Generative AI mit SAP

Maßgeschneiderte KI-Anwendungen entwickeln

Kapitel 4

Beispiele für die Anwendung von KI im Unternehmenskontext

Um ihre volle Macht zu entfalten, muss generative KI sinnvoll eingesetzt werden. Deshalb besprechen wir in diesem Kapitel geeignete Anwendungsfälle.

Sie haben nun einen guten Überblick über die KI-Strategie von SAP und die entsprechenden Architekturen und Tools erhalten. Nun zeigen wir Ihnen in diesem Kapitel, welche praktischen Anwendungsfälle im Unternehmenskontext möglich sind und wie Sie diese identifizieren und bewerten. In diesem Kapitel steigen wir tief in kundenindividuelle Anwendungsfälle ein und zeigen Ihnen nicht nur die Kombination verschiedener Technologien und mit der SAP BTP, sondern beleuchten auch KI-Quick-Wins. Sie lernen, wie eine solche Umsetzung im Detail aussehen könnte, welche Vorteile eine individuelle KI-Entwicklung hat und wie Sie KI-Anwendungsfälle von der Implementierung eines PoC hin zur skalierbaren Lösung effizient und umsetzbar steuern. In Abschnitt 4.1, »Praktische KI-Anwendungsfälle von SAP«, stellen wir Ihnen die von SAP integrierten KI-Funktionalitäten vor. In Abschnitt 4.2, »Best Practices für die Einführung eines KI-Assistenten«, gehen wir auf die Methoden zur Identifikation individueller Use Cases und der passenden KI-Architektur und -Services aus dem SAP-BTP-Service-Portfolio ein. In Abschnitt 4.3 erhalten Sie schließlich Beispiele für kundeneigenen Anwendungsfälle.

4.1 Praktische KI-Anwendungsfälle von SAP

Um belastbar entscheiden zu können, wo KI in SAP-Landschaften den größten Hebel hat, braucht es eine Taxonomie, die sich weniger an Algorithmen und mehr an Wertbeiträgen entlang der Geschäftsprozesse orientiert. In Kapitel 3, »Das SAP-KI-Portfolio«, haben Sie bereits die SAP-KI-Strategie kennengelernt, in der KI als Motor für das Unternehmen dient, weil durch KI schnellere und effizientere Geschäftsprozesse ermöglicht werden. Die folgenden Faktoren sind dabei ausschlaggebend für den Erfolg:

- schnelle und effiziente Ergebnisse durch KI-Integration in allen SAP-Geschäftsfunktionen und das Zusammenspiel von rollenbasierten KI-Agenten, um alle Arbeitsprozesse zu beschleunigen
- maximierter Nutzen durch eine einheitliche Benutzeroberfläche – den SAP-Copilot Joule

Sie haben die Interaktionsmuster von Joule kennengelernt, die auf die zentralen vier Felder von KI-Fähigkeiten zielen:

- Automatisierung
- Entscheidungsassistenz
- generative Assistenz
- Insights bzw. Analytics

Jede Fähigkeit lässt sich klar im Anwendungskontext verorten, mit messbaren Zielen hinterlegen und spiegelt zugleich die Art und Tiefe der Integration in SAP-Workflows wider. Betrachten wir nun die Felder im Detail:

- *Automatisierung* adressiert wiederkehrende, stark regelbasierte Arbeitsschritte mit hoher Transaktionszahl und klaren Eingangsdaten. Typische Ausprägungen sind *SAP Document AI* (siehe *http://s-prs.de/v10759014*), die Klassifikation von Vorgängen sowie das Routing von Fällen an die richtige Bearbeitungsstelle. Entscheidend ist hier, dass KI nicht als »Sonderweg« neben dem Prozess existiert, sondern als Eingriffspunkt im Standardfluss. Ein Vorgang wird automatisch angereichert, kategorisiert oder validiert, und der Mensch greift bei Abweichungen oder Unsicherheiten ein. Diese Muster finden sich heute u. a. in Procure-to-Pay (Einkaufsempfehlungen, Kategorisierung), Rechnungs- und Spesenprozessen (Belegerkennung, Betrugshinweise) sowie im Service (Ticketklassifikation und -priorisierung). In SAP-Produkten wird dieses Muster beispielsweise in *Ariba Guided Buying* (KI-gestützte Item-Empfehlungen), *SAP Concur* (ExpenseIt/OCR und KI-gestützte Prüfungen) und in der *SAP Service Cloud* (automatische Ticketkategorisierung) sichtbar.
- *Entscheidungsassistenz* zielt auf Situationen, in denen der Prozess nicht voll automatisiert werden kann, KI aber kontextualisierte Hinweise und Priorisierungen liefert. Das Paradebeispiel in der ERP-Welt ist z. B. das Situation Handling in SAP S/4HANA Cloud mit Joule (siehe *http://s-prs.de/v10759015*). Das System erkennt aus Konstellationen von Stammdaten, Bewegungsdaten und Terminen »Situationen«, die Aufmerksamkeit erfordern (z. B. zu späte Fakturierung, drohende Lieferverzögerungen, Rechnungsabweichungen), benachrichtigt gezielt die zuständigen Rollen und führt mit Handlungsvorschlägen durch die Behebung. In CRM-Szenarien finden wir analoge Muster mit Opportunity- und Lead-Scoring, die Verkäuferinnen und Verkäufern helfen, Pipeline und Aktivitäten nach Erfolgswahrscheinlichkeit zu priorisieren und so die knappe Zeit dort einzusetzen, wo sie den größten Effekt hat.
- *Generative Assistenz* umfasst KI-Funktionen, die natürliche Sprache und Inhalte erzeugen, zusammenfassen oder übersetzen. Der SAP-weit eingeführte Copilot Joule steht exemplarisch dafür. Anwender und Anwenderinnen formulieren Ziele oder Fragen in natürlicher Sprache, Joule gleicht sie gegen Geschäftsobjekte und (je nach Konfiguration) Unternehmenswissen ab, erzeugt Vorschläge, etwa Textentwürfe für Stellenanzeigen, Zielvereinbarungen oder E-Mails, erklärt Kennzahlen oder leitet passende nächste Schritten an. Der Mehrwert entsteht, weil generative Funktionen im Businesskontext geerdet werden (*Grounding*) und nicht als generische Chat-Gadgets daneben laufen.

- *Insights/Analytics* beschreibt das Spektrum von Augmented Analytics bis hin zur semantischen Suche über Berichte und Modelle. In SAP Analytics Cloud materialisiert sich das in Funktionen wie z. B. Search to Insight, Smart Insights und Smart Discovery. Fachanwenderinnen und Fachanwender stellen Fragen in natürlicher Sprache, erhalten automatisch generierte Visualisierungen, Erklärungen und Hypothesen und können daraus Stories erstellen, ohne den Umweg über manuelle Modellierungsschritte. In der Supply-Chain-Planung liefert SAP IBP ML-gestützte Prognosen (z. B. Demand Sensing) und Anomalie-Hinweise, die Planungsteams in iterativen Zyklen verproben und korrigieren. Auch hier ist eine wiederkehrende Gemeinsamkeit: KI wird als Assistenz in die Analyse-Umgebung eingebettet, nicht als externer »Data-Science-Handoff«.

Diese vier Felder sind hinreichend trennscharf, aber durchaus miteinander kombinierbar. Ein modernes Service-Szenario kann mit Ticketklassifikation (Automatisierung) beginnen, durch Opportunity-/Prioritätsempfehlungen (Entscheidungsassistenz) ergänzt werden, bei der Antworterstellung generative Assistenz bieten und die Ergebnisdaten in Analytics-Stories zurückspiegeln (Insights). Für die Implementierung bedeutet das, dass nicht nach »dem einen« KI-Use-Case gesucht wird, sondern kettenförmige Wertbeiträge im Prozess identifiziert und orchestriert werden. Zusammengefasst ist wichtig, wie Produktreife und Governance in die Taxonomie einfließen. Eingebettete Funktionen in Standard-LoBs sind typischerweise »opinionated«. Sie liefern eine solide Default-Qualität, sind wartungsarm und auditierbar. Offene Plattform-Bausteine wie z. B. SAP HANA Cloud Vector Engine, SAP AI Core und Orchestration Workflows geben Freiheit für Spezialfälle, verlangen aber mehr Engineering (z. B. Datenkuratierung, Guardrails, Telemetrie). Wer Use Cases verankert, sollte also stets »Buy for Base, Build for Edge« denken: eingebettete Business AI für den Mainstream, erweiterbar mit SAP-BTP-Bausteinen dort, wo Differenzierung zählt. Wer täglich in Finance, Logistik oder Fertigung arbeitet, hat selten Zeit, sich Berichte erstellen zu lassen und Anomalien händisch zu suchen.

Intelligentes Situation Handling dreht den Spieß um. SAP S/4HANA beobachtet definierte Geschäftskonstellationen, etwa aus offenen Posten, Lieferplänen, Toleranzregeln und erzeugt Situationen, wenn Handlungsbedarf erkennbar ist. Nutzerinnen und Nutzer erhalten Benachrichtigungen (über das SAP Fiori Launchpad oder die Inbox), eine kompakte Zusammenfassung und kontextuelle Aktionen (z. B. klären, eskalieren, terminieren). Viele Szenarien sind als vorkonfigurierte Templates verfügbar (z. B. Fakturaprüfung, Terminabweichungen), die Administration passt Schwellenwerte und Zuständigkeiten an. Der Effekt ist, dass weniger Zeit bis zur Kenntnisnahme und weniger Zeit bis zur Behebung verstreicht, zwei Kern-Metriken für reaktionsfähige Back-Office-Prozesse. Für Beraterinnen und Berater ist wichtig, dass die Assistenz von Datenqualität und Prozesstreue lebt. Gute Ergebnisse entstehen, wenn Stammdaten gepflegt, Ausnahmen sauber modelliert und Benachrichtigungswege eindeutig sind. In der Einführungspraxis hat es sich bewährt, mit ein bis zwei Situationsarten in einem Bereich zu beginnen, die Wirkung mit Metriken zu belegen und erst dann weitere Situationen zu aktivieren.

4.1.1 Einkauf

Der Einkauf ist oft ein Long-Tail mit großem Volumen. Viele Mitarbeitende bestellen gelegentlich, sind regelunsicher und brauchen Orientierung. Guided Buying adressiert dieses Dilemma mit KI-gestützten Empfehlungen und Guardrails direkt in der Bestelloberfläche. Was die Nutzerin oder der Nutzer sieht, ist eine kuratierte, benutzerfreundliche Suche samt »ähnliche Artikel«-Vorschlägen, bevorzugten Lieferanten und Compliance-Hinweisen im Fluss. Die KI lernt aus Einkaufshistorie und Kontext (Kategorie, Standort, Rolle) und schlägt Artikel vor, die kosteneffizient und regelkonform sind. Für das Procurement-Team stehen Nutzungsreports und Konfigurationen bereit, um die Empfehlungslogik geschäftstauglich zu halten. Das Ergebnis sind kürzere Bestellzeiten und eine höhere Katalogtreue. Für Beraterinnen und Berater lohnt es sich, Guided Buying als Change-Hebel zu begreifen. Regeln sichtbar machen, Preferred-Supplier-Sets aktiv pflegen und Feedback aus der Linie rasch einarbeiten.

4.1.2 Spesen

In Spesenprozessen entstehen Reibungsverluste, wenn Nutzerinnen und Nutzer Belege manuell übertragen und Auditoren alles nachkontrollieren. SAP Concur ExpenseIt (siehe *https://www.concur.de/products/expenseit*) nimmt hier den ersten Schritt ab: Das Belegfoto wird aufgenommen (auch offline), OCR extrahiert relevante Felder, KI ordnet Zahlungsart/Expense-Typ zu und erkennt u. a. auch Kreditkarten-Endziffern. Ergänzend überprüft SAP Concur Detect Ausgabenmuster, erkennt Duplikate, Ausreißer und Regelverstöße, wodurch die Prüfung vom 100-%-Check zur risikobasierten Prüfung wird. Nutzerinnen und Nutzer müssen weniger tippen, haben weniger Rückfragen und sind schneller fertig. Finance-Mitarbeitende finden bei potenziellen Problemen mehr Treffer bei höherer Prozessgeschwindigkeit.

Moderne Pipelines in SAP Concur koppeln OCR, PII-Filter (zum Schutz personenbezogener Daten im extrahierten Text) und Klassifikatoren. Beachten Sie bei Einführungen, dass die Qualität der Lösung von Fotoqualität und Belegvarianz abhängt. Die Governance regelt, wie streng Ausreißer gehandhabt werden und wo eine menschliche Freigabe zwingend bleibt.

4.1.3 Vertrieb

Im Vertrieb zeigt KI ihren Nutzen dort, wo Zeit am knappsten ist, also in der Pipeline-Pflege bzw. beim Nachfassen. Lead- und Opportunity-Scoring in *SAP Sales Cloud* mit *SAP CX Business AI* (siehe *http://s-prs.de/v10759016*) nutzt historische Erfolgsdaten, um Wahrscheinlichkeiten zu berechnen und Listen automatisch zu priorisieren; Vertriebsmitarbeitende fokussieren sich auf Vorgänge mit hohem Closing-Potenzial und erhalten folgende Schritte bzw. je nach Ausbaustufe generierte Antwortentwürfe direkt im Arbeitskontext.

4.1.4 Service

In Service-Szenarien analysieren Klassifikations-Agenten neue Tickets, erkennen Thema/Priorität und leiten sie an die richtige Gruppe weiter; aus Wissensartikeln und ähnlichen Fällen entstehen Lösungsvorschläge und (wo aktiviert) Zusammenfassungen für Agenten. Joule bindet das als Copilot in Sales/Service ein, sodass Anwender natürlichsprachlich nach Kontohistorie, Risiken oder nächsten Aktionen fragen können. Beachten Sie bei der Einführung, dass die Akzeptanz auf Nutzerseite gegeben ist. Ein initiales Scoring muss erklärbar und nachvollziehbar sein (z. B. Einflussfaktoren zeigen), damit Teams Vertrauen fassen. In Servicecentern empfiehlt sich, Routing-Automatisierung zuerst als Vorschlag zu pilotieren und erst nach stabilen Trefferraten ins Auto-Assign zu wechseln. Generative Funktionen (Mail-Entwürfe, Ticket-Zusammenfassungen) sollten mit Ton-/Compliance-Vorgaben (Guardrails) konfiguriert werden.

4.1.5 HR

HR-Teams arbeiten stark textgetrieben, von Stellenbeschreibungen über Zielvereinbarungen bis hin zu Feedback. SAP bietet hier generative Assistenten, die HR-Verantwortlichen und Führungskräften erste Entwürfe liefern, sprachlich variieren und unternehmensspezifisch rahmen (z. B. Skill-Taxonomien, Diversity-Hinweise). In der Recruiting-Suite von *SAP Fieldglass* unterstützt etwa ein Feature zum Job Description Enhancement. Aus groben Eckdaten werden zielgruppengerechte Ausschreibungen, die Fachanforderungen, Ton und Arbeitgebermarke widerspiegeln; Nutzerinnen und Nutzer vergleichen Original und KI-Vorschlag und übernehmen ganz oder teilweise. Auch die Zielerstellung (Performance/Development) lässt sich generativ anstoßen, wobei die HR-Abteilung die Konsistenz über Vorlagen und Review-Schleifen sicherstellt.

Joule dient als Einstieg über natürliche Sprache, etwa für Fragen zu Policies, Kennzahlen oder nächste Schritte im HR-Prozess. Bedenken Sie beim Rollout, generative Features kuratiert zu launchen. Beispiele guter und schlechter Prompts zeigen, wann Nutzerinnen und Nutzer generieren sollten (Leerseite, Ideenfindung) und wann nicht (verbindliche Rechts-/Policytexte). Die HR-Governance definiert Review-Pflichten, Bias-Kontrollen und Sprachleitfäden.

In der Steuerung externer Arbeitskräfte zählt die Zeit bis zur Besetzung. SAP Fieldglass unterstützt mit KI ganzheitlich und integriert, von KI-gestützter Lebenslaufanalyse, Skill-Extraktion oder Ranking. Hiring-Manager sehen früh qualifizierte Shortlists und Hinweise auf Matching-Lücken. In der Oberfläche treten diese Funktionen als Empfehlungen und Filter in Erscheinung; Administratorinnen und Administratoren steuern, welche Felder stärker gewichtet, welche Skills vorgeschlagen und welche Kriterien transparent gemacht werden. Gute Implementierungen begleiten dies mit Trainings, die erklären, was das Scoring bedeutet und was nicht.

4.1.6 Prozessanalyse

SAP Signavio bringt KI an zwei Stellen ein: in der Prozessanalyse (Process Intelligence, Process Mining) und im Prozessdesign. Neuere Process-AI-Funktionen liefern sofortige Hinweise auf Problemfelder und schlagen vorkonfigurierte Prozessmodelle aus einer umfangreichen Best-Practice-Bibliothek vor; zusätzlich werden passende KPIs/PPIs empfohlen. Für Prozessverantwortliche fühlt sich das an wie ein Copilot fürs Prozessdesign. Statt mit einem leeren Diagramm zu starten, beginnt sie mit passenden Vorlagen und fokussiert sich auf Anpassung und Validierung. Das beschleunigt nicht nur den Start, sondern hält auch Konsistenz über Bereiche hinweg.

4.2 Best Practices für die Einführung eines KI-Assistenten

Was alle diese Beispiele eint, ist die Tendenz zur vereinheitlichten Copilot-Erfahrung. Joule dient in mehreren Szenarien als Einstieg über natürliche Sprache, mit Grounding in Geschäftsobjekten und (wo freigeschaltet) in unternehmensspezifischen Dokumenten. Für Anwenderinnen und Anwender bedeutet das eine konsistente Interaktion wie fragen, vorschlagen lassen, prüfen, übernehmen. Für IT/Governance entsteht ein zentraler Hebel für Guardrails, Zugriffsmodelle und Telemetrie über Anwendungen hinweg. SAP erweitert dieses Prinzip mit *Joule Agents*, die klassifizieren, Antworten vorbereiten und Wissen aus gelösten Fällen heben – ein Schritt in Richtung arbeitsteiliger KI-Agenten, der dennoch in klaren Prozessgrenzen bleibt.

Über Produktgrenzen hinweg empfiehlt sich insgesamt ein Einführungsmuster, das sich an folgenden Leitsätzen orientiert:

- **Beginnen Sie im Prozess, nicht im Modell**

  Wählen Sie einen sichtbaren Engpass (z. B. Ticket-Backlog, Maverick Buying, Spesen-Durchlaufzeit) und docken Sie KI im Hauptprozessfluss an, nicht in einem nebenläufigen Tool. Auf unsere o. g. Beispiele angewendet, bedeutet dies beispielsweise, in SAP S/4HANA eine konkrete Situation zu aktivieren, in SAP Ariba Guided-Buying-Regeln mit Empfehlungen zu kombinieren oder in SAP Concur die ExpenseIt standardmäßig zu aktivieren.

- **Vorschlag vor dem Autopilot priorisieren**

  Lassen Sie Klassifikation bzw. Scoring zunächst im Hintergrund laufen und spielen Sie diese nicht zurück ins produktive System. Messen Sie Qualität, sammeln Sie Feedback, erklären Sie Einflussfaktoren. Schalten Sie Automatisierung inkrementell schärfer, wenn die Trefferraten stabil sind, das erhöht die Akzeptanz und schützt vor negativen Lernkurven im Livebetrieb.

- **Grounding und Guardrails ernst nehmen**

  Generative Funktionen entfalten ihren Nutzen erst mit Kontext (Geschäftsobjekte, Richtlinien, Dokumente) und Leitplanken (Ton, Compliance, PII-Schutz), in SAP Con-

cur also beispielweise in der Kette aus OCR, PII-Filter und Klassifikation, in SAP SuccessFactors definieren Styleguides und Freigaben den Rahmen.

- **Metriken ab Tag eins einsetzen**

  Definieren Sie Business-KPIs (z. B. die Zeit bis zur Freigabe, die Katalogtreue oder die Erstlösungsquote) und technische KPIs (Latenz, Trefferquoten, Drift). Nutzen Sie A/B-Vergleiche da, wo es möglich ist, und etablieren Sie Review-Routinen mit den Fachbereichen. Die bei SAP IBP üblichen Genauigkeits- und Priorisierungsmetriken sind gute Vorbilder.

- **Change Management als Produktbestandteil verstehen**

  Schulungen sollten nicht nur erklären, wie ein Tool bedient wird, sondern folgende Fragen klären:

  - Wann akzeptiere ich den Vorschlag?
  - Wann überschreibe ich ihn?
  - Wie gebe ich Feedback?

  Gerade bei Scoring/Empfehlungen ist Erklärbarkeit ein Katalysator für Adoption.

Die Anwendungsbeispiele haben Ihnen gezeigt, dass Business AI in SAP dort am stärksten ist, wo sie mit dem Prozess verschmilzt. Im Beratungs- bzw. Entwicklungsumfeld bedeutet das, den Nutzungsfluss in den Mittelpunkt zu rücken, Grounding/Guardrails sorgfältig zu konfigurieren, Wirkung zu messen und dort, wo der Standard nicht reicht, mit individuell ausgewählten SAP-BTP-KI-Services gezielt zu erweitern.

4.3 Kundeneigene KI-Anwendungsfälle

Unternehmen setzen nicht immer nur SAP-Anwendungen im Standard ein, sondern Geschäftsprozesse wachsen historisch und sind Transformationen unterworfen. Es gibt nicht nur viele Anwendungen außerhalb des SAP-Ökosystems, sondern auch unternehmensspezifische Besonderheiten, was Softwarepartner oder bereits bestehende Lizenzmodelle betrifft. Ein weiterer wichtiger Punkt ist, dass digitale Transformationen, insbesondere im SAP-Umfeld, immer sehr aufwendig sind. Historisch über Jahrzehnte gewachsene SAP-ECC-Systeme müssen mit großem Aufwand nach SAP S/4HANA und in die Cloud migriert werden. Für viele Unternehmen sind Cloud-Transformationen mit RISE with SAP oder GROW with SAP in eine Private oder Public Cloud teilweise noch nicht einmal Zukunftsmusik, vielleicht bereits angedacht, aber noch weit von einer kurz- oder mittelfristigen Implementierung entfernt. Da SAP Innovationen nur noch in der Cloud zur Verfügung stellt, sind heute für viele Unternehmen Lösungen aus dem Embedded-AI-Umfeld, wie in Kapitel 3, »Das SAP-KI-Portfolio«, beschrieben, nicht möglich. Um den Anschluss nicht zu verpassen, können diese jedoch bereits selbst mit ihren On-Premise-SAP-ECC- oder -SAP-S/4HANA-System mithilfe der SAP Business Technology Platform (SAP BTP, siehe Abbildung 4.1) und Custom AI mit AI Foundation umgesetzt werden.

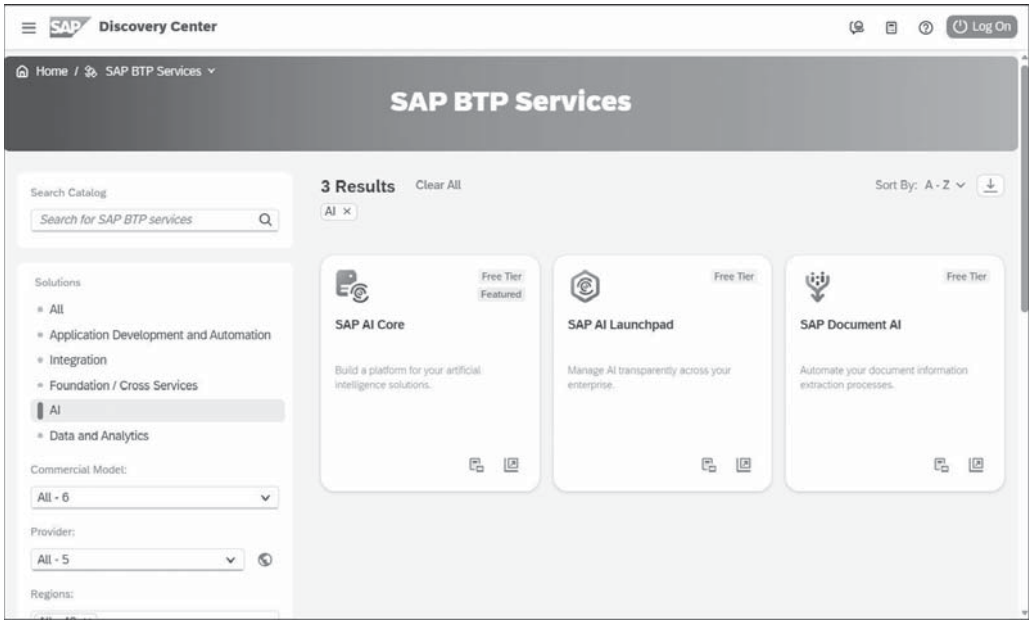

Abbildung 4.1: AI-Services auf der SAP BTP

Sie haben bereits in Kapitel 3 eine wichtige Quelle kennengelernt – das SAP Discovery Center und die Services der SAP BTP (siehe *https://discovery-center.cloud.sap/viewServices*). Hier finden sich alle Services, u. a. auch die AI Foundation Services: SAP AI Core, SAP AI Launchpad und SAP Document AI, die jeder SAP-Kunde unabhängig vom Status einer Cloud-Transformation mit einer SAP BTP im Einsatz nutzen und konsumieren kann.

Sie haben bereits die Funktionalitäten der Basis-Services der AI Foundation kennengelernt, d. h., die Etablierung einer sicheren KI-Basis im Unternehmen ist damit sichergestellt, und die Möglichkeit, mit Sprachmodellen aus dem SAP AI Core mit eigenen Prompts zu experimentieren, ist nur ein Setup entfernt. Dieses Setup lernen Sie in Kapitel 5, »Ihr Handwerkszeug für die KI-Entwicklung: SAP AI Core und SAP AI Launchpad«, näher kennen. Bevor wir allerdings in die einzelnen KI-Anwendungsfälle eintauchen, ist es wichtig, dass Sie die Möglichkeiten einer Use-Case-Bewertung kennenlernen, damit Sie für die Use Cases die richtige Technologie auswählen können. Daher starten wir zunächst mit einer einfachen Kategorisierung von KI-Use-Cases.

Betrachten wir KI-Anwendungsfälle etwas allgemeiner, können wir vier sogenannte *KI-Building-Blocks* identifizieren:

- **Steigerung operativer Effizienz**
  - Wissen, Prozesse und Onboarding
  - Intelligente Suche
  - Prozessautomatisierung minimiert repetitive Aufgaben, entlastet Mitarbeitende und senkt Fehlerquoten.

- **Personalisierte Angebote und Services**
  - Passgenaue Kundenerlebnisse – vom personalisierten Newsletter bis hin zu individuellen Produktempfehlungen
  - Customer (Self) Service und Support (ITSM-extern)
  - Guided Applications und Service Added Value (Up-/Cross-Selling)
- **Schnellere Entscheidungsfindung**
  - Echtzeit-Datenanalysen verkürzen Entscheidungswege und schaffen Transparenz über komplexe Zusammenhänge.
- **Neue Geschäftsmodelle und Kundenbindung**
  - Basis für innovative, skalierbare Geschäftsmodelle, die in bestehenden Märkten neue Umsatzquellen erschließen
  - Wettbewerbsdifferenzierung

Wenn Sie die Anwendungsfelder in Unternehmen im Beispiel der Building Blocks genauer betrachten, fällt Ihnen vielleicht auf, dass sie immer auf zwei der wichtigsten Wirk-Ebenen zielen:

- Nach innen, d. h. auf die Arbeitsumfelder der Mitarbeitenden und Firmenführung, beispielsweise in der Optimierung effizienter Suchen in Bezug auf Wissen und Dokumentenumgang, bzw. unterstützend im Sinne einer Automatisierung manueller repetitiver Aufgaben. Zum anderen ist die Entscheidungsfindung gerade in komplexen Zusammenhängen einer der großen Mehrwerte. KI als Innovationstreiber, nicht nur in der Außenwirkung am Markt, sondern vor allem auch in der Erschließung neuer Geschäftsmodelle, ist heutzutage ein wichtiges Momentum für eine Wettbewerbsdifferenzierung.
- Nach außen hin zielt KI zum einen – gerade im Handel – auf Kundenerlebnisse ab, d. h., digitale Erlebnisse über Angebote oder innerhalb des Service-Portfolios bringen Innovation zum Kunden und können das eigene Image stärken. Insbesondere im Support kann KI die oft ungeliebten und mühsamen Ticket-Prozesse nicht nur für die Kunden-, sondern auch für die interne Administrationsebene deutlich optimieren – in Effizienz, Komfort oder auch in Ansätzen zum Self-Service oder optimierter Ticket-Lösungen schlummern hier schnell zu hebende Potenziale.

Jetzt fehlen Ihnen nur noch zwei Puzzlestücke, um die Methode für die Identifizierung passender KI-Use-Cases in Ihrem Unternehmen vollständig kennenzulernen: der Anwendungsbereich und die Technologie, die Sie verwenden möchten. Für das erste Puzzlestück sollten Sie sich die Frage stellen, an welcher Stelle in Ihren Prozessen ein hoher Aufwand entsteht.

Anhand der Leitfragen aus dem Cheat Sheet in Abbildung 4.2 können Sie leicht identifizieren, in welchen Arbeitsabläufen oder Prozessen Potenziale für KI stecken. Sie finden das Cheat Sheet auch im Download-Material unter: *www.sap-press.de/6087*.

Abbildung 4.2: Cheat Sheet für die Identifikation eines Anwendungsfalls

Nehmen wir beispielsweise an, Sie identifizieren einen Bereich, in dem viel manuelle Arbeit entsteht, weil viele Dokumente noch nicht digitalisiert sind, die manuelle Bearbeitung daher nicht nur zeitaufwendig, sondern auch fehleranfällig ist. Solch ein Prozess wäre ideal durch eine KI-Unterstützung im Sinne der operativen Effizienz zu optimieren. Aber – um zum zweiten Puzzleteil zu kommen – welche Technologie könnte hier Anwendung finden?

Abbildung 4.3 gibt Ihnen hier eine Orientierung, ob generative KI oder Machine Learning (ML) das Mittel der Wahl für Ihren Anwendungsfall ist, je nachdem, welche Art von Aufwand, Ziel und Daten vorliegt. Sie können mit dieser Grafik bewerten, ob Sie einen Use Case basierend auf generativer KI und LLMs durchführen sollten oder lieber selbst ein ML-Modell trainieren sollten. Der Übergang ist hierbei fließend, und verschiedene Use Cases können auf beiden Wegen umgesetzt werden.

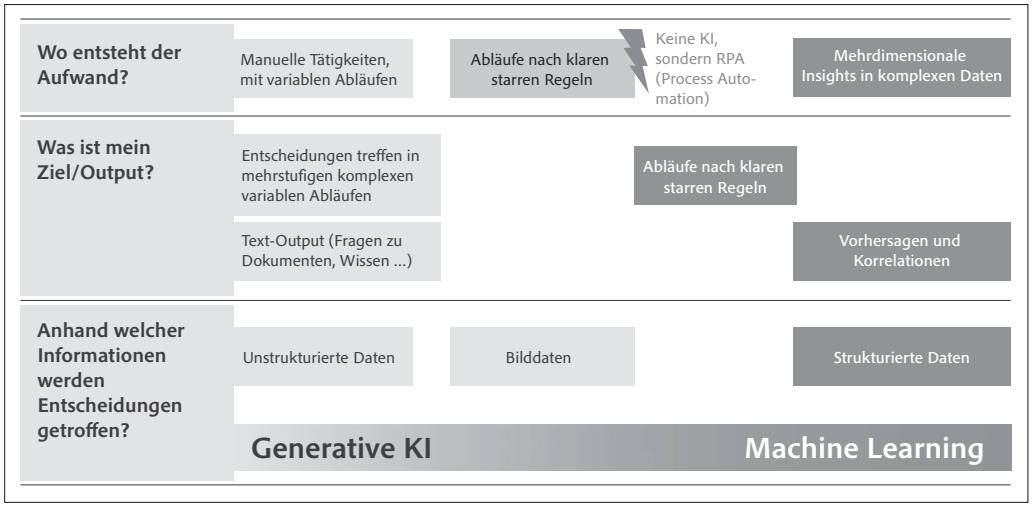

Abbildung 4.3: Leitidee von den Daten zur Lösung

4.3.1 Bearbeitung von Angeboten im Einkauf

Da Sie nun sowohl die Wirk-Ebenen verstehen als auch die vier Felder der KI-Fähigkeiten: Automatisierung, Entscheidungsassistenz, generative Assistenz und Insights bzw. Analytics kennengelernt haben, kennen Sie jetzt das vollständige Handwerkszeug für die Umsetzung Ihrer eigenen individuellen KI-Use-Cases. Lassen Sie uns nun einmal gemeinsam anhand der Anwendungsfelder und verfügbaren SAP-KI-Services entsprechende KI-Use-Cases im Detail betrachten. Betrachten wir dazu einmal die erste Kategorie »operative Effizienz«. Wenn Unternehmen von KI im SAP-Umfeld sprechen, meinen sie in der Praxis fast immer dieselbe Grundidee: Fachliches Wissen, das in Dokumenten, Wikis, E-Mails, Anlagenhandbüchern oder Richtlinien steckt, soll bei einer konkreten Aufgabe schnell, korrekt und nachvollziehbar nutzbar werden. Repetitive, ineffiziente und oft fehleranfällige Prozesse können mit den richtigen Services und Werkzeugen in der SAP BTP einfach und skalierbar umgesetzt werden. Stellen Sie sich vor, in Ihrem Unternehmen haben die Kollegen im Einkauf immer noch viel manuelle Aufwände durch die Bearbeitung von Angeboten. Angebote kommen per E-Mail an, werden manuell im SAP erfasst, und bis die Purchase Order im System verbucht wurde, vergeht viel Zeit mit manueller Arbeit im Abgleich des Angebots mit dem eigenen Produktkatalog, den Einkaufsrichtlinien, Besonderheiten der eigenen Prozesse, und möglicherweise passieren bei der Übertragung zwischen Angebot und SAP auch Eingabefehler. Dies ist ein idealer Use Case für SAP Document AI und ein echter KI-Quick-Win. Skizzieren wir kurz die Ausgangssituation und die Herausforderungen:

- Bestellanforderungen werden manuell auf Basis der erhaltenen Angebotsdokumente in SAP S/4HANA angelegt.
- Die Bestellanforderungen werden manuell geprüft, gegen den bestehenden Produktkatalog validiert und freigegeben.
- Es gibt uneinheitliche Dokumentenformate (Angebots- und Produktinformationen sind uneinheitlich, teilweise fehlende EANs, abweichende Produkttexte etc.).
- Die Datenübertragung ist ineffizient und zeitintensiv.
- Es gibt einen hohen manuellen Prüfungsaufwand und eine hohe Ablehnungs- und Fehlerquote.

Um diese Probleme zu lösen, ist SAP Document AI das richtige Tool. Es bietet GenAI-gestützte Extraktion von Daten aus Dokumenten und erkennt Kopf- und Positionsdaten in verschiedenen Formaten. Für diesen Anwendungsfall sollte es im Zusammenspiel mit den folgenden Komponenten eingesetzt werden:

- effizienter digitaler Workflow mit SAP Build Process Automation (zentrale Workflow-Steuerung vom UI-Upload der Dokumente über die Extraktion und Genehmigung bis hin zur Integration von SAP S/4HANA)
- automatisierte Validierung mit SAP AI Core (Bereitstellen und Verwalten von KI-Modellen als Basis für die Bereitstellung eines intelligenten Service zur Datenvalidierung)

SAP Build Process Automation liefert eine No-Code-Entwicklungsumgebung für Workflows und umfasst:

- Fluss-Steuerung, Request-/Approval-Formulare sowie Custom-Formulare
- Integration von externen Diensten (externe APIs, IFlows, E-Mail, Sub-Workflows)
- Automatisierung Entscheidungslogik und Regelframework
- Skripte
- verschiedene Triggermechanismen (API, Formular, Events)

Visibility-Scenarios erlauben Ihnen die Implementierung von eigenen KPI-Dashboards für maximale Prozesstransparenz. Im Workflow definieren Sie zunächst den fachlichen Prozess vom Angebots-Upload bis hin in die Extraktion, Freigabe und finale Verbuchung im SAP-System. SAP Document AI sorgt für die KI-unterstützte Erkennung der Angebotsdokument-Inhalte (siehe Abbildung 4.4).

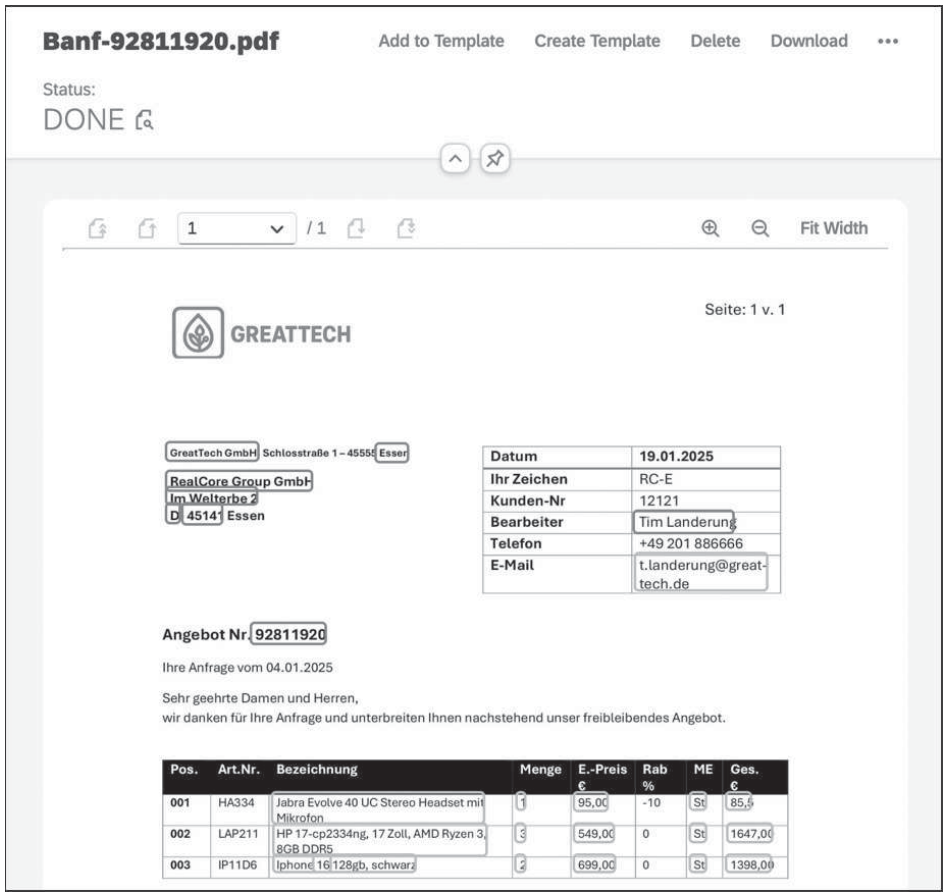

Abbildung 4.4: Beispiel-Extraktion von Informationen aus Angebotsdokumenten mit SAP Document AI

Der Premium-Service-Plan von SAP Document AI unterstützt die flexible Extraktion von Daten auf Basis von generativer KI, inklusive Möglichkeiten für die Klassifizierung von Dokumenten und das Anreichern der extrahierten Daten (beispielsweise als automatisierte Prüfung von Katalogartikeln oder interner Einkaufsrichtlinien). In Verbindung mit einer SAP-Fiori-App für den Einkauf lässt sich hier eine leichtgewichtige Applikation implementieren, über die die Einkäufer schnell und automatisiert ihre Angebote hochladen, extrahieren und validieren können. Auch ein »human in the middle«-Konzept mit einer finalen Überprüfung der Extraktion vor Verbuchung in SAP lässt sich leicht in den SAP Process Automation Workflow integrieren. Sie haben hier die SAP BTP Services in ihrer Anwendung und Wechselwirkung kennengelernt und das generelle Ziel – die operative Effizienz im Sinne von Automatisierung manueller Prozesse – in der Anwendung verstanden. Anhand der Methodik können Sie nun ähnliche Use Cases analog skalieren und die Automatisierung im Sinne weiterer Anwendungs- und Fachbereiche im Unternehmen skalieren.

4.3.2 KI-Assistent mit Unternehmenswissen

Nachdem Sie mit dem ersten Anwendungsfall das Feld »Automatisierung« kennengelernt haben, lassen wir nun SAP Build Process Automation hinter uns und betrachten einen weiteren spannenden Use Case aus dem Feld der »generativen Assistenz«. Möglicherweise stellen Sie sich die Frage, was in einzelnen Bereichen Ihres Unternehmens passiert, wenn aufgrund von Generationenwechsel Wissen verschwindet. Vielleicht nutzt auch Ihr Unternehmen bereits ChatGPT, aber Fragen zu eigenen Wissensquellen im Unternehmen werden dort nicht beantwortet und müssen wieder durch langwierige Suchen in diversen Quellen beantwortet werden. Hier kommen die Leitsätze aus Abschnitt 4.2, »Best Practices für die Einführung eines KI-Assistenten«, zum Tragen und führen Sie direkt zum nächsten Use Case. Generative Assistenz bedeutet eine Unterstützung der Fachanwenderin oder des Fachanwenders innerhalb des direkten Arbeitsumfelds. Die Möglichkeiten der Implementierung eines KI-Assistenten sind deutlich flexibler und mannigfaltiger – insbesondere im Bereich der User Interfaces: Es muss nicht immer Joule sein. SAP-Frontends wie SAP Fiori oder auch SAP Workzone lassen eine Vielzahl von Implementierungsmöglichkeiten zu. Betrachten wir hier zunächst den klassischen Chatbot, wie man ihn auch von ChatGPT kennt.

Das Herzstück einer generativen Assistenz mit Zugriff auf unternehmensinterne Wissensinformationen ist eine Vector-Datenbank. Im Gegensatz zu rein generativen Sprachmodellen wie ChatGPT, die ihre Antworten ausschließlich aus ihrem Trainingswissen ableiten, kombiniert RAG die Fähigkeiten eines LLM mit einem externen, dynamisch aktualisierbaren Wissensspeicher. RAG und seine Funktionsweise werden Sie in Kapitel 8, »Personalisierung Ihres KI-Systems mit Retrieval-Augmented Generation«, genauer kennenlernen. Dieser Wissensspeicher besteht typischerweise aus einer Vektor-Datenbank, in der relevante Inhalte in Form semantischer Embeddings abgelegt sind. Bei jeder Nut-

zerfrage erfolgt eine semantische Suche innerhalb dieser Vektor-Datenbank, um die relevantesten Dokumente zu extrahieren. Diese Inhalte werden dem LLM zur Verfügung gestellt, das auf dieser Grundlage eine maßgeschneiderte, kontextuell angereicherte Antwort formuliert. Die Umsetzung einer solchen Lösung im SAP-Kontext stützt sich auf die folgenden zentralen Komponenten der SAP BTP:

1. Den Ausgangspunkt bildet eine Benutzeroberfläche, z. B. basierend auf SAP Fiori, Benutzerinnen und Benutzer können sich darüber am KI-Assistenten-Chatinterface anmelden, ihre Fragen stellen sowie ihren Chatverlauf verwalten (anlegen, ändern, löschen).
2. Die Eingaben (also die Fragen) werden an eine auf dem SAP Cloud Application Programming Model basierende Backend-Anwendung weitergeleitet, die als Orchestrierungseinheit fungiert. Hier erfolgt die semantische Vektorisierung der Eingabe, beispielsweise mittels eines Foundation Model, das über den Generative AI Hub zugänglich gemacht wird.
3. Anschließend wird die Anfrage an eine Vektor-Datenbank weitergereicht, (z. B. SAP HANA Cloud Vector Engine), in der die zuvor eingebetteten Dokumente gespeichert sind. Über Konnektoren können diverse Datenquellen (hier im Beispiel Dokumente aus Confluence) angebunden werden, deren Informationen extrahiert und über die Vektor-Datenbank als Wissenskontext für das LLM zur Verfügung gestellt werden.
4. Die relevanten Wissensinhalte werden daraufhin extrahiert und zur Generierung einer Antwort an das LLM übergeben, die Benutzerinnen und Benutzer erhalten hiermit kontextuell korrekte Antworten auf Basis des Unternehmenswissens.

Die Sicherheit und Steuerung des Zugriffs erfolgt durch zentrale Identity-Services wie SAP Authorization and Trust Management Service, sodass unternehmensspezifische Berechtigungen eingehalten werden können. Dadurch lassen sich nicht nur die Zugriffe auf Datenquellen generell steuern, sondern durch die Anmeldung der Benutzer an der SAP-Fiori-Oberfläche lassen sich die Berechtigungen bis in die Endsysteme hinein mit übernehmen, sodass sich über eine standardisierte SAP-Fiori-Oberfläche für einen Chat-Assistenten auch leicht weitere Skalierungen, u. a. für Transaktions- oder Serviceaufrufe, in ein SAP-Backend compliant und sicher realisieren lassen. Abbildung 4.5 zeigt ein Beispiel für eine solche SAP-Fiori-Chat-Oberfläche.

KI-Assistenten dieser Art eignen sich hervorragend für die Integration in bereits bestehende digitale Arbeitsplätze von SAP-Fiori-Umgebungen oder SAP Build Workzone. Durch die Flexibilität im Rahmen der Custom-AI-Strukturen sind einer Skalierung im Sinne von weiteren Funktionen, weiteren Wissensquellen keine Grenzen gesetzt. Wie aber lässt sich generative Assistenz sinnvoll integrieren, wenn entweder das Anwendungsumfeld oder die verwendeten technischen Geräte die Limitierung darstellen? Nicht jedes Arbeitsumfeld ist von einem Desktop-zentrierten Arbeitsplatz geprägt.

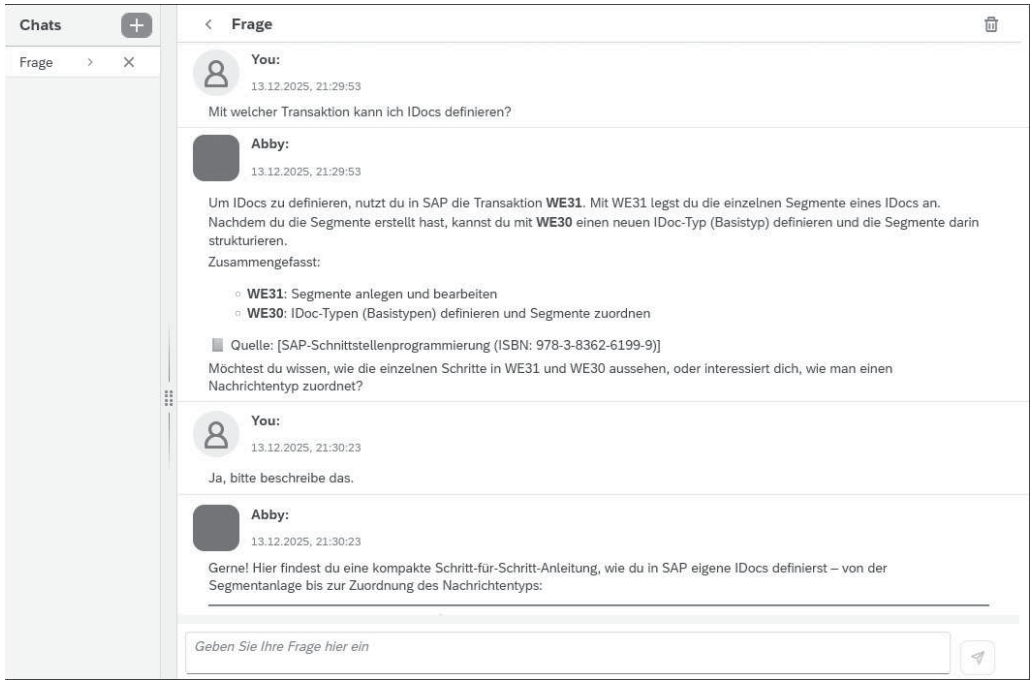

Abbildung 4.5: Beispiel für eine SAP-Fiori-Chatoberfläche

Jede Industrie hat ihre Besonderheiten, beispielsweise die folgenden:

- Im Handel sind viele Mitarbeitende auf der Fläche oder im Lager unterwegs, ihre mobilen Endgeräte sind oft Android-basiert, mit einer Auflösung kaum größer als 640×320 Pixeln, Fragen auf einer so kleinen Tastatur einzutippen, wäre suboptimal.
- In Industrie oder Forschung werden teils Handschuhe getragen und erschweren gegebenenfalls die Eingabe.
- Im Handwerk und bei Außeneinsätzen sind Desktopgeräte nicht verfügbar, mobile Endgeräte stehen hier im Fokus.

Generell gilt für eine erfolgreiche KI-Strategie im Unternehmen: Eine User-Akzeptanz wird nur dann erfolgreich erreicht, wenn die KI zum User gebracht wird, nicht der User zur KI. Das heißt, Sie müssen zuerst verstehen, welche Anforderungen das Arbeitsumfeld an die UI stellt, dann kann die richtige Technologie gewählt werden, alles weitere – das KI-Backend – kann flexibel und wiederverwendbar gestaltet sein mit dem Ziel, jedes mögliche User Interface gleichermaßen zu unterstützen. In Abbildung 4.6 sehen Sie eine weitere Option für ein User Interface – basierend auf derselben Basisarchitektur wie vorne beschrieben.

Abbildung 4.6: Mobiler KI-Assistant

Mobile Endgeräte unterliegen anderen Anforderungen als Desktop-Anwendungen, insbesondere wenn der digitale Arbeitsplatz selbst aus einer Vielzahl von Anwendungen besteht. Über sogenannte *Floating Bubbles* lässt sich einfach ein stets verfügbarer Zugang für die Benutzerinnen und Benutzer realisieren, eine Eingabeoption über ein Mikrofon als Speech-to-Text-Erkennung ersetzt mühsame Eingaben auf kleinen Mobiltelefon-Tastaturen, und Mehrsprachigkeit durch die Spracheingabe ist ein weiteres Plus für die Akzeptant der Benutzerinnen und Benutzer. Gerade in KI-Anwendungsfeldern für generative Assistenz sind die generellen Skalierungsmöglichkeiten der Lösung ein großer Vorteil.

Durch die standardisierte SAP-Fiori-Oberfläche mit ihren integrierten nativen SAP-Funktionen gerade im Hinblick auf Berechtigungen und User-/Rollenmanagement lassen sich kundenindividuelle User Interfaces besonders gut skalieren, ob in der Einbettung eines digitalen Arbeitsplatzes in Chat- oder Bubble-Form oder auf der Browser-Oberfläche als User Interface mit einem Ticket-System, sie lassen sich flexibel implementieren, durch ein Rollen- und Berechtigungskonzept in vielerlei Ausprägung für Anwendungen im gesamten Unternehmenskontext verwenden und schaffen so eine einheitliche Benutzererfahrung.

Durch die Flexibilität in der Einbindung des CAP-Frameworks für die Orchestrierung mit allen verfügbaren SAP BTP Services, lassen sich sowohl Fragen an LLMs zu unternehmensindividuellem Wissen (via RAG, SAP HANA Cloud Vector Engine) als auch Workflows (über SAP Build Process Automation) flexibel und skalierbar einbinden.

Somit bietet sich gerade im Umfeld von SAP-Custom-AI-Anwendungen ein großes Potenzial außerhalb von Joule bzw. schließt es die Lücken zwischen SAP- und Nicht-SAP-Anwendungen im KI-Kontext.

4.3.3 Personaleinsatzplanung

Nachdem Sie nun einen umfassenden Einblick in die Anwendungsfelder »Automatisierung« und »generative Assistenz« erhalten und deren Anwendung im Detail auch auf das eigene Unternehmen verstehen und anwenden gelernt haben, lassen Sie uns abschließend noch das Feld »Entscheidungsassistenz« beleuchten. Wie der Begriff bereits impliziert, geht es hier um datengetriebene Entscheidungen. Wenn Sie sich an den Leitsätzen aus Abschnitt 4.2, »Best Practices für die Einführung eines KI-Assistenten«, orientieren, geht es hier um Fragen an Daten, die wir als Menschen kognitiv nicht mehr im zweidimensionalen Raum wahrnehmen bzw. verarbeiten können. Diese Thematik wird von ML abgedeckt. Vielleicht gibt es in Ihrem Unternehmen Planungsprozesse im Lager oder auch in Produktion bzw. Fertigung. Es gibt Arbeitspläne und möglicherweise Besonderheiten in der Planung, wie saisonale Peaks oder Ähnliches. Neben den üblichen Schichtplänen (Soll-Zustand) gibt es noch die Realität (Ist-Zustand), in der häufig noch auf Basis von Excel versucht wird, die Realität der Verfügbarkeit von Mitarbeitenden mit ihren Arbeitszeitmodellen, Schichten, krankheitsbedingten Ausfällen etc. in einen realistisch umsetzbaren Planungsprozess zu bringen. Betrachten wir diese Thematik am Beispiel des folgenden Use Case – der Personaleinsatzplanung.

Auf Basis von Excel-Daten erfolgt eine Mitarbeiterplanung auf Wochenbasis, jeweils für die Folgewoche. Excel kommt dort an seine Grenzen, wo keine Formel mehr einen Zusammenhang zwischen den aktuellen und den Zahlen der Vorjahre ziehen kann. Prognosen sind ein idealer KI-Use-Case für die Anwendung von ML – es benötigt allerdings eine Vielzahl an Daten. Auf Basis von Planzahlen über zwei und mehr Jahre hinweg lässt sich ein Modell trainieren. Auf Basis der Daten lässt sich dann eine Vorhersage für die Wochenplanung vollautomatisiert durchführen. Hierzu ist die SAP BTP die ideale Plattform, da sich durch die ML-Partner von SAP – beispielsweise AWS – die Daten harmonisieren und standardisieren lassen und sie über SAP Datasphere in ein ML-Modell gebracht werden können. Das Modell wird auf diesem ML-Partner trainiert und kann anschließend auf dem SAP AI Core deployt werden. Der Modell-Output, die Vorhersage, kann über eine entsprechende UI entweder als SAP-Fiori-Applikation oder als Story/Dashboard in der SAP Analytics Cloud visualisiert werden. Sie können auch einem LLM aus dem Generative AI Hub Zugriff auf dieses Modell geben, um eine Interaktion in natürlicher Sprache zu ermöglichen. Alternativ könnten Sie auch mit der *SAP Business Data Cloud* eine inte-

grierte ML-Funktionalität mit SAP Databricks realisieren. Nachdem wir mit generativer Assistenz eine sehr sichtbare Form von KI kennengelernt haben, gibt dieser Use Case einen guten Einblick darin, wie viel KI auch im Hintergrund laufen kann, indem die Entscheidungsassistenz einen wichtigen Bestandteil im Verknüpfen von Prozessen durch die Verarbeitung von Daten und Bereitstellung passender Prognosen bietet. KI fügt sich dort am besten nahtlos in Prozesse ein, wo sie Prozesslücken nicht nur verbindet, sondern mit Mehrwert im Sinne von Datenverständnis und Entscheidungs-Output bietet. Überall dort, wo Transaktionen im System echten Einfluss auf die Gestaltung von Business-Mehrwert nehmen, ist KI ein wertvoller Motor im Geschäftsprozess. Das Anwendungsfeld, das Sie hier kennengerlernt haben, lässt sich damit auf viele ähnliche Geschäftsprozesse anwenden, u. a. die folgenden:

- im Einkauf, z. B. in der Anwendung für Vorschläge für Konditionen im Rahmen von Vertragsverhandlungen (auf Basis historischer Daten in Bezug auf die besten Konditionen oder die lohnendsten Rahmenverträge)
- im Verkauf, in Bezug auf Preisvorschläge im Vergleich zu internen Kalkulationen (und der eigenen Margenoptimierung)

Gerade im SAP-Umfeld mit seiner Vielzahl an Prozessen lohnt es sich, genau hinzuschauen, insbesondere im Bereich Analytics, d. h. ganz konkret: Welche Datenfelder werden derzeit gegebenenfalls gar nicht analysiert oder in Korrelation gesetzt? Hier bietet ML große Potenziale als Integration in die Wertschöpfungskette.

SAP Architecture Center

Alle vorgestellten KI-Use-Cases basieren auf den Referenzarchitekturen von SAP, siehe dazu auch: *https://architecture.learning.sap.com/*. Hier finden Sie die passende Referenzarchitektur für jeden Anwendungsfall und können im SAP Architecture Center auch entsprechend filtern, z. B. nach »Generative AI on SAP BTP«, aber auch vielen weiteren Anwendungsfeldern. Die Solution-Diagramme lassen sich als Drawio-Datei oder Powerpoint-Dateien speichern und wiederverwenden.

4.4 Zusammenfassung

In diesem Kapitel haben Sie nun einen Überblick über das SAP-KI-Portfolio im Bereich der integrierten KI-Funktionalitäten mit Embedded AI erhalten. Dabei haben Sie viele Anwendungsfelder kennengelernt und gelernt, wie Sie SAP-Tools in unterschiedlichen Bereichen, wie Logistik, Finanzwesen, Kundenservice, HR und Softwareentwicklung, nutzen können, um Ihre Prozesse mithilfe von generativer KI zu optimieren. In Abschnitt 4.2, »Best Practices für die Einführung eines KI-Assistenten«, dieses Kapitels haben Sie unsere Methodik für die Bewertung und Identifizierung individueller Anwendungsfälle kennengelernt und mit dieser ein eigenes Toolset gewonnen, mit dem Sie in Zukunft sicher Ihre

unternehmenseigenen Anwendungsfälle identifizieren, bewerten und bzgl. der richtigen Technologie die passende Architektur im SAP-Umfeld auswählen können. Gerade, wenn Sie sich noch nicht im Cloud-ERP-Umfeld von RISE oder GROW bewegen, haben Sie mit den Möglichkeiten von Custom AI auf der SAP BTP eine Vielzahl flexibler Varianten, mit generativer KI einen Mehrwert zu erzeugen.

Kapitel 5

Ihr Handwerkszeug für die KI-Entwicklung: SAP AI Core und SAP AI Launchpad

In diesem Kapitel stellen wir Ihnen die SAP-Umgebung für die KI-Entwicklung im Detail vor. Sie lernen die Architektur, ihre Komponenten und deren Zusammenspiel kennen.

Nachdem Sie nun einen Überblick über das KI-Portfolio von SAP bekommen haben und sowohl integrierte als auch individuelle KI-Anwendungen kennengelernt haben, betrachten wir in diesem Kapitel die Kernkomponenten, den SAP AI Core und das SAP AI Launchpad. In Abschnitt 5.1, »Architektur: Das SAP Generative AI Reference Model«, stellen wir Ihnen die SAP-Referenzarchitektur für generative KI auf der SAP BTP vor. In Abschnitt 5.2, »Funktionen von SAP AI Core und dem SAP AI Launchpad«, lernen Sie die Kernaufgaben von SAP AI Core und SAP AI Launchpad kennen. Wir vergleichen die verschiedenen Service-Pläne in Abschnitt 5.3 und beleuchten in Abschnitt 5.4 die Funktionen der Laufzeitumgebungen auf der SAP BTP. Abschließend gehen wir in Abschnitt 5.5 auf das Sizing und die Lizenzen ein.

5.1 Architektur: Das SAP Generative AI Reference Model

SAP hat als Grundlage für die strategische, technische und betriebliche Integration generativer KI in die SAP BTP eine klar definierte Referenzarchitektur entworfen (siehe Abbildung 5.1), das *SAP Generative AI Reference Model*, das Sie im vorherigen Kapitel schon als Framework des *SAP Architecture Center* kennengelernt haben (siehe *http://s-prs.de/v10759017*).

Mit der Einführung von SAP Business AI und der kontinuierlichen Erweiterung der SAP BTP um KI-spezifische Services verfolgt SAP das Ziel, KI als integralen Bestandteil des digitalen Kerns ihrer Produkte zu verankern. Dabei geht es nicht nur um das Einbetten einzelner Modelle, sondern um den Aufbau einer nachhaltigen, wiederverwendbaren und regelkonformen Architektur, die es erlaubt, generative KI systematisch und skalierbar in Geschäftsprozesse einzubinden Das SAP Generative AI Reference Model bildet die architektonische Basis aller KI-Anwendungen auf der SAP BTP. Sie bietet einen standardisierten

Rahmen zur Modellintegration, Datenanbindung, Prozesssteuerung und Ausgestaltung der User-Experience. Dabei integriert sie sowohl SAP-eigene als auch Drittanbietermodelle wie beispielsweise OpenAI, Llama oder Mistral über den Generative AI Hub und erlaubt deren sichere und kontextuelle Nutzung in Kombination mit SAP-Kerndaten. Das SAP Generative AI Reference Model (siehe Abbildung 5.1) folgt einem modularen Schichtenmodell, das entlang der typischen IT-Systemarchitektur in vier Hauptbereiche gegliedert ist:

1. Präsentation (UI/UX)
2. Orchestrierung und Integration
3. Verarbeitung durch KI-Modelle
4. Datenhaltung

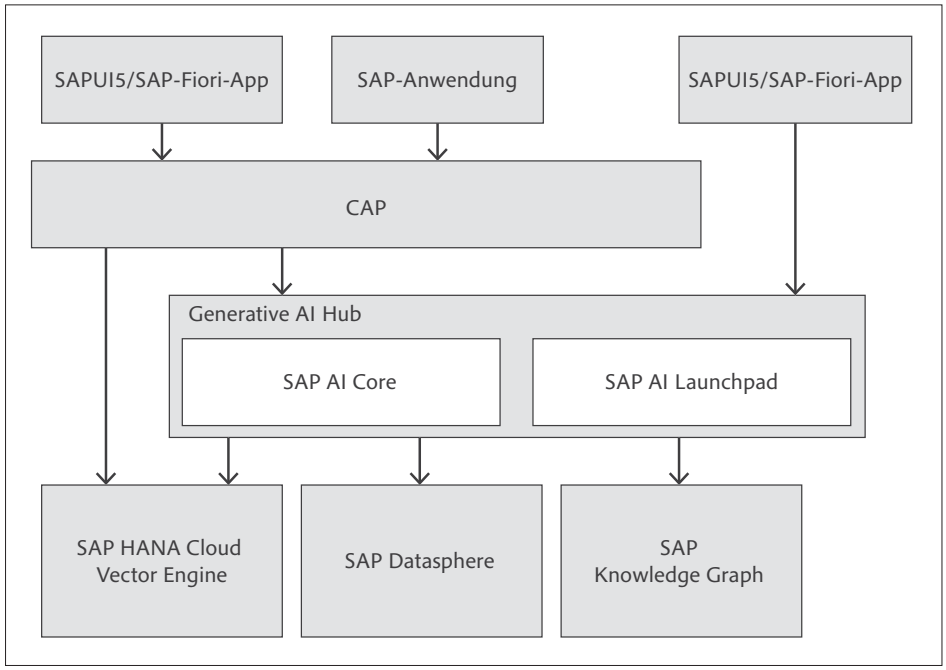

Abbildung 5.1: Das SAP Generative AI Reference Model für generative KI auf der SAP BTP

Dieses strukturierte Vorgehen ermöglicht es, unterschiedliche Technologien, Datenquellen und Modelle miteinander zu kombinieren, ohne auf monolithische Ansätze zurückzugreifen. Stattdessen steht die Interoperabilität im Vordergrund. Wie in der Abbildung dargestellt, besteht die Präsentationsebene (in der Grafik der obere Bereich) aus den, bereits in Kapitel 4 beschrieben User Interface Optionen, wie z. B. SAPUI5 oder einem SAP-Fiori-Chatinterface (für »Custom AI«). Da die Präsentationsebene eine Vielzahl möglicher Benutzeroberflächen umfassen kann, u. a. auch externe Drittanbieter-Oberflächen, in die die generierten Inhalte eingebunden werden können, lassen sich dadurch sowohl klassische

transaktionale Prozesse als auch moderne, konversationsbasierte Interaktionen in einer intuitiven Benutzererfahrung umsetzen. Die Orchestrierungs- und Integrationsschicht basiert in der Regel auf dem SAP Cloud Application Programming Model (CAP) sowie auf Erweiterungen durch externe Frameworks wie LangChain (siehe: *https://www.langchain.com/*). Sie dienen dazu, die Interaktionsflüsse zwischen Nutzeranfragen, Daten-Retrieval und Modellantwort zu steuern und die Prompts vor ihrer Weiterleitung an das jeweilige Modell aufzubereiten. Die Entwicklung generativer Anwendungen im SAP-Umfeld erfolgt bevorzugt im SAP Cloud Application Programming Model. Es bietet eine strukturierte Umgebung für die Definition von Datenmodellen, Services und Authentifizierungsmechanismen. In Kombination mit generativen Komponenten wird das SAP Cloud Application Programming Model um Funktionen wie Prompt-Generierung, Inferenzsteuerung und Dokumenten-Retrieval erweitert. Entwicklerinnen und Entwickler nutzen dabei zumeist Open-Source-Komponenten wie LangChain, vektorbasierte Libraries sowie das von SAP bereitgestellte SDK (Software Development Kit) für den Generative AI Hub.

SAP SDKs

SAP veröffentlicht im Rahmen der SAP-Help-Dokumentation alle aktuellen Entwicklerframeworks unter *http://s-prs.de/v10759018*. Für den SAP AI Core gibt es sowohl Python- als auch Java-Bibliotheken.

Ein besonderes Augenmerk des SAP Cloud Application Programming Models liegt auf der Mandantenfähigkeit (Multitenancy), da viele generative Lösungen im Kontext von Partnerentwicklungen oder als Erweiterung für SAP-Kunden bereitgestellt werden. Das SAP Cloud Application Programming Model bietet hier mit tenant-spezifischer Isolation und dynamischer Datenmodellierung eine flexible Grundlage, um unterschiedliche Kundenkontexte innerhalb einer Architektur zu bedienen. Dies wird durch für das SAP Cloud Application Programming Model typische Konzepte wie Extensible Entities und Service-Bindings unterstützt.

Im Bereich der KI-Verarbeitung kommen Foundation Models zum Einsatz, die über den Generative AI Hub orchestriert und verwaltet werden. Die Komponente der Foundation Models ist vollständig modular aufgebaut und erlaubt es, verschiedene Sprachmodelle je nach Anwendungsfall flexibel auszutauschen oder auch parallel zu betreiben. Der Zugriff auf diese Modelle erfolgt dabei meist über REST-basierte Schnittstellen oder vordefinierte SDKs, die SAP zur Verfügung stellt. Eine zentrale Rolle nimmt hier der SAP AI Core ein, der als Laufzeitumgebung für die Durchführung von Inferenzaufträgen dient und darüber hinaus für das Deployment, Monitoring und die Rückverfolgbarkeit der KI-Modelle verantwortlich ist. Das SAP AI Launchpad dient als Benutzeroberfläche zur Bedienung des SAP AI Core.

Zu guter Letzt umfasst die Ebene der Datenhaltung (ganz unten in Abbildung 5.1) typischerweise die SAP HANA Cloud Vector Engine, die SAP Datasphere sowie für semanti-

sche Repräsentationen den SAP Knowledge Graph. Die SAP HANA Cloud Vector Engine bietet den Kontext für die KI-Modelle, die darin enthaltenen Informationen werden mithilfe moderner Embedding-Technologien in Vektoren überführt, um sie für die spätere semantische Suche mittels Retrieval-Augmented Generation (RAG) nutzbar zu machen. Dieser Mechanismus ist essenziell, um generative Modelle mit kontextrelevanten, geschäftlichen Informationen zu versorgen, ohne diese direkt im Modell trainieren zu müssen. Zudem bietet die Data Fabric mit der SAP Business Data Cloud und der SAP Datasphere ein weiteres wichtiges Daten-Fundament auf der SAP BTP als semantische Ebene für alle KI- und Agenten-Szenarien.

5.1.1 Sicherheits- und Governance-Mechanismen im SAP Generative AI Reference Model

In das SAP Generative AI Reference Model sind umfassende Sicherheits- und Governance-Mechanismen eingebunden. Außerdem hat SAP mit dem *AI Governance Framework* eine strukturierte Grundlage zur Bewertung, Kontrolle und Auditierung von KI-Einsätzen gelegt. Das Framework umfasst die Verwaltung von Modellzugriffen, die Protokollierung von Inferenzanfragen sowie die Durchsetzung von Richtlinien zur Erklärbarkeit und Fairness. Gerade im europäischen Raum ist die Einhaltung regulatorischer Vorgaben wie der Datenschutz-Grundverordnung (DSGVO) oder des EU AI Acts (siehe Kapitel 13, »Rechtliche Rahmenbedingungen«) entscheidend. Das SAP Generative AI Reference Model berücksichtigt diese Anforderungen durch eine klare Trennung zwischen Inferenz und Retrieval, sodass keine sensiblen Daten an externe Foundation Models weitergegeben werden müssen. Zudem ermöglicht die Nutzung von On-Premise- oder Private-Cloud-Modellen eine zusätzliche Sicherheitsstufe. Mit SAP Cloud Identity Services, Logging- und Monitoring-Komponenten sowie Audit-Trails bietet die Architektur zudem eine umfangreiche Grundlage zur technischen Umsetzung von Nachvollziehbarkeit und Verantwortlichkeit im KI-Einsatz. Ergänzt wird dies durch ethische Leitlinien, die SAP im Rahmen von *Responsible AI* definiert hat und die im Modell sowie in der SAP KI Ethics Policy (siehe *http://s-prs.de/v10759019*) explizit reflektiert sind. Das SAP Generative AI Reference Model bildet somit eine tragfähige, flexible und sichere Grundlage für den unternehmensweiten Einsatz generativer KI in SAP-Landschaften. Sie vereint die Stärken der SAP BTP mit leistungsfähigen Foundation Models, semantischer Datenerschließung und modernen Entwicklungsparadigmen wie SAP Cloud Application Programming Model und LangChain. Die Architektur des SAP Generative AI Reference Model ist skalierbar, mandantenfähig und anpassbar an unterschiedliche regulatorische Anforderungen, wodurch sie eine zentrale Rolle für die nächste Phase der digitalen Transformation einnimmt.

5.1.2 Technischer Ablauf einer Interaktion mit dem KI-Modell

In diesem Abschnitt möchten wir Ihnen den technischen Ablauf, mit dem generative KI-Anwendungen innerhalb der SAP BTP orchestriert werden, vorstellen.

Dieser Prozess beginnt stets mit einer Nutzereingabe über eine entsprechende Benutzeroberfläche, typischerweise eine SAP-Fiori-App oder eine Anwendung, die aus SAP Build Apps initiiert wird. Die Eingabe, die in der Regel in natürlicher Sprache erfolgt, wird über standardisierte Schnittstellen an ein Backend-System auf der SAP BTP übergeben, das auf dem SAP Cloud Application Programming Model basiert.

Schritt 1: Vorverarbeitung der Anfrage

Innerhalb dieses Backends erfolgt zunächst eine Vorverarbeitung der Anfrage. Diese umfasst unter anderem eine strukturelle Validierung, das Logging der Anfrage sowie die Erzeugung eines initialen Prompts. Dieser Prompt wird mithilfe definierter Vorlagen (*Prompt Templates*) erstellt, die in der Lage sind, sowohl die Nutzereingabe als auch zusätzliche Kontextinformationen dynamisch zu integrieren. Anschließend wird die vorbereitete Anfrage an die Steuerungskomponente für RAG weitergeleitet. Diese Steuerung wird in der Praxis häufig durch das Framework LangChain realisiert, das innerhalb der Logik des SAP Cloud Application Programming Model eingebettet ist.

Schritt 2: Semantische Anreicherung des Prompts

Der nächste Schritt ist die semantische Anreicherung des Prompts mit unternehmensinternen Kontextdaten. Hierzu wird eine vektorbasierte Suche auf der SAP HANA Cloud Vector Engine durchgeführt. Diese Vektorsuche basiert auf zuvor eingebetteten Unternehmensdokumenten, wie etwa Verträgen, Produktbeschreibungen oder historischen Transaktionen. Die semantisch relevantesten Ergebnisse werden als Kontextbausteine extrahiert und dem ursprünglichen Prompt hinzugefügt.

Schritt 3: Weiterleitung an das Foundation Model

Die nun vollständig kontextualisierte Anfrage wird anschließend an ein Foundation Model weitergeleitet. Diese Modelle, etwa GPT-4 von OpenAI, Llama von Meta oder SAP-eigene Sprachmodelle, sind über den Generative AI Hub angebunden und können wahlweise in einer öffentlichen Cloud oder innerhalb einer geschützten, privaten Umgebung betrieben werden. Die Übergabe an das Modell erfolgt über eine standardisierte Schnittstelle und wird durch den SAP AI Core orchestriert, der für das technische Inferenzmanagement, die Ressourcensteuerung und das Monitoring verantwortlich ist.

Schritt 4: Verarbeitung im Foundation Model und Übermittlung zurück ans Backend

Nach erfolgreicher Inferenz, d. h. nach der Verarbeitung der Anfrage durch das gewählte Modell, wird die Antwort zurück an das SAP-Cloud-Application-Programming-Model-Backend übermittelt. Dort findet ein optionales Post-Processing statt, in dem beispielsweise die Einhaltung von Formatvorgaben überprüft, systeminterne Links oder Metadaten ergänzt oder technische Validierungen vorgenommen werden. Je nach Szenario kann die Antwort auch persistiert, versioniert oder in weiterführende Workflows überführt werden.