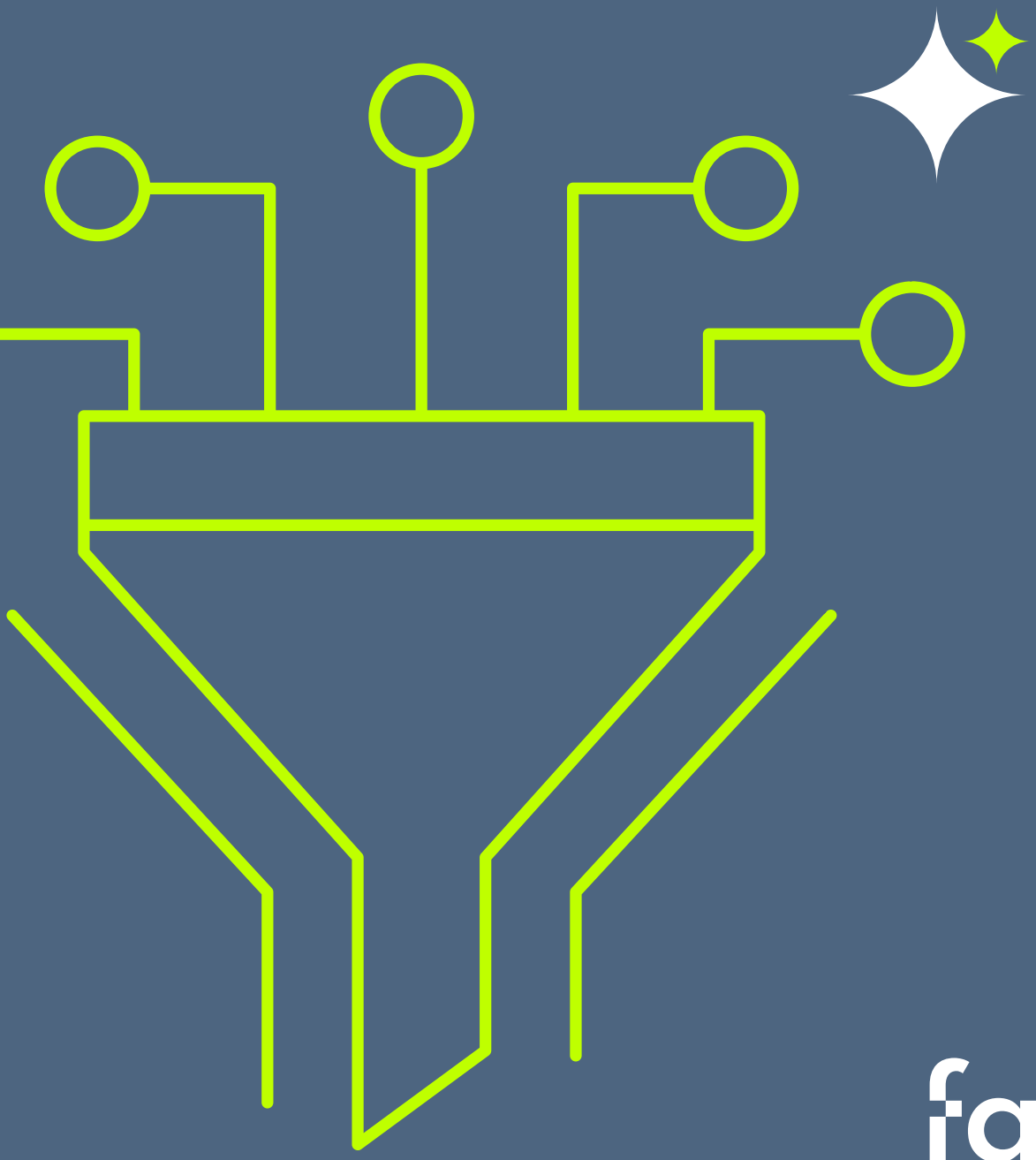
DAS BESTE AUS ZWEI WELTEN: DIE RAG-PIPELINE IM COMPANION FOR FIRST

Wo Sprachintelligenz und Datengenauigkeit zusammenkommen



fact digital uncomplicators

Wenn die Software zum Sparringspartner wird

Bitte anschnallen, die Reise Richtung Zukunft ist in vollem Gange.

Derzeit entwickeln wir für First Cloud ein optionales Modul, das das Beste aus der KI-Welt in den Arbeitsalltag der institutionellen Kapitalanlageverwaltung bringen wird. Mit dem Companion for First können unsere Kunden ihre First Cloud-Instanz auf Wunsch um das Verständnis von Sprache und leistungsfähige, agentische KI-Funktionen ergänzen. Erstmals wird unser Erfolgsprodukt damit in die Lage versetzt, schriftliche Fragen in natürlicher Sprache entgegenzunehmen und zu beantworten.

Sie möchten wissen, welche Aktien in Ihren Kundenportfolios derzeit unter Einstandsniveau notieren, aktuelle Bewertungen für jede dieser Gesellschaften erhalten, oder benötigen einen Tipp, wo Sie eine bestimmte Funktion in der Benutzeroberfläche finden? Dann fragen Sie den Companion doch einfach, wie es Ihnen gerade auf der Zunge liegt! Langwierige Suchen in der Menüstruktur oder das Aufsetzen komplexer SQL-Abfragen für den Datenabruf gehören damit der Vergangenheit an.

Mit dem Companion for First wird Ihnen ein intelligenter, dialogfähiger Assistent und Sparringspartner zur Seite stehen – immer bereit, Nutzer bei der Verwaltung von Kapitalanlagen in First Cloud fundiert, regulatorikkonform und zukunftssicher zu unterstützen.

FIRST CLOUD

First Cloud ist die bewährte Branchenlösung der Fact für die umfassende Verwaltung von Kapitalanlagen durch institutionelle Anleger.

Bestandsführung, Buchhaltung, Solvency II, Meldewesen, Risikocontrolling, ESG – alles inklusive und dank modernster Cloud-Technik nicht an spezifische Endgeräte, speziell eingerichtete Arbeitsplätze oder enge Betriebszeiten gebunden.

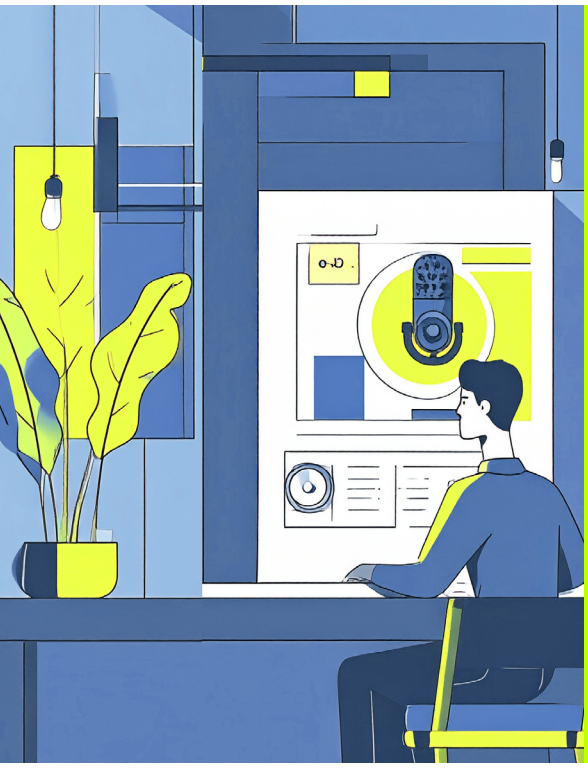
Alles, was man für die Nutzung von First Cloud benötigt, ist ein herkömmlicher Internet -Zugang und ein moderner Web-Browser.

Als Agentische KI hat der Companion das Ziel, seine Nutzer bei wiederkehrenden Aufgaben ebenso zu entlasten wie bei strategischen Bewertungen.

Zum Aufgabenfeld des Companions werden beispielsweise gezielte Analysen, verständlich aufbereitete Visualisierungen und automatisierte Berichte gehören. In der nächsten Ausbaustufe wird der Companion dann auch aktuelle Fachinformationen aus Finanznachrichtenquellen auswerten und automatisch zur Verfügung stellen, sowie kontextbezogene Hinweise und systemgestützte Erklärungen zu komplexen Funktionen in First Cloud liefern.

Mit diesem Whitepaper setzen wir unsere Reihe über KI-Technik und ihren Einsatz im Companion for First fort. Der Blick hinter die Kulissen ist dabei kein Selbstzweck, sondern will vermitteln, wie der Companion essenzielle Anforderungen der Kapitalanlageverwaltung von Grund auf berücksichtigt: etwa den Wunsch nach Präzision, Verlässlichkeit und Wiederholbarkeit, den Schutz von Kundendaten, die Erfüllung regulatorischer Anforderungen und vieles mehr.

Wir beginnen mit dem Rückgrat des Companion, der sogenannten „RAG-Pipeline“. Sie orchestriert den Informationsfluss von der umgangssprachlichen Anfrage an den Companion bis hin zum gewünschten Output.



KI VERSTEHEN

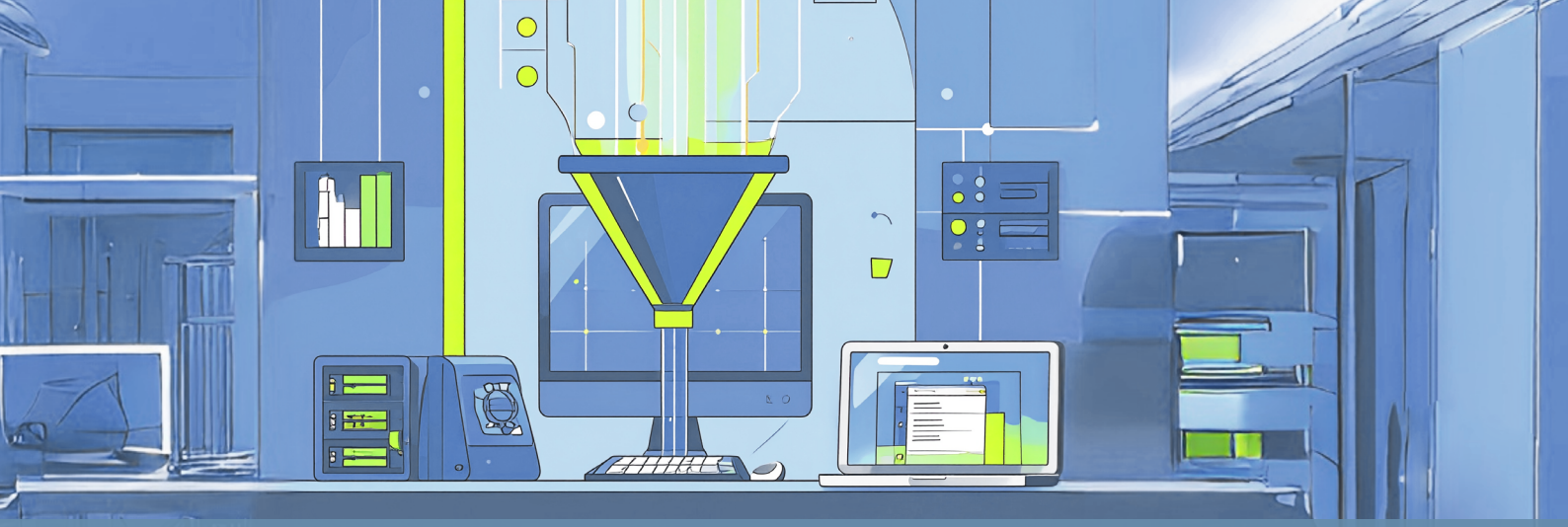
In dieser Reihe bisher erschienen:

→ **Grundlagen der KI-Technik**

→ **Die RAG-Pipeline im Companion for First**

Sie möchten mehr erfahren?
Alle Infos zum Companion for First auf einen Blick gibt es hier.

Besuchen Sie auch den Fact Blog.
Dort finden Sie stets aktuelle Artikel zum Thema, etwa ein Interview mit dem technischen Projektleiter für die Entwicklung des Companion for First, Norman Janert.



Aufgaben einer RAG-Pipeline

Was haben Chatbots, Übersetzungssysteme, Sentiment-Analyse (Ermittlung von Nutzerstimmungen) und Assistenten wie der Companion for First gemeinsam? Sie alle sind auf die intelligente Verarbeitung von Sprache durch Large Language Models (LLM) angewiesen. So nennt sich die KI-Funktion, die hinter Systemen wie ChatGPT von OpenAI, Claude von Anthropic oder Googles DeepMind steckt.

Diese Modelle kombinieren die Fähigkeit zur Sprachanalyse mit antrainiertem Weltwissen und können dadurch Antworten auf vielfältige Fragestellungen liefern. Allerdings nicht auf Wissen, das in spezifischen Anwendungen wie First Cloud verborgen liegt oder aus anderen Gründen nicht für das Training dieser Modelle zur Verfügung steht. Denn aus sich selbst heraus wissen diese Systeme zunächst einmal gar nichts.

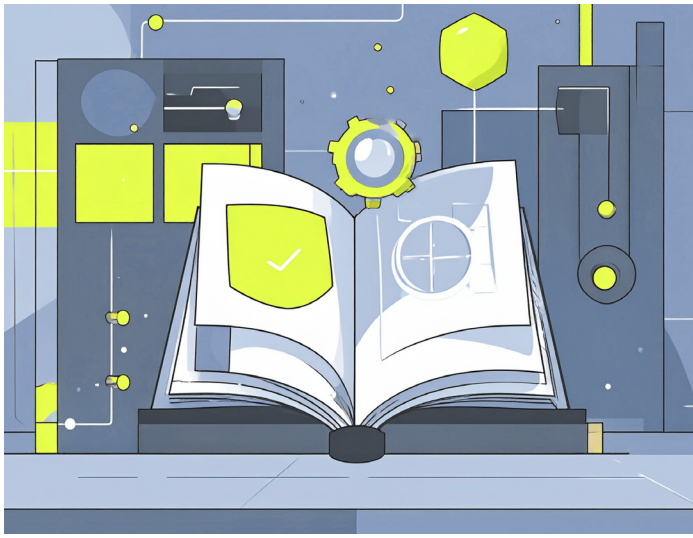
RAG steht für „Retrieval-Augmented Generation“ und beschreibt ein Software-System, das den Informationsabruf (Information Retrieval) aus einer Spezialanwendung wie First Cloud mit einem Large Language Model kombiniert (augmentiert). Von einer Pipeline spricht man in diesem Zusammenhang, weil die Verarbeitung der Nutzeranfragen in klar abgegrenzten Prozessschritten erfolgt, die einander zuarbeiten und entlang einer fest vorgegebenen Verarbeitungskette nacheinander aufgerufen werden. Ganz am Ende steht dann der vom Nutzer gewünschte Output.

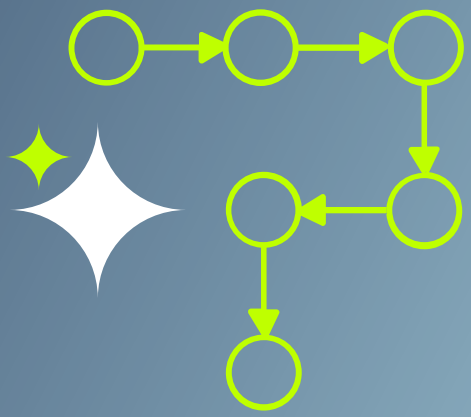
Das LLM hat dabei die Aufgabe, am Beginn der Pipeline die Anfrage des Nutzers in Bedeutungsinhalte aufzubrechen. Was wird hier verlangt, worum geht es, welche Zusatzinformationen über Art, Inhalt, Zeitraum etc. werden geliefert? Dazu wird das LLM aus der Pipeline heraus mit Spezialwissen aus dem jeweiligen Anwendungsbereich versorgt – nicht jedoch mit konkretem Detailwissen zu spezifischen Kundendaten oder anderen Interna, das ist ganz wichtig. Diese verbleiben stets innerhalb von First Cloud.

Herausforderungen an die RAG-Pipeline im Companion for First

Im Aufbau einer RAG-Pipeline spiegeln sich die spezifischen Herausforderungen und Merkmale einer Anwendung wider, die LLMs für die Sprachverarbeitung nutzt. Bei First Cloud sind dies:

- **Die deutsche Finanzsprache:** Es geht darum, dem LLM das Wissen über verschiedene Begriffe für identische Konzepte („Wertpapier", „Gattung", „Asset", „Kapitalanlage") und deutsche Sprachbesonderheiten wie Komposita („Apple-Aktien", „Öko-Fonds") mitzugeben – letztendlich alles, was die Nutzer von First Cloud an fachspezifischen Termini und Formulierungen einbringen könnten und über das Grundwissen einer LLM hinausgeht.
- **Komplexe Datenstrukturen:** First Cloud verwaltet die Kundendaten in tiefverschachtelten Strukturen mit hunderten von Datenbanktabellen, spezifischer Fachterminologie (Asset, Position, Movement, SCR und viele weitere) sowie individuellen, kundenspezifischen Konfigurationen. Eine externe LLM kann und soll diese Strukturen gar nicht kennen, deshalb muss der Zugriff darauf vollständig innerhalb des Companion for First erfolgen.
- **Regulatorische Anforderungen:** In der Finanzbranche können falsche Zahlen erhebliche juristische Konsequenzen und sogar Strafen nach sich ziehen. Der Companion for First muss über seine RAG-Pipeline deshalb höchste Genauigkeit, Nachvollziehbarkeit und Compliance (DORA, BaFin) gewährleisten. Das von manchen KI-Anwendungen bekannte gelegentliche „Halluziniere" hat hier keinen Platz.
- **Nachvollziehbarkeit:** Jede Antwort des Companion muss auf identifizierbaren und dokumentierten Quellen basieren (Tabellen, Dokumentation, Konfigurationen). Für Audit-Trails ist dies essentiell.
- **Performance und Skalierung**: Auch bei komplexen Analysen und gleichzeitiger Nutzung durch zahlreiche Anwender soll der Companion Antworten in kürzester Zeit liefern. Da innerhalb der RAG-Pipeline mehrere Module zu durchlaufen sind, spielen Performance und Optimierung dabei eine wichtige Rolle.

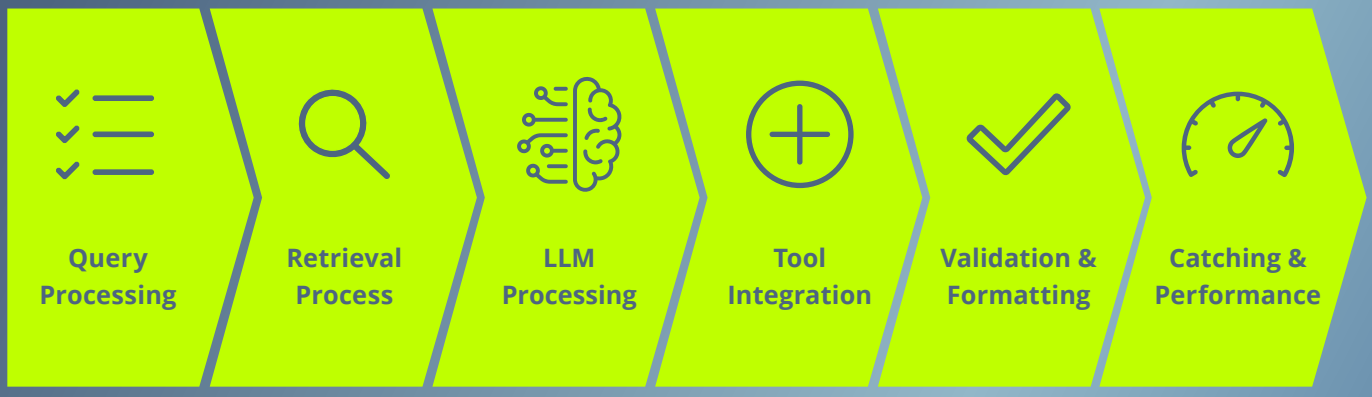




Komponenten der RAG-Pipeline im Companion for First

Sechs zentrale Komponenten bilden die RAG-Pipeline im Companion for First. Sie werden stufenweise durchlaufen und beliefern einander mit Informationen. Aus diesem Grund ist es nicht möglich, die verschiedenen Komponenten parallel auszuführen.

Es können aber sehr wohl mehrere RAG-Pipelines für unterschiedliche Nutzeranfragen und in mehreren Instanzen von First Cloud gleichzeitig aktiv sein. Zur Orchestrierung dieser Abläufe werden etablierte Software-Bibliotheken (Funktionssammlungen) wie LangGraph und LangChain genutzt, die sich für diesen Zweck bewährt haben.



1

Der Query-Prozess bildet den Einstieg in die Verarbeitungskette.
Darin wird die natürlichsprachige Benutzeranfrage in eine strukturierte und maschinenlesbare Form für die Weiterverarbeitung transformiert. Dieser Vorgang ist vollständig deterministisch, das heißt, er läuft immer nach denselben Regeln ab und liefert bei gleichem Input auch dieselben Ergebnisse (siehe unten).

2

Der Retrieval-Prozess dient der Wissensbeschaffung.
Basierend auf den strukturierten Queries aus dem vorhergehenden Verarbeitungsschritt werden Informationen aus den jeweils passenden Datenquellen ermittelt. Für die Suche nach Wortbedeutungen kommt beispielsweise eine interne Vektordatenbank mit bekannten Begriffen aus der Finanzwelt zum Einsatz. Solche Vektordatenbanken sind das Mittel der Wahl, wenn es um kontextbezogene semantische Ähnlichkeiten von Begriffen geht. „Aktien“, „Shares“ oder „Wertpapiere“ werden so beispielsweise auf das Konzept „Aktie“ reduziert, wenn der Kontext dies sinnvoll erscheinen lässt. Damit wird klar, worauf sich die weitere Verarbeitung in den nächsten Stufen der Pipeline beziehen soll.

Als Rückfallebene für nicht aufgelöste Begriffe dient ein externes LLM, beispielsweise wenn der Anwender abstrakt nach den „Top-DAX-Unternehmen“ fragt. Das LLM liefert dann im besten Falle eine aktuelle Liste dieser Unternehmen. Ist auch damit keine Auflösung möglich, muss vor der weiteren Verarbeitung beim Nutzer nachgefragt werden.

3

Den LLM-Prozess kann man sich wie das Gehirn der RAG-Pipeline vorstellen.
Aus dem zuvor gesammelten Kontext und der ursprünglichen Benutzeranfrage wird eine Frage an das LLM formuliert. Das LLM generiert daraufhin eine intelligente, handlungsorientierte Antwort. Sie beschreibt, was in den nachfolgenden Verarbeitungsstufen der Pipeline genau zu tun ist, um die Nutzeranfrage zu beantworten. Diese „To-Dos“ beziehen sich auf die Tools aus dem internen Werkzeugkasten von First Cloud. Für jede Aufgabe gibt es das passende Tool, und das LLM weiß durch den vom Companion übermittelten Kontext, welche dieser Werkzeuge für die jeweilige Anfrage genutzt werden können.

4

Mit der Tool-Orchestrierung gelangt man in die Ausführungsebene der RAG-Pipeline.
Die Antworten des LLM aus der vorherigen Stufe werden anhand der darin genannten Tools in konkrete Aktionen umgesetzt. Zunächst werden aber potenzielle Abhängigkeiten zwischen den angestrebten Tool-Aufrufen geprüft. Daraus wird die beste Reihenfolge für deren Ausführung ermittelt. Je nach Tool werden anschließend mittels passender SQL-Befehle beispielsweise Kennzahlen, Verlaufs- und Stammdaten aus der First Cloud-Datenbank extrahiert oder etwa aktuelle Kursdaten über externe Börsensysteme abgerufen.

5

Validierung und Formatierung dienen der Ergebnisaufbereitung und Qualitätssicherung innerhalb der RAG-Pipeline.
Die rohen Tool-Ergebnisse werden in verlässliche, fachlich korrekte und für den Nutzer verständliche Antworten in Form von Texten, Grafiken oder Tabellen umgesetzt.

6

Bei Caching und Performance geht es um die Effizienz der RAG-Pipeline.
Dabei werden Systemressourcen optimiert und akzeptable Antwortzeiten für den produktiven Betrieb sichergestellt. So wird beispielsweise geschaut, ob Teilergebnisse aus der Verarbeitungskette gespeichert (gecached) werden können, damit man bei zukünftigen ähnlichen Anfragen unmittelbar auf diese Ergebnisse zurückgreifen kann. Das spart Zeit und Ressourcen.

Jede dieser Komponenten besteht wiederum aus diversen Einzelschritten, die wir in den kommenden Whitepapern noch detaillierter vorstellen werden.

Gut zu wissen:

Der Unterschied zwischen Deterministisch versus Stochastisch

Betrachtet man die verschiedenen Komponenten innerhalb der RAG-Pipeline des Companions, erkennt man die Trennlinie von externer Funktionalität (LLM-Zugriff) und interner Funktionalität (alle weiteren Module). Diese Trennlinie bildet auch die Grenze zwischen klassischer, deterministischer Informatik und KI-Technik, wo Wahrscheinlichkeiten eine bedeutende Rolle spielen.

Deterministisch bedeutet in diesem Zusammenhang, dass eine Programmfunktion immer den gleichen Output liefert, wenn sie mit dem identischen Input konfrontiert wird. Nun mögen beispielsweise Aktienkurse von Minute zu Minute schwanken, aber wenn man eine entsprechende Programmfunktion nach dem Kurs der Apple-Aktie zu einem gegebenen Zeitpunkt mit Datum und Uhrzeit fragt, dann sollte immer derselbe Betrag in US-Dollar zurückgeliefert werden. Alles andere wäre ein Programmfehler.

Bei KI-Funktionen ist die Situation nicht immer so eindeutig, insbesondere wenn der Input kontextabhängige Unschärfen besitzt, wie dies bei Sprache häufig auftritt. Hier geht es mehr um **stochastische Wahrscheinlichkeiten**, als zu 100 %-vorbestimmte, deterministische Resultate.

Wie soll eine KI beispielsweise wissen, ob mit „Bank“ eine Sitzbank oder ein Geldinstitut gemeint ist, wie soll sie bei „Rente“ zwischen Altersversorgung und einem festverzinslichen Wertpapier trennen? Dies ist – wenn überhaupt – immer nur über den Kontext einer Frage oder Ansage möglich. Aber auch dieser Kontext kann zuweilen zweideutig sein, weil Sprache eben kein vollkommen durchdekliniertes System darstellt. Sonst wäre sie vermutlich gar nicht so breit und flexibel einsetzbar.

Wie soll eine KI wissen, ob mit „Bank“ eine Sitzbank oder ein Geldinstitut gemeint ist?

Bei der Sprachanalyse durch ein LLM als Teil der RAG-Pipeline ist es deshalb entscheidend, dem **LLM Kontextinformationen** mitzugeben, und dies bedeutet im Falle des Companion for First, dass „Rente“ mit hoher Wahrscheinlichkeit für eine Form von Asset steht und Bank nicht für ein Möbelstück. Insgesamt lässt sich festhalten, dass deterministische Verfahren und KI-Funktionen innerhalb der RAG-Pipeline des Companion for First Hand in Hand gehen und einander ergänzen. Erstgenannten wird aufgrund ihrer Exaktheit und Vorbestimmtheit allerdings der Vorzug gegeben, wo immer dies möglich ist.

FAQ

Was bedeutet RAG?

RAG steht für „Retrieval-Augmented Generation“. Dahinter steht das Konzept, das Wissen aus Spezialanwendungen mit generischen Large Language KI-Modellen zu verbinden (die KI-Modelle zu augmentieren).

Was ist eine RAG-Pipeline?

Die Bearbeitung einer Nutzeranfrage von der Eingabe bis zum gewünschten Output erfolgt in mehreren klar abgrenzbaren Stufen, die einander zuarbeiten. Gemeinsam bilden sie die RAG-Pipeline.

Wo liegt der Unterschied zu traditionellen KI-Systemen?

Traditionelle KI-Systeme wissen nur, worauf sie trainiert wurden. Ihr enormes Weltwissen stammt zumeist aus vielfältigen, öffentlich zugänglichen Quellen im Internet. Von Spezialwissen, wie es in geschlossenen Anwendungen vorliegt, verstehen sie nichts.

Was sind die Vorteile von RAG-Systemen?

RAG-Systeme nutzen das Sprachverständnis von Large Language Modellen, geben diesen aber zusätzliches Wissen an die Hand, damit sie ihre Stärken beispielsweise als fachspezifische Chatbots, Übersetzungssysteme oder spezialisierte Programmassistenten ausspielen können.

Warum spielen Large Language Modelle beim Companion eine wichtige Rolle?

Nur moderne Large Language Modelle (LLM) verfügen über die Fähigkeit, komplexe Sprachaussagen aufzuschlüsseln und Sinnzusammenhänge zu erkennen. Sie bilden sozusagen die Speerspitze der modernen KI-Technologie.

Wie schützt der Companion die Kundendaten im Hinblick auf den Einsatz externer LLMs?

Die genutzten LLMs werden durch den Companion zwar mit Spezialwissen aus dem Finanzbereich versorgt – nicht jedoch mit konkreten Details zu spezifischen Kundendaten oder anderen Interna. Diese verbleiben stets innerhalb der RAG-Pipeline von First Cloud und dringen nicht nach außen.

Die nächsten Stationen unserer Reise

Mit diesem Whitepaper haben wir die Grundlagen der RAG-Pipeline im Companion for First beleuchtet. In den kommenden Ausgaben der Reihe werfen wir den Blick tiefer ins Innere des Systems: Wie das Query-Processing funktioniert, welche Rolle Tooling und Training bei der Weiterentwicklung spielen und wie der Companion nahtlos in die Fact Cloud eingebunden ist. Schritt für Schritt entsteht so ein vollständiges Bild, das zeigt, wie agentische KI-Technologie in einem hochregulierten Umfeld zuverlässig Mehrwert stiftet.

Neugierig geworden?

Sie möchten mehr über First Cloud, den **Companion for First** und den Betrieb in der Fact Cloud erfahren? Dann schreiben Sie uns oder sprechen Sie direkt mit unserem Experten Aleksandar Ivezić. Er freut sich auf Ihre Anfrage:

Aleksandar Ivezić

a.ivezic@fact.de

+49 2131 777 238

fact digital uncomplicators

Hauptsitz Neuss
Fact Informationssysteme & Consulting GmbH
Hellersbergstraße 11 | 41460 Neuss | Tel.: +49 2131 777-0

Standort Frankfurt am Main
Fact Informationssysteme & Consulting GmbH
Wilhelm-Leuschner-Straße 81 | 60329 Frankfurt am Main | Tel.: +49 69 8740313-121

E-Mail: mail@fact.de

Stand September 2025

Folgen Sie uns auf LinkedIn

Alle im Whitepaper verwendeten Bilder wurden mit Unterstützung des KI-Tools Adobe Firefly generiert. Sie dienen ausschließlich der visuellen Veranschaulichung und haben keinen dokumentarischen Charakter.