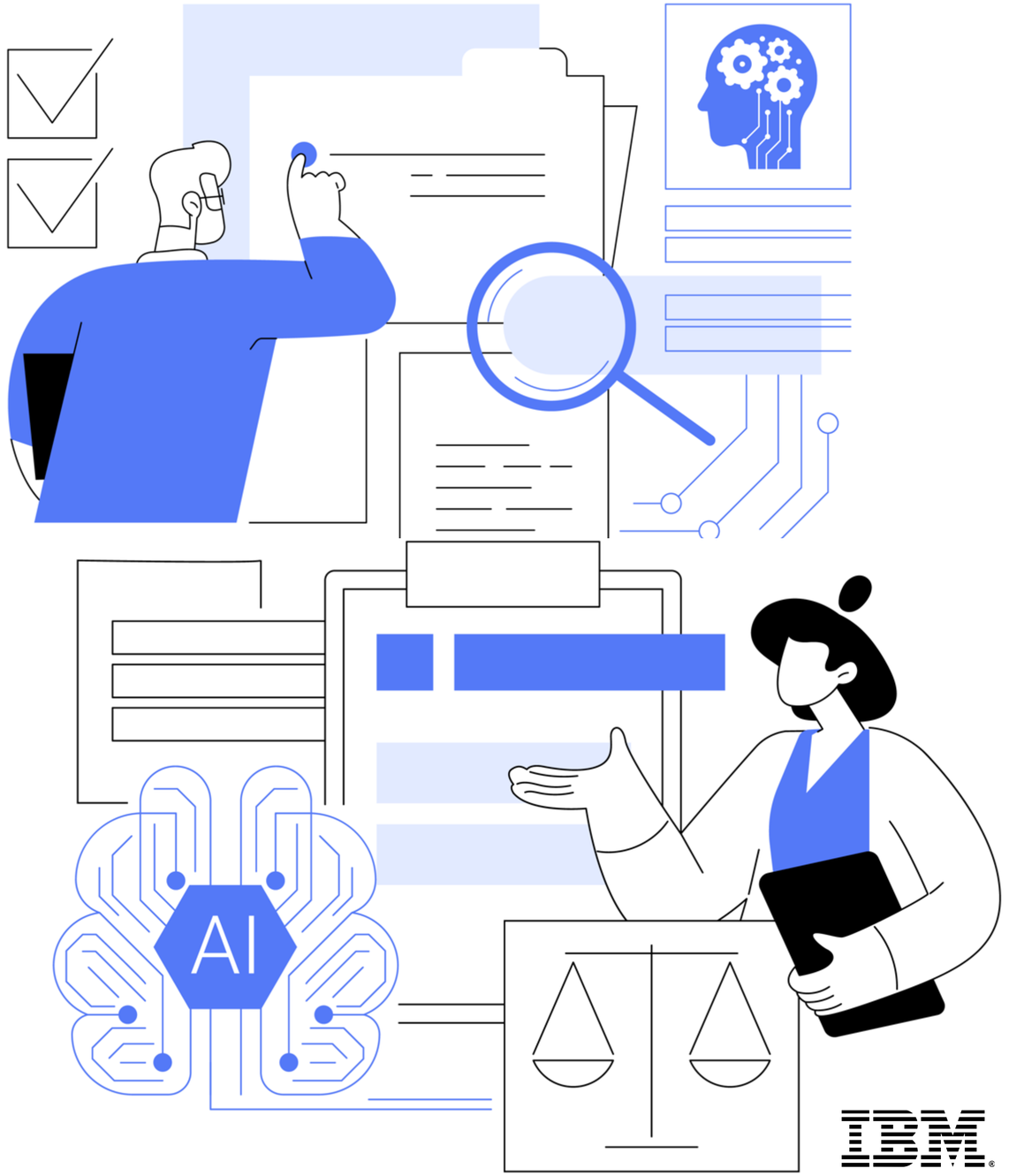
Einfach Steuern: Mit GenAI und LLMs durch Steuerrecht und Entscheidungen navigieren

DI Marco Köck, BMF

DI Thomas Jirku, IBM



Bundesministerium
Finanzen

IBM

DI Marco Köck

Bundesministerium für Finanzen

Warum habt ihr euch entschieden ein Projekt zu generativer KI zu machen ?

Generative KI – Metaziele für das BMF/BRZ

BMF/BRZ interne Verprobung einer generativen KI-Anwendung

Finden eines passenden Use Cases im Fachbereich
Um rasch einen Use Case in der Cloud (d.h. ohne lokale Installation) umsetzen zu können müssen Daten unkritisch und offen sein

Verprobung der IBM watsonx Plattform
Gemeinsame Entwicklung unter Nutzung der verfügbaren KI & Governance Werkzeuge um BMF/BRZ seitig alternative, zukünftige Use Cases auswählen und bewerten zu können

Wie seid ihr zu dem passenden Usecase gekommen?

GenAI Brainstorming

Classify

- Fallberichte für Prüfer klassifizieren
- Überprüfung der Plausibilität durch Bild Klassifizierung (Piraterie, Falsche Zeugnisse, Falsche Dokumente)
- Analyse von Röntgenbildern von Lastwägen
- Erkennung von Doppelförderungen

Extract

- Attribute Extraction (Name, Geburtsdatum, Subject ID)
- Analyse von Bescheiden
- Automatisierte Extraction von Formulardaten

Summarize

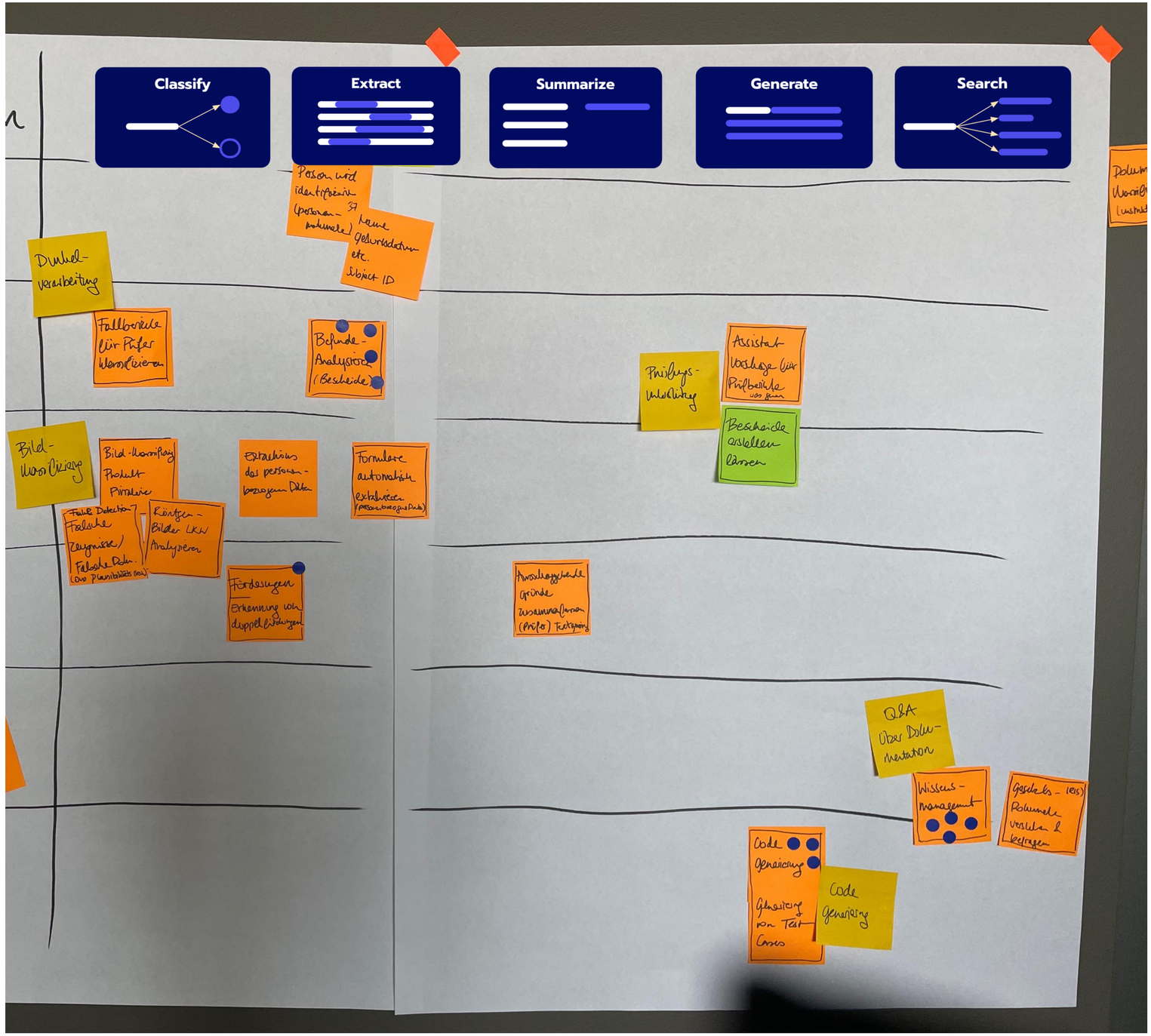
- Zusammenfassung der maßgeblichen Gründe

Generate

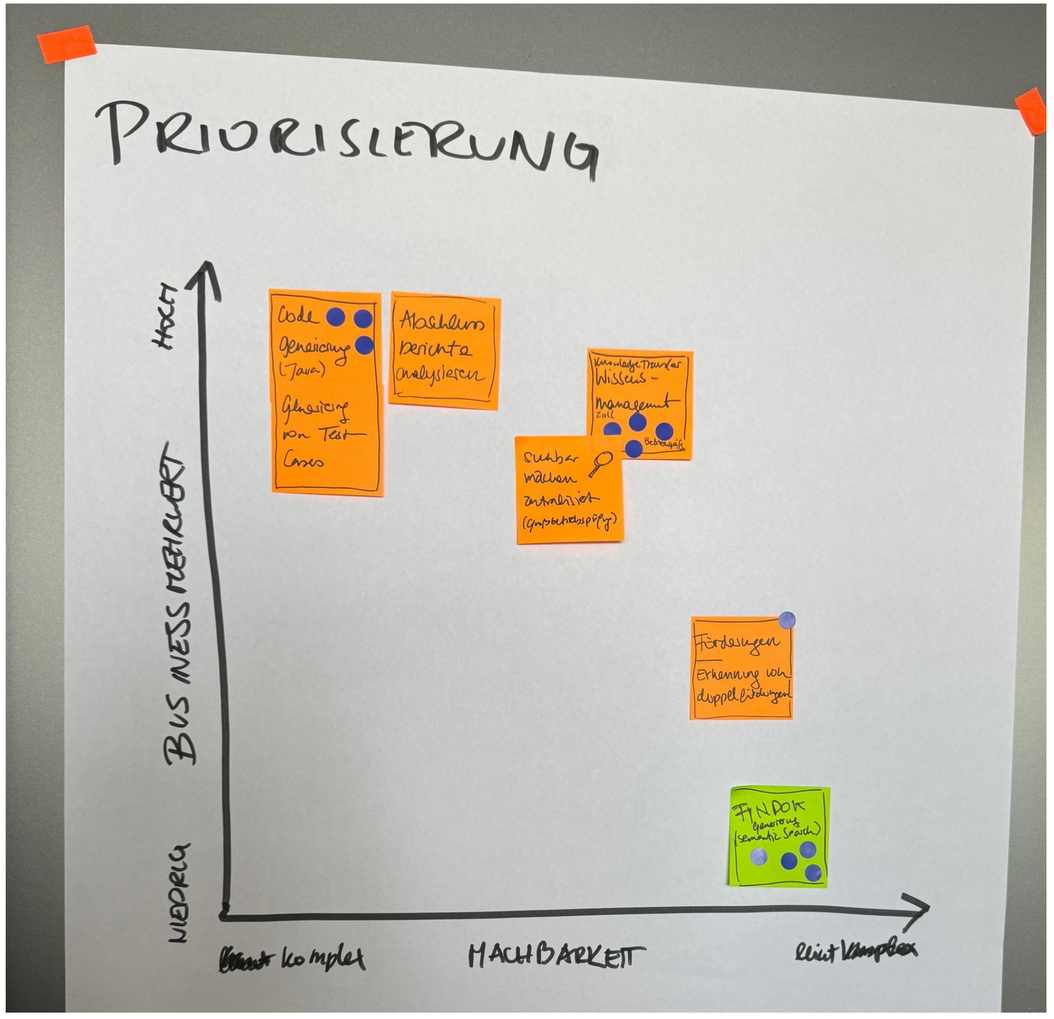
- Prüfungsunterstützung (Assistant zur Vorschlags-erstellung für Prüfberichte)
- Bescheide erstellen lassen
- Code Generierung
- Erzeugung von Testfällen

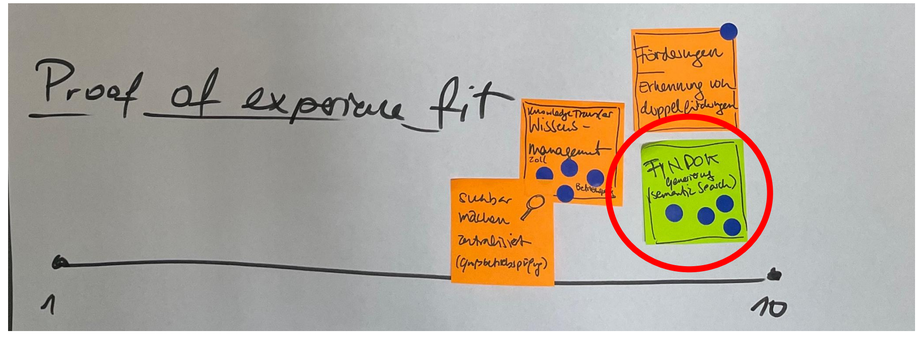
Q&A / Search

- Wissensmanagement
- Q&A über Dokumentation
- Verständnis und Befragung von Gesetzestexten




Classify
Extract
Summarize
Generate
Q&A
Dunkel-verarbeitung
Fallbeispiele für Prüfer verdichten
Befunde-Analysieren (Bescheide)
Bild-Verarbeitung
Bild-Verarbeitung Produkt Prinzipien
Fraud Detection Falsche Zeugnisse/ Falsche Daten
Röntgen-Bilder LKW Analysieren
Förderungen Erkennung von doppelförderungen
Prüfungs-Unterlagen
Assistat Vorschläge für Prüfberichte
Bescheide erstellen lassen
Q&A Über Dokumentation
Wissens-management
Code Generierung
Code Generierung


PRIORISIERUNG
HOCH
BUSINESS MEHRWERT
NIEDRIG
HACHBARKEIT
Code Generierung (Java)
Abschluss Berichte analysieren
Wissens-management
Förderungen Erkennung von doppelförderungen
FINDOK


Proof of exposure fit
Förderungen Erkennung von doppelförderungen
Wissens-management
FINDOK
1
10

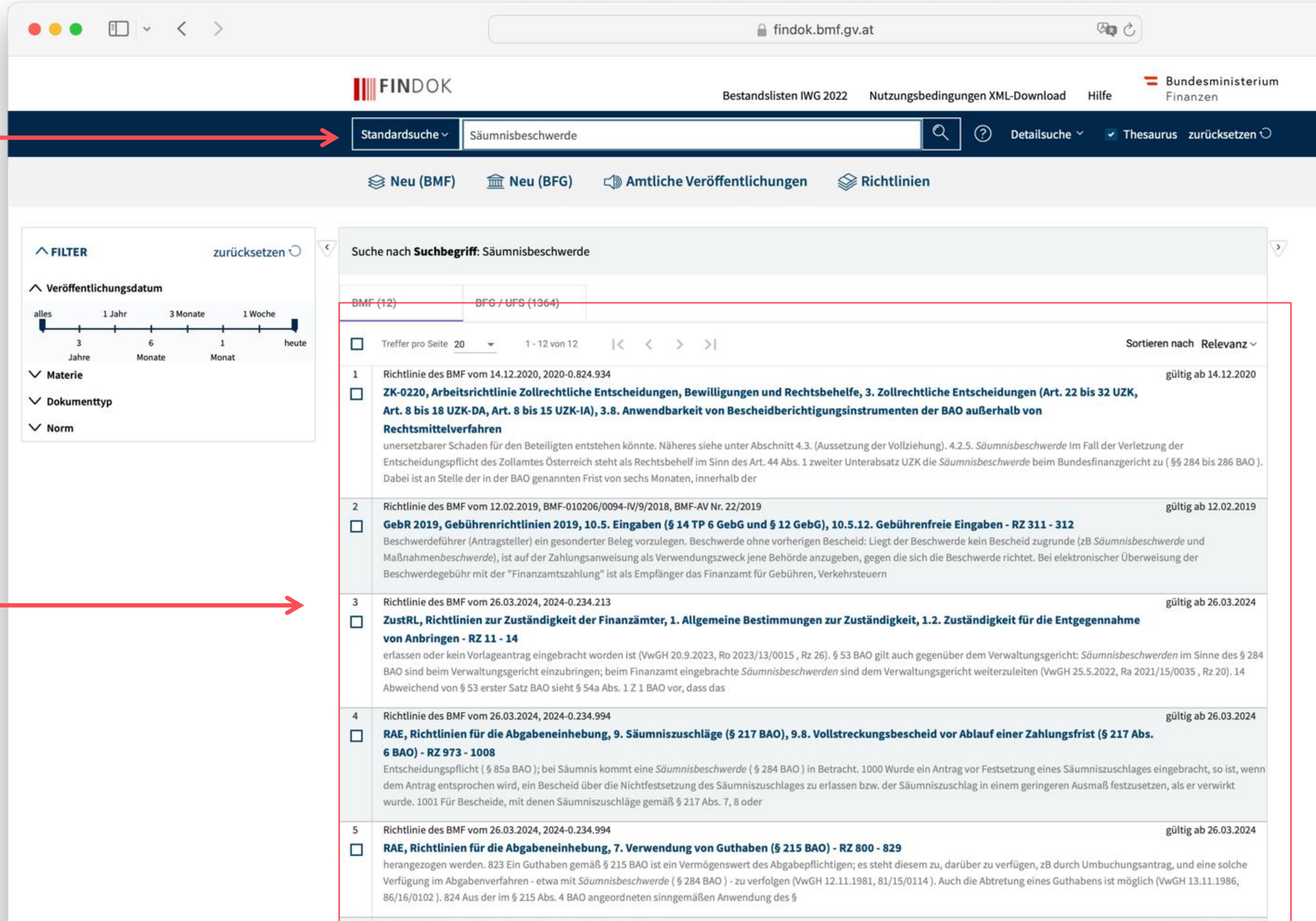

IDEA
Zusammen-fassung der Suchergebnisse
Erstellung der Urteile (Next Step)
Iterative Suche
Detail-Filterung der Suche
Generierung der Ergebnisse
Verfeinerung der Suche
Zusammenfassung pro Suchergebnis (abhängig vom Input)
natürliche Sprache?
<Suchanfrage>
Zusammenfassung Erg 1?
Zusammenfassung Erg 2?
Step 1: Finden "Suchenfinden"
Step 2: "Erfassen/Dokument"
Übersichts Liste
1. Thema X
2. Thema Y
Thema Y Zusammenfassung
Link 1
Link 2
Gesetz
Suche 1
Suche 2
Suche 3
Art Qu. A
mit Suche kombiniert
Anders:
Suchverhalten sehr komplex u. keine mit Sicherheit eine Frage nicht gut genug dargestellt werden
Richterin
Fall
Ähnliche Fälle
10 Min
Zusammenfassung v. Beschlüssen
TEXTEINGABE
RICHTER:IN RICHTIGE NÄCHTE WISSEN, OB EIN SKILEHRER SKI ABSETZEN KANN
EINGABE: "SWO SKI FÜR SKILEHRER ABSETZBAR?"
ES GIBT CA. 20 RELEVANTE BFG-ENTSCHEIDUNGEN.
JA, WENN
BEDINGUNG 1,
BEDINGUNG 2
NEIN, WENN
BEDINGUNG 1
BEDINGUNG 2
WEITERE RELEVANTE ENTSCHEIDUNGEN:
...RESULTAT
EXPERTS
Query
Docs
Prompt
Result

Was ist „FinDok“ - das Webportal des BMF

Eine Ausschließlich auf **Keyword basierte Abfrage**

Eine **Vielzahl von Quellenangaben**, die manuell durchsucht werden muss

Hypothese:
Wie können wir mithilfe generativer KI, basierend auf FinDok Daten unseren Nutzenden eine natürlich-sprachliche Abfrage bereitstellen?



MVP Scope / Statement

Wenn wir für Justicia (Richterin)

Folgendes bereitstellen:

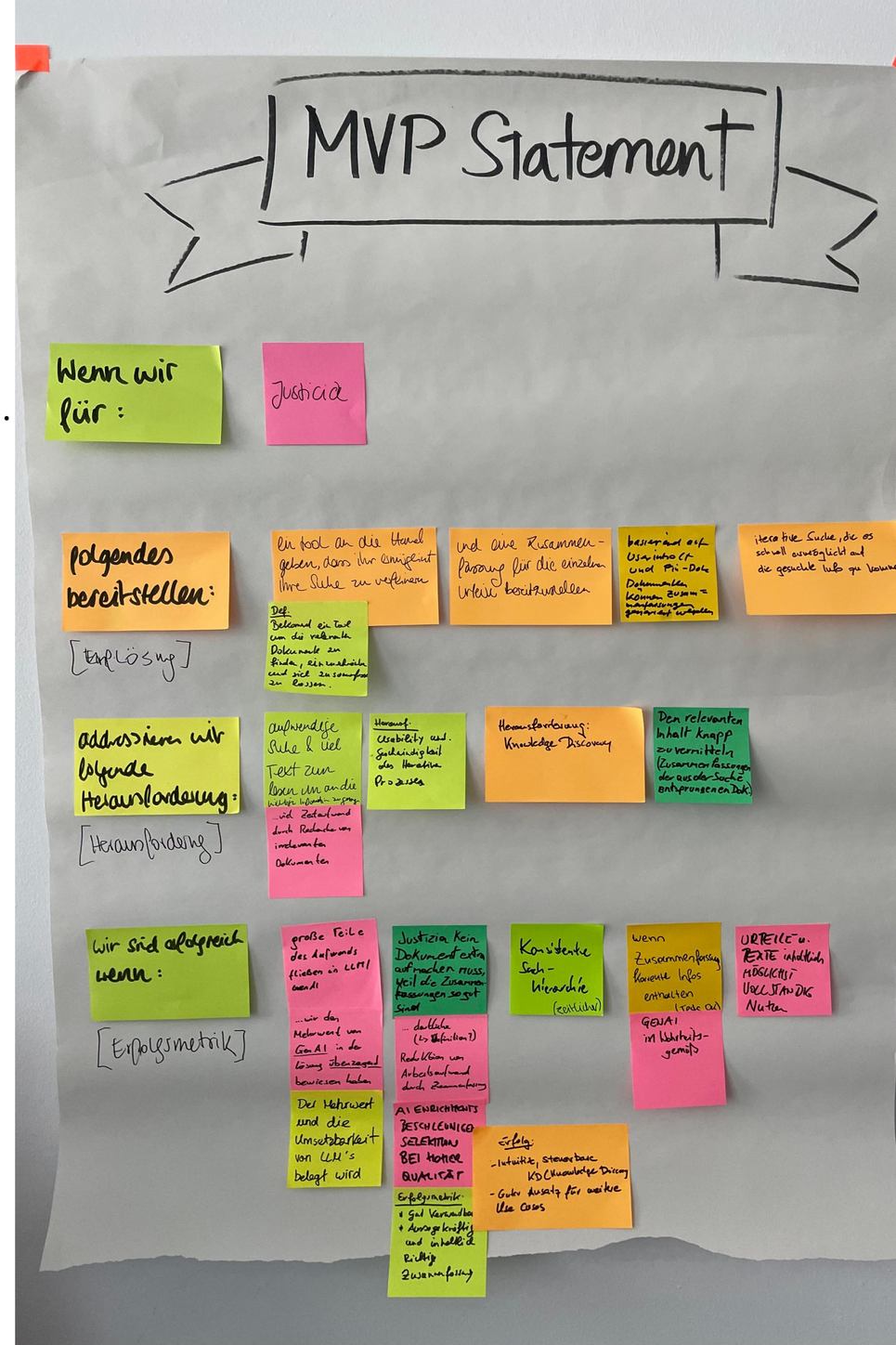
Ein Tool an die Hand geben, dass die Suche verfeinert, um die einschlägigen Dokumente zu identifizieren und zusammenfassen zu lassen. Dies erfolgt basierend auf der Nutzereingabe in natürlicher Sprache

Dies ermöglicht es uns zu adressieren (Wo gibt es aktuell Herausforderungen?)

- Viel Zeitaufwand durch Recherche von irrelevanten Dokumenten
- Knowledge Discovery

Wir sind erfolgreich, wenn...

1) der Mehrwert von GenAI in der Lösung überzeugend bewiesen und die Umsetzbarkeit belegt wurde
2) Justizia die richtigen/wichtigen Dokumente auf den ersten Blick sieht und diese wahrheitsgemäß und vollständig zusammengefasst wurden
3) Intuitive, steuerbare Knowledge discovery
4) Aussagekräftig und inhaltlich richtig
5) Die Kenntnis erlangt wurde, wie weitere Use Cases auf einem ähnlichen Prinzip erstellt, werden können



Generative KI - Ziele für das BMF/BRZ

Let’s create [together].

Konkrete UseCase Ziele

Praktische Erfahrungen mit KI (watsonx als Plattform)
BMF/BRZ Entwicklern zeigen, wie man internes Wissen/Dokumente nutzen und mithilfe von LLMs schnell mit einer intelligenten Suche verfügbar machen kann

LEARNING: Capabilities und Funktionen der watsonx-Plattform
Kennenlernen der Funktionen der watsonx-Plattform; Verständnis der Möglichkeiten und Limitationen der Plattform. Nutzung von Agile Programming mit Shadowing und Pairing sessions

IMPLEMENTIERUNG: FinDok Use Case
Umsetzung einer generativen FinDok Anwendung mit Echtdaten (100.000 Dokumente). Verprobung der zwei *Bereiche Semantische Suche und KI / LLM generierter Fragen-Beantwortung*

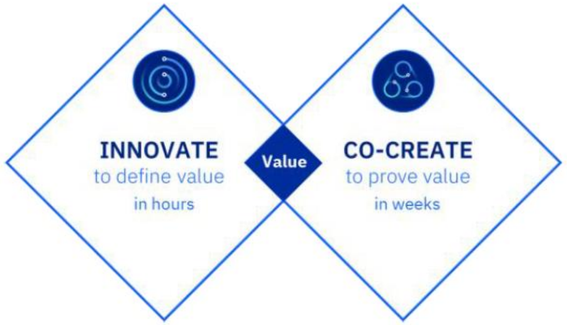
EXPERIMENTIEREN mit Sentence Embeddings & LLMs
Experimenteller Vergleich und Bewertung von Alternativen unter Nutzung der .ai und .gov Werkzeuge wie Watson Studio, Prompt Studio, Deployment Spaces, Governance etc.

Wie kann die Lösung einen Fachanwender oder eine Richterin bei der Recherche unterstützen?

Wie sah konkret die Umsetzung aus?

Let‘s create

IBM Value Engineering Method



Innovate
in 4 – 10 hours

“We **align** on a solution to a compelling business opportunity and **define** a scope to **prove value"**

Prepare
in 1 – 5 days

“We **prepare, plan,** and **commit** ourselves to be ready to rapidly build.”

Co-Create
in ~4 weeks

“We **build, validate,** and **improve** over a series of sprints and playbacks.”

Transition

“We **understand** how to get on a path to **scaled** production.”

Überblick Teaming

2 Tage

2 Design-Thinking-Workshops, um das Verständnis zu vertiefen und das MVP Statement zu definieren

---

6 Wochen

Intensive Co-Creation -
MVP build phase mit Enablement Sessions für das BMF & BRZ im Umgang mit watsonx.ai und LLMs

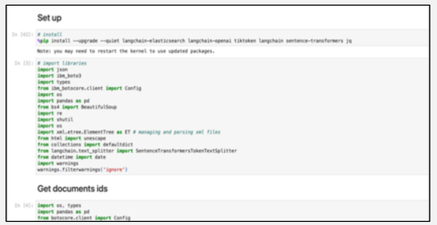
---

10h

Pairing Sessions in Stunden gemeinsam mit BMF & BRZ Kollegen um FinAI zu entwickeln und das Enablement für watsonx zu bieten

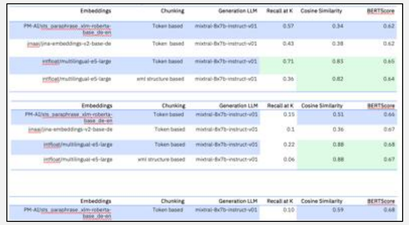
Methodische Herangehensweise im Rahmen des MVP

VERSTÄNDNIS UND VERARBEITUNG VON DATEN



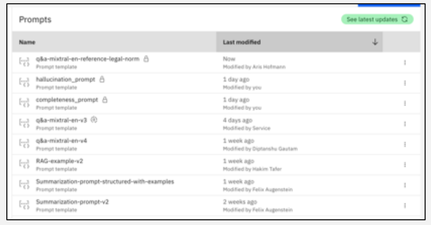
- Untersuchung von Daten, die aus ~83.000 Dokumenten bestehen
- Verwendung der XML-Struktur des Dokuments zur Extraktion von Text und Metadaten

RETRIEVAL EXPERIMENTATION



- Testen von 4 verschiedenen Einbettungsmodellen
- Testen verschiedener Chunking-Strategien
- Bewertung der Qualität des Abrufs mit der bereitgestellten Grundwahrheit
- Auswahl des Einbettungsmodells mit der besten Leistung

PROMPT ENGINEERING



- Entwicklung von Prompts für RAG und Bewertung der Qualität der Antwort
- Bewertung der Qualität der Generierung mit bereitgestellter Ground Truth

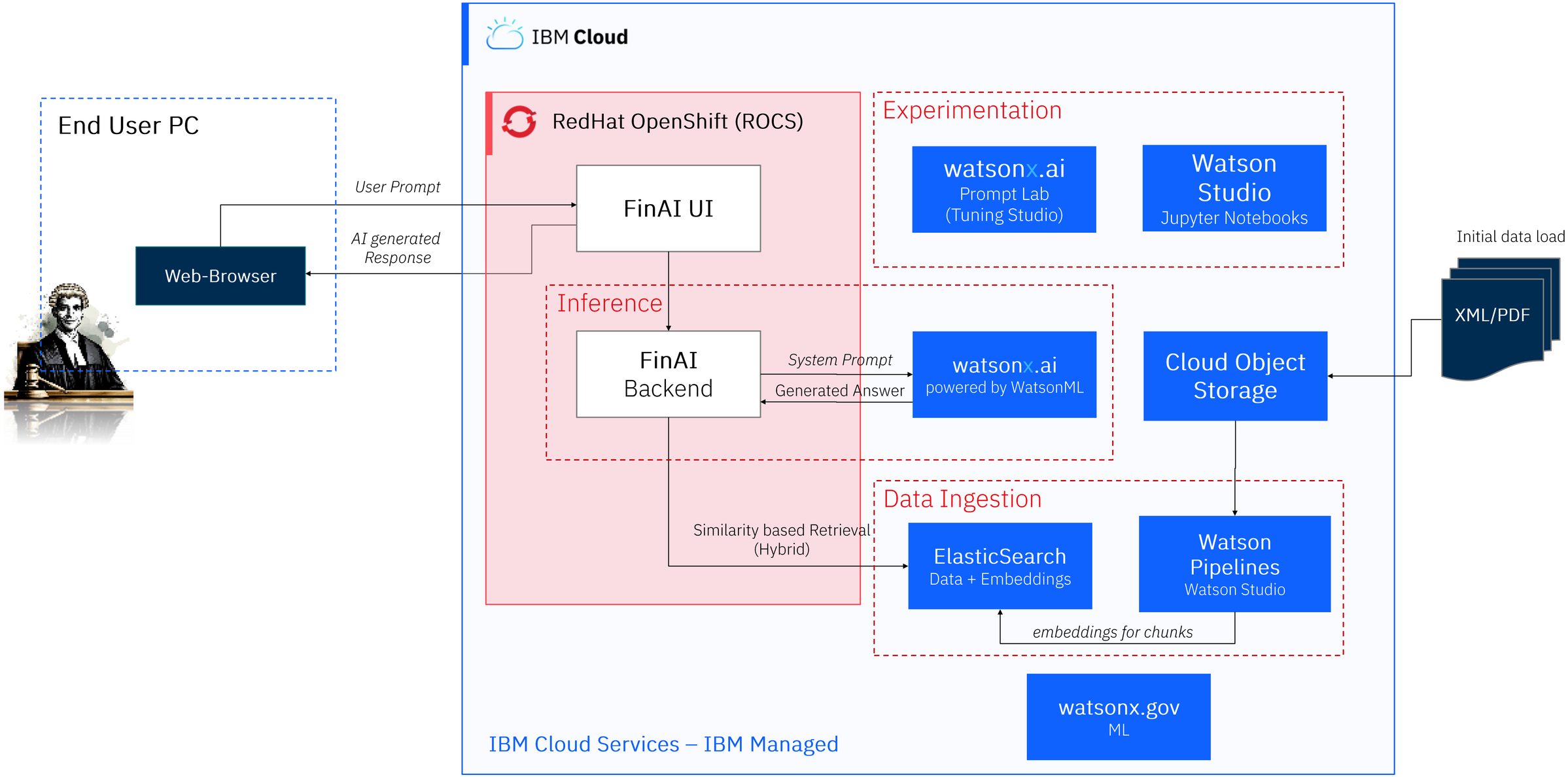
FEEDBACK VON EXPERTEN

- Austausch von Ergebnissen mit Experten und Durchführung von Feedback-Sitzungen

VERBESSERUNG DER LÖSUNG

- Verwendung von Expertenfeedback zur Verfeinerung der Lösung

MVP Architektur Highlevel



finai.ibm-co-created.com/search

https://finai.ibm-co-created.com/search

RTC&Wiki | Viya Ressources | SAS | SAS Ressourcen | BMF-Portal | BRZ | Gmail | Google Kalender | Google | Transaction Networ... | TNA Support - TNA... | Online | Search result - TED... | RH Prüfung KI in de... | Weitere Favoriten

FINDOK

FINAI SUCHE (NEU)

Stellen Sie Ihre Frage...

Alle Ergebnisse

Rows per page: 5 0–0 of 0

DOKUMENTE ZUSAMMENFASSEN

finai.ibm-co-create... | *Unbenannt - Editor | Converting screen:...

DEU 15:21 04.07.2024

Disclaimer: Dies stellt keine produktive Lösung dar, die Teil des FinDok-Systems ist. Es handelt sich vielmehr um eine Erprobung im Rahmen eines Pilotprojekts zur Nutzung von GenAI im Rechtsbereich für die Befragung von Dokumenten.

FinAI

Smarte Suche für FinDok

Use case: eine **GenAI Suche** für die umfangreiche **Rechtsliteratur-Datenbasis** von **FinDok**, ideal für Richter:innen, Steuerberater:innen und viele weitere Fachnutzer:innen, um natürlichsprachliche Fragen an die Wissensdatenbank zu stellen.

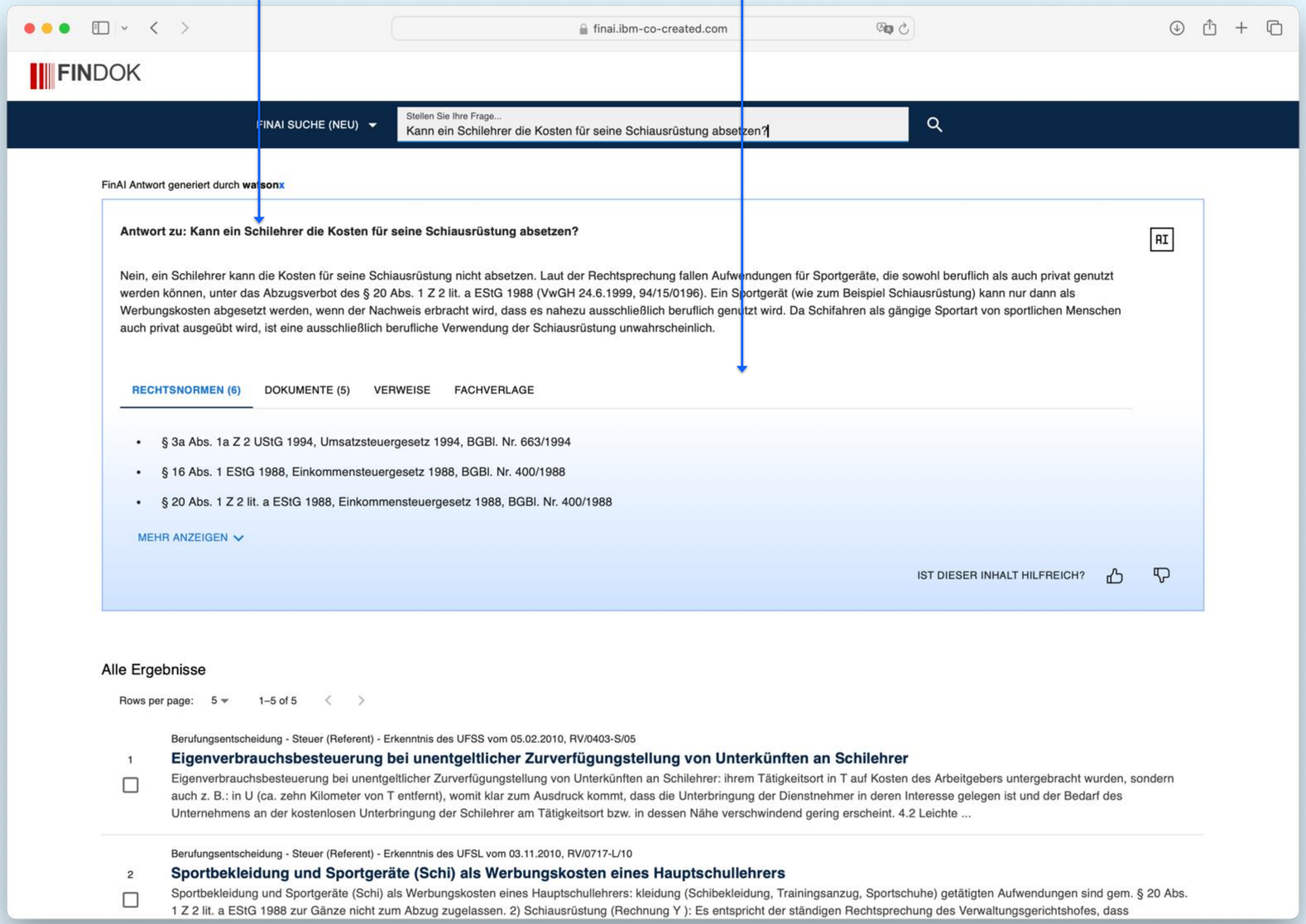
KEY FEATURES

Zurückverfolgung zur Quelle – das Dokument, aus dem eine Erkenntnis gewonnen wurde, wird angezeigt und nach Relevanz gerankt.

Die gesamte Datenbasis des FinDOk – eine Bibliothek von **100.000 FinDok Dokumenten** wurde für den RAG-Use Case genutzt.

Anzeige der generieten Antwort

Quellen werden angezeigt, die sich auf den vom LLM bereitgestellten Text beziehen



Frage 05

Ergebnisse, Evaluierung,Feedback

User Bewertung der FinAI Lösung durch den Fachbereich (Steuer)

„Schnelligkeit“

Welche neuen Potenziale birgt diese KI-gestützte Lösung?

9/10

Wie bewerten Sie das Textverständnis?

100%

Ist FinAI in diesem Umfang für den täglichen Gebrauch nützlich und geeignet?

75%

Erzeugt das Tool auf Grundlage Ihrer Eingaben präzise und korrekte Antworten?

8,5 /10

Wie würden Sie die Inhaltliche Richtigkeit bewerten?

"Genug"

Wie viel Zeit würde Ihnen diese Lösung pro Anfrage einsparen?

Evaluierung

Ground truth
Golden Questions set 1: Informations- & Dokumentationssysteme

Golden Questions set 2: Zoll

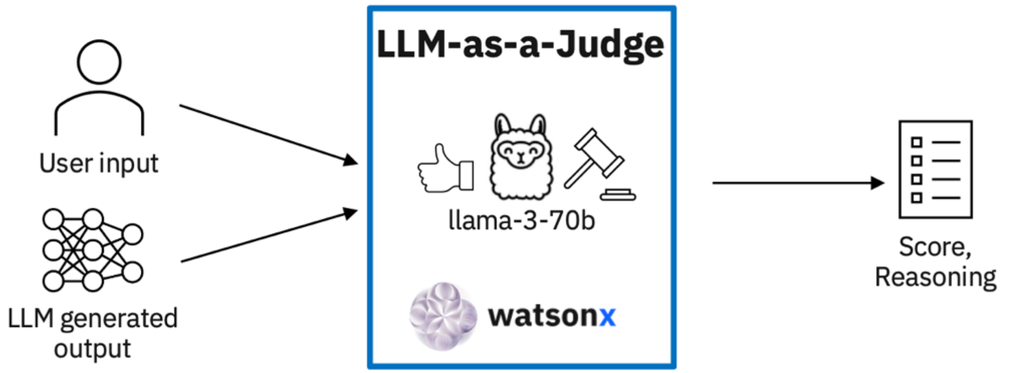
Golden Questions set 3: Predictive Analytics Competence Center (PACC)

Metrics
Recall at K gibt den Prozentsatz der goldenen Dokumente an, die in den oberen K abgerufenen Ergebnissen enthalten sind

Cosine Similarity berechnet die semantische Ähnlichkeit zwischen goldenen Antworten und generierten Antworten

BERTScore nutzt kontextuelle Embeddings um die semantische Ähnlichkeit zu bewerten.

LLM-as-a-judge (ground truth nicht notwendig)



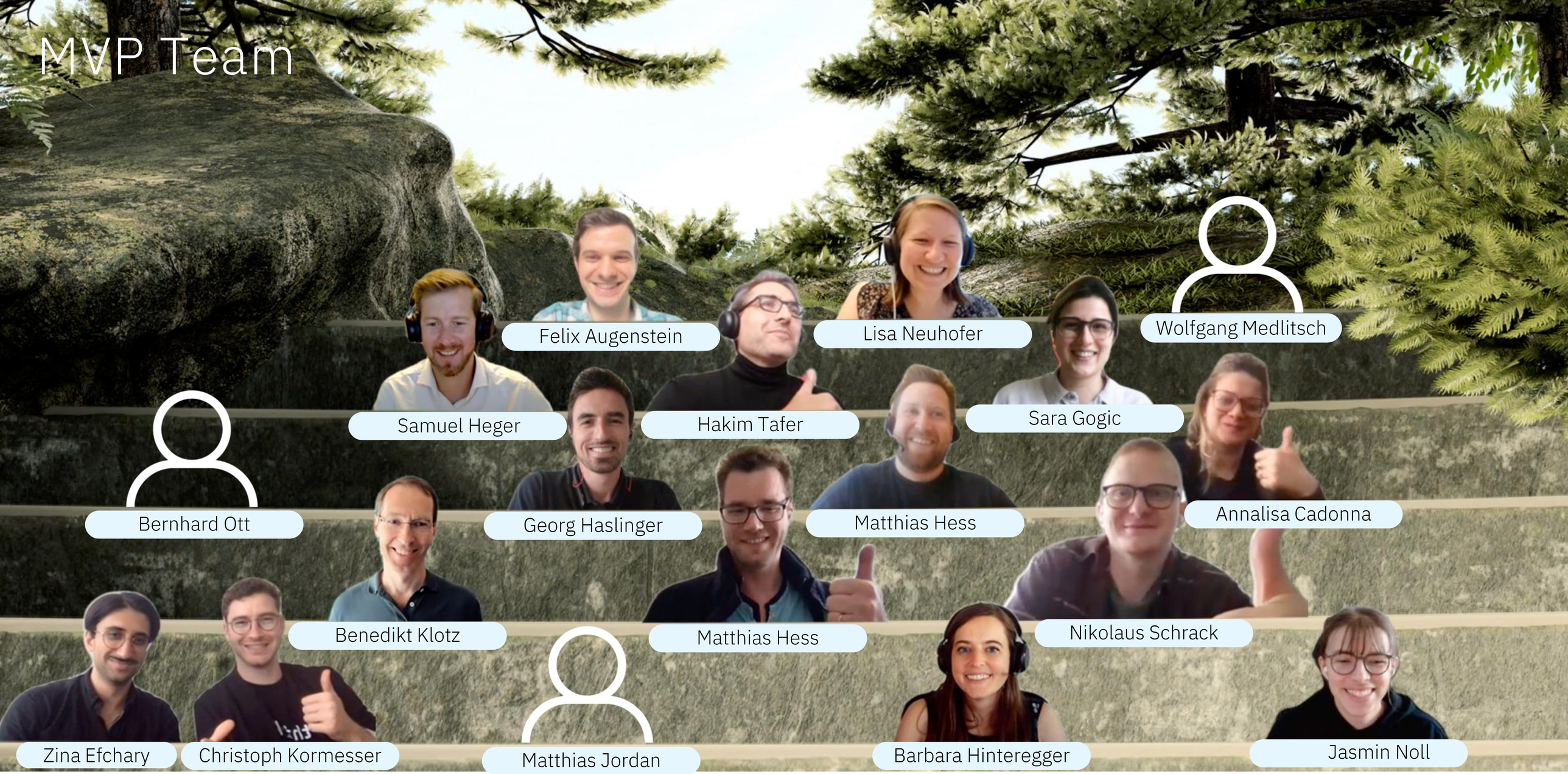
Metriken

Vollständigkeit der Antwort beurteilt, wie vollständig die Antwort auf die Frage ist

Halluzination score bewertet, wie relevant die Informationen in der Antwort in Bezug auf den Kontext sind

Wie beurteilst du die Zusammenarbeit mit IBM Client Engineering?

MVP Team



Ein gemeinsames Team

BRZ Bundesministerium Finanzen IBM

Ein Blick in die Zukunft... oder was würdet ihr heute anders machen?

Potenzielle Ausbaustufen und Erweiterungen des MVPs

Datenpool und KI („*Präzision*“)

Erweiterung Datenpools, Filtering, hierarchy
Erweiterung des Datenpools um neue Dokumente. Wählen der Dokumente aus, die zur Beantwortung der Frage herangezogen werden sollen und mit welcher Hiererachie.

Conversational AI
Möglichkeit, Folgefragen zu stellen und die Antwort zu verfeinern

Integration neuer Sprachmodelle
Verwendung eines deutschen Modells (auf der Roadmap von watsonx.ai, auch über BYOM)

Benutzerführung („*Attraktivität*“)

UI-Verbesserung
Verbesserung der Benutzerführung basierend auf dem Input des BMF-BRZ

Optimierung basierend auf User Feedback
Bessere Bewertbarkeit der Antwortqualitität durch Feedback und Erkennen der Nutzerbedürfnisse

Feedbackloop
User Feedback für die Optimierung der Antwortqualität einbeziehen und verarbeiten

Stabilität und Administration („Skalierung“)

CI / CD Pipeline und Umgebungen
Instandsetzen mehrerer Umgebungen und verbesserte, kontrollierte Aktivierung von neuen Funktionen

User Management
Einbau einer Nutzerverwaltung, die auch auf einem onprem Betrieb ausgerichtet ist.

Logging von Fragen & Antworten
Übersicht der gestellten Fragen und Antworten je Nutzer zur Verbesserung der Qualität