es wurden JSON-Schema-basierte Outputs eingeführt, sodass jede Antwort ein strukturiertes Antwortfeld und eine explizite Liste der Quelldokumente-Kennungen enthält. Damit werden die Outputs audittierbar und eignen sich für die Integration in Berichte, Dashboards oder nachgelagerte Workflows.



Abbildung 1: Entwicklungsstruktur des RAG-Prototyps, Baseline-Pipeline und vier Feature-Module, KI-Programm für KMU des Kantons Zürich (Februar–März 2026).

Modul 3: Fortgeschrittenes Retrieval

Drei Retrieval-Strategien wurden umgesetzt und verglichen: Vektor-Ähnlichkeitssuche (semantisches Matching), BM25-Stichwortsuche (exakter Begriffs-Match) und hybrides Retrieval, das beide mittels Reciprocal Rank Fusion kombiniert. Ergänzt wurden Query-Intelligence-Techniken wie Query Expansion (Erzeugen mehrerer Formulierungen derselben Frage vor der Suche) und HyDE (Erzeugen einer hypothetischen Antwort und deren Verwendung als Suchanfrage). Eine Komponente zur Verwaltung des Dialogkontexts wurde für die Kohärenz über mehrere Dialogschritte hinweg ergänzt.

Modul 4: Agentische Workflows

Das System wurde in Richtung mehrstufiger Reasoning-Prozesse erweitert: Dem Sprachmodell wurde Retrieval als expliziter Werkzeugaufruf zugänglich gemacht, mit einer Agent-Schleife, die es ihm erlaubt zu überlegen, Werkzeuge aufzurufen, Ergebnisse zu beobachten und weiterzuarbeiten, bevor es eine endgültige Antwort liefert. Damit kann das System eine komplexe Frage in Teilanfragen zerlegen, mehrere Quellen unabhängig prüfen und eine fundierte Antwort zusammenführen. Ein RAG-as-Subagent-Muster wurde ebenfalls zur modularen Integration demonstriert.

C. Technologie-Stack

Der Prototyp wurde vollständig auf etablierten Open-Source-Komponenten aufgebaut, um Zugänglichkeit, Reproduzierbarkeit und Freiheit von Anbieterbindung sicherzustellen:

- **Embedding-Modelle:** Umwandlung von Text (und Bildern) in Vektordarstellungen für das Retrieval
- **Vektordatenbanken:** Indexierung und Speicherung der Dokumenten-Embeddings
- **Large Language Models (LLMs):** es wurden sowohl lokal als auch in der Cloud gehostete Modelle getestet und verglichen
- **Renku:** die kollaborative Computing-Plattform des SDSC für gemeinsam genutzte Notebooks und Demo-Hosting
- **GitHub:** Versionskontrolle, Zusammenarbeit und öffentliche Verteilung von Code und Dokumentation

Die Codebasis ist als bearbeitbare Python-Bibliothek (conversational-toolkit) neben einer Anwendungsschicht organisiert, die sie an einen konkreten Anwendungsfall anpasst. Öffentlich gepflegt unter:

github.com/SwissDataScienceCenter/sme-kt-zh-collaboration-rag.

D. Zentrale Design-Prinzipien

Drei Prinzipien haben jede Architektur-Entscheidung geleitet und gelten unabhängig von Fachgebiet und Technologie-Stack:

- **Evidenzstufen trennen.** Verifizierte Fakten (z. B. von Dritten zertifizierte EPDs) von unbestätigten Aussagen (z. B. Lieferanten-Selbstdeklarationen) unterscheiden. In regulierten oder Compliance-Kontexten ist diese Unterscheidung nicht optional.
- **Evaluations-Infrastruktur früh aufbauen.** Qualität zu messen, bevor Funktionen hinzugefügt werden, ermöglicht schrittweise Verbesserungen und die frühe Erkennung von Fehlerfällen. Evaluation nachträglich aufzubauen, ist deutlich schwieriger.
- **Outputs audittierbar strukturieren.** Freitext-Antworten sind schwer zu verifizieren, strukturierte Outputs mit expliziten Quellenangaben ermöglichen nachgelagerte Validierung, Integration in Arbeitsabläufe und das Vertrauen, das aus Nachvollziehbarkeit entsteht.

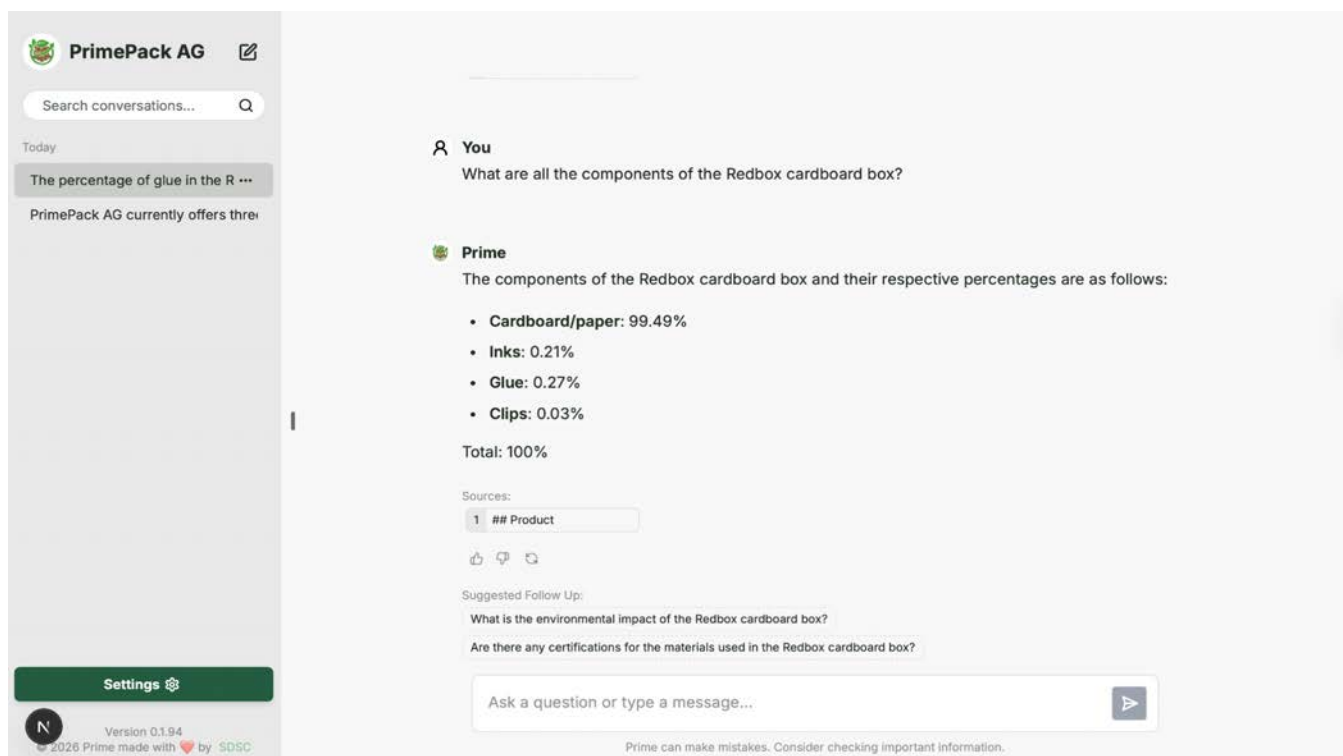


Abbildung 2: Benutzeroberfläche des RAG-Prototyps (Anwendungsfall PrimePack AG), KI-Programm für KMU des Kantons Zürich (Februar–März 2026).

Open-Source-Prototyp

Der Prototyp ist Open Source verfügbar und wird vom SDSC gepflegt:

<https://github.com/SwissDataScienceCenter/sme-kt-zh-collaboration-rag>

SwissDataScienceCenter / sme-kt-zh-collaboration-ragPublic

<> CodeIssuesPull requestsActionsProjectsSecurity and qualityInsights

main3 Branches0 Tags

Go to file

<> Code

Paulina Koernercleaned initial commit ✓acbba30 · 4 months ago🕒 1 Commit

github/workflows	cleaned initial commit	4 months ago
backend	cleaned initial commit	4 months ago
conversational-toolkit	cleaned initial commit	4 months ago
data	cleaned initial commit	4 months ago
frontend	cleaned initial commit	4 months ago
tests	cleaned initial commit	4 months ago
.gitattributes	cleaned initial commit	4 months ago
.gitignore	cleaned initial commit	4 months ago
.pre-commit-config.yaml	cleaned initial commit	4 months ago
LICENSE	cleaned initial commit	4 months ago

16 von 18

Weiterführende Informationen & Links

RAG Systeme für KMU – Video ansehen (Englisch):

https://www.youtube.com/watch?v=wtDOt9v_ngs



Weitere Links:

[KMU KI-Programm im Kanton Zürich – SDSC](#)

[Open-source RAG Prototyp auf GitHub](#)

zh.ch/standort

innovation.zuerich

ahead-zh.ch

Über die Standortförderung im Amt für Wirtschaft, Kanton Zürich

Wir fördern Innovationen und unterstützen Unternehmen – von der Gründung bis zur Vernetzung in den Zürcher Schlüsselbranchen. Mit unserer Plattform «Innovation Zurich» bieten wir Orientierung sowie Informationen zum Innovationsstandort. Ansässige Unternehmen unterstützen wir mit dem Angebot von “ahead” bei ihren Innovationsvorhaben. Zudem sind wir auch Anlaufstelle für Anliegen von ansiedlungsinteressierten Unternehmen. Weitere Informationen: zh.ch/standort | innovation.zuerich | ahead-zh.ch

Über das Swiss Data Science Center (SDSC)

Das Swiss Data Science Center (SDSC) ist eine von EPFL und ETH Zürich gegründete, nationale Forschungsinfrastruktur spezialisiert auf Datenwissenschaft und künstliche Intelligenz (KI). Seine Mission – datengesteuerte Wissenschaft und Innovationen mit gesellschaftlicher Wirkung zu ermöglichen – treibt seine Initiativen in Forschungsprojekten, im Wissens- und Technologietransfer sowie in der Bildung an. Mit einem grossen multidisziplinären Team von Fachleuten in Lausanne, Zürich und Villigen stellt das SDSC Fachwissen und Dienstleistungen für verschiedene Bereiche wie Gesundheitswesen und Biomedizin, Energie und Nachhaltigkeit, Klima und Umwelt, Digitale Gesellschaft sowie für grossformatige wissenschaftliche Infrastrukturen bereit. Zudem unterstützt das SDSC Erst- und Weiterbildungsprogramme an der EPFL und der ETH Zürich. Mehr Informationen: www.datascience.ch

Kontakt:

Für Fragen zum Programm kontaktieren Sie bitte:

Dr. Anna Fournier

Principal Data Scientist

Swiss Data Science Center

anna.fournier@datascience.ch

Wir danken den Zürcher KMU des KI-Innovationsprogramms 2025:



cisra.ch



databiomix.com



econetta.com



gonina.com



languagemasters.ch



microdul.com



papillon.ch



soort.com



stratoserv.ch



safetycenter.ch



vonlanthen.tv

Das KI-Innovationsprogramm für KMU im Kanton Zürich ist eine Initiative von:



**Kanton Zürich
Volkswirtschaftsdirektion
Amt für Wirtschaft**

zh.ch/standort



Swiss Data Science Center

datascience.ch