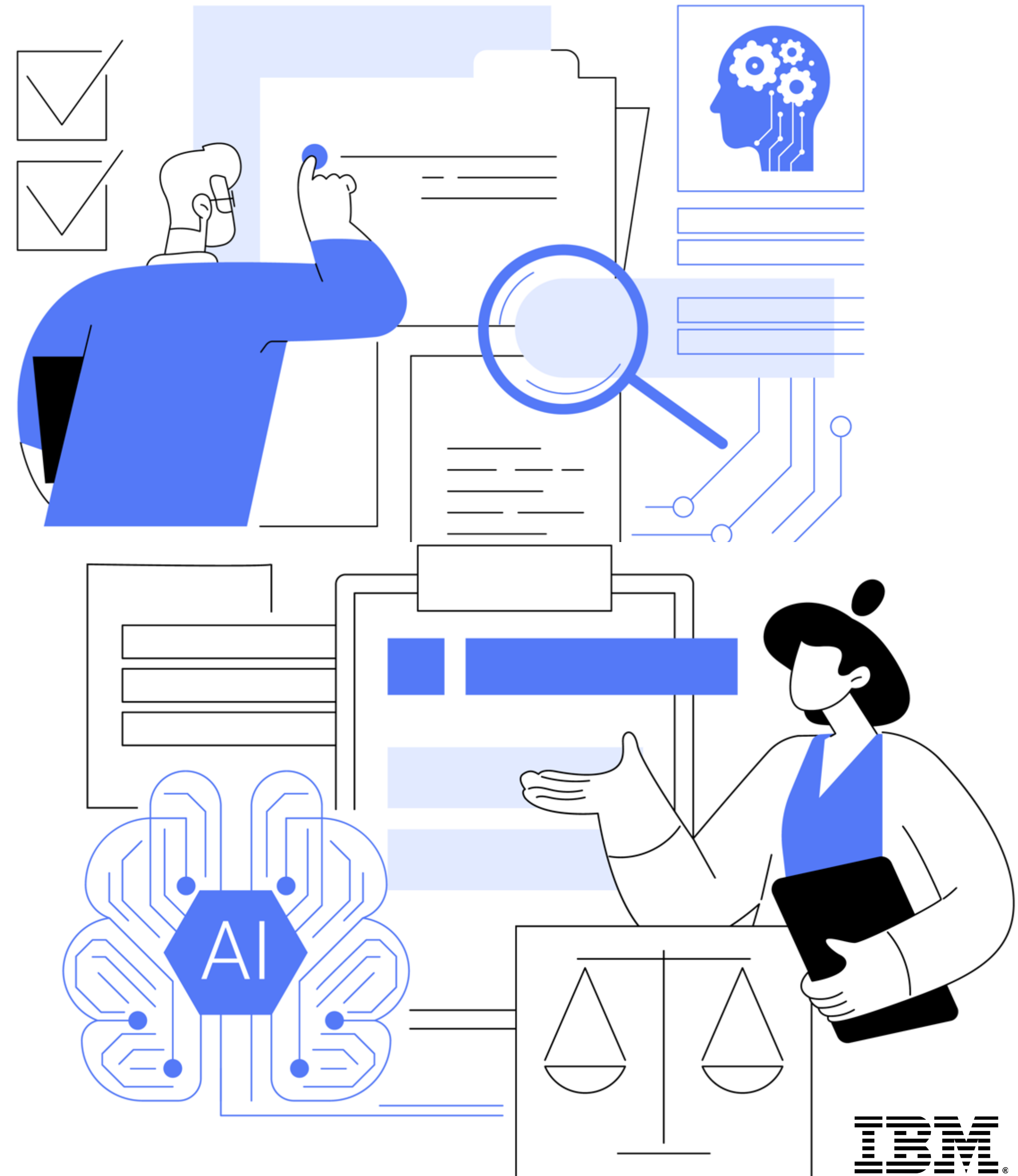


Einfach Steuern: Mit GenAI und LLMs durch Steuerrecht und Entscheidungen navigieren

DI Marco Köck, BMF

DI Thomas Jirku, IBM



DI Marco Köck

Bundesministerium für Finanzen



Warum habt ihr
euch entschieden
ein Projekt zu
generativem 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
Use Case
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

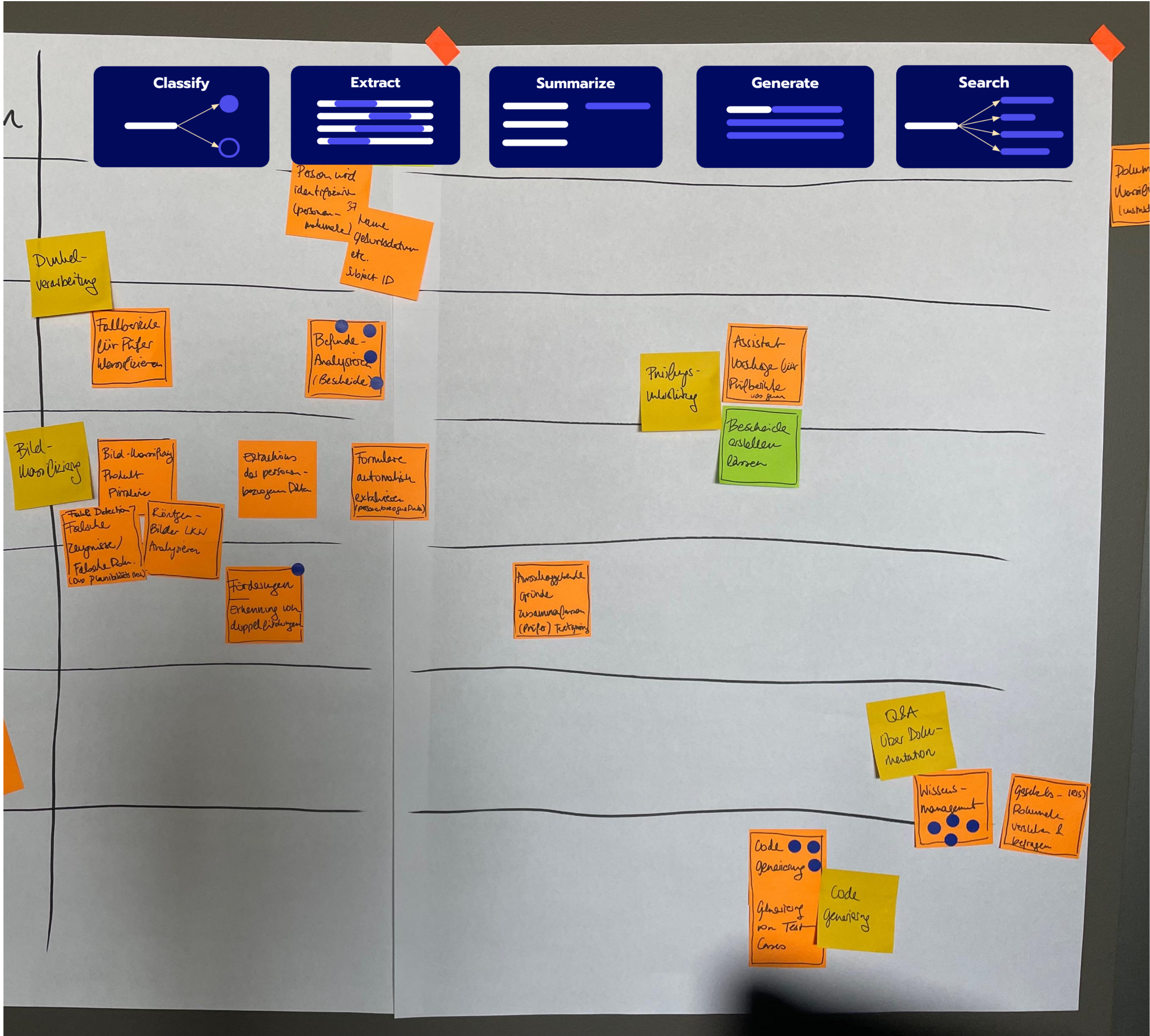
- Zusammenfassung der maßgeblichen Gründe

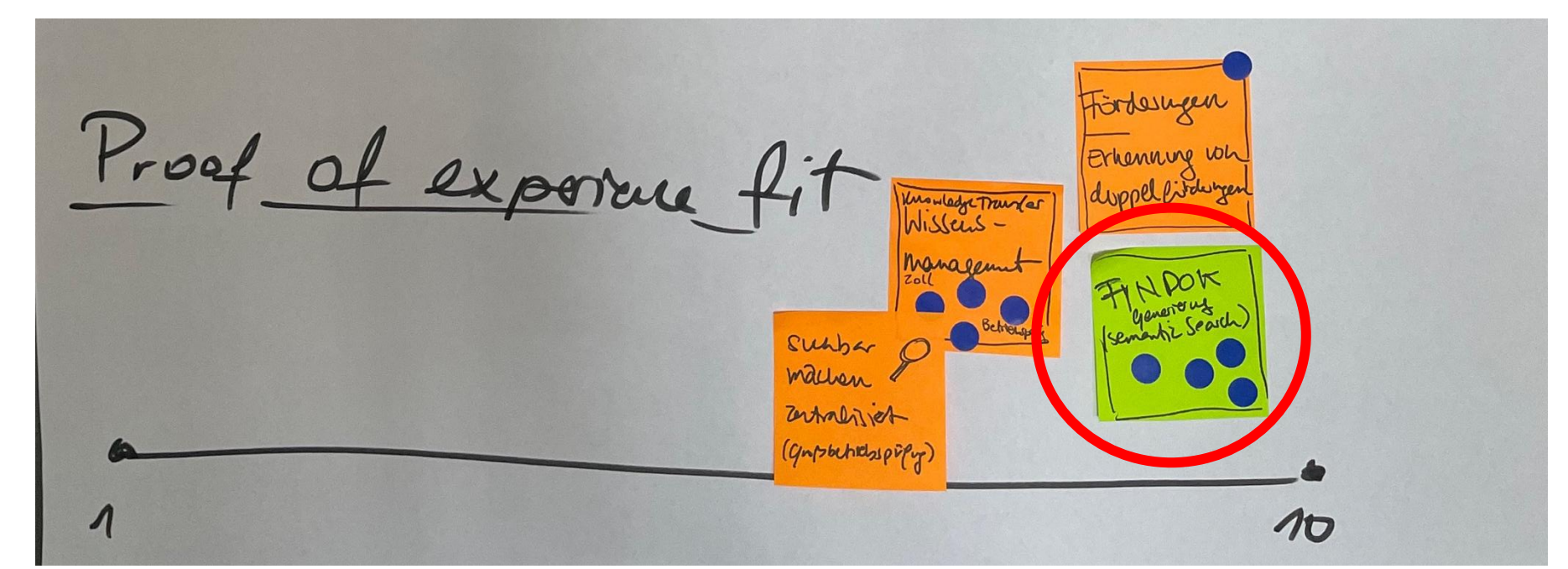
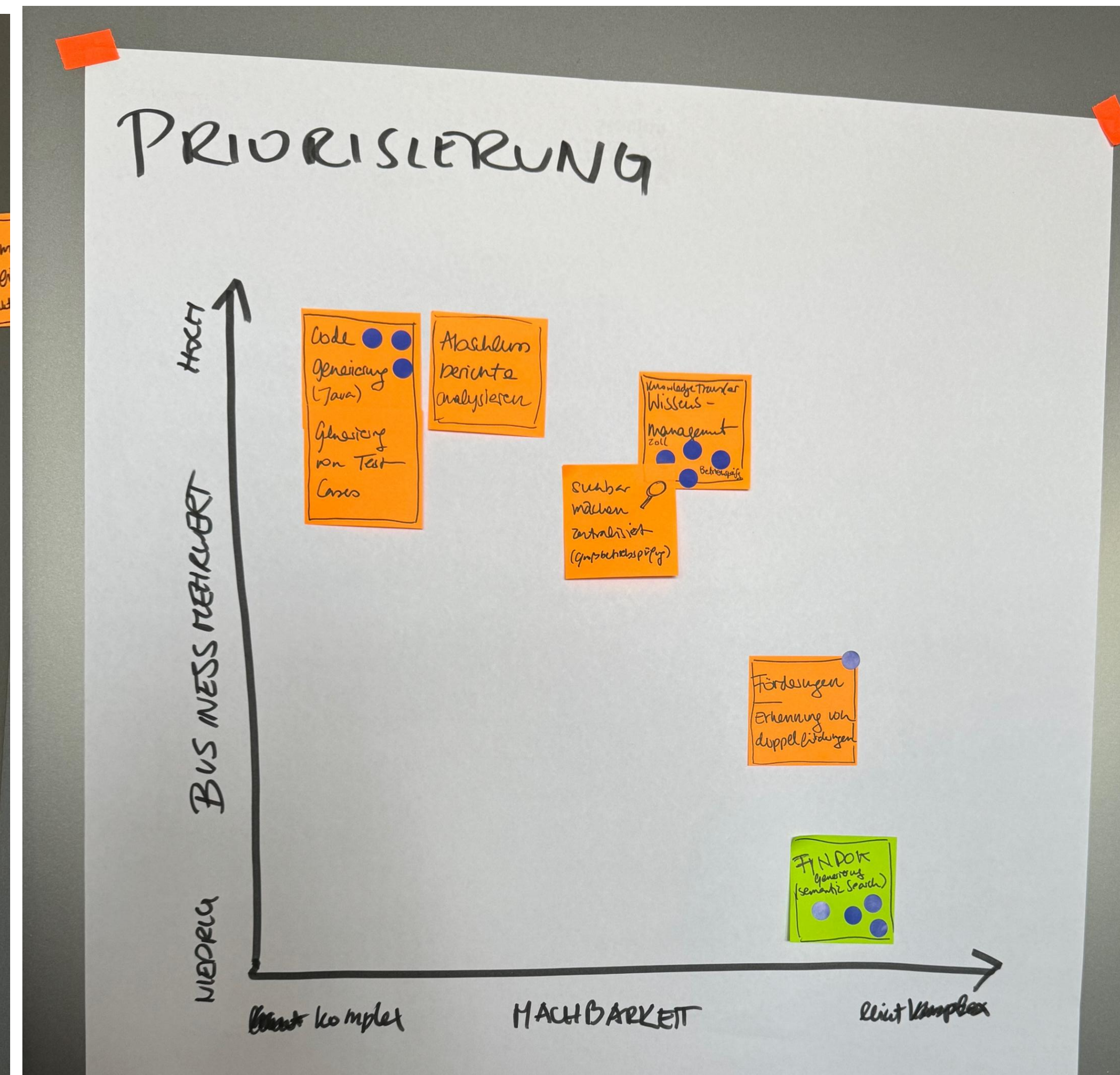
Generate

- 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







BI IDEA

Zusammenfassung der Suchergebnisse

Zusammenfassung pro Suchergebnis (abhängig vom Input)

```

        Findet: <Suchanfrage>
        Zusammenfassung Erg 17
        Zusammenfassung Erg 27
        ...
      
```

← natürliche Sprache?

Erstellung der Urteile (next step)

Step 1: Filter "Guide" "Suchen/Finden" "Erklärung/Dokumente"

Step 2: Filter "Erklärung" "Erklärung/Dokumente"

Chancen/Erwartung

Verbessert Suche

RIS/Redman, Probleme

Hoative suche

Detail-Filtering der Suche

SUCHE	WURDE	ANTWORT
☐	☑	Σ
☐	○	C1
☐	☑	C2
☐		C3

+SCORE Feedback

Generierung der Ergebnisse

Verfeinerung der Suche

GENERAL USER

Chat with Findet

Übersichtsliste (Thema 1, Thema 2, ...)

Thema Y Zusammenfassung

• Link 1

• Link 2 → Gesetz

RECHTEN/IN

TEXTEINGABE

RICHTER/IN NICHT DIE RICHTIGE WISSEN, OB EIN SKI ABSETZER KANN

EWGABE: "SIND SKI FÜR SKIABSETZER ABSETZBAR?"

"SIP OF COFFEE" (O.S.P. SEH)

Ähnliche Fälle

Zusammenfassung v. Beschlüsse

30 Min

EXPERTS

Using loc nonreport to generate report?

Query

Docs

Drop-down for reference with text passage

RESULTAT

Suche 1

Suche 2

Suche 3

Art Qu. A

reit Suche kombiniert

Andere: Sachverhalte sehr komplex u. keine mit Skizzen, eine Frage nicht gut genug dargestellt werden

Was ist „FinDok“ - das Webportal des BMF

Eine Ausschließung 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?

The screenshot shows the FinDok web portal interface. At the top, there is a search bar with the text 'Säumnisbeschwerde' and a search icon. To the left of the search bar, there is a dropdown menu labeled 'Standardsuche'. Below the search bar, there are several navigation links: 'Neu (BMF)', 'Neu (BFG)', 'Amtliche Veröffentlichungen', and 'Richtlinien'. On the right side, there are links for 'Bestandslisten IWG 2022', 'Nutzungsbedingungen XML-Download', 'Hilfe', and 'Bundesministerium Finanzen'. Below the navigation links, there is a filter section on the left with the title 'FILTER' and a 'zurücksetzen' button. The filter section includes a 'Veröffentlichungsdatum' section with a timeline from 'alles' to 'heute', and sections for 'Materie', 'Dokumenttyp', and 'Norm'. The main content area shows search results for 'Suchbegriff: Säumnisbeschwerde'. There are two tabs: 'BMF (12)' and 'BFG / UFS (1364)'. The results are sorted by 'Relevanz'. The first result is a guideline from the BMF dated 14.12.2020, titled 'ZK-0220, Arbeitsrichtlinie Zollrechtliche Entscheidungen, Bewilligungen und Rechtsbehelfe, 3. Zollrechtliche Entscheidungen (Art. 22 bis 32 UZK, Art. 8 bis 18 UZK-DA, Art. 8 bis 15 UZK-IA), 3.8. Anwendbarkeit von Bescheidberichtigungsinstrumenten der BAO außerhalb von Rechtsmittelverfahren'. The second result is a guideline from the BMF dated 12.02.2019, titled 'GebR 2019, Gebührenrichtlinien 2019, 10.5. Eingaben (§ 14 TP 6 GebG und § 12 GebG), 10.5.12. Gebührenfreie Eingaben - RZ 311 - 312'. The third result is a guideline from the BMF dated 26.03.2024, titled 'ZustRL, Richtlinien zur Zuständigkeit der Finanzämter, 1. Allgemeine Bestimmungen zur Zuständigkeit, 1.2. Zuständigkeit für die Entgegennahme von Anbringen - RZ 11 - 14'. The fourth result is a guideline from the BMF dated 26.03.2024, titled 'RAE, Richtlinien für die Abgabeneinhebung, 9. Säumniszuschläge (§ 217 BAO), 9.8. Vollstreckungsbescheid vor Ablauf einer Zahlungsfrist (§ 217 Abs. 6 BAO) - RZ 973 - 1008'. The fifth result is a guideline from the BMF dated 26.03.2024, titled 'RAE, Richtlinien für die Abgabeneinhebung, 7. Verwendung von Guthaben (§ 215 BAO) - RZ 800 - 829'. A red arrow points from the text 'Eine Ausschließung auf Keyword basierte Abfrage' to the search bar. Another red arrow points from the text 'Hypothese: Wie können wir mithilfe generativer KI, basierend auf FinDok Daten unseren Nutzenden eine natürlich-sprachliche Abfrage bereitstellen?' to the search results area.

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) Justicia 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 Echtdateien (100.000 Dokumente). Verprobung der zwei *Bereiche Semantische Suche und KI / LLM generierter Fragen-Beantwortung*

EXPERIMENTIEREN mit Sentence Embeddings & LLMs

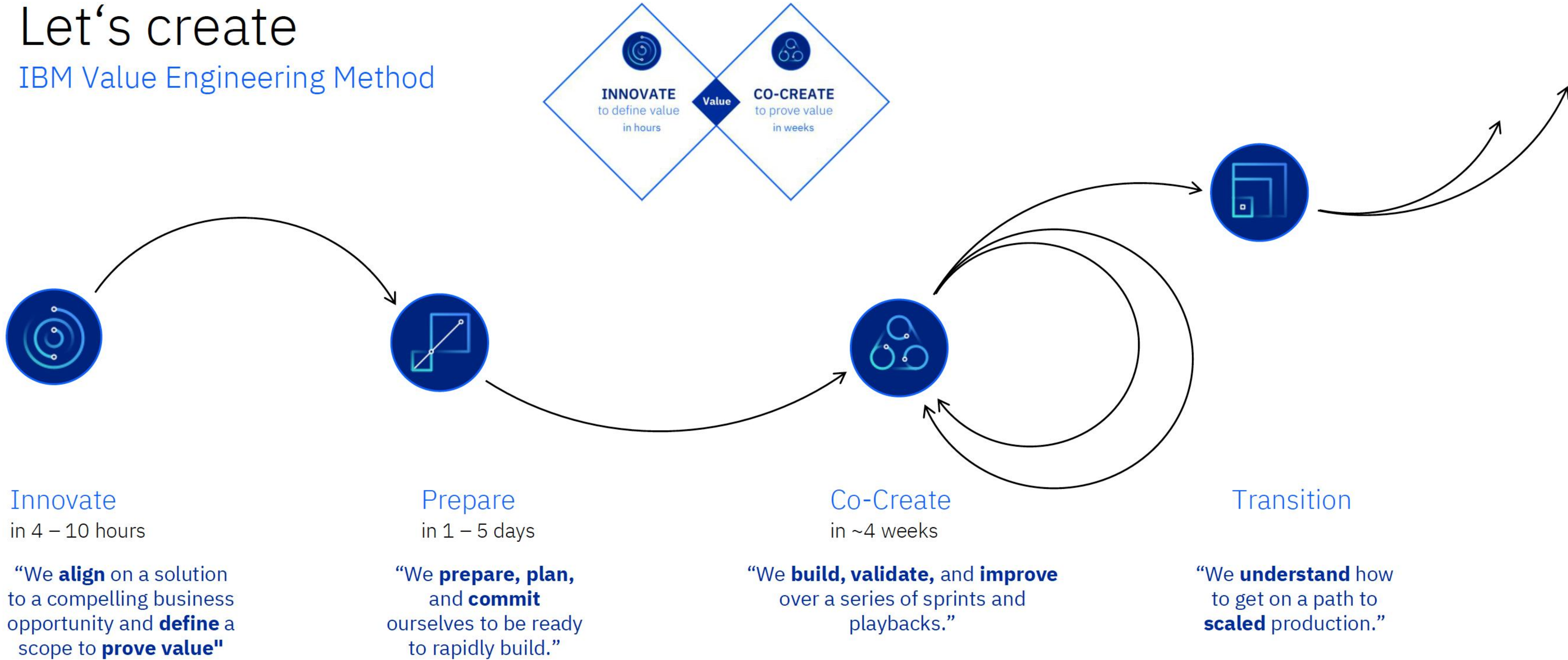
Experimenteller Vergleich und Bewertung von Alternativen unter Nutzung der .ai und .gov Werkzeuge wie Watson Studio, Prompt Studio, Deployment Spaces, Governance etc.

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



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

RETRIEVAL EXPERIMENTATION

PROMPT ENGINEERING

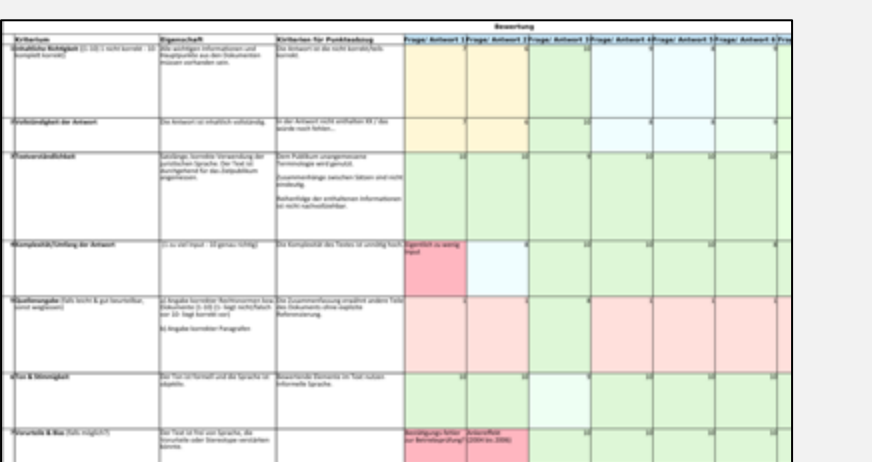
FEEDBACK VON EXPERTEN

VERBESSERUNG DER LÖSUNG



Embeddings	Chunking	Generation LLM	Recall at K	Cosine Similarity	BERTScore
PH-Alltags-embeddings-v0-mlu-mlu	Token based	mistral-8x7b-instruct-v0.1	0.57	0.34	0.62
PH-Alltags-embeddings-v2-base-de	Token based	mistral-8x7b-instruct-v0.1	0.43	0.38	0.62
PH-Alltags-embeddings-v3-large	Token based	mistral-8x7b-instruct-v0.1	0.71	0.83	0.65
PH-Alltags-embeddings-v4-large	chunk structure based	mistral-8x7b-instruct-v0.1	0.36	0.82	0.64

Name	Last modified
oqa-mistral-en-reference-legal-norm	None
hallucination_prompt	1 day ago
completeness_prompt	3 day ago
oqa-mistral-en-v3	4 days ago
oqa-mistral-en-v4	1 week ago
RAG-example-v2	1 week ago
Summarization-prompt-structured-with-examples	1 week ago
Summarization-prompt-v2	2 weeks ago



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

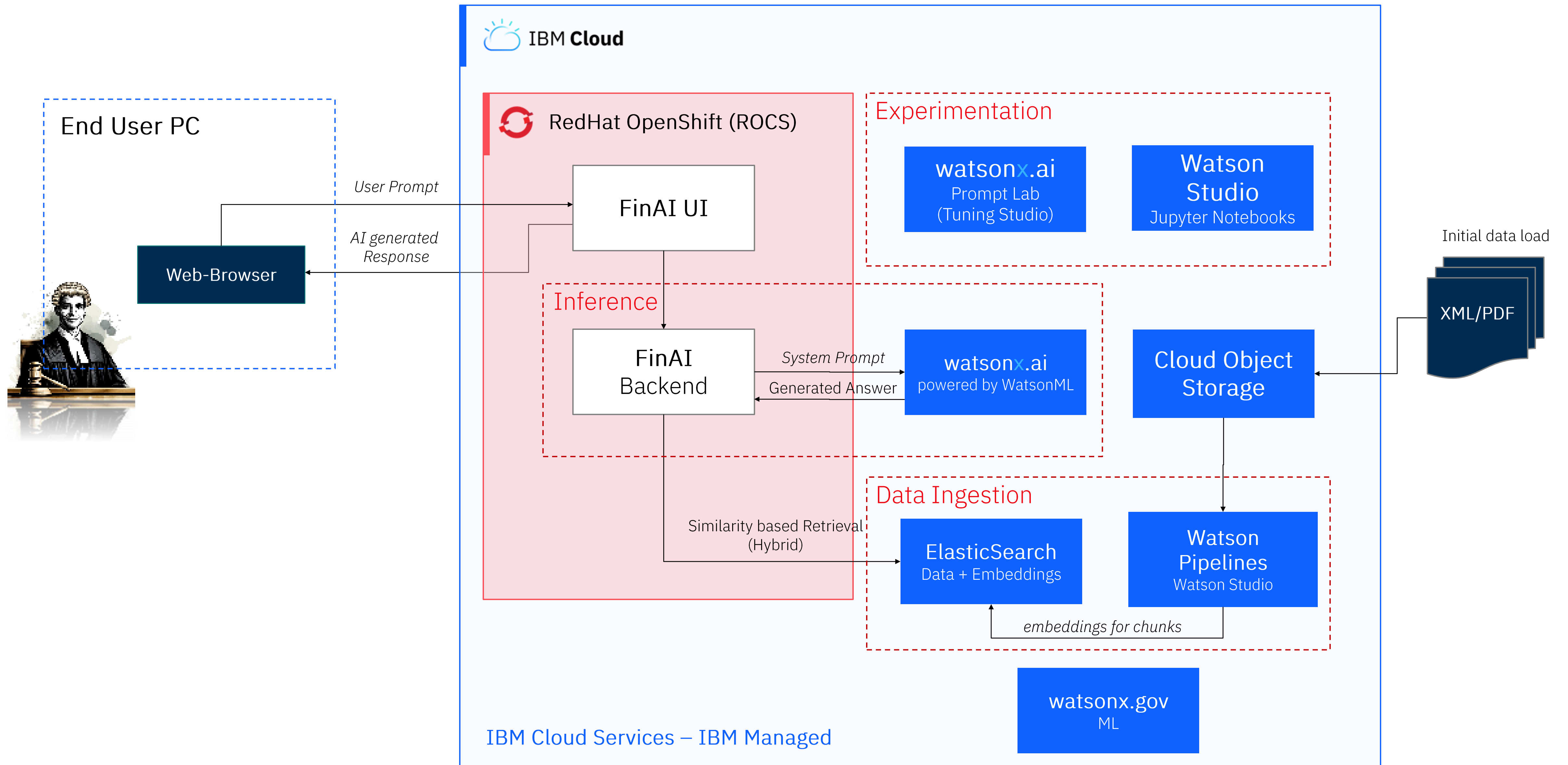
- Testen von 4 verschiedenen Einbettungsmodellen
- Testen verschiedener Chunking-Strategien
- Bewertung der Qualität des Abrufs mit der bereitgestellten Grundwahrheit
- Auswahl des Einbettungsmodells mit der besten Leistung

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

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

- Verwendung von Expertenfeedback zur Verfeinerung der Lösung

MVP Architektur Highlevel





FINAI SUCHE (NEU) ▾

Stellen Sie Ihre Frage...



Alle Ergebnisse

Rows per page: 5 ▾ 0-0 of 0 < >

DOKUMENTE ZUSAMMENFASSEN


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.

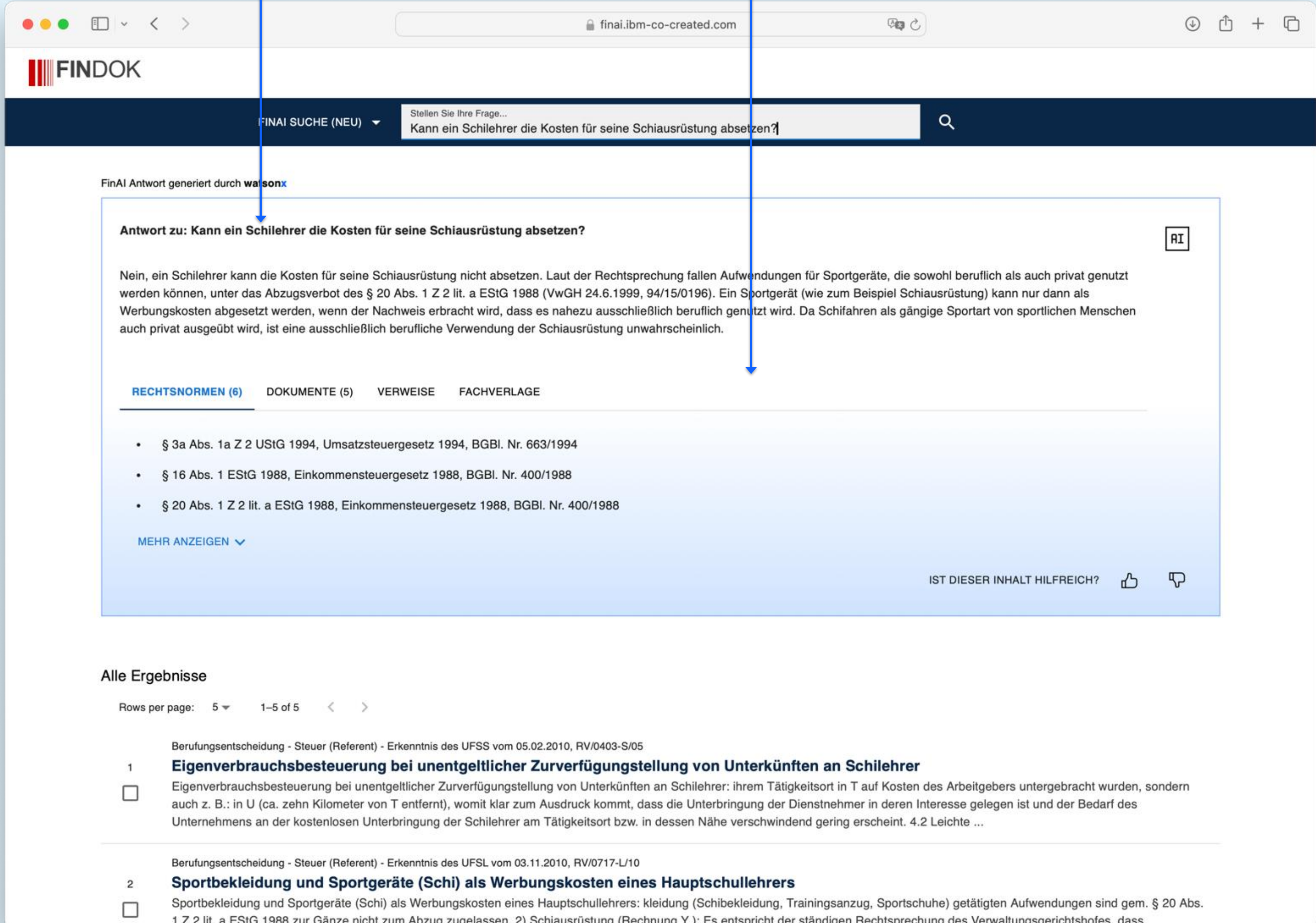
KEY FEATURES

 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 generierten Antwort

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



The screenshot shows a web browser window with the URL `finai.ibm-co-created.com`. The page header includes the **FINDOK** logo and a search bar containing the question: "Kann ein Schilehrer die Kosten für seine Schiausrüstung absetzen?". Below the search bar, the text "FinAI Antwort generiert durch waissonx" is displayed. The main content area shows the generated answer: "Antwort zu: Kann ein Schilehrer die Kosten für seine Schiausrüstung absetzen? Nein, ein Schilehrer kann die Kosten für seine Schiausrüstung nicht absetzen. Laut der Rechtsprechung fallen Aufwendungen für Sportgeräte, die sowohl beruflich als auch privat genutzt werden können, unter das Abzugsverbot des § 20 Abs. 1 Z 2 lit. a EStG 1988 (VwGH 24.6.1999, 94/15/0196). Ein Sportgerät (wie zum Beispiel Schiausrüstung) kann nur dann als Werbungskosten abgesetzt werden, wenn der Nachweis erbracht wird, dass es nahezu ausschließlich beruflich genutzt wird. Da Schifahren als gängige Sportart von sportlichen Menschen auch privat ausgeübt wird, ist eine ausschließlich berufliche Verwendung der Schiausrüstung unwahrscheinlich." Below the answer, there are tabs for "RECHTSNORMEN (6)", "DOKUMENTE (5)", "VERWEISE", and "FACHVERLAGE". A list of legal references is shown, including § 3a Abs. 1a Z 2 UStG 1994, § 16 Abs. 1 EStG 1988, and § 20 Abs. 1 Z 2 lit. a EStG 1988. At the bottom, there is a section titled "Alle Ergebnisse" with a table of search results, including a result about "Eigenverbrauchsbesteuerung bei unentgeltlicher Zurverfügungstellung von Unterkünften an Schilehrer" and another about "Sportbekleidung und Sportgeräte (Schi) als Werbungskosten eines Hauptschullehrers".

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 \(PACC\)](#)

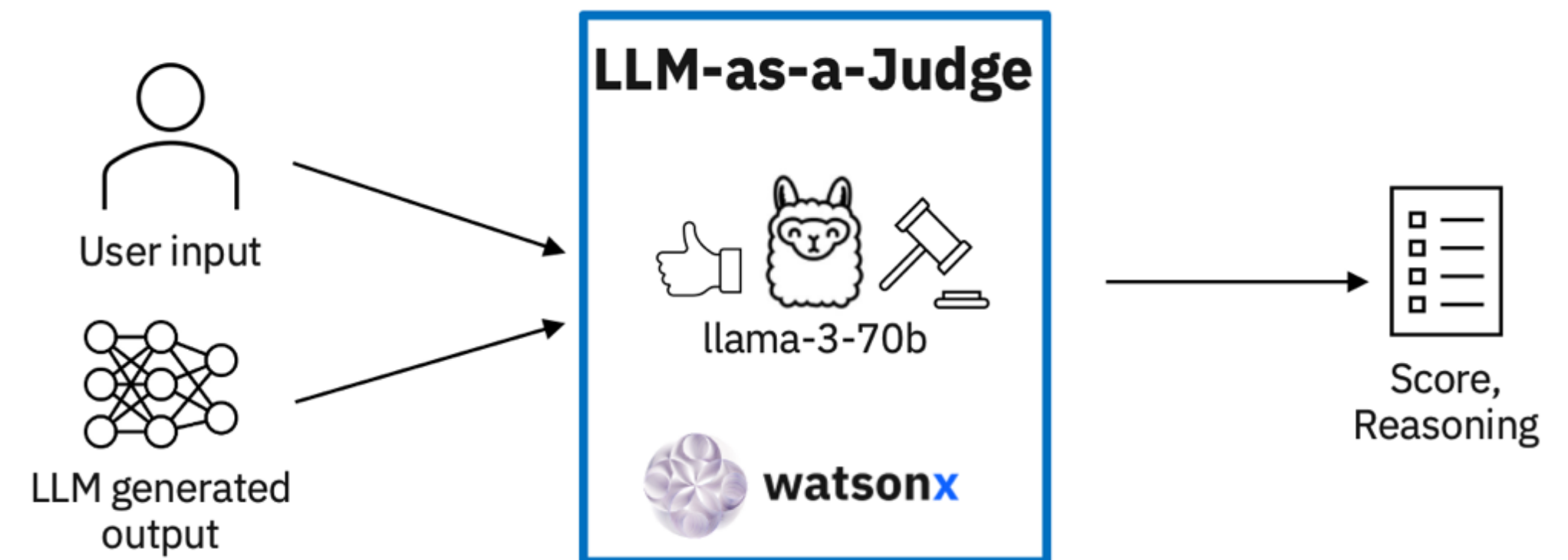
Metrics

[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)



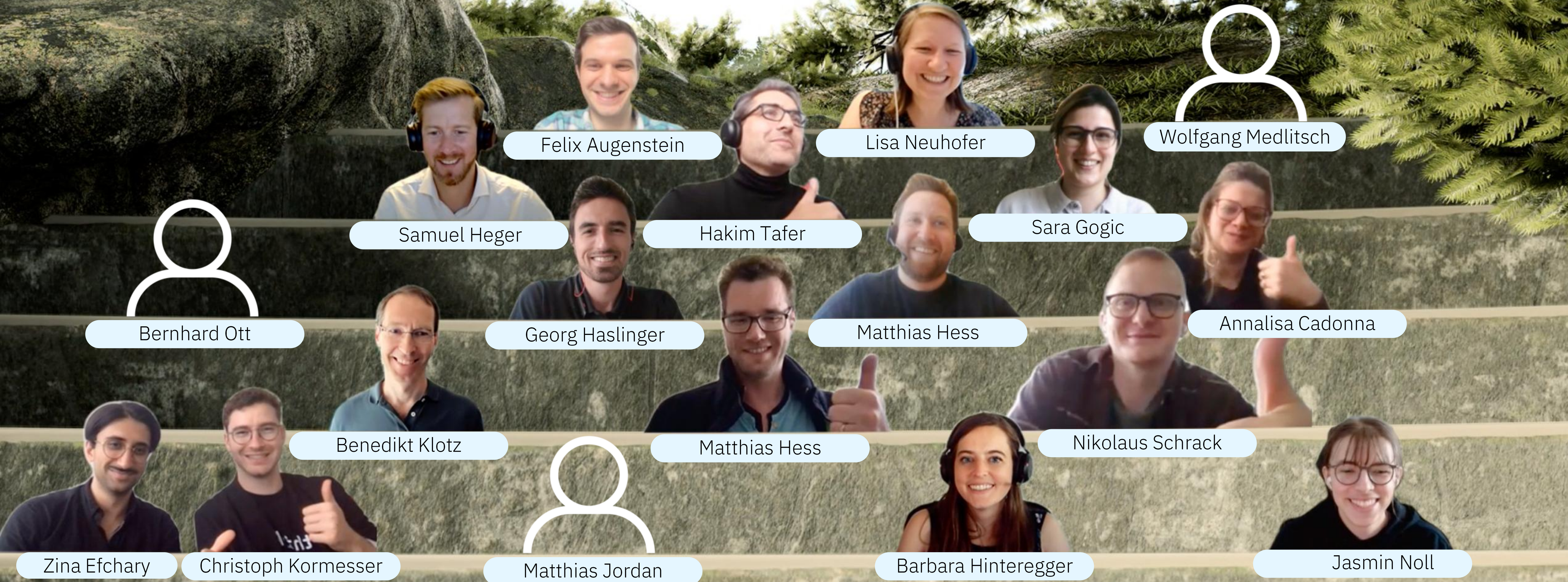
Metriken

[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?

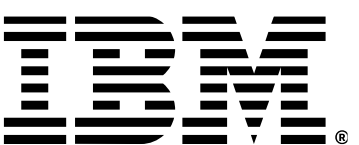
MVP Team



Ein gemeinsames Team



Bundesministerium
Finanzen



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 Hierarchie.

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 Antwortqualität 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

IBM