Bundesministerium
Europäische und
internationale Angelegenheiten

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY

CIB

KowHow: Studie zum Einsatz von KI-basierten Wissensmanagementsystemen im BMEIA

2026-03-31

K. Grosser	CIB solutions GmbH
R. King	AIT Austrian Institute of Technology GmbH
S. König	AIT Austrian Institute of Technology GmbH
D. Liakhovets	AIT Austrian Institute of Technology GmbH

Ausschreibung: KIRAS F&E-Dienstleistungen (FED) 2023

Projektnummer: FO999914199

Projekttitel: Wissensmanagement für den diplomatischen Dienst

Österreichisches Sicherheitsforschungs-Förderprogramm KIRAS – eine Initiative des Bundesministeriums für Finanzen (BMF)

Bundesministerium
Finanzen

FFG Forschung wirkt.

KIRAS Sicherheitsforschung

Inhalt

Einleitung

Die Motivation für das Projekt KnowHow ist die Herausforderung für Behörden, große Mengen an strukturierten und unstrukturierten Daten zu verwalten, um wertvolle Erkenntnisse zu gewinnen. Derzeit verfügt das Bundesministerium für Europäische und internationale Angelegenheiten (BMEIA) über kein übergreifendes internes Wissensmanagementsystem und wichtige digitale Objekte sind in verschiedenen Dateifreigaben in unterschiedlichen Formaten gespeichert. Herkömmliche Informationsmanagementsysteme haben oft Schwierigkeiten, relevante Informationen effizient zu verarbeiten und abzurufen, was zu Ineffizienz und verpassten Analysemöglichkeiten führt.

Ziele der Studie

Ein Ziel dieser Studie ist es, zu untersuchen, wie die Anwendung von modernen maschinellen Lerntechniken, semantischer Suche und großen Sprachmodellen (Large Language Models, oder LLMs) ein effizientes Wissensmanagementsystem für das BMEIA unterstützen könnte. Ein weiteres Ziel ist es, die Effizienzgewinne bei ausgewählten Wissensaufgaben, die mit fortschrittlichen KI-Techniken erzielt werden können, quantitativ nachzuweisen.

Zur Durchführung dieser Evaluierung und als Grundlage für die Studie wurde ein Demonstrator eines Wissensmanagementsystems auf Basis neuester Ansätze aus dem Bereich der Künstlichen Intelligenz entwickelt, implementiert und von Wissensarbeiterinnen und Wissensarbeitern im BMEIA getestet. Darüber hinaus wurde ein Tool zur Verwaltung und Auswertung der Bewertungsübungen entwickelt und implementiert.

Der Demonstrator konzentriert sich auf drei Hauptbereiche der KI-Anwendung: 1) die Verwendung von Transformationsmodellen in Kombination mit semantischer Indexierung, um eine semantische Suche und Abfrage zu ermöglichen und so die Genauigkeit und Relevanz der Suche, die Benutzerfreundlichkeit und den Zugang zu relevanten Informationen zu verbessern; 2) die Erstellung von Wissensstrukturen aus unstrukturierten Informationen durch die Anwendung von Modellen des maschinellen Lernens zur Extraktion von Entitäten und Beziehungen, was zu einer verbesserten Organisation, Analyse und Abfrage von Informationen führt; und 3) die Bewertung von Technologien zur Entscheidungsunterstützung durch die Implementierung semantischer Anreicherungsansätze wie Retrieval Augmented Generation (RAG), die bei der Durchführung von Analyseaufgaben Abfragen und Antworten in natürlicher Sprache ermöglichen.

Aufbau der Studie

Die Studie besteht aus mehreren Abschnitten, inklusive ein Literaturüberblick und eine Marktanalyse, eine Gap-Analyse, eine Anforderungsanalyse, eine Systemspezifikation, eine Technologieevaluierung, das Evaluierungsdesign und die Evaluierungsergebnisse.

Ein umfassender Literaturüberblick zur Anwendung von KI im Dokumenten- und Wissensmanagement – mit besonderem Fokus auf die öffentliche Verwaltung und den diplomatischen Dienst – wurde durchgeführt. Dabei konnten bestehende Forschungsergebnisse systematisch erfasst und Forschungslücken identifiziert werden. Parallel dazu wurde eine Marktanalyse des aktuellen Stands der Technik und der verfügbaren Lösungen durchgeführt. Es wurden Hardware-Empfehlungen erarbeitet und an Bedarfsträger übermittelt (Abschnitt 1).

Expertinnen und Experten des BMEIA wurden für die Gap-Analyse ausgewählt und mit ihnen Workshops vorbereitet und durchgeführt. Die Auswertung zeigte bestehende Lücken und Grenzen der aktuellen Informationsmanagementsysteme im BMEIA, insbesondere im Hinblick auf die effiziente Entdeckung und Abfrage von Wissen. Darauf aufbauend wurden Lösungsvorschläge zur Skalierung der Integration von KI in große, heterogene und multimodale Datensätze erarbeitet (Abschnitt 2).

Für die Anforderungsanalyse erfolgte zusätzliche technische Abstimmung in bilateralen Meetings mit Projektpartnern. Auf Basis der Workshops- und Recherche-Ergebnisse wurden Anforderungen an das prototypische KI-unterstützte KnowHow-Tool spezifiziert und User Stories formuliert welche eine Grundlage für die technische Implementierung des KnowHow-Tools bilden (Abschnitt 3, 4).

Forschungsarbeiten wurden durchgeführt, um die Technologien zu ermitteln, die für die Erfüllung der festgelegten Anforderungen am besten geeignet sind.

Es wurden Methoden zur variablen und Szenario-unabhängigen Wissensabfrage entwickelt, die flexible Informationszugriffe erlauben. Moderne Benutzerschnittstellen wie konversationsbasierte Dialogsysteme und explorative Suchansätze wurden erprobt. Dokumentenextraktions- und Retrieval-Pipelines wurden recherchiert, die eine effiziente Vorverarbeitung und Indexierung der Dokumenteninhalte gewährleisten.

Die Ergebnisse bildeten eine solide Grundlage für die Weiterverwendung in der Entwicklung des Demonstrators (Abschnitt 5, 6).

In Abstimmung mit der Spezifikation des KnowHow-Tools wurden im Rahmen des Anforderungsworkshops sowie in bilateralen Meetings die Anforderungen an das Evaluationsdesign definiert und dokumentiert. Diese inkludieren den Ablauf der Evaluationsphase, Anforderungen bzgl. Zusammensetzungen der Testgruppen, Anforderungen bzgl. der Evaluierungsaufgaben (Test-Cases) und der Evaluierungsumgebung (Abschnitt 7).

Die Evaluierungsumgebung wurde entsprechend den Anforderungen entwickelt und in der Zielumgebung aufgesetzt (Abschnitt 8), um die Durchführung der Studie zu ermöglichen. Im Verlauf der Evaluierungsphase wurden Daten im abgestimmten Umfang automatisiert gesammelt und für anschließende Auswertung persistiert. Die Evaluierungsphase wurde von BMEIA- und AIT-Expertinnen und -Experten laufend betreut, wobei Anfragen der Studienteilnehmerinnen und Studienteilnehmer anonymisiert bearbeitet wurden.

Nach dem Abschluss der Evaluierungsphase wurden Ergebnisse evaluiert und im Abschnitt 9 dokumentiert.

1 Literaturüberblick und Marktanalyse

1.1 Literaturrecherche

Eine umfassende Durchsicht der Literatur zu KI im Dokumenten- und Wissensmanagement, insbesondere im Kontext der öffentlichen Verwaltung, vor allem im diplomatischen Dienst, um vorhandenes Wissen und Forschungslücken zu identifizieren wird mit dieser Literatur Recherche vorgelegt. Die Analyse befasst sich mit der Anwendung von Wissensmanagement Systemen, verschiedenen Methoden im Bereich der KI gestützten Informationsspeicherung und Abfrage (z.B. RAG) sowie zugehörigen semantischen Datenbanken.

Die Problemstellung betrifft einen Wissensbestand, der sich hauptsächlich PDF- Dokumenten aus dem Intranet des BMEIA zusammengesetzt ist. Diese semi-strukturierten Daten sind nicht nach Inhalt, sondern überwiegend nach der publizierenden Stelle organisiert. Es liegen einfache Metadaten vor, und die Sprache der Dokumente ist Deutsch. Die zentralen Herausforderungen dieses Informationsbestands umfassen die Vielzahl von Abkürzungen, die große Anzahl ähnlicher Dokumente, zahlreiche Verweise auf andere Dokumente, sowie das Fehlen einer einheitlichen Struktur innerhalb der Dokumente.

Im Allgemeinen sind die meisten benötigten Informationen zwar vorhanden, jedoch ohne umfangreiche Erfahrung und ein tiefes Verständnis des Intranets nur schwer auffindbar.

1.1.1 Informations-Abfrage, Semantische Datenbanken und RAG

Dieser Abschnitt gibt einen Überblick über die jüngsten Fortschritte in KI gestützter Informations-Speicherung und Abfrage Systemen, insbesondere über sogenannte Retrieval Augmented Generation (RAG) Systeme und deren Anwendungen. Wichtige Erkenntnisse im Hinblick auf das vorliegende Problem werden dargestellt. Für semi-strukturierte Datensätze, wie PDF-Dokumente im KnowHow-Bereich, können effektive Vorverarbeitungstechniken wie Chunking-Strategien und Metadaten-Anreicherung (z. B. Seitenzusammenfassungen, reverse HyDE-Methoden [hypothetische Fragen, die ein Chunk beantworten könnte], Zeitstempel und Inhaltsverzeichnisse) die Leistung erheblich verbessern. Ein

"Chunk" ist ein Abschnitt des gesamten Wissenskorpus. Im Kontext eines PDF ist das beispielsweise ein Text-Absatz. Darüber hinaus sind strukturierte und hierarchische Index-Formate sowie Wissensgraph-basierte Indexierungs-Ansätze effiziente Methoden für Wissens Speicherung und Organisation [1].

Wissensgraph-basierte RAG-Methoden gewinnen zunehmend an Bedeutung, beispielsweise durch Techniken wie "Knowledge-Graph-Prompting". Hierbei erstellt ein LLM, ohne vordefinierte Ontologie, aus einem Datenkorpus automatisch einen Wissensgraph. Bei der Informations-Abfrage werden Anfragen in Graph-Abfragen (z. B. über SparQL) umgewandelt, um relevante Informationen aus dem Wissensgraphen abzurufen, die anschließend an ein LLM weitergeleitet werden. Dieser Ansatz wird durch Frameworks wie LlamaIndex oder LangChain unterstützt und kann mit Open-Source-Graph-Datenbanken wie Neo4j oder Nebula integriert werden [2].

Ein weiterer graphbasierter Ansatz, Graph RAG, nutzt die Modularität von Graphen, um Daten für die Zusammenfassung zu partitionieren und so globale Perspektiven für Multi-Dokument-Datensätze zu ermöglichen. Diese Strategie wurde von Microsoft als Open-Source-Projekt unter der MIT-Lizenz implementiert [3].

Weitere Daten-Indexierung Methoden umfassen RAPTOR, das hierarchische Baumstrukturen mit Zusammenfassungen und Ähnlichkeits-Clustern verwendet, um die Dokumentenrepräsentation zu optimieren [4]. Re2G führt domänenspezifisches Fine-Tuning von Re-Ranking-Mechanismen ein, um die Relevanz der abgerufenen Dokumente zu verbessern [5]. Studien wie RAGGED [6] und BERGEN geben Designrichtlinien für RAG Systeme und bieten Einblicke in die Anwendbarkeit von Datensätzen zur Evaluation [7]. Sie bieten somit Ansätze und Ideen für effektivere und anpassungsfähigere Implementierungen in verschiedenen Domänen.

Ein umfassender Überblick über die derzeit verwendeten semantischen Datenbanken, inklusive ihrer Eigenschaften, Spezifikationen sowie Vor- und Nachteile beschreiben Pan et al. [8].

1.1.2 KI im Wissensmanagement

Wissensmanagement - Systeme (Knowledge Management Systems, KMS) stellen eine Verbindung aus Technologie, Prozessen, Menschen und organisatorischen Kontexten dar, wie von Chouikha Zouari et al. (2018) beschrieben [9]. Diese vier Komponenten bilden das

Rückgrat effektiver KMS-Strukturen. Diese strukturelle Perspektive unterstreicht die Bedeutung einer Abstimmung technischer Werkzeuge mit menschlichen und organisatorischen Bedürfnissen, um sicherzustellen, dass das System nicht nur funktional, sondern auch in den Arbeitsalltag integriert ist.

Aufbauend auf dieser Grundlage untersucht Jarrahi (2023) [10] die potenzielle Synergie zwischen künstlicher Intelligenz (KI) und Wissensmanagement (KM). Die Studie betont, dass KI als ergänzender Partner für Menschen in KM-Aktivitäten agieren sollte, anstatt diese zu ersetzen. Es werden praktische Strategien für die Förderung dieser Partnerschaft vorgeschlagen, einschließlich der Rolle der KI bei der Unterstützung von Entscheidungsprozessen, der Automatisierung routinemäßiger Aufgaben und der Verbesserung des Zugriffs auf organisatorisches Wissen. Diese Anwendungen verdeutlichen das Potenzial von KI, die Effektivität von KMS zu steigern, sofern sie mit einem menschenzentrierten Ansatz gestaltet werden.

Die praktischen Herausforderungen bei der Implementierung von KI in Unternehmensumgebungen werden von Packowski et al. weiter beleuchtet [11]. Hier werden die Erfahrungen von IBM von in Verwendung befindlichen Retrieval-Augmented Generation (RAG)-Systemen dokumentiert. Der Artikel kritisiert traditionelle Evaluierungsmethoden und argumentiert, dass diese bei der Bearbeitung „neuer Fragen“ unzureichend sind, da solche Bewertungen menschliches Urteilsvermögen erfordern. Packowski liefert Richtlinien zur Strukturierung und Verbesserung Daten und Dokumenten, um deren Eignung für RAG-Anwendungen zu verbessern. Jiang et al. [12] liefern eine Untersuchung zur Akzeptanz von KI-Tools durch Anwender. Technische Lösungen funktionieren im Einklang mit menschlichen und organisatorischen Überlegungen.

1.1.3 Lösungen für domänenspezifische Dokumente

Abkürzungen, Akronyme

Die Verarbeitung von Abkürzungen in domänenspezifischen Dokumenten kann in den Phasen vor dem Indexieren, während des Indexierens und bei der Abfrage erfolgen. In der Vorindexierungsphase können Einbettungsmodelle anhand eines domänenspezifischen Glossars feinabgestimmt werden, um das Verständnis zu verbessern. Während des Indexierens werden Abkürzungen gemäß vordefinierten Glossaren oder Regelwerken umgewandelt. In der Abfragephase werden Techniken des Prompt Engineerings verwendet, bei denen das Glossar als Few-Shot-Prompt in das Kontextfenster integriert wird [13].

Duplikate, Stark ähnliche Dokumente

Die Verwaltung von Duplikaten und nahezu duplizierten Dokumenten kann während der Erstellung der Wissensbasis oder der Inferenz erfolgen.

Während des Indexierens werden exakte Duplikate entfernt, um den Korpus zu optimieren. In der Retrieval-Phase kann eine Filterung nach Zeitstempeln neuere Versionen von nahezu duplizierten Dokumenten priorisieren und als Aktualisierungen behandeln. Eine innovative Methode, wie von Weiss vorgeschlagen beinhaltet das Subtrahieren der Vektoreinbettungen der abgerufenen Dokumente vom Einbettungsvektor der Abfrage [14]. Dies führt zu einem „reduzierten“ Abfragevektor, der weniger erforschte Suchräume erkundet, die Ressourcendiversität erhöht und redundante Top-Ergebnisse beim Abruf minimiert.

1.1.4 Software-Aufzählung

Open-Source RAG Frameworks:

- Langchain
- Llamaindex
- Huggingface
- Haystack

Open-Source Vector Index Frameworks:

- Marqo
- Weaviate
- Pinecone
- Qdrant
- Chroma
- Milvus

Open-Source Evaluation Frameworks:

- RAGAS
- DeepEval

Vergleich:

- RAG tools comparison comparison: rag-tools
- Vector database comparison: vector-db-comparison

1.1.5 Fazit

Die Literaturrecherche schlägt eine Brücke zwischen KI gestutzter Informationsabfrage und Wissensmanagement. Sie identifiziert Retrieval-Augmented Generation (RAG) als einen zentralen Ansatz und untersucht Methoden zur Verbesserung solcher Systeme. Die Vorverarbeitung (pre-processing) von Daten sowie der Einsatz von Metadaten können die Leistung von RAG-Systemen erheblich steigern. Die Evaluierung komplexer Wissensmanagement-Systeme ist nicht trivial. Die Anwenderin oder der Anwender steht im Mittelpunkt und das erwartete Produkt muss mit diesem Anspruch designt werden.

Die Recherche hat keine expliziten Implementierungen fortschrittlicher Wissenssysteme im diplomatischen Dienst gefunden. Daher zielt die folgende Studie darauf ab, neue Ansätze in diesem Bereich bereitzustellen. Der Fokus des angeforderten Systems liegt auf der effizienten Umsetzung einer fortgeschrittenen Suchumgebung. Benutzer sollen Zugriff auf alle Informationen im Intranet haben. Ein RAG-System ist eine Weiterentwicklung der Suche in dem die abgerufenen Suchergebnisse durch ein LLM zu einer menschen-lesbare Zusammenfassungen kombiniert werden. Das angestrebte Ergebnis ist ein System, das Wissen über eine einfache Suchschnittstelle bereitstellen kann.

1.2 Marktanalyse

RAG (Retrieval Augmented Generation) ist eine neuartige Methode zur Erweiterung eines Netzwerkes um natürliche Daten. In diesem Text wird angenommen, dass die Daten in Dokumentform vorliegen. Dabei kann ein Suchindex über einen Datensatz erzeugt werden und relevante Suchtreffer können von einem großen generativem Sprachmodell eingeordnet und bewertet werden, welches daraufhin eine natursprachliche Antwort gibt. Diese Antwort bezieht sich aber auf konkrete inhaltliche Verweise auf den Datensatz. Naive RAG-Systeme können nur Bedeutungen innerhalb einzelner Dokumente finden, das kann mit GraphRAG-Systemen erweitert werden. Die Integration von Knowledge Graphen in RAG-Prozesse erweitert die Funktionalität erheblich, indem es dokumentübergreifende Bedeutungen durchsuchbar macht. Diese Fähigkeit, den gesamten Datensatz in die Analyse einzubeziehen, macht RAG in Kombination mit Knowledge Graphen zu einem mächtigen Werkzeug für datenintensive Geschäftsprozesse.

Zusätzlich zu der Leistung als Suchabfrage müssen RAG-Systeme in einer Softwareumgebung eingebettet werden, die praktische Nutzung durch Schriftgutbearbeiter möglich macht. Hier sind die Anforderungen im Außenministerium insbesondere eine genaue Abfrage der Gültigkeitszeiträume und zusätzlich Beachtung der Sicherheitsklassifizierung von Dokumenten.

1.2.1 Methodik und Vorgehensweise

Auf der Suchmaschine Google wurde nach folgenden Keywords gesucht um RAG-Anbieter zu finden:

- Chat with your data
- Chat with your document

Die Firmen in den ersten 10 Treffern wurden zusammengetragen und untersucht. Die Firmenliste wurde um den Author bekannte Firmen erweitert.

Anforderungen und Eigenschaften

Die für das RAG-Produkt nötigen Anforderungen wurden im Laufe des Arbeitspaket 2 erbrachten GAP-analyse und darauffolgenden ermittelt. Sie werden in drei Zusatzanforderungen operationalisiert, welche über das Standardangebot von RAG-Produkten (fortgeschrittenes semantisches DocumenQA über einen ganzen Datensatz) hinausgehen und die gefundenen Anbieter werden hiergehend untersucht.

- Sicherheit: aufgrund von vertraulichen Daten hat BMEIA den Anspruch die Datenhaltung intern zu halten. Es hat den Zusatzanspruch, dass einige Dokumente klassifiziert sind.
- CrossDocument Search: Das BMEIA verfügt über ein komplexes Dokumentökosystem, in dem komplexe Interaktionen zwischen Inhalten in unterschiedlichen Dokumenten. Dokumentübergreifende Bedeutungen zu identifizieren ist ein aktives Forschungsthema.
- Advanced Inputmanagment: Während der Indexierung eines Bestanddokumentensatz, oder beim Hinzufügen einzelner Dokumente können eine Reihe von Dokumentveredelungen teilautomatisch durchgeführt werden. Das inkludiert die Anwendung von:

- Hochleistungs-OCR: Dokumenteinhalte werden im Fall von Scans durchsuchbar gemacht
- Bildaufwertung im Fall von Scans
- Anreicherung der suchrelevanten [1]Metadaten durch fachliche Dokumentklassifikation
- Klassifikation über Semantik
- Multimodale Klassifikation
- Klassifikation von Barcodes
- Anreicherung von suchrelevanten Metadaten durch Dokumentextraktion

Tabelle 1 - Übersicht Marktrecherche

Anbieter	Produkt	Modalität	Sicherheit (MAC+Intern)	CrossDocument Search	Advanced InputManagment
CHATdoc [15]	chat-DOC webUI	Webservice	Nein	Nein	Nein
Sharly AI [16]	sharly webUI	Webservice	Nein	Ja	Nein
askyourpdf [17]	askyour-pdf API	Webservice	Nein	Nein	Nein
PDF.ai [18]	PDF.ai API	Webservice	Nein	Nein	Nein
MATHEMA GmbH [19]		Lokale Suche	Ja	Nein	Nein
TextCortex [20]	ZenoChat	Webservice & Lockale Suche	Ja	Ja	Nein
AI For Verticals, Inc [21]	YES/NO Agent	Webservice	Nein	Nein	Nein
acto GmbH [22]		Lokale Suche	Ja	Nein	Nein
inovex GmbH [23]	YourGPT	Lokale Suche	Ja	Nein	Nein

1.2.2 Fazit und Handlungsempfehlung

Fazit:

Keine derzeitigen identifizierten Lösungen deckt die Anforderungen des BMEIA vollumfänglich ab. Insbesondere gibt es Lücken beim vollumfänglichen Inputmanagement der angebotenen Lösungen, die somit die komplexe Dokumentstruktur des BMEIA nichtausreichend einlesen und wiederverwerten können. Weiters bieten viele der angebotenen Lösungen nur eingeschränkte Anpassungsfähigkeit. Dokumentübergreifende Suche wird meistens nicht angeboten, ist aber ein zentrales Feature für zueinander in Bezug stehenden Erlässen.

Handlungsempfehlung

Um die spezifischen Anforderungen des BMEIA an Datenhaltung, dokumentübergreifende Suche und fortgeschrittenes Inputmanagement zu erfüllen, empfiehlt es sich, im Rahmen des KnowHow-Projektes einen maßgeschneiderten Prototypen auf Llamaindex-Basis anzubinden. Dieser Prototyp sollte speziell auf die sichere Handhabung interner und klassifizierter Dokumente ausgerichtet sein und fortschrittliche Technologien für das Management und die Analyse komplexer Dokumentenstrukturen integrieren. Durch die Eigenentwicklung kann zudem sichergestellt werden, dass die Lösung flexibel an zukünftige Anforderungen angepasst werden kann und die Effektivität und Effizienz der Dokumentenverarbeitung signifikant verbessert wird.

1.3 Zusammenfassung

Die Literaturrecherche untersucht den Einsatz von KI im Dokumenten- und Wissensmanagement, insbesondere für die öffentliche Verwaltung und den diplomatischen Dienst. Zentrale Herausforderungen ergeben sich aus semi-strukturierten PDF-Dokumenten mit wenigen Metadaten und einer uneinheitlichen Struktur (1.1.1). Moderne Retrieval-Augmented Generation (RAG)-Systeme bieten Lösungen zur Informationsabfrage durch Techniken wie Chunking, Metadaten-Anreicherung und Graph-basierte Indexierung (1.1.2). Wissensmanagement-Systeme (KMS) müssen sowohl technische als auch organisatorische Aspekte integrieren, wobei KI als unterstützender Faktor betrachtet wird (1.1.3). Domänenspezifische Herausforderungen wie Abkürzungen, Duplikate und stark ähnliche Dokumente erfordern spezialisierte Vorverarbeitungstechniken (1.1.4).

Die Analyse zeigt, dass RAG-Systeme durch effektive Vorverarbeitung und semantische Indexierung eine vielversprechende Lösung für das Wissensmanagement in der öffentli-chen Verwaltung darstellen. Die vorhandene Forschung liefert zwar Methoden zur Opti-mierung solcher Systeme, jedoch existieren keine spezifischen Implementierungen für den diplomatischen Dienst (1.1.6).

Im Rahmen der Marktrecherche (1.2) wurden RAG-Lösungen in Bezug auf die Anforderungen des BMEIA untersucht. Keine der identifizierten Lösungen kann die Anforderungen vollumfänglich erfüllen, insbesondere betreffend das Inputmanagement aber auch mangels Anpassungsfähigkeit und erforderlicher dokumentenübergreifenden Such-Funktionalität. Aus diesem Grund wird eine Entwicklung des maßgeschneiderten prototy-pischen

Systems empfohlen, welches spezifische Anforderungen des Wissensmanage-ments im diplomatischen Dienst berücksichtigt und ausreichende Flexibilität für eine zu-künftige Weiterentwicklung bietet (1.2.3).

2 Gap-Analyse

2.1 Technologischer Status Quo und Zukunft des Wissensmanagement-Systems

Das Intranet ist das zentrale Informationsmanagement-System des BMEIA und wurde full-stack intern entwickelt. Aktuell funktioniert das System dokumentbasiert, sodass Inhalte als einzelne Dateien ins System eingepflegt werden und bei einem Suchvorgang eine Auflistung der Dokumente gefunden werden kann. Zukünftig soll das Intranet zu einem Content-basierten Wissensmanagement-System transformiert werden. Während derzeit das Intranet nur von der Zentrale in Wien befüllt wird, wäre es in der Zukunft denkbar, eigene Redaktionsbereiche für andere Organisationseinheiten (OE) bereitzustellen. Der Zugriff auf die Dokumente/ Inhalte im Intranet wird je nach Berechtigung eingeschränkt; die Zugriffsberechtigungen sollten in der Zukunft ggf. weiter verfeinert werden.

Aktuelle Herausforderungen betreffen insbesondere die Auffindbarkeit der relevanten Dokumente auf Basis der Volltext- oder Metadaten-Suche; ein Schwerpunkt der Intranet-Transformation liegt daher auf einer intelligenten, benutzerfreundlichen Suche. Die vorhandene Volltextsuche funktioniert prinzipiell, wird aber nur von einem beschränkten, in der Anwendung geübten Benutzerkreis verwendet. Die Suche ist derzeit nicht ausreichend benutzerfreundlich, hat keine alias-Suche und liefert ggf. zu viele Suchergebnisse. Die verbesserte Suche soll daher semantische Zusammenhänge berücksichtigen, wie bspw. unterschiedliche Fachbegriffe mit derselben Bedeutung oder Abkürzungen (diesbezüglich wäre auch ein Glossar denkbar), sowie einen Relevanz-Schwellenwert der Ergebnisse für die Suchanfrage. Eine Inhaltsbeschreibung von Dokumenten könnte helfen, diese gezielt mit semantischer Suche zu finden bzw. zu strukturieren.

Metadaten von den Dokumenten sind hilfreich bei der Suche, werden jedoch nicht einheitlich gepflegt. Aktuell werden Dokumente von vielen Redakteuren veröffentlicht, die dazugehörige Metadaten (bspw. Datum, Relevanz, Themen) uneinheitlich befüllen, was die Auffindbarkeit der Dokumente auf Basis von Metadaten erschwert. Dem soll in der Zukunft mit einer zentralen Redaktion und mit einer einheitlichen Metadaten-Struktur und Format entgegengewirkt werden.

Dokumente werden außerdem den Organisationsbereichen zugeordnet, da die Intranet-Struktur der Aufbauorganisation des Ministeriums folgt. Eine spezifische Herausforderung ist jedoch dabei, dass sich die Organisationsstruktur und die Zuständigkeiten in Bezug auf Themenbereiche im Laufe der Zeit ändern. Folgende Lösung wäre hierfür möglich: a) eine Entwicklung der Themenstruktur für Inhalte (ggf. halbautomatisiert möglich) und b) eine Zuordnung der OE den Themenbereichen (diese muss manuell erfolgen und gepflegt werden). Bei diesem Ablauf würden die Inhalte den Themen und nicht den OE zugeordnet und leichter gefunden werden können.

(Nahe-)Duplikate sowie unterschiedliche Versionen eines Dokumentes stellen eine weitere Herausforderung dar. Oftmals wird auch ein neuer Erlass über eine Änderung im Referenz-Erlass herausgegeben. All das erschwert das Beurteilen der Dokument-Gültigkeit und das Finden des aktuell gültigen Dokumentes. Ein Content-basiertes Informationssystem würde dieses Problem lösen, indem immer nur die aktuelle Version des jeweiligen Inhaltes verfügbar ist. Aus der Perspektive des dokumentenbasierten Systems wäre einerseits ggf. eine Strukturierung der ähnlichen Dokumente möglich, sodass die Zusammenhänge zwischen unterschiedlichen Versionen klar nachvollziehbar sind (bspw. Gültigkeitszeitraum oder -bereich, Änderungen etc.) Dafür bedarf es Anpassungen in den redaktionellen Prozessen, wie etwa die bereits erwähnte Verwendung von einheitlichen Metadaten-Strukturen. Gleichzeitig kann ein Hinweis auf ähnliche Dokumente seitens des Systems die Benutzerin oder den Benutzer auf ggf. widersprüchliche oder nicht eindeutig als gültig/ungültig gekennzeichnete Versionen aufmerksam machen.

Den BMEIA-Mitarbeiterinnen und -Mitarbeitern stehen heterogene Datenquellen zur Verfügung: Erlässe und Runderlässe (RE) (Anm.: In der Zentrale werden keine Erlässe, sondern Dienstzettel eingesetzt, um Weisungen mitzuteilen), RE-Sammlung und Handbuch für den auswärtigen Dienst (HAD). Erlässe und Runderlässe (RE geht an alle Botschaften) haben einen Weisungscharakter. Das HAD wird in vielen Dokumenten referenziert und ist ein wichtiger Arbeitsbehelf, hat jedoch keinen Weisungscharakter. Seine Struktur folgt nicht der Aufbauorganisation, sondern mehr einer fachlich-inhaltlichen Gruppierung. Die RE-Sammlung ist schlecht strukturiert und wird nicht konsequent gepflegt. Sie inkludiert zum Teil alte analoge RE, die als gescannte Dokumente ohne OCR-Verarbeitung abgelegt sind und nur nach Metadaten durchsuchbar sind (vorwiegend sehr alte Dokumente). Die RE-Sammlung hat aber noch rechtliche Relevanz. Für manche Dokumente sind auch digitale Versionen als PDF vorhanden, die nicht gescannt, sondern getippt sind. Eine Konsolidierung der Datenquellen ist im Laufe der angestrebten Intranet-Transformation erforderlich.

Neben der Dokument-Suche steht den BMEIA-Mitarbeiterinnen und -Mitarbeitern Personen- und OE-Suche zur Verfügung. Diese ist gut strukturiert (bspw. Suche nach Name, Telefon, Funktion, IKT-Beauftragte, Sektion, Vertretungsbehörde, Ort, Land etc.), implementiert eine alias-Suche und liefert auch zusätzliche Informationen wie Lokalzeit, Karte, Metadaten. Die Suche wird viel genutzt und soll in der gleichen Qualität bestehen bleiben.

2.2 Einsatzmöglichkeiten von KI-basierten Lösungen

Verbesserte Suche

Eine optimierte Suche nach Dokumenten oder Inhalten, implementiert als semantische bzw. hybride Suche, soll User Experience (UX) verbessern und das Auffinden der relevanten Inhalte erleichtern. Hierfür wird eine semantische Indexierung der Daten durchgeführt, die in weiterer Folge Anfragen in natürlicher Sprache ermöglicht, ohne dass es einer kompletten Übereinstimmung bedarf (wie derzeit bei der Volltextsuche). Diese Vorgehensweise erlaubt eine Alias-Suche, da der Abgleich auf Basis der semantischen Ähnlichkeit stattfindet. Kombiniert mit den Metadaten-basierten Restriktionen, wie bspw. Zeitraum oder Themenbereich, ermöglicht die hybride Suche das präzise Filtern von relevanten Inhalten. Unter Anwendung von angepassten bzw. anpassbaren Schwellenwerten für die Ähnlichkeit und Diversität der Inhalte soll eine übersichtliche Ergebnisliste erstellt werden.

Chat-Funktionalität

Neben der Suche nach Inhalten kann auch eine Automatisierung für weitere Aufgaben, wie bspw. das Erstellen der Zusammenfassungen oder die Extraktion von bestimmten Daten aus einem Dokument, hilfreich sein. Hierfür können Large Language Models (LLMs) eingesetzt werden, um Antworten in natürlicher Sprache zu generieren. Anschließend an einen Such-Vorgang oder durch das Formulieren einer Frage in natürlicher Sprache kann die Benutzerin oder der Benutzer einen Dialog starten, um auf Basis der relevanten Inhalte weitere Aufgaben zu bearbeiten.

Extraktion der Metadaten

Metadaten bilden eine wesentliche Grundlage für eine optimierte Suche. Zugleich wird jedoch Redaktionsarbeit durch das konsequente Befüllen von vielen Metadaten-Feldern erschwert. Um diese zu erleichtern, können manche Metadaten aus einem ins System einzupflegenden Dokument automatisiert extrahiert und den Benutzerinnen und Benutzern vorgeschlagen werden. Hierfür bieten sich bspw. NER oder LLM-basierte strukturierte Extraktion an.

Neustrukturierung der Dokumente/ Inhalte

Wie in 2.1 beschrieben, ist die aktuell vorhandene Strukturierung der Dokumente hinsichtlich der Suche mangelhaft. Eine Kategorisierung nach (ggf. feingranularen) Themenbereichen könnte unter Verwendung von Unsupervised Learning Methoden, wie Clustering von Dokumenten, vereinfacht werden. Die gefundenen Kategorien müssten von Expertinnen und Experten geprüft und benannt werden.

Zuordnung der Dokumente zu den Themenbereichen

Ähnlich wie die Vorschläge für andere Metadaten-Felder könnten Vorschläge für Themen-Zuordnung generiert werden. Eine genaue Klassifizierung kann jedoch je nach Anzahl der Kategorien schwierig sein und sollte den Redakteurinnen und Redakteuren überlassen werden.

Dokument-Vergleich

KI-basierter Vergleich der Dokumente könnte helfen, eine übersichtliche Darstellung der Dokument-Versionen und -Änderungen zu generieren.

OCR-Verarbeitung

Durch die OCR-Verarbeitung können die Inhalte von gescannten Dokumenten (TIFF, JPEG, PDF) für die Suche verfügbar gemacht werden.

2.3 Zusammenfassung

Der Abschnitt 2 beschrieb die Erkenntnisse aus der Gap-Analyse. Der Schwerpunkt des KnowHow-Projektes, die Entwicklung einer prototypischen Lösung für eine optimierte Suche für das BMEIA-Wissensmanagementsystem, wurde dabei im Kontext einer (Gesamt-)Transformation des Intranets betrachtet. Diese bedeutet sowohl Veränderungen im Informationsmanagementsystem (content-basiert anstatt dokument-basiert) als auch auf der Prozess-Ebene, insbesondere eine einheitliche Erfassung der erforderlichen Metadaten und ggf. eine zentrale Redaktion.

Die identifizierten Herausforderungen inkludieren Optimierung der Suche, Harmonisierung der heterogenen Datenquellen, Neustrukturierung der Inhalte nach Themen anstatt Organisationsbereichen, Vereinheitlichung der Metadatenerfassung, Vermeidung von Nahe-Duplikaten, Identifikation und Abgleich von unterschiedlichen Dokument-Versionen, eindeutige Kennzeichnung der Gültigkeit. KI-basierte Lösungen können dabei für die semantische Suche und Chat-Funktionalität eingesetzt werden, aber auch unterstützend bei der Neustrukturierung der Inhalte (bspw. Clustering der Dokumente) und Erfassung der Metadaten (Generierung der Vorschläge für Redakteurinnen und Redakteure) wirken. Mit dem Dokument-Vergleich könnten Nahe-Duplikate erkannt sowie ggf. verschiedene Versionen konsolidiert werden.

3 Anforderungsanalyse

Aufbauend auf Ergebnissen der Lückenanalyse werden in diesem Abschnitt Anforderungen an das prototypische System als User Stories beschrieben. Die Schwerpunkte sind dabei die optimierte Suche und deren Voraussetzungen, wie der Prozess der Dokument-Einreichung ins System. User Stories beschreiben Benutzerinteraktionen mit dem System und das notwendige In- und Output sowie Voraussetzungen und Risiken.

Die spezifischen Voraussetzungen werden in der jeweiligen User Story definiert, während die allgemein gültigen Voraussetzungen als allgemeine Anforderungen angeführt sind.

3.1 Allgemeine Anforderungen

- Folgende Anforderungen gelten für alle User Stories:
- Zugang zum System haben nur Personen, die sich gegenüber dem System authentifiziert haben.
- Den Personen sind direkt oder indirekt über die Organisationsstruktur Rollen zugewiesen, die die verfügbaren User Stories einschränken. Siehe erste Zeile der jeweiligen User Story.
- Für den vom System verwalteten Inhalt (Dokumente) gelten Sichtbarkeitsregeln und individuelle Rechte, die bestimmen, was eine Person sehen oder tun kann. (Autorisierung)
- Angemessene Antwort-Zeit: Wenn eine Antwort des Systems länger als 2 Sekunden dauert, wird diese Tatsache optisch signalisiert. Nach Möglichkeit wird ständig signalisiert, dass das System noch arbeitet.
- Wenn das Ergebnis einer User Story nicht erreicht werden kann, wird der fachliche Grund in der Anwendung dem Benutzer mitgeteilt.
- Technische Gründe werden als solche ohne technische Details mitgeteilt und weitere Informationen für den Support der Anwendung zentral abgelegt.
- In jedem Fall werden weitere Handlungsmöglichkeiten mitgeteilt. (Zum Beispiel Suche genauer spezifizieren, wenn zu viele relevante Treffer oder später erneut versuchen, wenn ein Basissystem temporär nicht verfügbar ist.)

3.2 User Stories

US 1: Dokumente hinzufügen	Rolle: Redakteurin/ Redakteur (BMEIA)	Priorität: Hoch

Beschreibung:

Userin oder User fügt ein Dokument in das Suchsystem ein und reicht dabei suchrelevante Metadaten ein. Falls bereits eins oder mehrere Dokumente mit sehr ähnlichen Inhalten vorhanden sind, bekommt Userin oder User einen Hinweis sowie eine Liste dieser Dokumente. Außerdem wird eine Textzusammenfassung (Abstract) des eingefügten Dokuments hinterlegt.

Voraussetzungen:

a) Dokument
b) suchrelevante Metadaten. Suchrelevante Metadaten werden in eine Maske eingegeben. Im Projektverlauf wird geprüft welche der Metadaten sich automatisch ermitteln lassen. Diese werden dann in der Maske vorbelegt.
c) Datum des Dokumentes (für das Alter des Dokumentes und der Information; eventuell automatisch aus dem Dokument zu ermitteln; falls das Alter einer Information für die Relevanz keine Rolle spielt, sollte das Feld nicht erforderlich sein, höchstens als Suchkriterium)
d) Gültig ab/seit, wenn abweichend, falls benötigt (falls die Suche alle relevanten Stellen inklusive veralteter Stellen und den Überarbeitungen finden kann und LLM in der Lage sind daraus abzuleiten, was gilt, dann sollte das Feld nicht nötig sein.)
e) Gültig bis, falls vorhanden und benötigt (dito)
f) Ist ein gültiges Dokument (Falls die Eigenschaft nicht über andere Wege, zum Beispiel über neuere „Versionen“, ermittelt werden kann. Vgl. dito)
g) Sicherheitsstufe
h) Liste von Organisationseinheiten und Personen, die Stellen finden dürfen. (erforderlich, falls die Sicherheitsstufe nicht reicht.)
i) Organisationseinheit als Eigentümer (falls benötigt, um Entstehung zu hinterfragen oder als Suchkriterium)
j) Dokumentklasse, falls man zum Beispiel Erlasse und anderes klar unterscheiden möchte (Suchkriterium)
k) Themenbereiche
l) Weitere Suchkriterien

Ergebnisse:

1. Das Dokument ist durch die Suche auffindbar. Die vollständige Indizierung kann bis zu einer Stunde dauern. Jedoch sind Dokumente über ihren Link sofort nach Erfassung abrufbar. Der Abstract kann als Begleitinfo zum Dokument im Intranet eingepflegt werden (vorausgesetzt, ein entsprechendes Feld in den Datenstrukturen ist vorgesehen).

Risiken:

1. Userin oder User kann die richtigen Metadaten nicht bereitstellen. Das Dokument kann schlechter gefunden werden.

Eintrittswahrscheinlichkeit: niedrig
Auswirkung: hoch
Mitigation: Die Metadaten müssen die für die Suche nötige Werte enthalten. Um die Akzeptanz der Benutzer bei der Erfassung zu haben, sollten die Metadaten leicht verständlich und nicht zu viele sein. Einige Metadaten könnten automatisch vorgeschlagen werden.
Fall-back: Prüfungen zur Vollständigkeit von Metadaten bei der Erfassung von Dokumenten.

2. Der direkt extrahierbare Text des Dokumentes entspricht nicht dem lesbaren Inhalt. Das Dokument kann schlechter gefunden werden.

Eintrittswahrscheinlichkeit: niedrig
Auswirkung: hoch
Mitigation: Der Textinhalt muss möglichst vollumfänglich in die Suche einfließen. Das kann eine Texterkennung (OCR) erfordern.
Fall-back: Prüfungen zur Textqualität bei der Erfassung von Dokumenten.

3. Das Dokument ist ein Bild oder eine Grafik oder Skizze. Das Dokument hat keinen beschreibenden, lesbaren Text.

Eintrittswahrscheinlichkeit: niedrig
Auswirkung: hoch
Mitigation: Beschreibung als Teil der Metadaten
Fall-back: Verhindern solcher Inhalte.

Bewertungskriterien:
1. Dokumente müssen wenigstens über eindeutige Wörter durch die Volltextsuche gefunden werden.
2. Dokumente müssen wenigstens über die Suche nach den eingegebenen Metadaten gefunden werden.

Kommentar des Entwicklungsteams:

Der Text kann mit einer Texterkennung erschlossen werden. Die Qualität der Sprache kann automatisch ermittelt werden. Metadaten können über Klassifizierung oder Datenextraktion gewonnen werden. Die Dokumenterfassung ist somit ein ganzer Prozess.
Das neue Dokument kann andere Dokumente, die bereits erfasst wurden, ungültig machen. Das sieht CIB als wichtige Anwendung, geschieht aber nach heutiger Planung nicht automatisch.

US 2: Suche nach Dokumenten	**Rolle: Mitarbeiterin/ Mitarbeiter (BMEIA)**	**Priorität: Hoch**

Beschreibung:

Userin oder User führt eine Suche nach relevanten Dokumenten aus und bekommt eine priorisierte Ergebnisliste.

Voraussetzungen:
a) Benutzer-Anfrage ist für die Suche geeignet.
b) Texteingabe-Feld
c) Metadaten-basierte Filter (bspw. Zeitrahmen)
d) Schwellenwerte für Relevanz und Diversität der Fundstellen

Ergebnisse:
1. Priorisierte Fundstellen-Liste (mit Verweisen auf Dokumente und Relevanz-Indikator)
2. Fundstellen sind so umfangreich, dass genügend Kontext zum Verständnis angezeigt wird, aber gleichzeitig so klein, dass sie möglichst schnell gelesen werden können und genügend viele auf einmal sichtbar sind. Fundstellenretrieval beinhält zusätzlichen Dokumentkontext, wenn vorhanden und möglich.

3. Nicht genügend diverse Fundstellen werden ungeachtet der Relevanz zusammengefasst als eine einzige angezeigt. Dazu wird die relevanteste der Fundstellen ausgewählt. Zusätzlich wird jeweils die Anzahl der unterdrückten Fundstellen angezeigt.
4. Die so zusammengefassten Fundstellen können pro Zusammenfassung eingeblendet werden.
5. Die Fundstellen-Liste ist in der Anzahl limitiert, umfasst maßgeblich relevante Fundstellen und unterdrückt durch die Limitierung keine relevanten Fundstellen.
6. In die Relevanz spielen Ähnlichkeit zum Volltext, semantische Nähe (auch mehrsprachig), Alter der Information und Gültigkeit hinein.
7. Wenn die Texteingabe wenige Stichwörter enthält, werden Fundstellen über Volltextsuche und Suche nach Synonymen gefunden.
8. Wenn die Texteingabe eine Aussage ist, werden semantisch ähnliche Fundstellen gefunden.
9. Wenn die Texteingabe eine Frage ist, werden Fundstellen gefunden, die die Frage beantworten helfen.

Risiken:

1. Limitierung zu klein,
   was häufig zu Meldungen wegen nicht angezeigter relevanter Fundstellen führt.

Eintrittswahrscheinlichkeit: gering
Auswirkung: mittel
Mitigation: Umfangreiche Tests auf einem genügenden Dokumentenkorpus
Fall-back: Limit erhöhen. Nachladen auf Wunsch des Bedieners.

2. Limitierung zu hoch,
   was häufig zu langwierigem Blättern in zu langen Fundstellenlisten führt.

Eintrittswahrscheinlichkeit: gering
Auswirkung: gering
Mitigation: Umfangreiche Tests auf einem genügenden Dokumentenkorpus
Fall-back: Limit verkleinern.

3. Diversitätsbewertung trennt unscharf.
   (Folge: Es kann kein Schwellwert eingestellt werden, der für alle Fundstellen die sehr ähnlichen korrekt ausblendet. Relevante und “gute” Fundstellen bleiben zunächst unsichtbar, können jedoch eingeblendet werden.)

Eintrittswahrscheinlichkeit: mittel
Auswirkung: mittel
Mitigation: Entwicklung guter Fundstellen-Cluster mit mehreren unterschiedlichen Merkmalen
Fall-back: Optional die Diversitätsbewertung und das Ausblenden deaktivieren.

4. Viele Fundstellen aus demselben Dokument
   führen dazu, dass weitere relevante Dokumente nicht sofort zu sehen sind.

Eintrittswahrscheinlichkeit: niedrig
Auswirkung: mittel
Mitigation: Die Relevanzbewertung sollte dazu führen, dass „bessere“ Fundstellen aus anderen Dokumenten vorrangig angezeigt werden. Die Warnung vor weiteren relevanten Ergebnissen weist den Benutzer darauf hin, die Suche zu verfeinern.
Fall-back: Optional Steuerung des Verhaltens bezüglich Aufwertung von Fundstellen nach Häufigkeit im selben Dokument. Optionales Gliedern nach Dokument analog zur Diversität.

5. Fundstellengröße reicht nicht für das Verständnis.
   (Benutzer müssen oft in das Dokument sehen; mehr Klicks)

Eintrittswahrscheinlichkeit: niedrig
Auswirkung: niedrig
Mitigation: Während der Entwicklung auf Erfahrungen mit den Daten setzen.

Fall-back: Nachkorrektur und Neuindizieren.

6. Relevanzbewertung
   führt häufig dazu, dass die gewünschten Fundstellen und Dokumente nicht an den ersten Stellen stehen.
   (Die Bewertung gemäß Semantik, Gültigkeit, Alter und Volltextvorkommnisse zu einer Liste hat viele Parameter.)

Eintrittswahrscheinlichkeit: mittel
Auswirkung: mittel
Mitigation: Während der Entwicklung auf Erfahrungen mit den Daten setzen.
Fall-back: Nachjustieren der Parameter. Optionale Steuerung der Parameter durch den Benutzer. Harte Kriterien für Gültigkeit und Alter als Filter.

7. Veraltete Informationen in der Fundstellenliste
   (Durch neue Versionen ungültig gewordene Fundstellen oder durch das Alter nicht mehr hilfreiche Daten. Analog könnten Dokumente mit zukünftiger Gültigkeit eine Rolle spielen.)

Eintrittswahrscheinlichkeit: mittel
Auswirkung: mittel
Mitigation: Die Verwendung der Fundstellen ist sowieso durch Benutzer im Einzelfall zu prüfen und zwar sowohl aus dem sachlichen Kontext als auch bezogen auf die Gültigkeit und das Alter. Gute Metadaten können harte Bedingungen ermöglichen und Fehler verringern. Dennoch verbleibt ein Restrisiko, dass Metadaten nicht gut erfasst werden und sich Benutzer auf die angebotenen Fundstellen ungeprüft verlassen.
Fall-back: Wenn das Risiko nicht eingegangen werden kann, ist möglicherweise ein Vieraugenprinzip für die dazu erforderlichen Metadaten hilfreich. (Zum Beispiel gegenseitiger Freigabeprozess der Redakteure für die Dokumentenerfassung.)

Bewertungskriterien:

Neben manueller, subjektiver Bewertung im Rahmen einer Evaluierung durch Benutzer mit realen (migrierten?) Dokumenten, kann ein maschineller, objektiver Test zu einem gewissen Grad die Qualität der Relevanz- und Diversitätsbewertung über Schwellwerte und Limits messen.

Zur maschinellen Auswertung kann ein umfangreicher und diverser Dokumentensatz als Basis verwendet werden. Maschinell können ein Prozentsatz von Duplikaten und Varianten, sowie Textpassagen direkt oder umformuliert als Suchanfragen erzeugt werden. Künstlich können Gültigkeit und Alter festgelegt werden. Die Erwartung ist, dass die Platzierung der entsprechenden Dokumente in der Ergebnisliste entsprechend hoch ist. Das kann wiederum maschinell ausgewertet und zu einer Qualitätsmaßzahl berechnet werden.

Kommentar des Entwicklungsteams:

Möglichkeiten zur Optimierung der Relevanz: Aufgrund Anzahl der Fundstellen pro Dokument die Relevanz nach oben oder unten korrigieren. Diversitätsbewertung durch Clusterbildung der Fundstellen.

Interner Ablauf:
1. Volltextsuche über extrahierte Stichwörter und Synonyme,
2. Semantikbewertung und Ähnlichkeitssuche,
3. Filtern gemäß Sichtbarkeitsregeln und Metadaten,
4. Filtern von Duplikaten,
5. Kontext anreichern,
6. Diversitätsbewertung über Cluster (Reranking zur Ermittlung des Stellvertreters?),
7. Relevanzberechnung für Indikator (Klassifikation in Ampelfarben?),

8. Reranking der genügend diversen Fundstellen (Hauptergebnisse/Stellvertreter),
9. Limitierung,
10. Warnungen,
11. Teile der Fundstellen hervorheben (Stichwörter und Synonyme oder der „beste“ Satz).
12. Anzeige mit Links auf Dokumente
13. Ggf. Dokumentenbetrachter

US 3: Anfrage in natürlicher Sprache	**Rolle: Mitarbeiterin/ Mitarbeiter (BMEIA)**	**Priorität: Mittel**

Beschreibung:

Userin oder User formuliert eine Anfrage in natürlicher Frage. Darauffolgend werden relevante Dokumente/ Chunks gefunden, priorisiert und darauf basierend wird eine Antwort formuliert. Ggf. wird die Antwort in einem Chat-Verlauf verfeinert.

Der Ablauf ist zunächst mit US 2 identisch und deswegen auch die Voraussetzungen, Ergebnisse und Risiken zuzüglich der folgenden:

Voraussetzungen:

a) Das Textfeld enthält eine Frage oder einen Auftrag.

Ergebnisse:

1. Antwort in natürlicher Sprache (mit Verweis auf die Fundstellen und Dokumente)
2. Fundstellenliste, die jetzt noch (von Userin/User) priorisiert werden kann. (Abschluss durch Aufforderung zur neuen Antwort.)
3. Hinweis(e) auf eine ggf. nicht oder nicht ganz korrekte Antwort.

Risiken:

1. Die Antwort ist zu knapp oder zu ausführlich.

Eintrittswahrscheinlichkeit: mittel
Auswirkung: niedrig
Mitigation: Promptengineering
Fall-back: Zunächst kurz antworten lassen mit Möglichkeiten zur Verlängerung, Parameter für Prompting

Bewertungskriterien:

Subjektive Bewertung im Rahmen einer Evaluierung oder Maschinelle Bewertung durch ein sehr großes Sprachmodell.

Kommentar des Entwicklungsteams:

Wenn man die User Stories 2 und 3 noch ausführlicher beschreiben möchte, indem man den Weg bis zum Eingabefeld für die Anfrage noch aufnimmt, dann stellt sich die Frage, ob der Benutzer wirklich vorher schon weiß oder sich Gedanken machen möchte, ob er mit einem kurzen Text nur Stellen suchen oder mit einer längeren Frage oder Formulierung eine Antwort bekommen möchte.
Vielleicht ist es einfacher für den Benutzer, wenn man US 3 als ausgebauten US 2 versteht und für diese Ausbaustufe nur das Eingabefeld vergrößert und im Ergebnis die Antwort ergänzt. Wenn überhaupt eine Fallunterscheidung nötig ist, dann wird sie im Hintergrund getan.
Im gleichen Sinne wären US 4, US 5 und US 6 auch nur weitere Schaltflächen und Ausbaustufen.

US 3a: Bürgerinnen- und Bürger-Anfrage in natürlicher Sprache	Rolle: Mitarbeiterin/ Mitarbeiter (BMEIA)	Priorität: Mittel

Beschreibung:

Userin/ User kopiert/formuliert eine Bürgerinnen- und Bürger-Anfrage (bspw. E-Mail) in natürlicher Frage in häufigen Anwendungssprachen. Anhand von diesem Text wird durch LLM eine mögliche Anfrage an das System formuliert. Darauffolgend werden relevante Dokumente/ Chunks gefunden, priorisiert und darauf basierend wird eine Antwort formuliert. Userin/ User kann die Klassifikationsstufen (public/low/high) der Dokumente wählen, die für die Antwortformulierung herangezogen werden. Userin/ User kann die Sprache des Antwortentwurfes vor der Generierung wählen. Ggf. wird die Antwort in einem Chat-Verlauf verfeinert.

Voraussetzungen:

a) Das Textfeld enthält eine Frage oder einen Auftrag.

Ergebnisse:

1. Antwort in natürlicher Sprache (mit Verweis auf die automatisch generierte Anfrage, Fundstellen und Dokumente)
1. Fundstellenliste, die jetzt noch (von Userin/ User) priorisiert werden kann. (Abschluss durch Aufforderung zur neuen Antwort.)
2. Hinweis(e) auf eine ggf. nicht oder nicht ganz korrekte Antwort.
3. Im Falle der Verwendung von klassifizierten Dokumenten (Sicherheitsstufe "hoch") bekommt die Userin/ User eine Warnung bzw. einen Hinweis darauf.

Risiken:

1. Die automatisch generierte Anfrage weicht von der Anfrage ab
   Eintrittswahrscheinlichkeit: mittel
   Auswirkung: mittel
   Mitigation: Promptengineering
   Fall-back: Zunächst kurz antworten lassen mit Möglichkeiten zur Verlängerung, Parameter für Prompting

(Wie US 3.)

Bewertungskriterien:

Wie US 3.

Kommentar des Entwicklungsteams:

Wie US 3.

US 4: Fundstellen-basierter Dialog	Rolle: Mitarbeiterin/ Mitarbeiter (BMEIA)	Priorität: Niedrig

Beschreibung:

Userin oder User hat bereits US 2 oder US 3 ausgeführt und erstellt in einem Dialog mit LLM einen verfeinerten (neu formulierten, angepassten) Text.

Voraussetzungen:

a) Die Fundstellenliste wurde nach US 2 oder US 3 priorisiert. (Abschluss durch Aufforderung zum Dialog.)

Ergebnisse:

1. Angepasster Text, bspw. für weitere Recherche und Verwendung außerhalb des Systems.
2. Möglichkeit, den Dialog in immer neuen Eingabefeldern fortzusetzen.

Risiken:

1. Normales Risiko bei Nutzung heutiger LLM.

Eintrittswahrscheinlichkeit: hoch
Auswirkung: niedrig
Mitigation: Klare Anweisung zur Prüfung der Ergebnisse anhand der Quellen und eigener Erfahrungen.
Fall-back: Kein bekannter Fallback zu LLM-Technologien.

Bewertungskriterien:

Subjektive Evaluierung (Feedback) von Userin/ User.

Kommentar des Entwicklungsteams:

US 5: Dokument-basierter Dialog	**Rolle: Mitarbeiterin/ Mitarbeiter (BMEIA)**	**Priorität: Mittel**

Beschreibung:

Userin/ User hat US 2 oder US 3 ausgeführt und wählt eines oder mehrere Dokumente (Quellen der Fundstellen) aus. Userin/ User lässt LLM die ausgewählten Dokumente weiterverarbeiten, z.B. zusammenfassen.

Voraussetzungen:

a) Wie US 4, nur werden nach der priorisierten Fundstellen die Inhalte der gesamten Dokumente als Eingabe für das LLM verwendet

Ergebnisse:

1. Wie US 4

Risiken:

Wie US 4

Bewertungskriterien:

Subjektive Evaluierung (Feedback) von Userin/ User.

Kommentar des Entwicklungsteams:

US 6: Dokument-Vergleich	**Rolle: Mitarbeiterin/ Mitarbeiter/ Redakteurin/ Redakteur (BMEIA)**	**Priorität: -**

Beschreibung:

Userin/ User führt eine Suche gemäß US 3 durch bis hin zur Priorisierung und wählt dann den Vergleich der Dokumente. Als Antwort kommt eine Zusammenfassung der gemeinsamen Inhalte und der Unterschiede unter Verwendung der gesamten Dokumentinhalte.

Voraussetzungen:

a) Die Fundstellenliste wurde priorisiert. (Abschluss durch Aufforderung zum Dokumentvergleich.)

Ergebnisse:

1. Zusammenfassung der gemeinsamen Inhalte und der Unterschiede unter Verwendung der gesamten Dokumentinhalte

Risiken:
Wie US 3

Bewertungskriterien:

Subjektive Evaluierung (Feedback) von Userin/ User.

Kommentar des Entwicklungsteams:

US 6 wird als Teil der US 1 gedeckt.
Wir gehen davon aus, dass die selektierten Dokumente in den unterstützten Kontext des LLM passen. Fallback wäre ein komplizierter Ansatz analog zu einem Vektor- oder Grafen-Diff-Algorithmus auf Basis der Chunks der Dokumente.

US 7: Anzeige von nützlichen und funktionierenden Anfragen	**Rolle: Mitarbeiterin/ Mitarbeiter (BMEIA)**	**Priorität: Mittel**

Beschreibung:

Userin/ User bekommt Beispiele/ Vorschläge von häufigen und gut funktionierenden Prompts bzw. Anfragen.
Userin/ User wird angeleitet wie textuelle Eingaben in das System zu erfolgen haben.

Voraussetzungen:

a) Ausreichende Datengrundlage (Userinnen- und User-Anfragen) mit Bewertungen hinsichtlich Nützlichkeit.

Ergebnisse:

1. Userinnen und User erhalten eine Trainingsmöglichkeit für den Umgang mit einem KI-Assistenten, indem sie funktionierende und nützliche Prompts sehen.

Risiken:

1. Datenschutzrechtliche Bedenken

Eintrittswahrscheinlichkeit: mittel
Auswirkung: mittel

Mitigation: Zustimmung der Userinnen und User der Datenspeicherung, Anonymisierung der Daten
Fall-back: Verwendung vorgefertigter Beispiele
2. Nicht ausreichende Datengrundlage
Eintrittswahrscheinlichkeit: mittel
Auswirkung: mittel
Mitigation: Nutzerinnen und Nutzern zur Verwendung des Systems animieren; Trainingsmöglichkeiten anbieten
Fall-back: Verwendung vorgefertigter Beispiele

3. Vorschläge sind nicht vielfältig genug
Eintrittswahrscheinlichkeit: mittel
Auswirkung: gering
Mitigation: Verschiedene Clustern-Algorithmen testen
Fall-back: Verwendung vorgefertigter Beispiele

Bewertungskriterien:

Subjektive Evaluierung (Feedback) von Userin oder User.

Kommentar des Entwicklungsteams:

Vorschläge werden gegebenenfalls anhand von "Häufigen Suchen" generiert. Userin oder User kann in einem Übungschat ausprobieren, eine Aufgabe zu üben.

3.3 Zusammenfassung

Im Abschnitt 3 wurden die Anforderungen an das prototypische System in Form von User Stories definiert, sowie allgemeine Voraussetzungen für alle User Stories wie bspw. Zugangsberechtigung oder angemessene Antwort-Zeit. Die User Stories fokussieren sich auf dem Inputmanagement, Such- und Chat-Funktionalität. Insgesamt wurden 8 User Stories definiert: "US1 Dokumente hinzufügen", "US2 Suche nach Dokumenten", "US3 Anfrage in natürlicher Sprache", "US3.a Bürgerinnen- und Bürger-Anfrage in natürlicher Sprache", "US4 Fundstellen-basierter Dialog", "US5 Dokument-basierter Dialog", "US6 Dokument-Vergleich", "US7 Anzeige von nützlichen und funktionierenden Anfragen". Die US3.a kann als Use-Case-spezifische Erweiterung von US3 betrachtet werden; US6 wird tlw. im Rahmen von US1 implementiert.

Die Erkenntnisse aus der Anforderungsanalyse bilden eine Grundlage für die Ausarbeitung der technischen Spezifikationen für die prototypischen Lösung sowie für das Evaluierungsdesign.

4 Systemspezifikation Demonstrator (KnowHow Tool)

Das Ziel der Spezifikation ist es, eine ausreichende Grundlage für die Entwicklung des Demonstrators (KnowHow Tools) bereitzustellen, sodass sowohl die technischen Aspekte für den Betrieb des Tools als auch die Evaluierungs-spezifischen Anforderungen bekannt sind und in der Umsetzung berücksichtigt werden.

Dieser Abschnitt enthält die Spezifikation der technischen Anforderungen, Anforderungen hinsichtlich Datensicherheit und Sicherheitsmaßnahmen sowie eine Beschreibung der Komponenten und deren Kommunikation. Ergänzend zur technischen Spezifikation wurden nicht-funktionale Anforderungen.

4.1 Systemanforderungen, Hardware- und Softwareanforderungen

4.1.1 Systemanforderungen

GPU-Anforderungen:

- Manche Docker-Container benötigen die GPU. Dafür muss die NVIDIA-Docker-Runtime sowie CUDA und CUDNN installiert werden.
- Das passende Linux-System muss ebenfalls installiert werden. Weitere Anforderungen an die Linux-Distribution gibt es nicht.

4.1.2 Hardwareanforderungen

- 1 × CPU ≥ 64 Kerne / 128 Threads
- 188 Gi ECC-RAM
- 1 × GPU ≥ 48 GB VRAM
- 1 × NVMe-SSD ≥ 1,7 TB für System & Containers

4.1.3 Softwareanforderungen

- Betriebssystem: 64-bit Linux (getestet mit Ubuntu 22.04 LTS, Kernel ≥ 5.15)
- Containerisierung: Docker ≥ 26 und docker-compose V2
- GPU-Unterstützung: NVIDIA Treiber≥570.124.06, Container Toolkit inkl. CUDA ≥ 12.6 und cuDNN passend zum Treiber

4.2 Sicherheitsmaßnahmen, Datensicherheit

4.2.1 Sicherheitsmaßnahmen

- Sicherheitsrelevante Authentifizierungs- und Zugriffskonzepte werden durch das BMEIA definiert und umgesetzt.
- Die Absicherung der Infrastruktur sowie die Definition von Zugriffsregelungen erfolgen durch das BMEIA.
- Zugriffe auf Anwendungen werden entsprechend organisatorischer und technischer Vorgaben gesteuert und kontrolliert.

4.2.2 Datensicherheit

- Die Datensicherheit ist durch den Einsatz von zwei gespiegelten SSDs gegeben.

4.3 Schnittstellen und Komponenten

- Installation der Suchsystemkomponenten:
- Alle Komponenten für das Suchsystem werden von CIB über Docker und Docker-Compose installiert.
- Teile werden auch vom AIT verändert und nachinstalliert.
- Docker-Container:
- Unter den mit Docker-Compose installierten Docker-Containern befinden sich API- und Webanwendungen, die sowohl von Entwicklerinnen und Entwicklern der AIT oder CIB als auch vom BMEIA genutzt werden müssen.
- Docker-Images:

- CIB installiert die Docker-Container über Docker-Images aus https://harbor.cib.de. Sollte dies aus dem BMEIA heraus nicht funktionieren, kann auch ein Downloadserver (FTP) verwendet werden oder die Docker Images als ZIP Files übertragen werden.

Tabelle 2 - Komponenten

Komponente	Funktion	Schnittstelle
Gateway-API	Zentraler REST-Endpunkt für Index, Ingest, Search und Generate.	HTTPS / OpenAPI 3.1
Workflow-Engine (LlamaDeploy)	Orchestriert Index-, Ingest-, Search- und RAG-Workflows, skaliert Tasks asynchron.	gRPC / interner Message-Bus
OpenSearch	Speicherung von Text- und Vektor-indizes; Hybrid-Suche (BM25 + k-NN).	HTTPS / REST
MinIO	Objektspeicher für Quelldokumente.	S3-kompatibel
Embedding-Service	Erzeugt Vektorembeddings (Sentence Transformers o. Ä.).	REST
Reranking-Service	Ordnet Trefferliste neu mittels Cross-Encoder.	REST
LLM-Service	Generiert Antworten, Zusammenfassungen und Rewrite-Prompts.	REST

4.3.1 Datenfluss - Übersicht

- Index → Gateway erstellt Index in OpenSearch (k-NN-Parameter, Synonyme, Pipeline-Config).
- Ingest → Dokumente werden in MinIO gespeichert, in Sätze zerlegt, vektorisiert und in OpenSearch abgelegt; optional wird eine Zusammenfassung der ersten Seiten generiert.
- Search → Gateway führt wahlweise lexikale, vektorbasierte oder hybride Suche durch; Ergebnisse werden rerankt und per MMR diversifiziert.
- Generate (RAG) → Workflow ruft LLM-Service auf und liefert Antwort inkl. Token- und Kostenstatistik.

4.3.2 Skalierbarkeit & Betrieb

- Stateless Gateway ermöglicht horizontales Scaling (Kubernetes, Docker Swarm)
- OpenSearch-Cluster: konfigurierbare Shards und Replikate; Vektorindizes optional produkt-quantisiert (HNSW + Faiss)

4.4 Nicht-funktionale Anforderungen

Tabelle 3 - Nicht-funktionale Anforderungen

Kategorie	Zielvorgabe
Datenschutz	Speicherung ausschließlich Text-Chunks; Originaldokumente liegen im Objektspeicher geschützt durch Username/Password
Internationalisierung	Unterstützung für DE, EN, ES, RU, ZH, AR in Suche und RAG
Wartbarkeit	Sämtliche Services laufen in Container-Images aus reproduzierbaren Builds (CI/CD-Pipeline)

5 Umsetzung und Dokumentation des Demonstrators

Dieser Abschnitt beschreibt die Umsetzung und Funktionalität des KnowHow-Demonstrators basierend auf der „CIB smartER“ Software. Der Demonstrator realisiert ein wiederverwendbares Retrieval-Augmented-Generation-(RAG)-System auf Basis von LlamaIndex und LlamaDeploy. Ziel ist eine skalierbare, modular erweiterbare Plattform für Indexaufbau, Dokument-Ingest, Suche (lexikalisch, semantisch, hybrid) und antwortgenerierende RAG-Workflows. Die Suchergebnisse werden erweitert mit dem KnowHow Knowledge Graph, welcher auf Grundlage einer angepassten Ontologie gebildet wird.

5.1 RAG-System

5.1.1 Implementierung

Architekturübersicht: CIB smartER kombiniert OpenSearch für Vektor- und Volltextsuche, einen Embedding-Service, einen (optionalen) Reranking-Service sowie einen LLM-Service. LlamaDeploy [24] orchestriert als Ausführungs- und Orchestrierungsumgebung die LlamaIndex-Workflows [25] (Index, Ingest, Search, RAG). Ein FastAPI-basiertes Gateway kapselt und publiziert die Workflows als konsistente REST-Schnittstellen.

OpenSearch: Für semantische Suche werden Embeddings als k-NN-Index [26] abgelegt. Standardmäßig wird FAISS als Engine mit HNSW und Cosine Similarity verwendet. Hybrid Search kombiniert lexikalische (BM25) und semantische Ergebnisse über Normalisierung (min-max bzw. L2) und Fusionsverfahren (arithmetisches/geometrisches/harmonisches Mittel; optional RRF seit OpenSearch 2.19).

Workflows: LlamaIndex-Workflows sind ereignisgetrieben und bestehen aus einzelnen Steps (z. B. Query-Parsing → Retrieval → Reranking → LLM-Antwort → Post-Processing). Die Orchestrierung erfolgt durch LlamaDeploy (API-Server, Control-Plane, Orchestrator, Services, Message-Queue).

5.1.2 Daten-Ingest

Der Ingest-Workflow übernimmt Upload und Verarbeitung von Dokumenten pro Index. Unterstützte Formate umfassen u. a. PDF und DOCX. Der Ablauf umfasst: Text-Extraktion, Node-Parsing (Satz-Splitter mit chunk_size=512/overlap=20), Embedding-Generierung und Speicherung der Nodes in OpenSearch. Bei bestehenden docId-Werten erfolgt ein Replace-Ingest.

Originaldokumente können in MinIO gesichert werden; in OpenSearch werden ausschließlich verarbeitete Text-Chunks (Nodes) und Metadaten persistiert.

5.2 Knowledge Graph

5.2.1 Implementierung

Die zweite Komponente des Know-How Tools ist ein Wissensgraph (vlg. Knowledge-Graph). Die Notwendigkeit ist durch komplizierte Verknüpfungen von Dokumenten in den BMEIA Daten gegeben. Das bedeutet, in der Datengrundlage bestehen Referenzen zwischen Dokumenten, welche durch das bestehende System nicht abgebildet werden. In Dokument A wird ein anderes Dokument B referenziert, welches im Korpus keine weitere Verbindung zu Dokument A aufweist, sie sind zum Beispiel nicht im selben Verzeichnis, oder haben ein ähnliches Publikationsdatum. Der Wissensgraph löst diese Situation auf, indem er diese Verbindungen herstellt und sichtbar macht.

Der Wissensgraph setzt Dokumente als Entitäten zueinander in Beziehung. Ein Wissensgraph besteht aus Knoten und Kanten zwischen den Knoten. Die Dokumente (zb. "Gesetz A") sind die Knoten (vlg. Nodes) im Graph und die Beziehungen zwischen den Dokumenten (zb. "referenziert") sind die Kanten im Graph. Die Aufgabe des Wissensgraph ist somit die Beziehungen zwischen BMEIA Dokumenten abzubilden und bei einer gegebenen Suchanfrage wiederzugeben.

Die Erstellung eines Wissensgraphen erfordert zwei Eingangsvariablen. Einerseits eine handgefertigte Ontologie, welche die vorhandene Domäne bestmöglich beschreibt. Andererseits die tatsächlichen Dokumente, also die gesamten Informationen, unterschiedlichster Dokumenttypen, des BMEIA Datensatzes.

Die Ontologie beschreibt alle relevanten Entitäten und Beziehungen zwischen Ihnen. Sie bezieht sich immer auf eine bestimmte Domäne, in diesem Fall die BMEIA Domäne. Die Ontologie bildet somit die d

Der Wissensgraph wird in eine Neo4J Datenbank eingespeist, dort erhalten und abgefragt. Neo4J ist eine Open-Source Graph Datenbank, die eine programmatische Ein- und Ausgabe ermöglicht. [1]

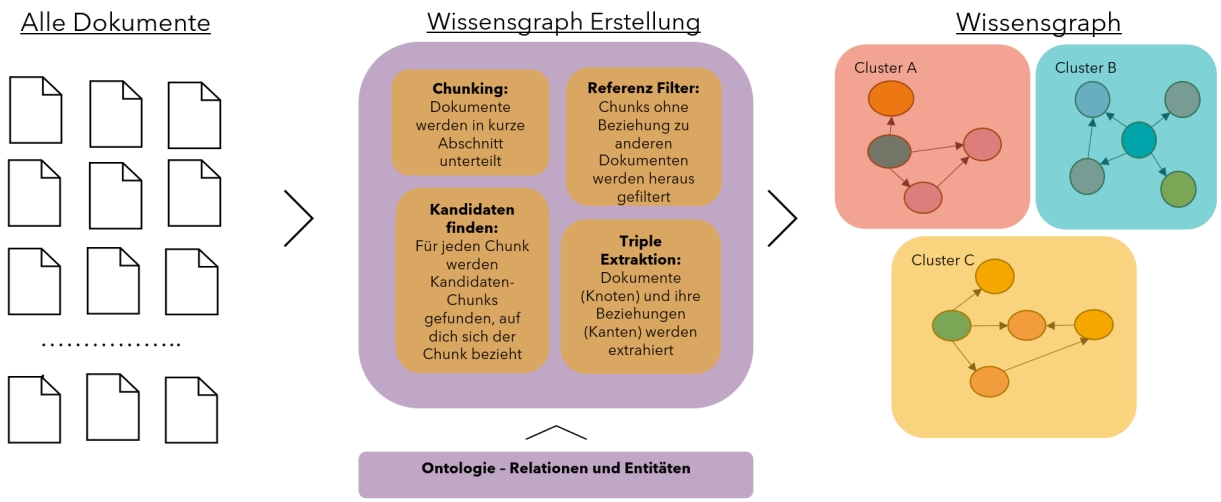

Abbildung 1 - Wissensgraph Erstellung - Detail

5.2.2 Daten-Ingest

Der Wissensgraph wird auf Basis einer Ontologie erstellt. Aus dem gesamten Datensatz werden zur Wissensgraph Erstellung, sogenannte Triples (Entität, Relation, Entiät) extrahiert. Ausschließlich Triples die in der vordefinierten Ontologie enthalten sind, sind zulässig. Im KnowHow Projekt passiert die Verarbeitung der Dokumente zu kleineren Einheiten (chunks) bereits in der RAG-Datenvorverarbeitung. Diese Sub-Dokumente werden, gemeinsam mit der Ontologie, als Input für die Wissensgraph Funktion verwendet. Der Output dieser Funktion sind die Wissens-Triples, die dann als Wissensgraph repräsentiert sind. Der Graph wird dann in einer Neo4J Datenbank gespeichert.

Die Wissensgraph-Erstellung dem Dokumentensatz (Chunks) besteht aus Input, Output und vier Verarbeitungs-Phasen:

b) Eingang: Gesamter BMEIA Datensatz, als Dokumentenabschnitte (Chunks)
c) Filter 1: Chunks werden nach Referenzen auf andere Dokumente gefiltert
d) Filter 2: Für die übergebliebenen Chunks aus Filter 1 werden, mittels semantischer Suche, zusammenhängende Chunks gesucht. Hier wird eine Kandidatenliste für jeden Eintrag gesucht und weiter zum nächsten Schritt gereicht
e) Extraktion der Triple aus Chunk und seiner Kandidatenliste an Chunks aus referenzierten Dokumenten. Ein Triple wir nur extrahiert, wenn es in der Ontologie erlaubt wird.
f) Eintragen der Triple in Wissensgraph Datenbank
g) Ausgang: Wissensgraph

2.

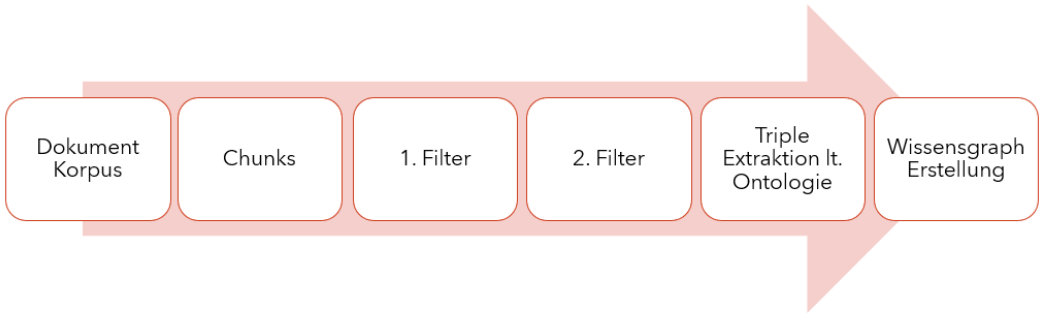

Abbildung 2 - Wissensgraph Erstellung - Überblick

5.2.3 Daten-Abfrage

Eine Anfrage an den Wissensgraph wird der ersten Komponente im Know-How Tool, dem RAG-System, nachgelagert. Die Vektordatenbank-Abfrage im RAG-Protokoll findet eine voreingestellte Anzahl an Dokumenten-Chunks. Jeder dieser Chunks hat eine ID. Auf Basis dieser "Chunk-ID" wird der Wissensgraph abgefragt, der ebenfalls zu jedem Eintrag eine "Chunk-ID" gespeichert hat. Eine Anfrage an den Wissensgraph returniert alle Chunks, die mit dem gesuchten Chunk in Verbindung stehen, sprich eine Relation zwischen ihnen existiert. Außerdem wird auch die Art der Verbindung wiedergegeben. Die Wissensgraph Abfrage erfolgt mit der Graph-Abfragen-Sprache "Cypher". [2]

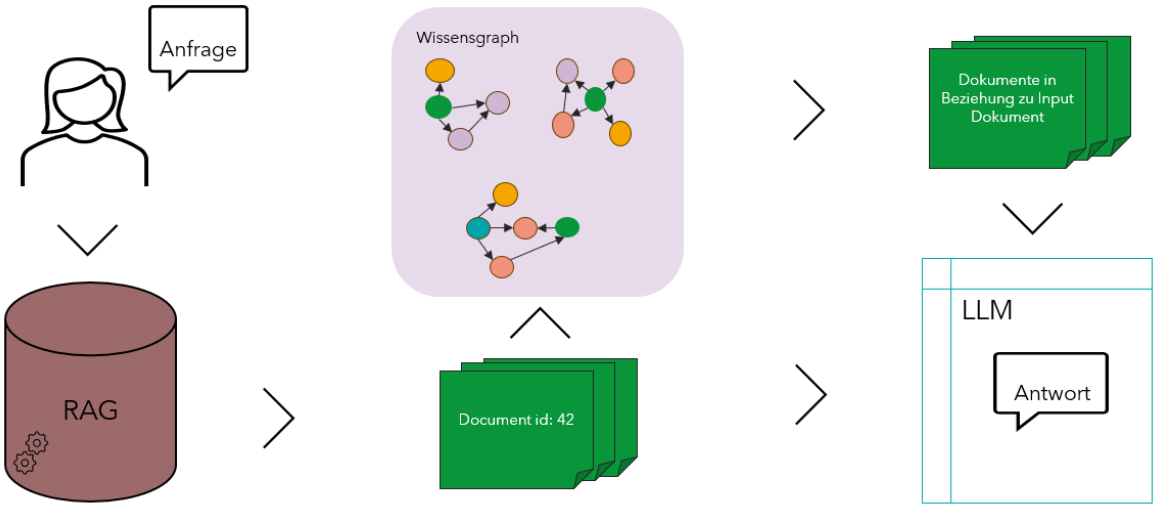

Abbildung 3 - Wissensgraph Abfrage

5.2.4 Verwendete Software

Tabelle 4 - Übersicht Software (Knowledge Graph)

Anwendung	Software
Deployment	Docker auf Ubuntu
Programmiersprachen	Python, Docker, Cypher
LLMs für Referenz-Filter, Kandidatengenerierung und Triple Extraktion	Llama3.1:8b, gemma3:12b, phi4-mini:3.8b
Wissensgraph Datenbank	Neo4J Community (open source)

5.3 Schnittstellen und Integration

Das CIB smartER Gateway stellt eine REST-API bereit und vermittelt zwischen Caller und LlamaDeploy-Workflows. Kernendpunkte sind:

• POST /index/{index_id} – Index anlegen (u. a. embeddingServiceConfig, Optional: Synonyme, knnMethod, searchPipelineConfig, indexSettings)

• DELETE /index/{index_id} – Index löschen

• POST /ingest/{index_id} – Dokument(e) einreichen (multipart/form-data; nodeParser, summarizeFirstPages, embedding/llm-Konfiguration)

• POST /search/{index_id} – Node-Suche (mode=fulltext|semantic|fulltextSemantic, Filter, Reranking)

• POST /generate/{index_id} – RAG-Antwort (optional nodeList/docList, Conversation-History, Context-Window, Qualitätsprüfungen)

6 Demonstrator

6.1 Umsetzung der User Stories

6.1.1 User Story 1:

Dokument wird in KnowHow Tool vom User hochgeladen. Dokument durchläuft Chunking Verfahren. Die produzierten Textabschnitte werden dann simultan in die semantische Datenbank des RAG-Systems und in den Wissensgraphen eingespeist.

- Ziel: Nutzerinnen und Nutzer können Dokumente hochladen; diese werden robust geparst, in inhaltliche Chunks zerlegt und indexiert.
- Umsetzung:
- UI: Upload-Flow in der Streamlit-Seite für Indexverwaltung.
- Backend: FastAPI-Gateway leitet an Ingest-Workflow weiter (LlamaIndex).
- Parsing/Chunking: LlamaIndex Node-Parser mit Chunk-Größe 512 Tokens und 20 Tokens Overlap; Formate: PDF, DOCX, MSG.
- Speicherung: Chunks in OpenSearch; Originale in MinIO.
- Metadaten: Optionaler Extraktionspfad über Prompt-Vorlagen (Dublin Core, Custom). Vorlagen werden versioniert im Index-_meta gespeichert und vom Ingest-Workflow konsumiert.
- Dedup/Qualität: Hash-basierte Duplikaterkennung; Protokollierung des Ingest-Pfads.
- Ergebnis: Stabiler, nachvollziehbarer Ingest mit reproduzierbaren Chunks und angereichertem Metadatenmodell.

6.1.2 User Story 2:

Relevante Dokumente werden vom RAG-System gefunden und dem User angezeigt.

- Ziel: Relevante Inhalte zu einer Anfrage effizient auffinden.
- Umsetzung:
- Hybrid-Suche: Kombination aus BM25 und semantischer k-NN-Suche auf BGE-M3-Embeddings (1024D).
- Reranking: BGE-reranker-base zur Verbesserung der Ergebnisqualität.

- REST-API und UI: Gateway-Endpunkte und Streamlit-Suchmaske, inkl. Trefferliste mit Metadaten und Zitaten.
- Ergebnis: Höhere Präzision durch semantische Suche und Reranking, merklich reduzierte Antwortzeiten.

6.1.3 User Story 3:

Anfrage in natürlicher Sprache werden vom KnowHow Tool verarbeitet und relevante Dokumente, eine textuelle Antwort und etwaige Beziehung zwischen den relevanten Dokumenten wiedergegeben.

- Ziel: Freitextfragen verständlich beantworten, inkl. Quellenangaben.
- Umsetzung:
- RAG-Pipeline: Retrieval über OpenSearch + LlamaIndex; Synthese mit custom-LLM.
- Wissensgraph Information zu Dokumenten die mittels Opensearch gefunden werden
- Prompting: System- und Rollenprompts; Logging von Kontextlänge, Knotenzahl und Antwortlänge zur Nachvollziehbarkeit.
- Ergebnis: Konsistente, quellengestützte Antworten mit transparenter Kontextbildung.

6.1.4 User Story 4:

Erneute Anfragen können problemlos an das KnowHow Tool gestellt werden. Die Antwort basiert dann auf einer erneuten Suche.

- Ziel: Folgefragen auf Basis des bisherigen Verlaufs ermöglichen.
- Umsetzung:
- Chat-Fortsetzung via „POST /generate/{index_id}“ mit „messages“ und zuvor priorisierter „nodeList“. Dialog bleibt streng an die Fundstellen gebunden (Grounding), inkl. Verfeinerung.
- Verlaufskontext: Übergabe der relevanten Historie an Retrieval und Synthese.
- UI/UX: Konversationsansicht mit Zitaten; erneutes Reranking je Iteration.
- Ergebnis: Natürlich wirkender Dialogfluss mit schrittweiser Verfeinerung und stabiler Ergebnisqualität.

6.1.5 User Story 5:

Anfrage an das LLM wie in US 2, nur werden zusätzlich vom User weiter Dokumente in die Anfrage an das KnowHow Tool hinzugefügt.

Umsetzung: analog zu US 4, jedoch mit „docList“ (anstelle/ergänzend zu „nodeList“). Die ausgewählten Quelldokumente werden vollständig in die Antwortgenerierung einbezogen (z. B. Zusammenfassungen).

6.1.6 User Story 6:

Wie im Entwicklerkommentar von US 6 angekündigt wird US6 als Teil der US 1 gedeckt.

Wir gehen davon aus, dass die selektierten Dokumente in den unterstützten Kontext des LLM passen. Fallback wäre ein komplizierter Ansatz analog zu einem Vektor- oder Grafen-Diff-Algorithmus auf Basis der Chunks der Dokumente. Ein solches Fallback wurde nicht verfolgt.

6.1.7 User Story 7:

Aufgrund der Implementierung des Evaluierungstools als separate Software kann eine automatisierte Speicherung von Benutzeranfragen zu Analysezwecken nicht durchgeführt werden. Es werden daher passende Anfragen-Beispiele vorbereitet, um Benutzerinnen und Benutzer in der Tool-Bedienung anzuleiten.

6.2 Benutzeroberfläche

Streamlit kommuniziert mit dem Gateway. Zentrale Bereiche: Index-Management, Ingestion, Suche, Konversation und Admin-Dashboard. Die Suche bietet Modus-Umschaltung (Fulltext/Semantic/Hybrid), Filterpanel, Rerank-Toggle, TopN-Regler und MMR-Slider. In der Konversation können Nodes/Dokumente angeheftet und an das LLM übergeben werden.

Die Benutzeroberfläche ist mit Streamlit umgesetzt und kommuniziert ausschließlich über das FastAPI-Gateway (REST). Die Frontend-Seiten sind modular aufgebaut (z.

B. cibai/smarter/frontend/pages/1_Index.py, 3_Search.py, 4_RAG.py) und verwenden zustandsbasierte Steuerung (Session State) für stabile Benutzerflüsse, Ladeanzeigen und Fehlermeldungen.

Architektur und Navigation

Streamlit ↔ Gateway: Alle Operationen (Index anlegen, Ingestion, Suche, Chat/RAG) erfolgen über das Gateway (z. B. cibai/smarter/gateway/index.py, gateway/search.py).

Hauptbereiche: Index-Management, Ingestion, Suche, Konversation (RAG+KG), Admin-Dashboard.

Bedienkonzept: Klar getrennte Seiten, konsistente Seitenleisten-Navigation, kontextabhängige Panels (Filter, Pinned Context, Upload).

Index-Management

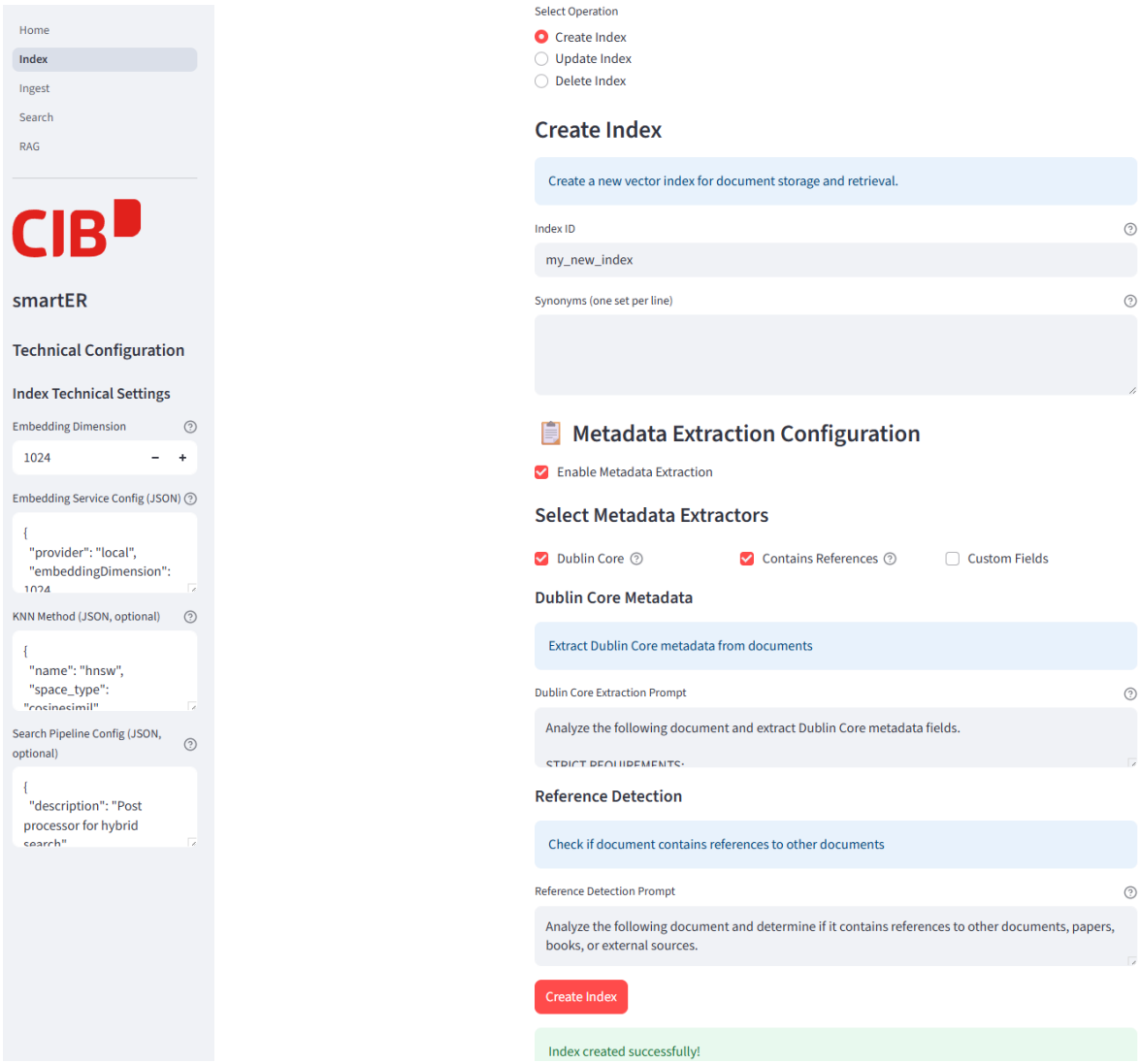

Abbildung 4 - Index-management

Indexanlage/-pflege: Erstellen, Auswählen, Löschen von Indizes inkl. Parameterisierung für Chunking (z. B. 512 Tokens, Overlap 20), Embedding-Modell (BAAI/bge-m3, 1024D) und optionalen Reranker (BAAI/bge-reranker-base).

Metadaten-Extraktion:

Aktivierbarer Bereich “Metadata Extraction Configuration”.

Vorlagenverwaltung (Dublin Core, Scientific Paper, Legal Document, etc.) mit Laden/Anpassen der Prompts.

Pflichtfelder-Auswahl und Validierung auf gültiges JSON; Prompts werden versioniert im OpenSearch-Index-_meta gespeichert und im Ingest automatisch genutzt.

Synonyme/Konfiguration: Indizes können (falls hinterlegt) Synonyme nutzen; die UI weist darauf hin, wenn Synonyme aktiv sind.

Ingestion

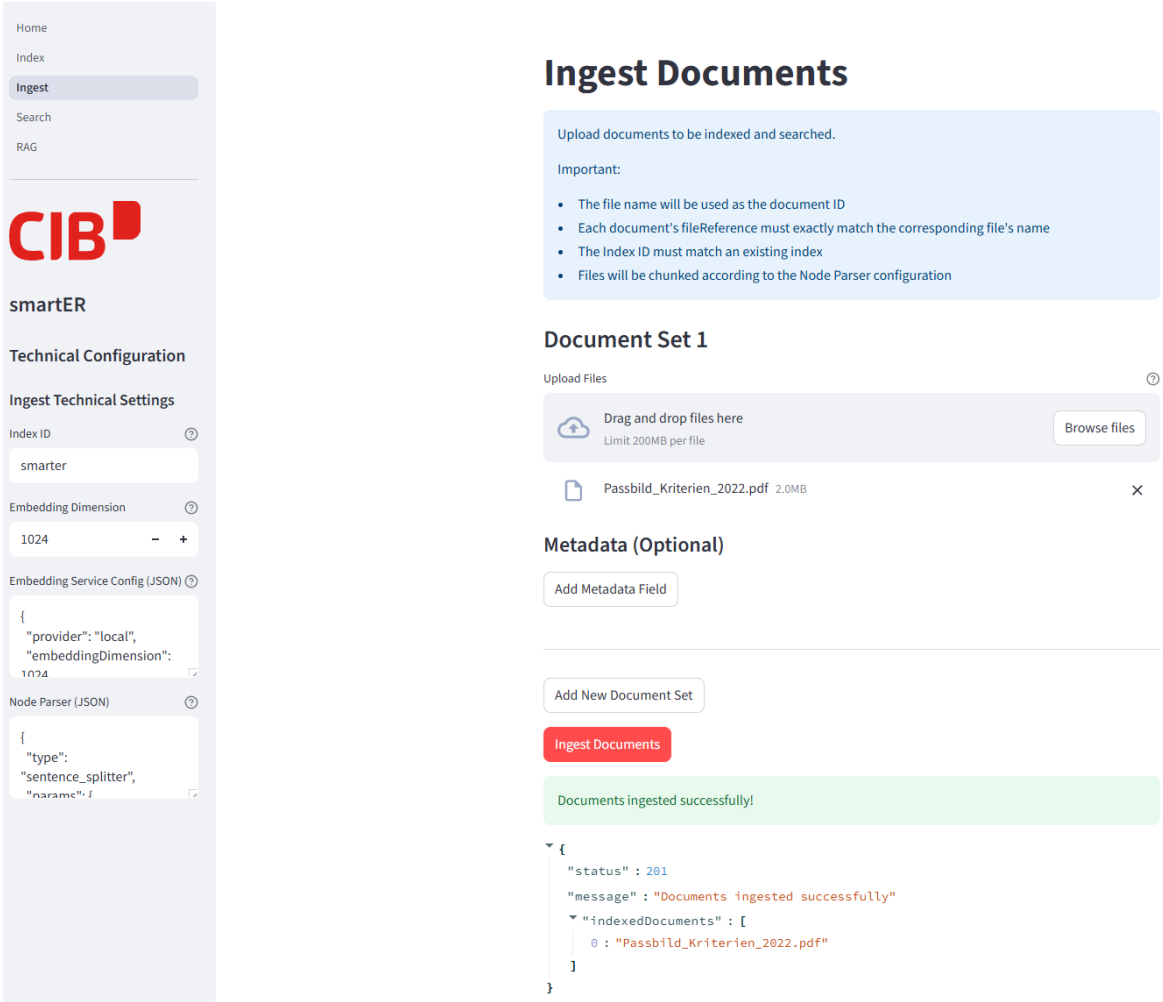

Abbildung 5 - Ingestion

Upload: Mehrfach-Upload (PDF, DOCX, MSG) mit Fortschrittsanzeigen. Dokumente werden geparst, in Chunks zerlegt, mit Embeddings versehen und in OpenSearch abgelegt; Originale in MinIO.

Qualität/Deduplikation: Hash-basierte Erkennung bereits verarbeiteter Inhalte; transparente Protokollierung (z. B. Anzahl Dokumente/Chunks, Fehlversuche).

Metadaten: Falls aktiviert, werden die ausgewählten Prompt-Vorlagen angewandt; die UI zeigt Status und Ergebnis (z. B. ausgefüllte Dublin-Core-Felder) an.

Wiederholbarkeit: Ingest-Runs können erneut gestartet werden, ohne bestehende Indizes zu verfälschen.

Chunks werden außerdem in den Wissensgraphen eingespeist. Hier laufen die Chunks durch eine Pipeline in der Zusammenhänge zwischen Dokumenten gefunden werden. Diese Zusammenhänge werden im Wissensgraph repräsentiert. Der Wissensgraph wird in in der Software Neo4J gespeichert, welche per Docker zur Verfügung gestellt wird.

Suche

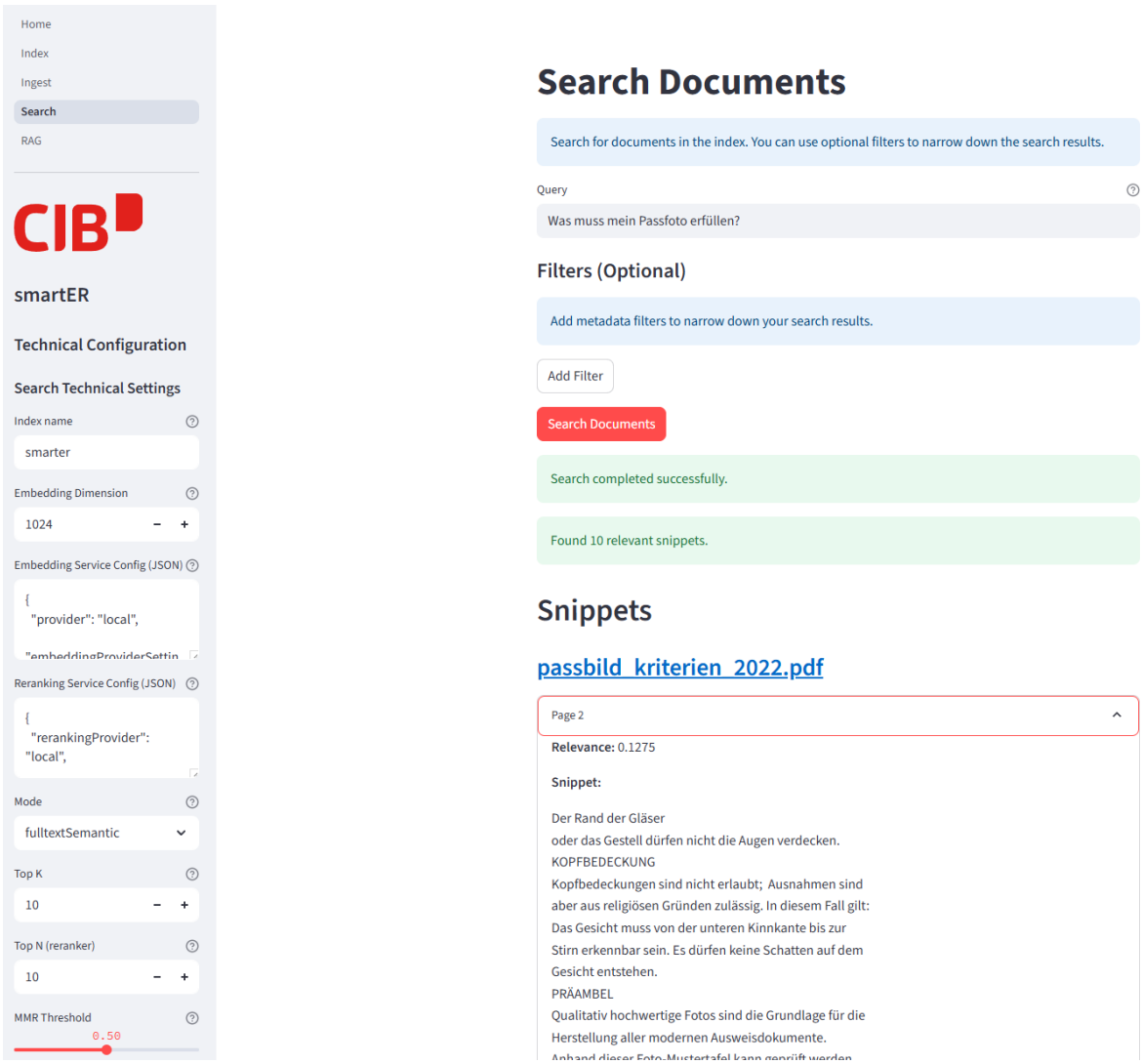

Abbildung 6 - Suche

Suchmodi:

Fulltext: BM25-basiert (klassische Schlüsselwortsuche).

Semantic: k-NN-Suche über Embeddings (BGE-M3).

Hybrid: Kombination aus BM25 und semantischer Suche.

Umschaltung erfolgt per Modus-Toggle in der UI.

Reranking: Optionaler Rerank-Toggle (BGE-reranker-base) zur Qualitätssteigerung; Hinweis auf leichte Mehrlatenz.

Ergebnissteuerung:

TopN-Regler: Anzahl der anfänglichen Treffer (k).

MMR-Slider: Diversität vs. Relevanz (Maximal Margin Relevance [27]); Nutzer justieren das Gleichgewicht interaktiv.

Filterpanel:

Facettierte Filter (z. B. Sprache, Dateityp, Zeitraum, Dublin-Core-Felder), frei kombinierbar.

Reset-Funktion und Anzeige aktiver Filterchips.

Ergebnisdarstellung:

Karten mit Snippets, Metadaten, konfigurierbaren Scores (z. B. BM25/Embedding/Rerank-Score).

Aktionen pro Treffer: “Anheften” für Konversation, “Original öffnen/anzeigen”, “Details” (voller Chunk, Metadaten).

Leistungsaspekte: Caching von Index-Checks und optimierte Abfragen reduzieren Latenz; die UI signalisiert lange laufende Operationen mit Spinners/Progress.

Konversation (RAG + KG)

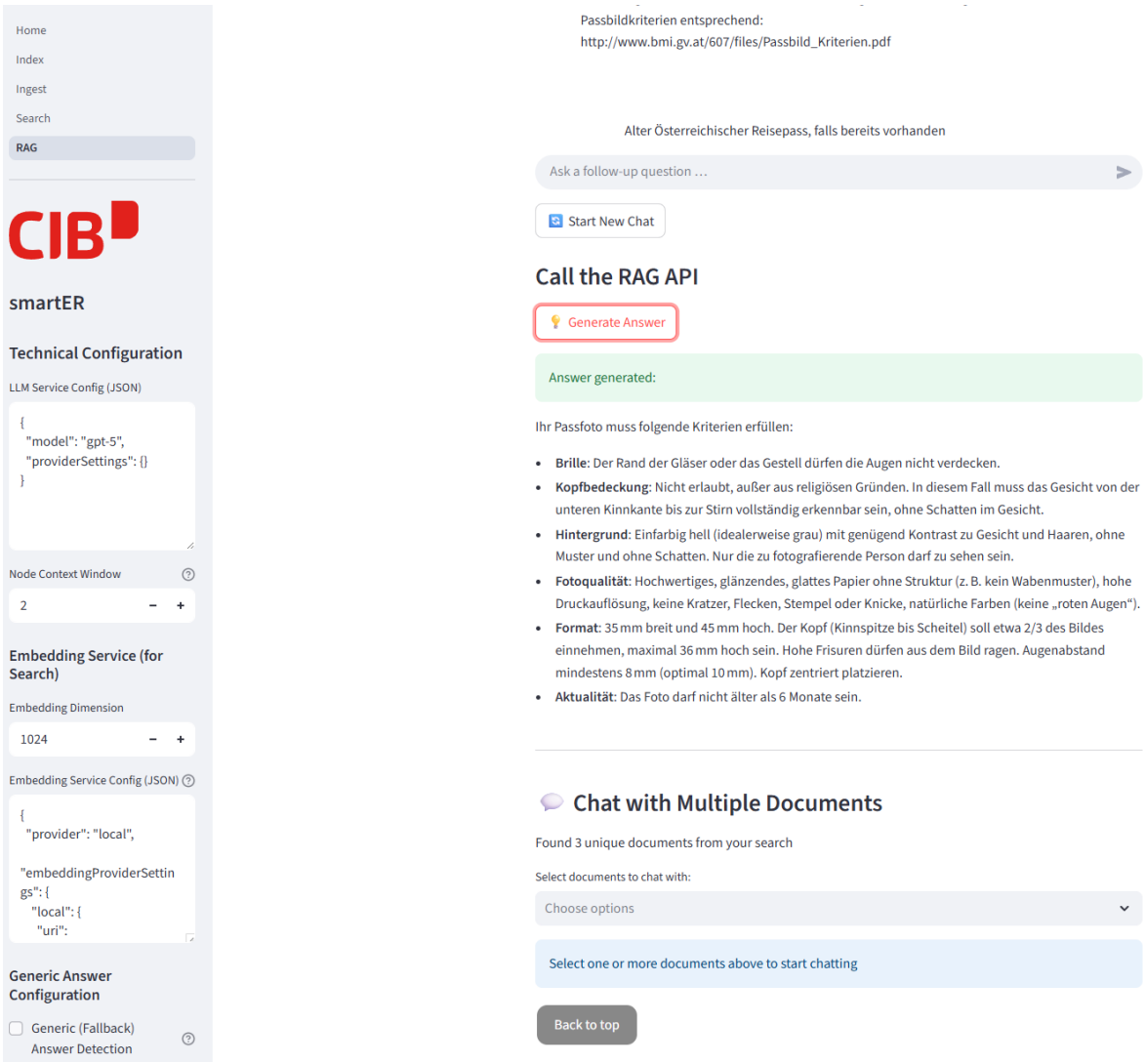

Abbildung 7 - RAG

Chat-Ansicht: Freitextfragen mit quellengestützten Antworten; Zitierliste und Quellen-Sprungmarken.

Kontext-Management:

Das Kontext-Management der Konversationsansicht ermöglicht es, einzelne Nodes oder Dokumente über ein Pinned-Context-Panel gezielt anzuheften und dem LLM als Grundlage für die Antwortgenerierung zu übergeben. Darüber hinaus können pro Anfrage zusätzliche Dokumente ad hoc hochgeladen werden, ohne dass diese dauerhaft im Index gespeichert werden. Der Nutzer wählt dabei zwischen zwei Modi: Entweder werden ausschließlich die

angehefteten Kontexte herangezogen, oder diese werden mit neu abgerufenen Treffern aus der Suche kombiniert.

Zur Steuerung der Antwortqualität stehen dieselben Mechanismen wie in der Suchansicht zur Verfügung – darunter der Rerank-Toggle, der TopN-Regler sowie der MMR-Slider [27] –, sodass sich die Kontextqualität für die Generierung gezielt justieren lässt. Antworten können jederzeit mit identischem oder angepasstem Kontext neu generiert werden. Das eingesetzte LLM ist frei konfigurierbar, wodurch sich unterschiedliche Modelle je nach Anwendungsfall nutzen lassen.

Ergänzend zur klassischen Retrievalsuche wird bei jeder Anfrage der Wissensgraph einbezogen. Zunächst werden über OpenSearch passende Chunks zur Benutzeranfrage identifiziert. Anhand der jeweiligen Chunk-IDs wird anschließend eine Anfrage an den Wissensgraphen gestellt, der Dokumente zurückliefert, die in einer relevanten Beziehung zu den gefundenen Chunks stehen – etwa weil sie ein Dokument ersetzen, außer Kraft setzen oder inhaltlich erweitern. Diese zusätzlichen Informationen werden vom KnowHow-Tool verarbeitet und in Beziehung zu den ursprünglich gefundenen Dokumenten gesetzt. Dem Nutzer wird schließlich eine vereinte Informationsbasis aus OpenSearch-Treffern und Wissensgraph-Erweiterungen zur Verfügung gestellt.

6.2.1 Beispiel-Interaktionsablauf

1) Index anlegen → 2) Dokument(e) hochladen → 3) Hybrid-Suche mit Reranking → 4) Relevante Nodes pinnen → 5) RAG-Antwort generieren (inkl. Kontextfenster und ggf. Qualitätscheck) und mit Wissensgraph Informationen erweitern → 6) Antwort prüfen und iterieren.

6.2.2 Modellauswahl:

smartER ist modellagnostisch konzipiert. Das System akzeptiert jede LLM-Schnittstelle, die dem OpenAI-API-Standard entspricht. Dadurch können sowohl proprietäre Cloud-Modelle (z. B. OpenAI GPT-4, Anthropic Claude) als auch lokal gehostete Open-Source-Modelle eingebunden werden, ohne Anpassungen am Anwendungscode vornehmen zu müssen. Die einzige Voraussetzung ist ein kompatibler API-Endpunkt.

Im Rahmen des KnowHow-Projekts fiel die Wahl auf GPT OSS 22B, ein Open-Source-Modell mit 22 Milliarden Parametern. Dieses wurde lokal über Ollama bereitgestellt, was eine vollständig selbstgehostete Lösung ohne Abhängigkeit von externen Cloud-Diensten ermöglicht. Die Entscheidung für ein lokales Deployment wurde vor allem durch Anforderungen an Datenschutz und Datensouveränität getrieben, da die verarbeiteten Dokumente sensible Unternehmensinformationen enthalten können.

Für das KnowHow-Evaluierungstool wurde hingegen Qwen3 32B eingesetzt, ein Modell mit 32 Milliarden Parametern. Die Wahl eines separaten, leistungsfähigeren Modells für die Evaluierung ermöglicht eine unabhängige Bewertung der Antwortqualität, da das bewertende Modell nicht identisch mit dem generierenden Modell ist. Aufgrund der jeweiligen Modellgrößen von 22B und 32B konnten beide Modelle gleichzeitig auf derselben Hardware über Ollama gehostet werden, sodass kein zusätzlicher Infrastrukturaufwand entstand. Dies erlaubte einen parallelen Betrieb von Antwortgenerierung und Qualitätsevaluierung ohne gegenseitige Beeinträchtigung.

Durch die modulare Anbindung kann jedes eingesetzte Modell jederzeit ausgetauscht werden – etwa um leistungsfähigere Modelle zu evaluieren oder um für unterschiedliche Anwendungsfälle spezialisierte Modelle einzusetzen –, ohne dass Änderungen an der übrigen Systemarchitektur erforderlich sind.

6.3 Zusammenfassung

CIB smartER liefert einen wiederverwendbaren, skalierbaren RAG-Baustein mit klaren Workflows (Index, Ingest, Search, RAG), einer sauberen Gateway-API und einer praxistauglichen UI. Initial basiert das System auf OpenSearch, MinIO (für Originale) und lokal betriebenen AI-Services; eine Neo4j-basierte Graph-Erweiterung ist vorgesehen. Die CIB smartER software und der Wissensgraph fusionieren zum KnowHow Tool.

7 Spezifikation des Evaluationsdesigns

7.1 A/B Testing

In der Evaluierungsphase wird das prototypische System, insbesondere im Hinblick auf die Such-Funktionalität, dem bestehenden System gegenübergestellt. Das Ziel der Evaluierung ist es, die Effizienz des neuen Systems bei der Lösung der für Benutzerinnen und Benutzer typischen Aufgaben zu bewerten. Das A/B Testing ist eine Testmethode, die zwei Varianten eines Systems vergleicht und eignet sich daher für die Evaluierungsphase. Es werden zwei Benutzer-Gruppen gebildet: Gruppe A (Kontrollgruppe) verwendet in der Evaluierungsphase das Original-System, Gruppe B (experimentelle Gruppe) hingegen das neue System. Die Aufteilung soll randomisiert erfolgen und keine der Gruppen soll eine Prävalenz hinsichtlich der (zu erwartenden) Erfahrung im Umgang mit einem der Systeme aufweisen.

Um einen Vergleich durchführen zu können, müssen Evaluierungsaufgaben definiert werden, die möglichst auch die typischen Suchvorgänge in der Organisation abbilden. In der Evaluierungsphase werden die Teilnehmerinnen und Teilnehmer aus den Testgruppen dieselben Aufgaben unter Anwendung der verschiedenen Systeme (Gruppe A – Original-System, Gruppe B – das KnowHow Tool) lösen. Als Zielmetrik wird dabei der Zeitbedarf für die jeweilige Test-Aufgabe gemessen. Zusätzlich kann auch das Feedback der Benutzerinnen und Benutzer (Benutzer-Zufriedenheit) abgefragt und evaluiert werden. Die zu testende Hypothese kann folgendermaßen formuliert werden: „Unter Anwendung des neuen experimentellen Systems brauchen Benutzerinnen und Benutzer durchschnittlich weniger Zeit für die Bearbeitung derselben Testaufgaben als unter Anwendung des Original-Systems“. Die Null-Hypothese wäre, dass es keinen signifikanten Unterschied in den durchschnittlichen Bearbeitungszeiten beider Gruppen gibt.

Die Auswertung der Ergebnisse kann mittels T-Test für unabhängige Stichproben, Welch-T-Test, U-Test (Rangsummen-Test) erfolgen.

7.1.1 Ablauf

In der Evaluierungsphase benutzen die Teilnehmerinnen und Teilnehmer aus den beiden Gruppen das jeweilige System, um eine oder mehrere vorgeschlagene Test-Aufgabe(n) aus dem Aufgaben-Pool (wie bspw. Suche nach Dokumenten mit bestimmten Inhalten) zu lösen. Die Bearbeitung einer Test-Aufgabe aus der Sicht der Benutzerinnen und Benutzer und des Systems ist in der Abbildung 8 skizziert. Zu den Evaluierungszwecken werden dabei a) Benutzer-ID, b) die Start- und End-Zeit, c) die Suchanfrage(n), d) Erfolg oder Misserfolg, e) die Zugehörigkeit zu der Test-Gruppe (A oder B), f) die Test-Case ID und zu Zwecken der Evaluierung der Gender-Aspekte des Tools außerdem g) das Geschlecht erfasst. Die Informationen wie e) und g) können durch die Benutzer-ID einem Benutzerprofil zugeordnet werden, sodass die Erfassung bei der Anmeldung im System erfolgen kann.

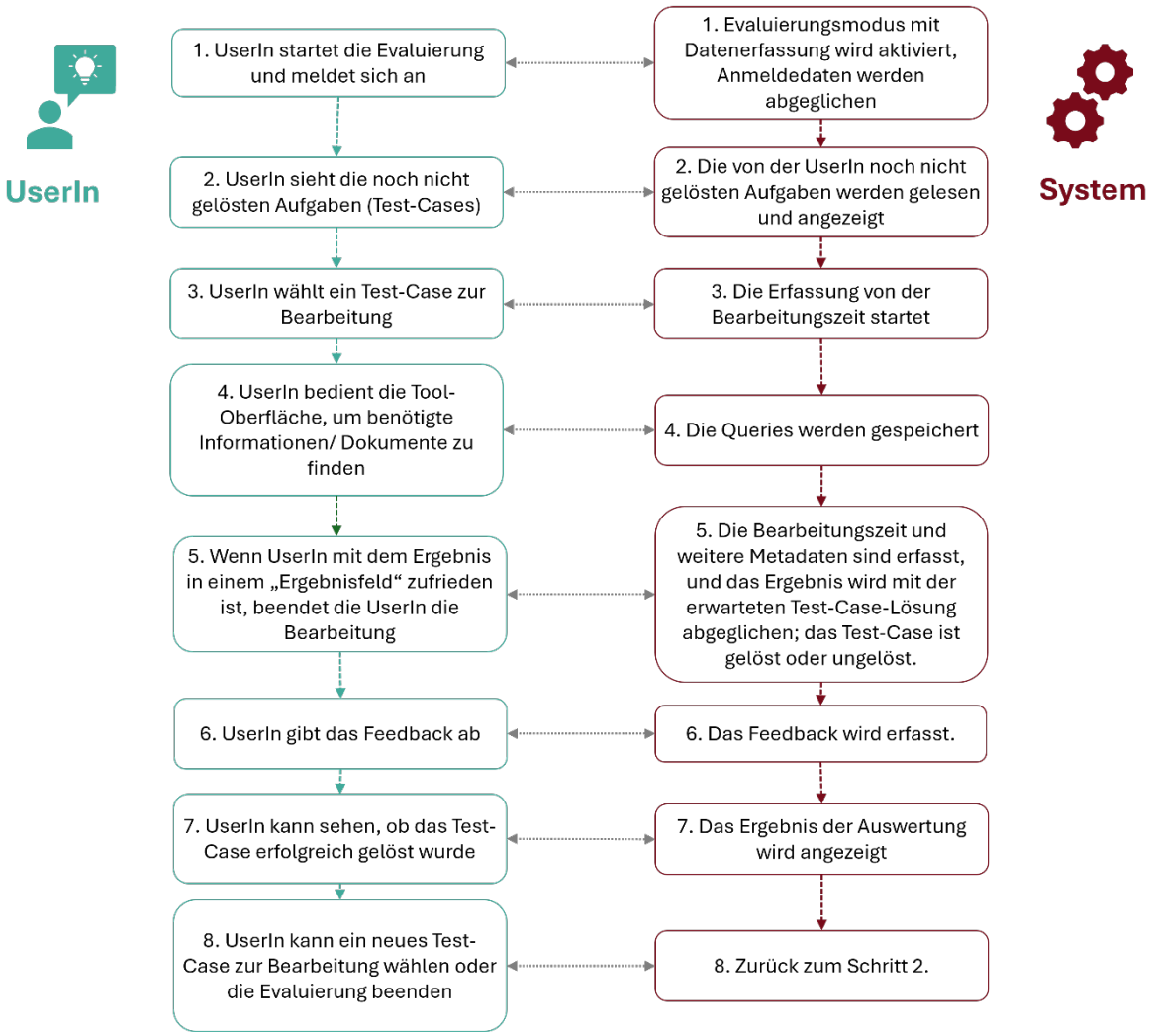

Abbildung 8 - Ablauf: Bearbeitung eines Test-Cases

Um die von der Benutzerin oder dem Benutzer bereits gelösten Aufgaben aus der Aufgabenliste entfernen bzw. als solche anzeigen zu können, sollten die Einträge der Evaluierungs-Abläufe in einer für das System abrufbaren Datenbank a) Benutzer-ID, b) Test-Case ID und c) Test-Case Status (gelöst/ ungelöst) enthalten. Falls mehrere Versuche für die erfolgreiche Lösung eines Test-Cases notwendig sind, wäre eine Aggregation der Versuchszeiten für die finale Evaluierung denkbar. Ob und wie die ungelösten Test-Cases in die Evaluierung miteinfließen sollen (sofern solche am Ende der Evaluierungsphase vorhanden sein sollen), muss im Rahmen einer Vorbereitung der Auswertungsphase diskutiert werden. Denkbar wäre ein Vergleich der Anzahl von gelösten/ ungelösten Test-Cases pro Gruppe. Hier kann bspw. der Fisher-Test durchgeführt werden, um die Signifikanz der Abweichungen zu prüfen.

7.1.2 Zusammensetzung der Testgruppen und Anzahl der Studienteilnehmerinnen und Studienteilnehmer

Das Ziel der Evaluierung ist es, einen statistisch signifikanten Unterschied bei dem Zeitaufwand festzustellen, die Teilnehmerinnen und Teilnehmer unter Verwendung des bestehenden Systems (Gruppe A) und des KnowHow-Prototypen (Gruppe B) benötigen, um vordefinierte Aufgaben zu lösen.

Es wird angenommen, dass die Bearbeitung einer typischen Aufgabe mit dem bestehenden System 90 Sekunden und mit dem KnowHow-System 60 Sekunden dauern würde, mit einer Standardabweichung (σ) von 10 Sekunden bei den Antwortzeiten beider Gruppen.

Die Stichprobengröße für jede Gruppe kann wie folgt berechnet werden:

$$n = 2\sigma^2 \cdot \frac{(Z_{\alpha/2} + Z_{\beta})^2}{\Delta^2}$$

Dabei gilt:

- σ: Standardabweichung der Aufgabenzeiten (10 s, bei gleicher Varianz).
- $Z_{\alpha/2}$: Z-Score entsprechend dem Signifikanzniveau (1,96 für α = 0,05).
  - α = 0,05 bedeutet, dass der Test eine 5%-ige Wahrscheinlichkeit (oder weniger) hat, zu dem Schluss zu kommen, dass es einen Zeitunterschied gibt, wenn kein solcher Unterschied besteht (ein falsch positives Ergebnis).

- Z_β: Z-Score entsprechend der Teststärke (1,24 für 80 % Teststärke).
  - Das bedeutet, dass der Test eine 80%-ige Wahrscheinlichkeit hat, einen echten Unterschied zwischen den beiden Gruppen (A und B) richtig zu erkennen, wenn ein solcher Unterschied besteht und mindestens so groß ist wie der minimal erkennbare Unterschied (Δ).

- Δ: Minimaler erkennbarer Unterschied in den durchschnittlichen Zeiten (wir nehmen 8s an).

Unter diesen Annahmen ergibt sich n = 2 x 102 x (1.96 + 1.24)2 / 82 = 32 Personen in jeder Gruppe.

Für weitere statistische Messungen in Bezug auf das Geschlecht empfehlen wir eine gleichmäßige Aufteilung der Geschlechter innerhalb jeder Gruppe.

Zusammenfassend werden folgende Gruppengrößen empfohlen:

Tabelle 5 - Zusammensetzung der Testgruppen

Gruppe	männlich	weiblich
Gruppe A (Kontrollgruppe)	16	16
Gruppe B (KnowHow Tool)	16	16

7.1.3 Anzahl der Test-Cases

Die Anzahl der Aufgaben (Test-Cases) beeinflusst die Streuung der Bearbeitungszeiten und die Teststärke.

Sei μ_1 der Mittelwert der Aufgaben-Bearbeitungszeiten in der Gruppe A und μ_2 der Mittelwert in der Gruppe B, dann sind die Mittelwerte der Bearbeitungszeiten bei k Aufgaben pro Person in Gruppen A, B:

$$\mu_{\text{person\_A}} = k\mu_1;\ \mu_{\text{person\_B}} = k\mu_2$$

Diese sind gleich den Mittelwerten der Bearbeitungszeiten in den Gruppen A, B:

$$\mu_{\text{group\_A}} = k\mu_1;\ \mu_{\text{group\_B}} = k\mu_2$$

Wenn die Standardabweichung der Aufgabenzeiten in den Gruppen ist gleich σ (für jede Aufgabe berechnet über alle Personen in der Gruppe), und die Anzahl der Aufgaben ist k, dann ist die Varianz der Bearbeitungszeiten x unter Annahme der Unabhängigkeit:

$$\mathrm{Var(person)} = \mathrm{Var(group)} = \sum_{k} \mathrm{Var}(x_k) = k\sigma^2$$

Unter Annahme der gleichen Varianz in den Gruppen ergibt sich die geschätzte Standardabweichung in den Gruppen:

$$\sigma_1 = \sigma_2 = \sigma\sqrt{k}$$

Im Falle der gleichen Varianzen in den gleich großen unabhängigen Stichproben wird der t-Wert folgendermaßen berechnet:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sigma\sqrt{\frac{2}{n}}},$$

wo $\overline{x_1}$, $\overline{x_2}$ - Stichprobenmittelwerte, σ - die Standardabweichung, und n - die Anzahl Beobachtungen (Teilnehmerinnen und Teilnehmer) je Gruppe. Für den Fall mit k Aufgaben ergibt sich:

$$t = \frac{k(\mu_1 - \mu_2)}{\sigma\sqrt{\frac{2k}{n}}} = \sqrt{\frac{kn}{2}} \cdot \frac{(\mu_1 - \mu_2)}{\sigma},$$

sodass $t \sim \sqrt{k}$.

Je höher die erwartete Varianz in den Aufgabenzeiten oder je geringer der minimale erkennbare Unterschied in den durchschnittlichen Gruppenzeiten, desto höher ist die Anzahl der Aufgaben k zu wählen bei gleichbleibender Anzahl Personen.

Ausgehend von den oben getroffenen Annahmen (7.1.2), wobei minimaler erkennbarer Unterschied in den durchschnittlichen Bearbeitungszeiten 8s beträgt, und die Standardabweichung der Aufgabenzeiten bei 10s liegt, ist der t-Wert bereits bei k = 1 ausreichend hoch:

$$t = \sqrt{\frac{1 * 32}{2}} \cdot \frac{8}{10} = 3.2$$

und entspricht einem p-Wert von 0.002 (bei 2*n - 2 = 62 Freiheitsgraden), was deutlich unter dem angenommenen Signifikanzniveau von 0.05 liegt. Der kritischer t-Wert, der dem Signifikanzniveau von 0.05 entspricht, liegt bei 1.999 ≈ 2.[1]

Steigt bspw. die Standardabweichung der Aufgabenzeiten auf 20s bei den gleichbleibenden anderen Parametern, sind k = 2 Aufgaben notwendig, um einen t-Wert über dem kritischen Wert zu erreichen. Bei 30s sind es 4 Aufgaben, bei 50s - 10 Aufgaben, und bei 100s mindestens 40 Aufgaben.

Cohen's d bestimmt die Effektgröße für Mittelwertunterschiede zwischen zwei Gruppen mit gleicher Gruppengrößen und gleichen Gruppenvarianzen und fließt in die Berechnung des t-Werts und der Teststärke ein. Cohen‘s d ist wie folgt definiert:

$$d = \frac{\overline{x_1} - \overline{x_2}}{\sigma},$$

sodass

$$d = \frac{k(\mu_1 - \mu_2)}{\sigma\sqrt{k}} = \frac{\sqrt{k}(\mu_1 - \mu_2)}{\sigma}, d \sim \sqrt{k}$$

Die Effektgröße beträgt bei der Anzahl der Aufgaben k = 1 bereits 0.8 (d > 0.8 entspricht einem starken Effekt). Steigt bspw. die Standardabweichung auf 20s, braucht es 4 Aufgaben für die gleiche Effektstärke.

Die Teststärke (power) beträgt bei der Effektgröße d = 0.8, n = 32 und dem Signifikanzniveau von 0.05: power = 0.883.[2]

---

[1] Berechnet mit scipy [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html]

[2] Berechnet mit statsmodels [https://www.statsmodels.org/stable/generated/statsmodels.stats.power.TTestIndPower.html]

Die tatsächlich notwendige Anzahl der Test-Cases ist auf Basis der zu erwartenden Streuung in den Bearbeitungszeiten zu bestimmen (bei den gleichbleibenden Anzahl Teilnehmerinnen und Teilnehmer und minimalem erkennbaren Unterschied). Da aufgrund des Evaluierungsdesigns voraussichtlich Aufgaben unterschiedlicher Arten formuliert werden, sollten k Aufgaben jeder Art ausgewählt werden. Grundsätzlich trägt die Erhöhung der Aufgabenanzahl zur Signifikanz und Stärke des Tests bei und kann eine geringere Anzahl Teilnehmerinnen und Teilnehmer ausgleichen.

7.2 Anforderungen hinsichtlich der Evaluierung des Prototypen

Die Konkretisierung der Anforderungen an Test-Cases (bspw. Anzahl, Komplexität, Bewertungskriterien etc.) sowie an Evaluierungsdesign wird im Rahmen eines vorbereitenden Workshops erfolgen. Nachfolgend sind die bereits bekannten Anforderungen angeführt.

1. Studienteilnehmerinnen und Studienteilnehmer

    1.1. Ausreichende Anzahl Teilnehmerinnen und Teilnehmer für die Evaluierung in beiden Test-Gruppen (s. Empfehlung) ist erforderlich.

    1.2. Test-Gruppen sollten möglichst gemischt hinsichtlich ihrer (zu erwartenden) Erfahrung im Umgang mit dem neuen/alten System zusammengestellt werden, sodass kein Bias entsteht (wie bspw. vorwiegend erfahrene Benutzerinnen und Benutzer des Originalsystems in der Kontrollgruppe).

    1.3. Teilnehmerinnen und Teilnehmer sollen darauf hingewiesen werden, dass ihre Anfragen zu Analysezwecken gespeichert werden (ggf. muss eine schriftliche Zusage erfolgen).

    1.4. Teilnehmerinnen und Teilnehmer sollten vor/am Anfang der Evaluierungsphase eine Trainingsmöglichkeit im Umgang mit dem KnowHow Tool erhalten, bspw. Im Rahmen eines Workshops.

2. Test-Cases

    2.1. Test-Cases (praktische Aufgaben für Teilnehmerinnen und Teilnehmer) sollen in ausreichender Anzahl definiert werden.

2.2. Test-Cases sollen repräsentativ für den Arbeitsablauf der Mitarbeiterinnen und Mitarbeiter und somit möglichst den tatsächlichen Anwendungsfällen ähnlich sein.

2.3. Jedes Test-Case soll über ein messbares Erfolgskriterium verfügen, sodass bei der Interaktion mit dem System das jeweilige Test-Case eindeutig als gelöst/ungelöst definiert werden kann.

3. Evaluierungsumgebung

3.1. Jeweilige Gruppe (A/ B) nutzt ausschließlich die vorgesehene System-Variante (Original-System/ KnowHow-Tool), um die Test-Cases zu bearbeiten.

3.2. Ausreichende Anzahl und Variabilität der Dokumente/ Inhalte in Test-Systemen ist erforderlich, um Test-Cases abzubilden.

3.3. Ggf. muss der Datenbestand für den Evaluierungsrahmen eingeschränkt werden, um die Vergleichbarkeit zu gewährleisten.

3.4. Die Benutzerinnen und Benutzer können sich in der Evaluierungsumgebung anmelden, sodass die Evaluierungsabläufe pro Test-Case der Benutzer-ID zugeordnet werden können.

3.5. Die Kontoinformationen der Benutzerinnen und Benutzer (ID, Anmeldedaten, Test-Gruppe sowie Geschlecht zu Zwecken der Evaluierung der gender-spezifischen Aspekte) werden im System gespeichert.

3.6. Zu Zwecken der Evaluierung soll ein „Evaluierungsmodus“ oder eine dezidierte Evaluierungsumgebung zur Verfügung stehen.

3.7. Im Evaluierungsmodus/ in der Evaluierungsumgebung kann ein Test-Case ausgewählt werden oder es wird den Benutzerinnen und Benutzer angeboten. Wird die Bearbeitung eines Test-Case gestartet, speichert das System folgende Daten: Start- und End-Zeit der Bearbeitung; Suchanfrage(n); Erfolgsstatus (gelöst/ungelöst); Test-Gruppe (A oder B); Test-Case ID.

7.3 Zusammenfassung

Abschnitt 7 beschrieb die technischen Anforderungen für den Betrieb des prototypischen KnowHow-Tools und die Spezifikation des Evaluationsdesigns. Das KnowHow-Tool wird in einer containerisierten Form (Docker, docker-compose) bereitgestellt. Sicherheitsmaßnahmen in Bezug auf den Zugriff und die Datensicherheit werden mitbedacht, außerdem werden Anforderungen und Lösungen bzgl. Skalierbarkeit betrachtet.

Die Evaluierungsphase hat das Ziel, die Unterschiede zwischen den zwei Systemen (Intranet-Suche und KnowHow-Tool) im Sinne der benötigten Zeit für bspw. Suchvorgänge festzustellen und ist als A/B-Testing konzipiert. Hierfür werden insgesamt 64 StudienteilnehmerInnen benötigt, die in zwei gleich große Gruppen aufgeteilt werden. In der Evaluierungsphase verwendet die Gruppe A das Intranet, und die Gruppe B das KnowHow-Tool, um die Evaluierungsaufgaben (Test-Cases) zu lösen. Die Datenerfassung erfolgt über einen Evaluierungsmodus bzw. eine dezidierte Evaluierungsumgebung.

Das Design zielt darauf ab, einen statistisch signifikanten Unterschied in den durchschnittlichen Bearbeitungszeiten in den Gruppen festzustellen. Die Auswertung kann mittels Student's T-Test für zwei unabhängige Stichproben bzw. nicht-parametrische Tests erfolgen. Darüber hinaus können gender-spezifische Aspekte evaluiert werden (vorausgesetzt ist die erforderliche Verteilung in den Gruppen) sowie Unterschiede in Bezug auf die Aufgaben-Typen.

8 Umsetzung und Dokumentation der Evaluierungsumgebung

8.1 Test-Cases

Test-Cases wurden von BMEIA-Expertinnen und -Experten ausgewählt und formuliert. In einem iterativen Prozess im Rahmen einer Workshop-Serie wurden die Test-Cases unter Mitwirkung des AIT-Teams bis zum erforderlichen Detailgrad ausgearbeitet. Die Test-Cases sollen möglichst die echten Aufgaben der BMEIA-Mitarbeiterinnen und -Mitarbeiter widerspiegeln und sind in drei Typen unterteilt:

a) **Dokument-Suche:** Studienteilnehmerinnen und Studienteilnehmer werden gebeten, ein bestimmtes Dokument zu finden. Die Antwort kann entweder als Dokument-Downloadlink oder als Dokument-ID eingegeben werden.
b) **Inhalts-Suche:** Studienteilnehmerinnen und Studienteilnehmer werden gebeten, bestimmte Inhalte eines Dokumentes wiederzugeben. Dies kann bspw. eine Aufzählung von Themen oder Begriffen sein. Die Antwort erfolgt im Form von Freitext und muss keinem Muster folgen.
c) **Text-Erstellung**: Studienteilnehmerinnen und Studienteilnehmer werden gebeten, eine E-Mail-Anfrage zu beantworten. Sie erhalten außerdem Hinweise zu den erforderlichen Informationsangaben. Die Antwort kann als Freitext formuliert werden, Studienteilnehmerinnen und Studienteilnehmer werden dabei angehalten, serviceorientierte und höfliche Formulierungen zu verwenden.

Insgesamt wurden 25 Test-Cases definiert, wobei 9 der Evaluierungsaufgaben aus je 2 Schritten bestehen: (1) Dokument-Suche und (2) Inhalts-Suche bzw. (1) Dokument-Suche und (2) Text-Erstellung. Diese Aufgaben sind in der Benutzer-Ansicht entsprechend gruppiert, um die Bearbeitung zu vereinfachen, da in solchen Fällen das Dokument aus dem Schritt (1) im Schritt (2) erforderlich ist.

Somit ergeben sich in der Benutzer-Ansicht 16 Aufgaben, wobei 9 davon je 2 Schritte enthalten. Tabelle 6 bietet die Übersicht über die Aufteilung der Test-Cases.

Tabelle 6 - Übersicht Test-Cases

Typ	Anzahl Test-Cases	davon mit Cross-Sprache-Fokus[3]
Dokument-Suche	16	3
Inhalts-Suche	7	-
Text-Erstellung	2	-
Gesamt	**25**	

Beispiel-Aufgabe (Dokument- und Inhalts-Suche):

1.1. "Finden Sie das Dokument zur Arbeitspsychologischen Betreeung."

1.2. "Führen Sie an, wobei (bei welchen Themen) die Arbeitspsychologie unterstützen kann."

Beispiel-Aufgabe (Dokument-Suche Cross-Sprache):

"An der ÖB Paris soll den französischsprachigen Lokalangestellten das Thema 360-Grad Feedback im BMEIA nähergebracht werden. Dafür möchten Sie das BMEIA-Führungsbild in französischer Sprache finden."

Die Test-Cases verfügen neben der ID, der Beschreibung und dem Task-Typ über eine oder mehrere korrekte Antworten, Informationskriterien und Evaluierungsbeispiel (für Inhalts-Suche und Text-Erstellung), Formkriterien (für Text-Erstellung), Schwellenwerte für Form- und Informationskriterien, einen Aufgabentitel für die Benutzer-Ansicht. Zusätzlich wurde ein Feld für die Sicherheitseinstufung vorgesehen, die Test-Cases wurden jedoch unter der Sicherheitseinstufung "keine" definiert. Verfügbare Datenfelder je nach Test-Case-Typ sind in der Tabelle 7 aufgelistet.

[3] In der Aufgabe wird ein Dokument in einer bestimmten Sprache abgefragt, bspw. eine Formularversion in Englischer Sprache.

Tabelle 7 - Verfügbare Datenfelder je nach Test-Case-Typ

Feld	Bedeutung	Dokument-Suche	Inhalts-Suche	Text-Erstellung
_id	Interner Identifikator	ja	ja	ja
task_id	Test-Case Identifikator. z.B. 3.1	ja	ja	ja
task_type	Typ, z.B. Dokument-Suche	ja	ja	ja
security_level	Sicherheitseinstufung, True/False	ja	ja	ja
title	Aufgabentitel	ja	ja	ja
description	Beschreibung der Aufgabe	ja	ja	ja
correct_answer	Eine oder mehrere korrekte Antworten	ja	ja	ja
criteria_info	Informationskriterien für die Auswertung	nein	ja	ja
criteria_form	Formkriterien für die Auswertung	nein	nein	ja
evaluation_example	Beispiel einer Benutzerantwort mit kompletter Auswertung	nein	ja	ja
instruction_message	Optionale Anleitung zur Kriterienbewertung für ein LLM-Judge	nein	ja	ja
criteria_threshold	Minimale erforderliche Anzahl erfüllter Informationskriterien für eine positive Endbewertung	nein	ja	ja
criteria_form_threshold	Minimale erforderliche Anzahl erfüllter Formkriterien für eine positive Endbewertung	nein	nein	ja

8.1.1 Auswertung

Bei der Auswertung wird festgestellt, ob ein Bearbeitungsvorgang eine/r Studienteilnehmerin oder Studienteilnehmer ein korrektes Ergebnis lieferte. Ein Bearbeitungsvorgang bezieht sich immer auf ein bestimmtes Test-Case. Die Auswertung der Korrektheit ist erforderlich, um die quantitativen Metriken wie Zeit bis zum Erfolg/ Anzahl Versuche bis zum Erfolg ableiten zu können.

Bei 62[4] Studienteilnehmerinnen und Studienteilnehmer und 25 Test-Cases sind im Optimalfall (Erfolg beim ersten Versuch, alle Studienteilnehmerinnen und Studienteilnehmer lösen alle Aufgaben) 1550 Bearbeitungsvorgänge zu erwarten. Unter den realistischen Annahmen, dass nicht alle teilnehmenden Personen alle 25 Aufgaben lösen werden, und aber eine korrekte Lösung im Schnitt mehr als einen Versuch benötigt, ist die zu erwartende Menge der Bearbeitungsvorgänge immer noch zu hoch für eine manuelle Auswertung. Aus diesem Grund wurde der Auswertungsprozess automatisiert. In Einzelfällen sind jedoch eine manuelle Überprüfung und Korrektur jederzeit möglich.

Die erfassten Teilnehmer-Daten enthalten nur die erforderlichen Informationen für die Studie und keine echten Personennamen, sodass der Auswertungsprozess anonym verläuft (s. Tabelle 8). Zugangsdaten inkl. Benutzernamen ("user1", … "user62") wurden generiert und erlauben allein keine Rückschlüsse auf die einzelnen Personen. Die BMEIA-Studienbetreuerinnen und -Betreuer verfügen jedoch über die Teilnehmer-Liste, die ein Mapping von Benutzernamen zu echten Personennamen ermöglicht. Diese Maßnahme erlaubt es, bei auftretenden Problemen möglichst gezielt auf diese zu reagieren. Die technische Studienbetreuung (AIT und CIB-Teams) erhält dabei nur den Benutzernamen und die Aufgaben-ID(s).

Tabelle 8 - Erfasste Benutzerdaten

Feld	Bedeutung
_id	Interner Identifikator
username	Benutzername (user1, user2, … user62)
hashed_password	Passwort-Hash
gender	Geschlecht (m/w/d)
length_of_service	Dienstalter in Jahren
test_group	Test-Gruppe (A/B)
solved_tasks	Liste der gelösten Test-Cases
failed_tasks	Ungelöste Test-Cases inkl. Hinweis zum letzten Versuch

[4] Die erwartete Anzahl der Teilnehmerinnen und Teilnehmer betrug nach der Empfehlung 64 Personen und wurde nach der Bekanntgabe der Teilnehmer-Liste auf 62 korrigiert.

tasks_in_evaluation	Liste der Test-Cases, die sich aktuell in der Auswertung befinden

Nachfolgend ist der Auswertungsprozess für den jeweiligen Test-Case-Typ beschrieben:

a) **Dokument-Suche**
   Der Download-Link des Dokumentes oder die Dokument-ID wird mit den korrekten Antworten abgeglichen. Ist die Benutzer-Eingabe unter den korrekten Antworten aufgelistet, wird die Lösung akzeptiert. Anderenfalls wird ein Hinweis generiert, ob die Eingabe nicht mit der erwarteten Form konform war oder ein falsches Dokument gefunden wurde. Der Hinweis sollte beim nächsten Versuch den Benutzerinnen und Benutzern helfen, die Aufgabe richtig zu lösen.
b) **Inhalts-Suche**
   Die vordefinierten Informationskriterien (wie bspw. Benennung eines bestimmten Geldbetrages) werden von einem sogenannten LLM-Judge (einem KI-gestützten System) geprüft, s. 8.1.2. Diese Methode ist darauf ausgerichtet, die Antworten in natürlicher Sprache so auszuwerten, dass unterschiedlichste Formulierungen, Paraphrasen und Synonyme akzeptiert werden. Die Antworten der Benutzerinnen und Benutzer sind daher an keine Form gebunden.
   Sollten die Kriterien nicht ausreichend erfüllt sein, wird die/der Benutzerin oder Benutzer darüber informiert, bspw. "2 von mindestens 4 Kriterien erfüllt".
c) **Text-Erstellung**
   Es werden Informationskriterien wie bei der Inhalts-Suche geprüft. Zusätzlich werden nach einem ähnlichen gleichen Schema die Formkriterien evaluiert. Für die Gesamtbewertung werden die beiden Kriterien-Sets mit den entsprechenden Schwellenwerten verglichen. Für eine positive Gesamtbewertung müssen sowohl die Informations- als auch die Formkriterien ausreichend erfüllt werden.
   Sollte die Aufgabe nicht gelöst worden sein, wird die/der Benutzerin oder Benutzer informiert, ob dies an dem Inhalt und/oder an der Form liegt und wie viele Kriterien bis zur Erfüllung fehlen.

Tabelle 9 bietet eine Übersicht der erfassten Daten pro Bearbeitungsvorgang.

Tabelle 9 - Erfasste Daten pro Bearbeitungsvorgang

Feld	Bedeutung

_id	Interner Identifikator
user_id	Benutzer-ID, erlaubt das Mapping mit Benutzerdaten
status	Status des Bearbeitungsvorgangs: "processed" - bearbeitet und in der Evaluierung; "cancelled" - abgebrochen, "solved" - korrekt gelöst, "failed" - ungelöst
result	Benutzerantwort
task_id	Test-Case-ID, erlaubt das Mapping mit den Test-Cases
timestamp	Zeitstempel
elapsed_seconds	Verstrichene Zeit in Sekunden
user_feedback	Benutzerfeedback, falls abgegeben
eval_reason	Detaillierte Bewertungsgrundlage, nur für Inhalts-Suche und Text-Erstellung, s. 8.1.2

8.1.2 LLM-Judge

Die Auswertung der Test-Cases zur Inhalts-Suche und Text-Erstellung erfordert im Gegensatz zur Dokument-Suche eine komplexere Vorgehensweise, da hier kein direkter Abgleich mit einer oder mehreren korrekten Antworten möglich ist. Das Auswertungssystem soll verschiedene Formulierungen akzeptieren, sofern eine Antwort bestimmte Kriterien erfüllt. Ebenso wenig hilfreich ist eine einfache Überprüfung der Distanzen zwischen den semantischen Repräsentationen (Embedding-Vektoren), da diese zwar semantische Ähnlichkeiten auch unter der Verwendung der Synonymen und Paraphrasen berücksichtigt, jedoch nicht sicherstellen kann, dass Fakten korrekt wiedergegeben wurden. Aus diesem Grund wurde ein LLM-basiertes Auswertungssystem implementiert, ein sogenanntes LLM-Judge.

Dieses System agiert in der konkreten Implementierung Kriterien-basiert. Jedes der manuell definierten Informations- oder Formkriterien wird einzeln geprüft und bewertet. Das LLM-Judge gibt außerdem eine Begründung zu jeder Bewertung in natürlicher Sprache ab. Diese Begründungen dienen einerseits dazu, dass das LLM-Judge eine Reasoning-Grundlage für die Bewertung erstellt, andererseits lassen sich dadurch mögliche Fehlerquellen leichter erkennen. Experimentell wird zusätzlich die Konfidenz der Bewertung geschätzt, d.h. wie sicher die Bewertung aus der Sicht des Systems ist. Niedrigere Werte sollten dabei helfen, unsichere Bewertungen für eine mögliche Nachprüfung oder auch generell vage formulierte Kriterien bei wiederholter Unsicherheit zu identifizieren.

Das LLM-Judge bekommt eine Anleitung zur Aufgaben-Bewertung, die sowohl allgemeine Hinweise (bspw., dass Synonyme zu akzeptieren sind) als auch Aufgaben-spezifische Informationen, wie ein manuell erstelltes detailliertes Bewertungsbeispiel (evaluation_example) enthält. Die LLM-Antwort ist an ein Schema gebunden (s. Tabelle 10), welches vordefiniert ist und dynamisch mit den Details zum jeweiligen Test-Case befüllt wird (Informations- und Formkriterien, optional eine spezifische Anleitung). Diese Vorgehensweise stellt sicher, dass die LLM-Antwort immer geparst und weiterverarbeitet werden kann. Außerdem sinkt die Wahrscheinlichkeit der Halluzinationen und Fehlbewertungen im Vergleich zu einer rein Prompt-basierten Bewertung, mögliche Fehlerquellen lassen sich leichter identifizieren und beheben.

Tabelle 10 - Bewertungsschema

Bewertungs-Schema: Feld/ Eigenschaft	**Bewertung_Kriterium**	**Begründung_Kriterium**	**Konfidenz_Kriterium**
description	Textuelle Beschreibung des Kriteriums <N> (individuell für Test-Case), bildet die Anleitung für LLM-Judge.	Textuelle kurze Begründung zum Kriterium: Warum es erfüllt oder nicht erfüllt ist.	Einschätzung der Bewertungskonfidenz (wie sicher ist die Bewertung), zwischen 0.0 und 1.0
title	Bewertung_Kriterium<N>	Begründung_Kriterium<N>	Konfidenz_Kriterium<N>
type	Boolean (True/False)	String (Text)	Float [0.0, 1.0]

Bewertungsbeispiel (vereinfacht):

Aufgabenbeschreibung: "Führen Sie an, wobei (bei welchen Themen) die Arbeitspsychologie unterstützen kann."

Benutzerantwort: "Überlastung, zeit-management."

Kriterien: 1. Überlastung, 2. Burnout, 3. Zeit- und Selbstmanagement, <…>

Bewertungsschema:

Bewertung_Kriterium1: description: "Benutzer hat folgendes Thema oder Begriff erwähnt (auch Paraphasen, Synonyme sind erlaubt): **Überlastung**"

Begründung_Kriterium1: description: "Kurze Begründung der Bewertung zum Kriterium: Warum ist es erfüllt oder nicht erfüllt?"

Konfidenz_Kriterium1: description: "Einschätzung der Bewertungs-Konfidenz: Wie sicher ist die Bewertung, zwischen 0.0 (unsicher, viel Spielraum in der Interpretation der Benutzer-Antwort) und 1.0 (Bewertung ist völlig sicher)"

Bewertung_Kriterium2: description: "Benutzer hat folgendes Thema oder Begriff erwähnt (auch Paraphasen, Synonyme sind erlaubt): **Burnout**"

Begründung_Kriterium2: <…>

Konfidenz_Kriterium2: <…>

Bewertung_Kriterium3: description: "Benutzer hat folgendes Thema oder Begriff erwähnt (auch Paraphasen, Synonyme sind erlaubt): **Zeit- und Selbstmanagement**"

Begründung_Kriterium3: <…>

Konfidenz_Kriterium3: <…>

LLM-Judge-Bewertung:

Bewertung_Kriterium1: True

Begründung_Kriterium1: "Benutzer erwähnt explizit Überlastung."

Konfidenz_Kriterium1: 1.0

Bewertung_Kriterium2: False

Begründung_Kriterium2: "Benutzer erwähnt 'Burnout' nicht."

Konfidenz_Kriterium2: 1.0

Bewertung_Kriterium3: True

Begründung_Kriterium3: "Benutzer erwähnt ‚zeit-management‘, was dem Kriterium entspricht."

Konfidenz_Kriterium3: 0.8

Die Formkriterien, relevant für die Text-Erstellung-Tasks, werden ähnlich ausgewertet. Diese inkludieren folgende Anforderungen: Höfliche und formelle Begrüßung am Anfang der Antwort, serviceorientierte und freundliche Formulierungen, eine Höflichkeitsformel am Ende der fiktiven E-Mail.

Anschließend wird die Anzahl der erfüllten Kriterien mit dem jeweiligen Schwellenwert verglichen und eine Gesamtbewertung des Bearbeitungsvorgangs als gelöst/ungelöst abgegeben. Dieser Prozess erfolgt ohne LLM-Judge.

Die Schwellenwerte sind manuell für jede Aufgabe bestimmt und erlauben bewusst einige Fehler, um ggf. abweichendes Verhalten des LLM-Judge bzw. Ungenauigkeiten in den Benutzer-Antworten oder Kriterien-Formulierungen abzufedern. Die Formkriterien werden separat von den Informationskriterien ausgewertet und haben einen eigenen Schwellenwert. Für eine positive Gesamtbewertung der Text-Erstellungs-Aufgaben müssen beide Kriterien-Sets ausreichend erfüllt werden.

Das LLM-Judge-System wurde auf Basis eines Development-Datensatzes mit 3 Beispiel-Aufgaben und 20 manuell erstellten fiktiven Antworten und deren Bewertungen entwickelt. Diese geringe Datenmenge erlaubt zwar keine statistisch signifikanten Aussagen in Bezug auf die verglichenen LLMs (Tabelle 11), trug jedoch durch die enthaltenen Edge-Cases und vorgefertigte "Problemfälle" zur Robustheit des Systems bei, die durch stichprobenartige Kontrollen während der Evaluierungsstudie bestätigt wurde. Insbesondere eine korrekte und möglichst konkrete Formulierung der Kriterien sowie ein Bewertungsbeispiel verringern die Fehleranfälligkeit.

Tabelle 11 - Modell-Vergleich

Modell	N (Korrekt)	N (Teils korrekt)	N (Falsch)
qwen3:32b (Q4_K_M)	16	3	1

qwen2.5:32b-instruct (Q4_K_M)	18	1	1
Qwen3-30B-A3B-Instruct-2507 (Q4_K_M)	17	2	1
gemma2:27b (Q4_0)	17	3	0
DeepSeek-R1-Distill-Qwen-32B:q4_k_m	18	1	1

In der deployten Tool-Version kommt gemma2:27b-instruct-q8_0 zum Einsatz. Dieses Modell konnte in der Entwicklungsumgebung aufgrund fehlender Hardware-Kapazitäten nicht getestet, jedoch ist dieses als q8-Version performanter als das getestete gemma2:27b in der q4-Version.

8.2 Evaluierungstool

Das Evaluierungstool wurde als eine vom KnowHow-Tool unabhängige Software implementiert. Das Evaluierungstool soll eine Funktionalität zur Auswertung der quantitativen (Bearbeitungszeit, Korrektheit) und qualitativen (Benutzerzufriedenheit) Merkmale der Bearbeitungsvorgänge der Test-Benutzerinnen und -Benutzer bereitstellen. Test-Personen aus den beiden Gruppen (A: Intranet, B: KnowHow-Tool) sollten auf dieselbe Art und Weise mit der Evaluierungsumgebung interagieren, sodass keine Verzerrung in den Bearbeitungszeiten entstehen kann. Aus diesem Grund wurde eine zuerst angedachte Integration mit dem KnowHow-Tool nicht durchgeführt, da diese möglicherweise einen Vorteil für die Gruppe B bringen würde.

Die Evaluierungsumgebung inkludiert ein Python-basiertes Backend bestehend aus mehreren Komponenten und eine Svelte-basierte Benutzeroberfläche und wurde als Docker-Compose Stack in der Zielumgebung deployt.

8.2.1 Backend

Das Evaluierungstool besteht aus den folgenden Komponenten (s. Abbildung 9):

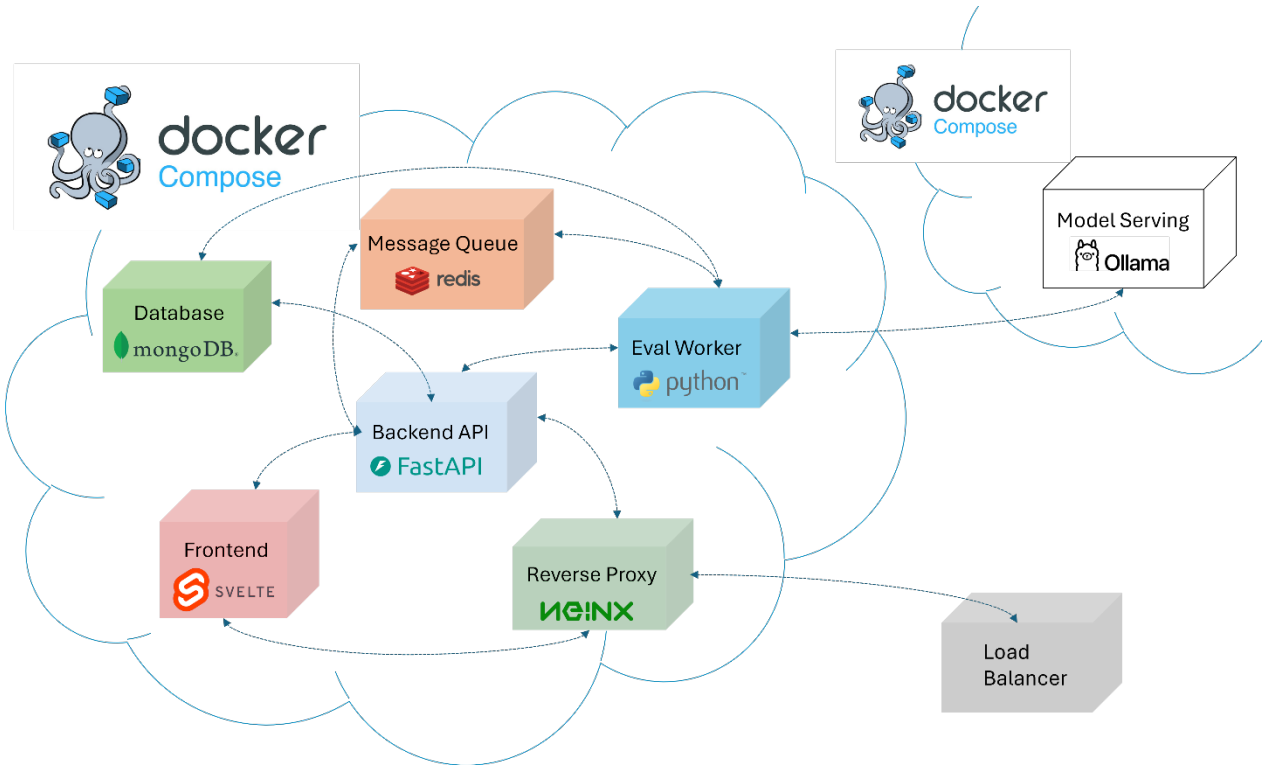

Abbildung 9 - Übersicht Evaluierungstool: Komponenten

Backend API: Schnittstellen für die Kommunikation zwischen dem Frontend und weiteren Modulen, implementiert mit FastAPI.

Database: Dokument-orientierte nicht-relationale MongoDB Datenbank dient der Persistenz der Benutzerdaten, Test-Cases, Bearbeitungsvorgänge und Feedback-Fragen. MongoDB verfügt über eine asynchrone Client-Implementierung und ist daher geeignet für den asynchronen Workflow des Tools.

Message Queue: Redis Server dient als Message Queue für die Auswertung der Bearbeitungsvorgänge. Diese werden vom Backend empfangen, in der Datenbank persistiert und in die Message Queue gepostet.

Evaluation Worker: Dieses Modul erlaubt eine asynchrone Auswertung der Bearbeitungsvorgänge aus der Message Queue. Es bekommt die Bearbeitungs-ID aus der Message Queue, fragt entsprechende Daten aus der Datenbank ab, führt die Auswertung aus und

schreibt das Auswertungsergebnis in die Datenbank. Evaluation Worker kommuniziert direkt mit der **Model Serving** Komponente, die in einem eigenen Docker Compose-Stack mit Ollama deployt ist.

Reverse Proxy: Umgesetzt mit Nginx, dient als Routing-Komponente. Diese wurde aufgrund der Deployment-Architektur erforderlich. Ein externer Load-Balancing System wurde für einen Port konfiguriert, an dem der Nginx-Service exposed ist.

Frontend: Benutzeroberfläche (WebUI), implementiert mit SvelteKit. Diese Komponente kommuniziert mit Backend API und ist beschrieben im nachfolgenden Abschnitt.

8.2.2 Frontend (WebUI)

Die Benutzeroberfläche des Evaluierungstools implementiert die Funktionen aus dem Workflow in der Abbildung 8 bis auf den Schritt 4. Die Erfassung der User-Anfragen ist aufgrund der abgekoppelten Implementierung des Tools nicht möglich. Die Bearbeitung eines Test-Cases läuft in einem anderen System (Intranet oder KnowHow Tool) ab, welches mit der Evaluierungsumgebung aufgrund der erfolgten Design-Entscheidungen nicht integriert wurde.

Die Benutzeroberfläche verfügt über eine Einstiegsseite, Anmelde- und Registrierungs-Seiten und einen Arbeitsbereich.

Die Einstiegsseite (Abbildung 10) bietet Informationen zum Studienablauf (bspw. Details zu Aufgabentypen und deren Auswertung) und zur Vorgehensweise bei Problemen (bspw. Schwierigkeiten bei der Lösung der Test-Cases). Insgesamt sind 10 Fragen und Antworten aufgelistet.

Mit der Anmeldemaske (Abbildung 11) können sich die Studienteilnehmerinnen und Studienteilnehmer mit den vergebenen Zugangsdaten anmelden. Die Benutzerdaten sind bereits in der Datenbank hinterlegt. Die Registrierungsmaske (Abbildung 12) kommt während der Studie nicht zur Anwendung, außer zu Testzwecken für interessierte BMEIA-Mitarbeiterinnen und -Mitarbeiter.

Abbildung 10 - Einstiegsseite

Abbildung 11 - Anmeldeseite

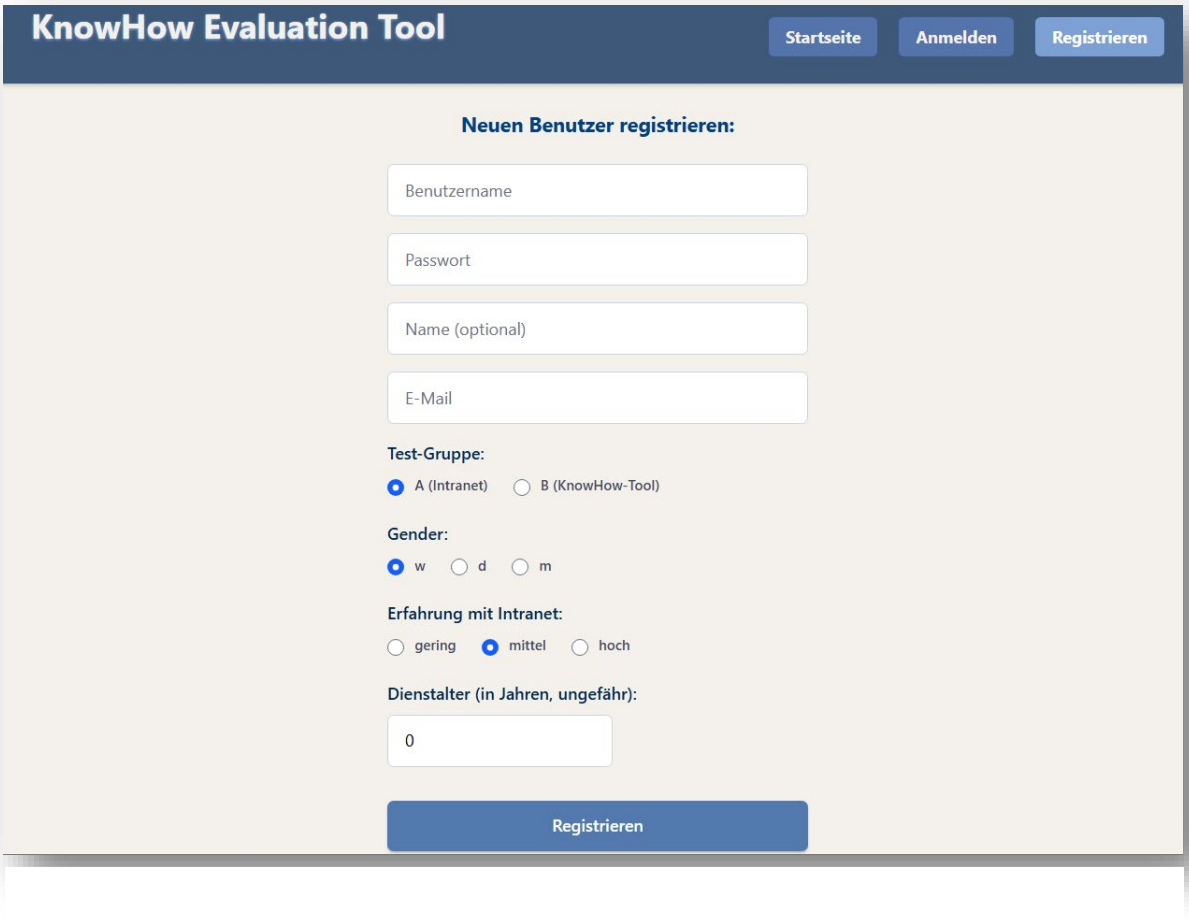

Abbildung 12 - Registrierung

Nach der erfolgten Anmeldung wird der Link bzw. Button "Arbeitsbereich" (**Error! Reference source not found.**) sichtbar. Dieser bietet eine Übersicht der Test-Cases. Die bereits von dem/der Benutzerin und Benutzer gelösten Test-Cases sind mit einem Bestätigungszeichen (Checkmark) markiert und können nicht mehr ausgewählt werden. Die noch nicht gelösten Test-Cases können zur Bearbeitung ausgewählt werden, dabei wird die Aufgabenbeschreibung angezeigt mit Details zum Aufgabentyp (**Error! Reference source not found.**). Falls die Aufgabe beim letzten Versuch nicht gelöst wurde, bekommt der/die Benutzerin oder Benutzer einen Hinweis, worauf zu achten ist. Dieser Hinweis bezieht sich immer auf die letzte Bearbeitung des Test-Cases.

Nach dem Click auf "Bearbeitung starten" wird die verstrichene Zeit gemessen (Abbildung 15). Studienteilnehmerinnen und Studienteilnehmer haben die Möglichkeit, die Bearbeitung abzubrechen, falls sie abgelenkt oder unterbrochen wurden, um die Zeit-Messungen nicht zu verfälschen.

Abbildung 13 - Arbeitsbereich

Abbildung 14 - Aufgabenansicht

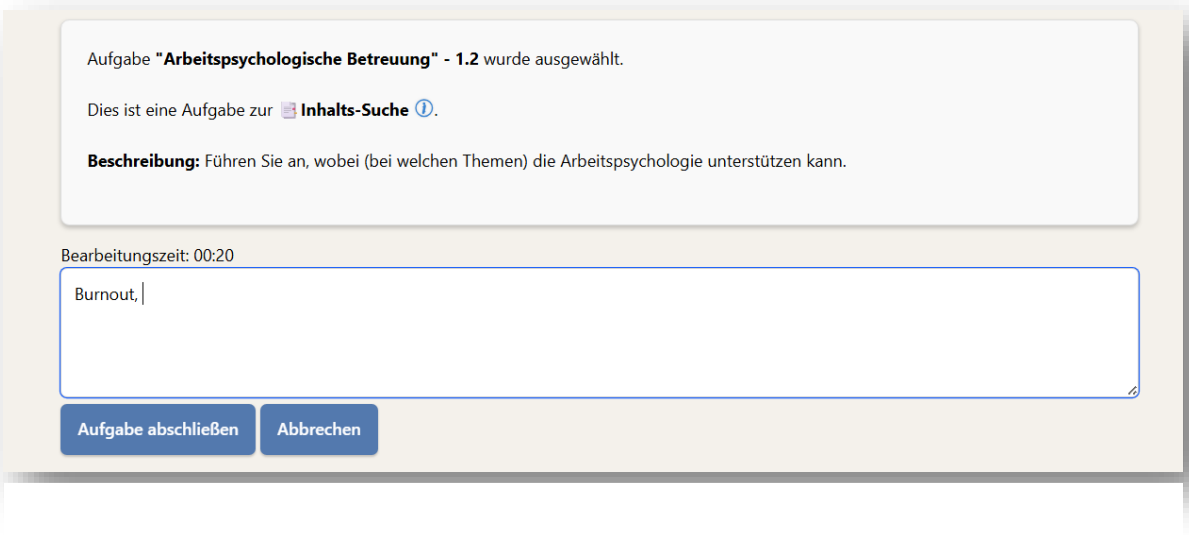

Abbildung 15 - Bearbeitungsvorgang

Nach dem Absenden einer Antwort ("Aufgabe abschließen") können Studienteilnehmerinnen und Studienteilnehmer Feedback zum Bearbeitungsvorgang abgeben (Abbildung 16). Es ist möglich, nur eine oder mehrere Fragen zu beantworten oder das Feedback ganz zu überspringen.

Abbildung 16 - Feedback

9 Evaluierungsergebnisse

9.1 Vorbemerkungen und Datenaufbereitung

Die tatsächliche **Studienbeteiligung** war etwas niedriger als angenommen mit je 28 Personen, jeweils 15 weiblich, 13 männlich, pro Test-Gruppe. Keine Personen haben ihr Geschlecht als "divers" angegeben. 6 Personen aus der ursprünglichen Teilnehmer-Liste haben an der Studie nicht teilgenommen.

Abbildung 17 und Tabelle 12 bieten einen Überblick über die Verteilung der Teilnehmerinnen und Teilnehmer nach Dienstalter und Gender in den Gruppen (A - Kontroll-Gruppe, Intranet; B - KnowHow-Tool). Verschiedene Dienstalter-Gruppen sind repräsentiert, wenn auch die Gruppe B etwas jünger ausfällt.

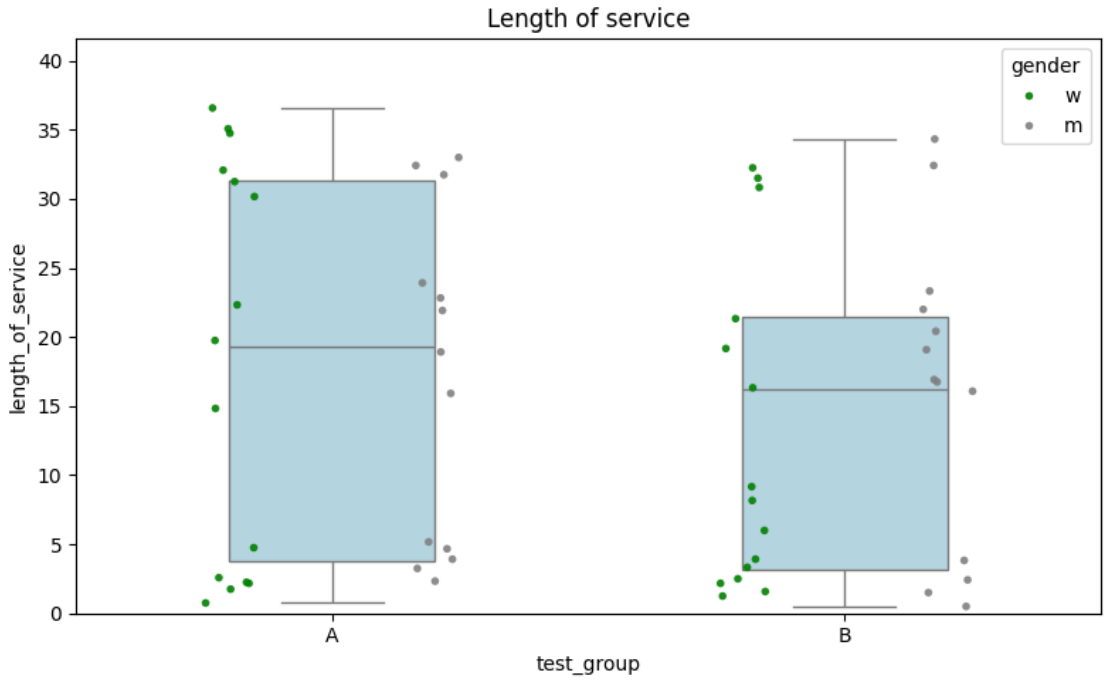

Abbildung 17 - Dienstalter und Gender der Studienteilnehmerinnen und Studienteilnehmer nach Test-Gruppe (Box-Plot)

Tabelle 12 - Dienstalter der Studienteilnehmerinnen und Studienteilnehmer

Parameter	A	B
count	28 (15w / 13m)	28 (15w / 13m)
mean	17.54	14.25
std	13.11	11.38
min	0.75	0.5
25%	3.75	3.12
50%	19.34	16.2
75%	31.38	21.5
max	36.58	34.33

Der **Studienzeitraum** wurde verlängert, um eine möglichst hohe Beteiligung zu erreichen, und lag zwischen dem 12.01.2026 und 22.02.2026. Den Studienteilnehmerinnen und Studienteilnehmer wurden insgesamt 5 Online-Einführungstermine angeboten, im deren Rahmen sie über den Studienablauf informiert und im Umgang mit Tools eingeschult wurden. Zusätzlich wurden Schulungsunterlagen und eine Videoaufzeichnung für Personen vorbereitet, die keinen der Termine wahrnehmen konnten. Der erste Schulungstermin fand am 08.01.2026 statt. Einige Studienteilnehmerinnen und Studienteilnehmer aus der Gruppe A konnten bereits anschließend und noch vor dem offiziellen Start der Studie am 12.01.2026 mit der Bearbeitung der Test-Cases anfangen; diese Ergebnisse wurden ebenfalls bei der Auswertung berücksichtigt.

Jede/r Studienteilnehmerin oder Studienteilnehmer konnte eine, mehrere oder alle der 25 Test-Cases bearbeiten. Die Anzahl Versuche pro Aufgabe war uneingeschränkt. Die maximale Bearbeitungszeit für einen Versuch wurde auf eine Stunde festgelegt, um etwa vergessene Bearbeitungsvorgänge zu vermeiden. Die abgebrochenen ("cancelled") Bearbeitungsvorgänge wurden in der Auswertung nicht berücksichtigt. S. Übersicht der Anzahl **Bearbeitungsvorgänge** in Abbildung 18 und Tabelle 13.

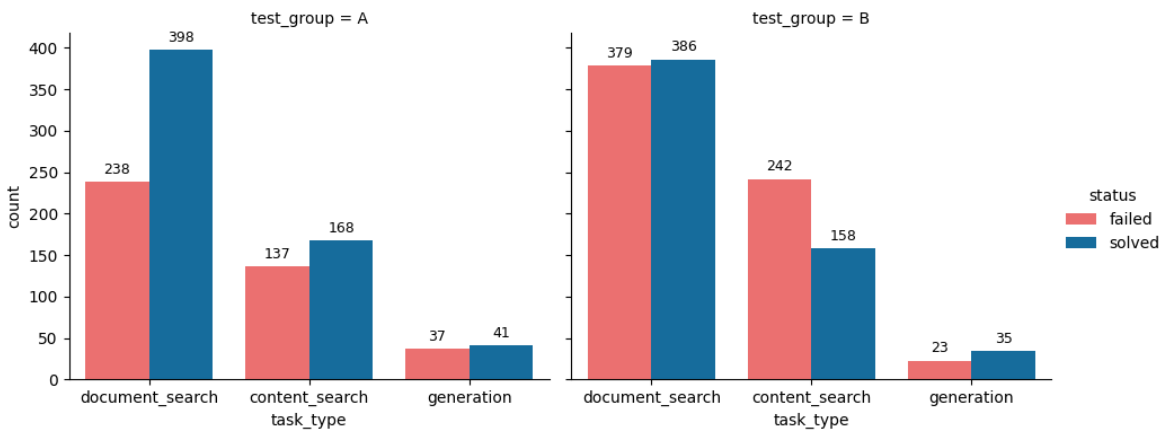

Abbildung 18 - Anzahl Bearbeitungsvorgänge nach Test-Case-Typ, Gruppe und Status (Bar-Plot)

Tabelle 13 - Anzahl Bearbeitungsvorgänge nach Gruppe und Status

Test-Gruppe	Status	Anzahl	Gesamt/Gruppe
A	solved	607	1019
A	failed	412	
B	solved	579	1223
B	failed	644	

In einigen Fällen konnten Studienteilnehmerinnen und Studienteilnehmer eine Aufgabe nie korrekt lösen, dies resultierte in einem Anteil der **unvollständigen, zensierten** (right-censored) **Beobachtungen**.

Nach jeder Bearbeitung eines Test-Cases konnten Studienteilnehmerinnen und Studienteilnehmer ihr Feedback bzgl. Such-Erfahrung, UX, Effizienz, Vertrauenswürdigkeit etc. abgeben. Diese Vorgehensweise erlaubte ein unmittelbares und differenzierteres Feedback je nach Aufgabe und Aufgabentyp, Erfolg oder Misserfolg, verwendete Tool-Funktionalität anstatt ein allgemeines Feedback am Ende der Studie.

Besondere Ereignisse im Laufe der Studie:

1. Aktualisierung der Aufgaben: Aufgrund der Intranet-Aktualisierung wurde eine entsprechende Aktualisierung einiger Aufgaben-Antworten notwendig. Diese wurde umgehend durchgeführt und die betroffenen Vorgänge neu evaluiert und berichtigt, sodass keine Verzerrung der erfassten Zeiten entstanden ist.
2. Test-Case 12.1 Dokument-Suche: Das für die Aufgabe 12.1 erforderliche Dokument "Visaliste" wurde aus technischen Gründen im Indexierungsprozess des KnowHow-Tools nicht berücksichtigt. Das Dokument war in weiterer Folge im KnowHow-Tool zuerst nicht auffindbar. Das Dokument wurde zwar nachträglich verfügbar gemacht, jedoch konnte ein möglicher Nachteil der Gruppe B hier nicht ausgeschlossen werden. Um eine Verzerrung in den gemessenen Zeiten zu vermeiden, wurden sämtliche Bearbeitungsergebnisse der Dokument-Suche Aufgabe 12.1 in der Auswertung nicht berücksichtigt.
3. Verbesserungen und Aktualisierung des KnowHow-Tools: In der ersten Phase der Studie wurden einige Updates des KnowHow-Tools betreffend die Benutzeroberfläche und die Performanz durchgeführt. Diese Updates wurden am 21.01.2026 abgeschlossen. Da die Aktualisierungen ggf. einen Einfluss auf die Bearbeitungszeiten hatten, wurden zusätzliche Tests auf der verringerten Datenmenge ab diesem Zeitpunkt durchgeführt.

9.1.1 Datenbereinigung und Datenaufbereitung

Aufgrund der nachträglichen Korrektur bei einigen Aufgaben (s. besondere Ereignisse (1)) sind Versuch-Abfolgen pro User und Test-Case entstanden, die mehrere als "gelöst" ausgewertete Versuche enthalten. Zur korrekten Messung der Zeit bis zum Erfolg wurden diese überflüssigen Bearbeitungsvorgänge entfernt. Es wurden (in chronologischer Reihenfolge) nur die Versuche bis zum ersten Erfolg behalten, pro User und Test-Case. **Die Bearbeitungen des Test-Case 12.1 (s. besondere Ereignisse (2)) wurden entfernt,** was die Anzahl der Bearbeitungsvorgänge um 117 verringerte (33 Gruppe A, 84 Gruppe B); s. Tabelle 14 und Abbildung 19 - Abbildung 22. Weitere Kennzahlen und Abbildungen beziehen sich auf die Daten ohne Test-Case 12.1.

Tabelle 14 - Anzahl Bearbeitungsvorgänge nach Gruppe und Status, ohne 12.1

Test Gruppe	Status	Anzahl ohne 12.1 (Differenz)	Gesamt/Gruppe ohne 12.1
A	solved	583 (-24)	986 (-33)

A	failed	403 (-9)	
B	solved	556 (-23)	1139 (-84)
B	failed	583 (-61)	

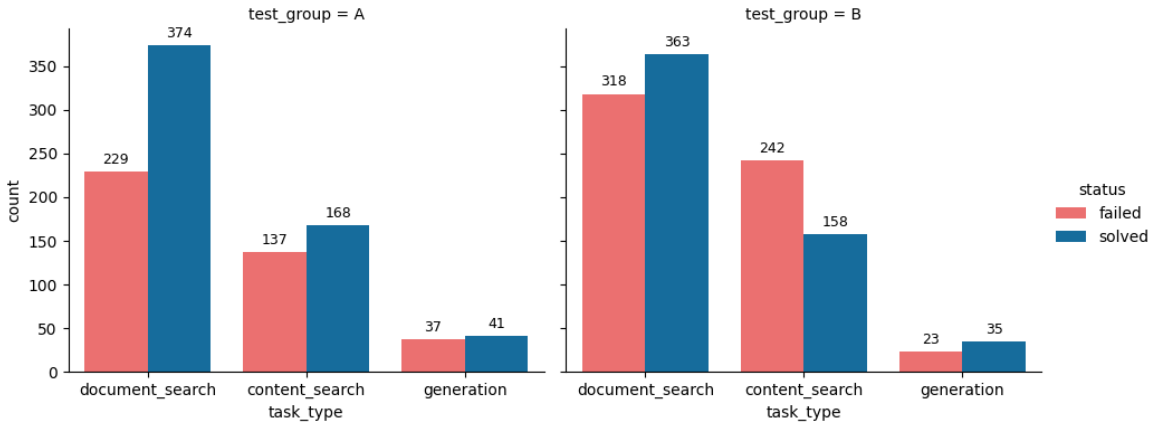

Abbildung 19 - Anzahl Bearbeitungsvorgänge nach Test-Case-Typ, Gruppe und Status ohne 12.1

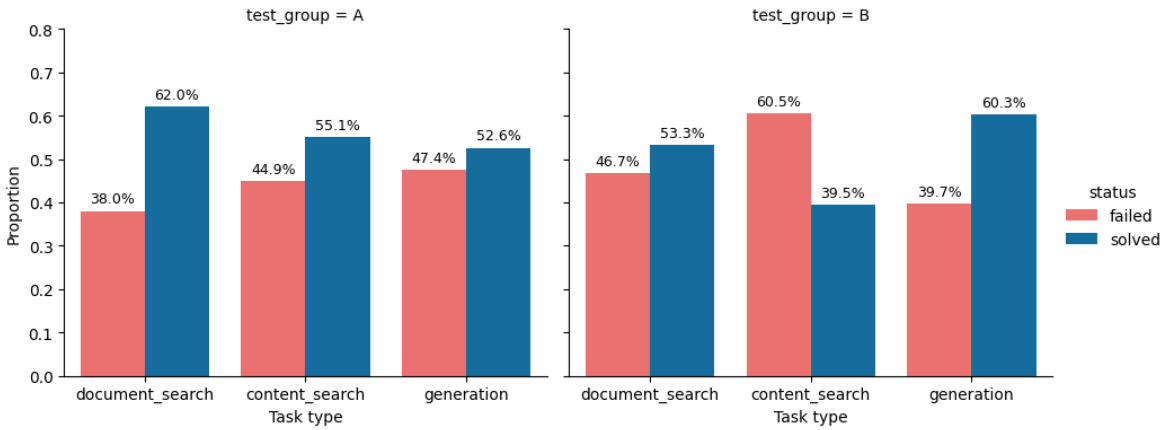

Abbildung 20 - Prozent korrekte/ inkorrekte Versuche in gesamten Versuchen pro Test-Case-Typ in Gruppen

Während Gruppe B insgesamt aktiver (1189 Bearbeitungsvorgänge) im Vergleich zur Gruppe A (986 Bearbeitungsvorgänge) war, liegt sie prozentuell bei den korrekt gelösten

("solved") Versuchen unter der Gruppe A in der Dokument-Suche (53,3 vs. 62,0) und Inhalts-Suche (39,5 vs. 55,1), allerdings über der Gruppe A in der Text-Generation (60,3 vs. 52,6), s. Abbildung 20.

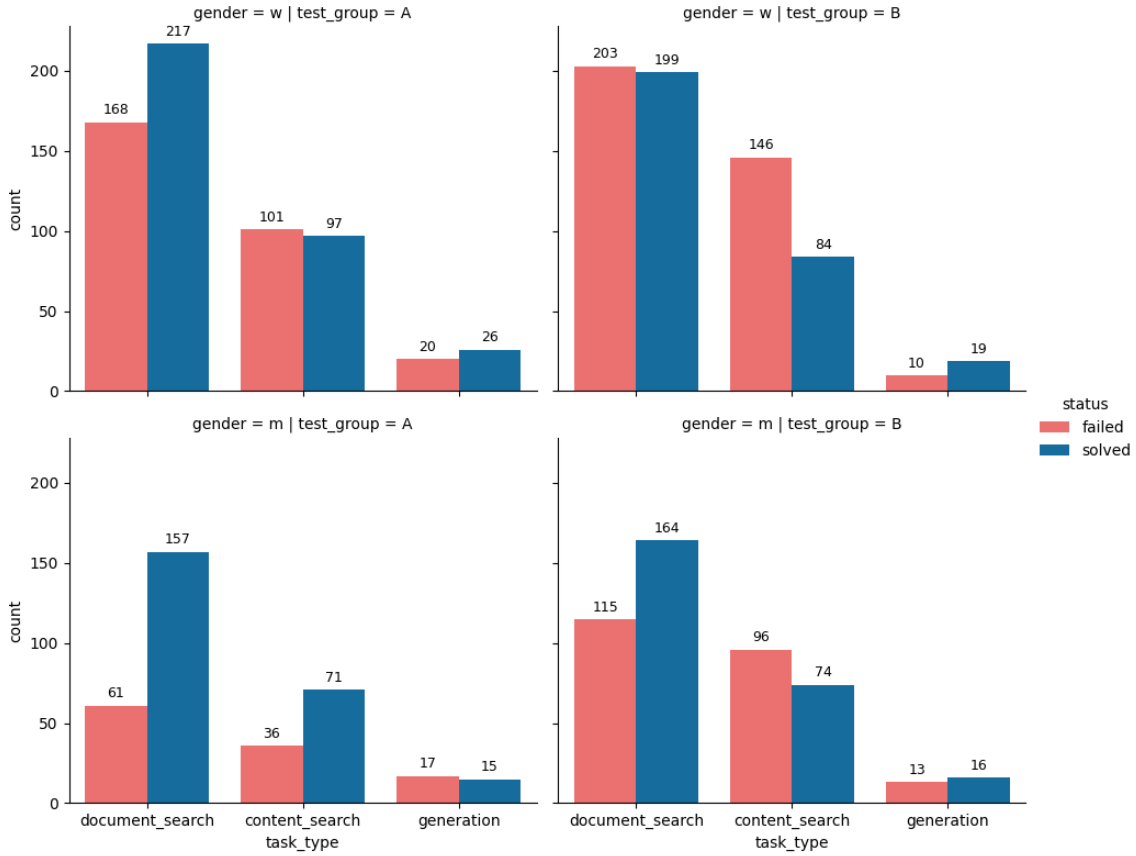

Abbildung 21 - Bearbeitungsvorgänge nach Test-Gruppe, Status, Aufgaben-Typ und Gender

Frauen in beiden Gruppen waren aktiver als Männer in Bezug auf die Anzahl Bearbeitungsvorgänge (Abbildung 21), was sich zum Teil durch höhere Frauenbeteiligung erklären lässt (15 zu 13 in beiden Gruppen). Betreffend den Anteil gelösten/ ungelösten Versuche nach Gender zeigt sich ein ähnliches Bild in den Gruppen: Bei den Dokument-Suche und Inhalts-Suche Aufgaben haben Männer in ihrer Test-Gruppe einen höheren Anteil der gelösten Versuche zu gesamten Versuchen im Vergleich zu Frauen, bei den Text-Generationsaufgaben liegen jeweils Frauen in ihrer Gruppe höher als Männer (Abbildung 22).

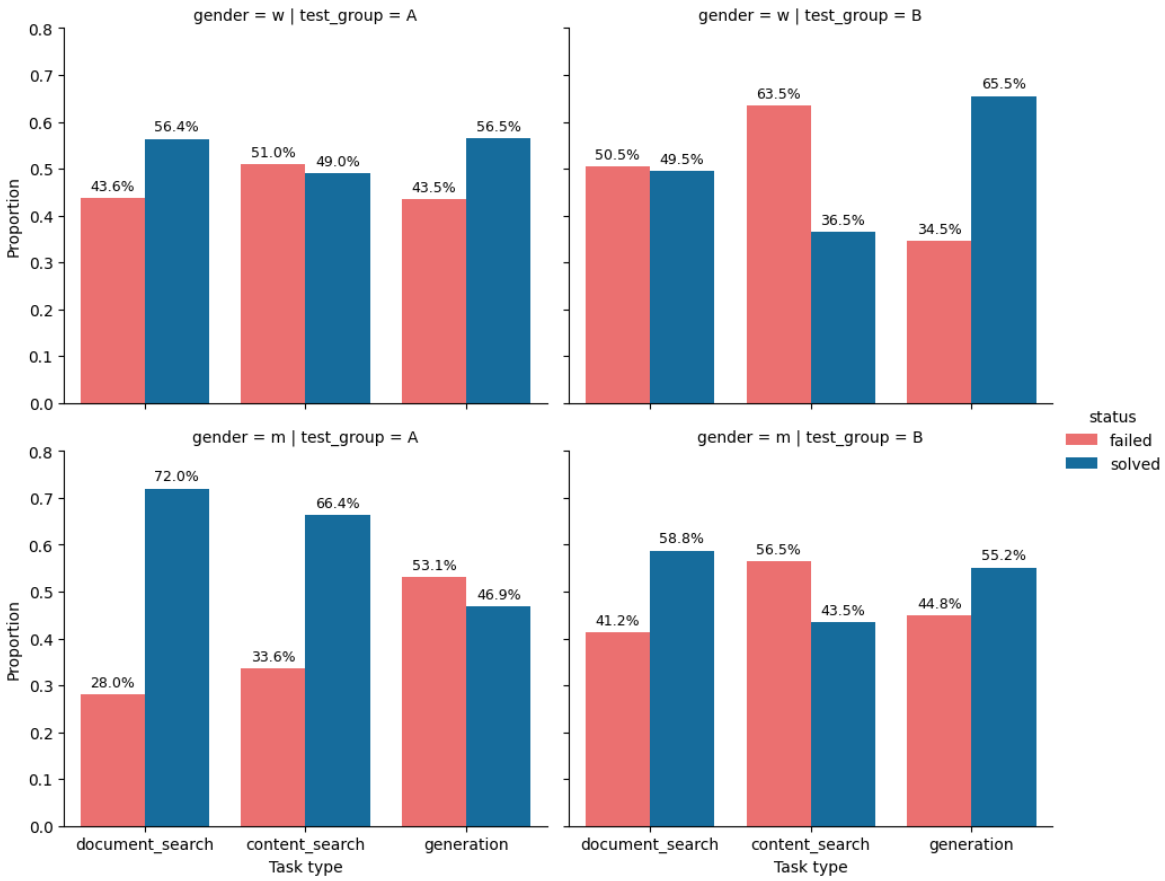

Abbildung 22 - Prozent korrekte/ inkorrekte Versuche in gesamten Versuchen pro Test-Case-Typ nach Test-Gruppe und Gender

Aggregierte Zeit pro User und Aufgabe: Verstrichene Zeit wurde nach User-ID und Aufgabe aggregiert (summiert), um die benötigten Zeiten der beiden Gruppen zu vergleichen. Einige Zeitmessungen enthalten kein Erfolgsereignis: Userinnen und User haben zwar versucht, eine Aufgabe zu lösen, haben aber nie eine korrekte Lösung gefunden. Solche Fälle spiegeln die Realität wider, können aber von statistischen Tests wie T-Test, Mann-Whitney U-Test o.ä. nicht korrekt behandelt werden, da die Beobachtung unvollständig ist (right-censored), und nur bekannt ist, dass die gemessene Zeit nicht ausgereicht hat, um den Erfolg zu erzielen. Solche Zeitmessungen bilden etwa 7% von den gesamten aggregierten Zeitmessungen (85 von 1224, s. Tabelle 15). Um Datenpunkte dieser Art korrekt berücksichtigen zu können, ist eine Ereigniszeitanalyse erforderlich.

Tabelle 15 - Aggregierte Zeitmessungen pro User und Test-Case und right-censored Messungen

Test-Gruppe	Anzahl right-censored	Anzahl gesamt	% right-censored
A	39	622	6,27
B	46	602	7,64
total	85	1224	6,94

18 Test-Cases wurden von mind. 25 Personen aus jeder Gruppe bearbeitet. Fast alle Test-Cases, bis auf 13.2, wurden von mehr als 20 Personen aus jeder Gruppe versucht (Abbildung 23).

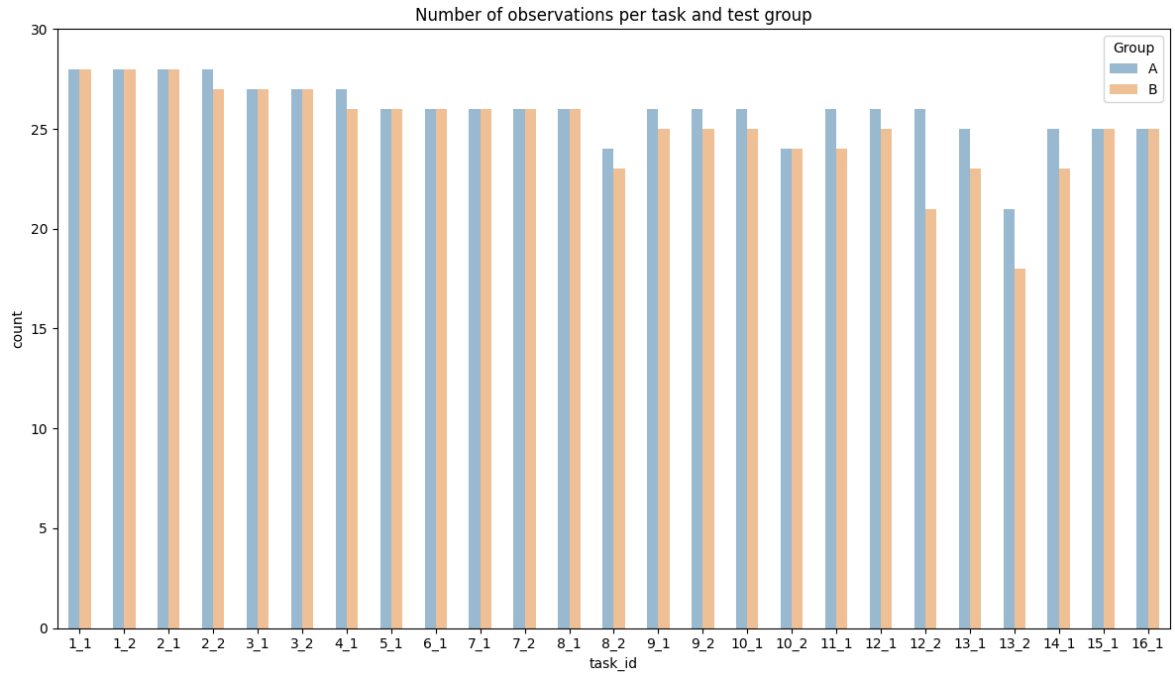

Abbildung 23 - Anzahl aggregierter Zeitmessungen pro Test-Case in den Test-Gruppen

Die Zeitmessungen sind positiv schief verteilt, weshalb in den Abbildungen die log-transformierte Form verwendet wird und für die Berechnung der Inferenzstatistiken der U-Test bzw. Welch T-Test auf log-transformierten Daten eingesetzt werden.

Die Studie hat primär das Ziel, die systematischen Unterschiede in den Bearbeitungszeiten der Gruppe zu evaluieren. Die Gender- und Dienstalter-Aspekte werden vorwiegend explorativ betrachtet und deren mögliche Effekte in der Ereigniszeitanalyse untersucht.

9.2 Explorative Analyse und deskriptive Statistik

9.2.1 Test-Gruppen

Aggregierte Zeitmessungen (Angaben in Sekunden, s) pro User und Aufgabe in Gruppen sind ähnlich verteilt, s. Tabelle 16 (alle Zeitmessungen) und Tabelle 17 (nur mit Erfolgsereignis). Die Abbildung 24 zeigt die Kerndichteschätzung der beiden Gruppen auf log-transformierten Zeiten.

Betrachtet man alle Zeitmessungen, liegt der Medianwert der Gruppe B mit 46,5s um 2,5s über dem Medianwert der Gruppe A (44s). Das erste Quartil beträgt jeweils 22s, das dritte liegt bei der Gruppe B um 1s höher (103,75s) als bei der Gruppe A (102,75s). Der Mittelwert der Gruppe B ist um etwa 7s höher (107,54s). Die Standardabweichungen der Gruppenzeiten unterscheiden sind um weniger als 1s und sind deutlich höher als in 7.1.2 angenommen.

Betrachtet man nur die Zeitmessungen mit einem Erfolgsereignis (gelöste Aufgaben), wird der Unterschied in den Zentralmaßen etwas höher: Der Medianwert der Gruppe B liegt mit 45s um 4s höher als in der Gruppe A (41s), der Mittelwert um ca. 9,5s. Nach wie vor wenig unterscheiden sich das erste (Gruppe B um 0,5s niedriger) und das dritte Quartil (Gruppe B um 2,5s höher). Die Standardabweichungen unterscheiden sich um etwa 4s.

Tabelle 16 - Deskriptive Kennzahlen für aggregierte Zeitmessungen inkl. right-censored Messungen

Parameter/ Test-Gruppe	count	mean	std	min	25%	median	75%	max

A	622	100.64	182.52	1.0	22.0	44.0	102.75	2225.0
B	602	107.54	183.23	1.0	22.0	46.5	103.75	1454.0

Tabelle 17 - Deskriptive Kennzahlen für aggregierte Zeitmessungen, nur gelöste (ohne right-censored)

Parameter/ Test-Gruppe	**count**	**mean**	**std**	**min**	**25%**	**median**	**75%**	**max**
A	583	97.16	183.28	1.0	21.5	41.0	96.00	2225.0
B	556	106.71	187.24	1.0	21.0	45.0	96.25	1454.0

Die Verteilungen überlappen sich größtenteils, die Gruppe B hat etwas höhere Dichte im Bereich der niedrigeren Zeitwerte im Vergleich zur Gruppe A (Abbildung 24). Der Modus der Gruppe B fällt höher aus als in der Gruppe A.

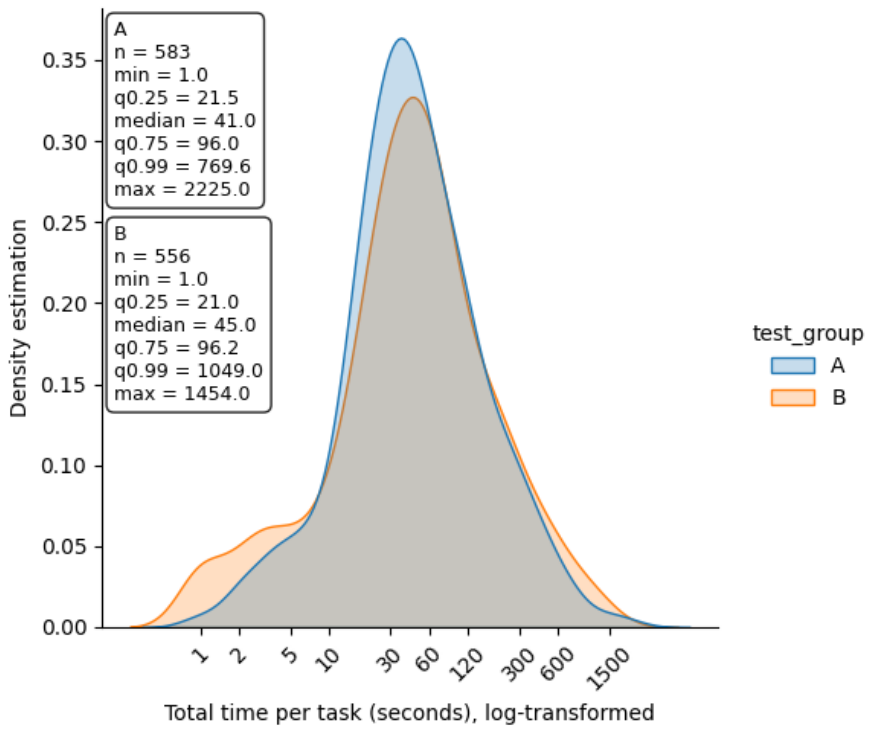

Abbildung 24 - Kerndichtediagramm (ohne right-censored)

9.2.2 Test-Gruppe und Gender

Die Verteilung der Zeitmessungen nach Gender-Gruppen ist Tabelle 18, Tabelle 19, Abbildung 25 dargestellt. In beiden Test-Gruppen überlappen sich die Verteilungen größtenteils. Frauen in der Gruppe B weisen einen niedrigeren Medianwert (42s alle Zeitmessungen bzw. 39,5s ohne right-censored) im Vergleich zu Männern (50s bzw. 49,5s), was einem Unterschied von 8s bzw. 10s entspricht. Gruppe A weist keinen (in allen Zeitmessungen, je 44s) bzw. geringeren (Männer mit 39s um 3s schneller als Frauen, ohne right-censored) Unterschied im Medianwert auf.

Während in der Gruppe A die geschätzte Kerndichte von Zeitmessungen der Frauen weiter in den höheren Wertbereich verschoben ist im Vergleich zu Männern, ist dies in der Gruppe B nicht der Fall, hier zeigt sich eine etwas höhere Dichte im niedrigeren Wertbereich bei Frauen. Der Modus ist in beiden Test-Gruppen niedriger bei Frauen als bei Männern derselben Gruppe.

Trotz eines hohen Überschneidungsgrades kann vermutet werden, dass gender-spezifische Unterschiede in den Test-Gruppen bestehen könnten: Frauen in der Gruppe B könnten einen Vorteil im Vergleich zu männlichen Kollegen aus derselben Gruppe haben, während dies in der Gruppe A nicht der Fall ist. Zugleich schneiden Männer der Gruppe B tendenziell langsamer ab als Männer der Gruppe A.

Tabelle 18 - Deskriptive Kennzahlen in Test-Gruppen nach Gender, alle Zeitmessungen

Test-Gruppe	gender	count	mean	std	min	25%	median	75%	max
A	**m**	267	82.15	110.16	1	23	44	93.5	748
	w	355	114.55	221.07	2	21	44	104	2225
	total	622	100.64	182.52	1	22	44	102.75	2225
B	**m**	281	106.57	182.19	1	25	50	102	1454
	w	321	108.4	184.42	1	19	42	106	1348
	total	602	107.54	183.23	1	22	46.5	103.75	1454

Tabelle 19 - Deskriptive Kennzahlen der Test-Gruppen nach Gender, ohne right-censored

Test-Gruppe	gender	count	mean	std	min	25%	median	75%	max
A	**m**	243	72.11	94	1	22.5	39	78.5	634
	w	340	115.06	224.92	2	21	42	104	2225
	total	583	97.16	183.28	1	21.5	41	96	2225
B	**m**	254	106.96	186.89	1	24.25	49.5	91	1454
	w	302	106.51	187.85	1	17	39.5	96.75	1348
	total	556	106.71	187.24	1	21	45	96.25	1454

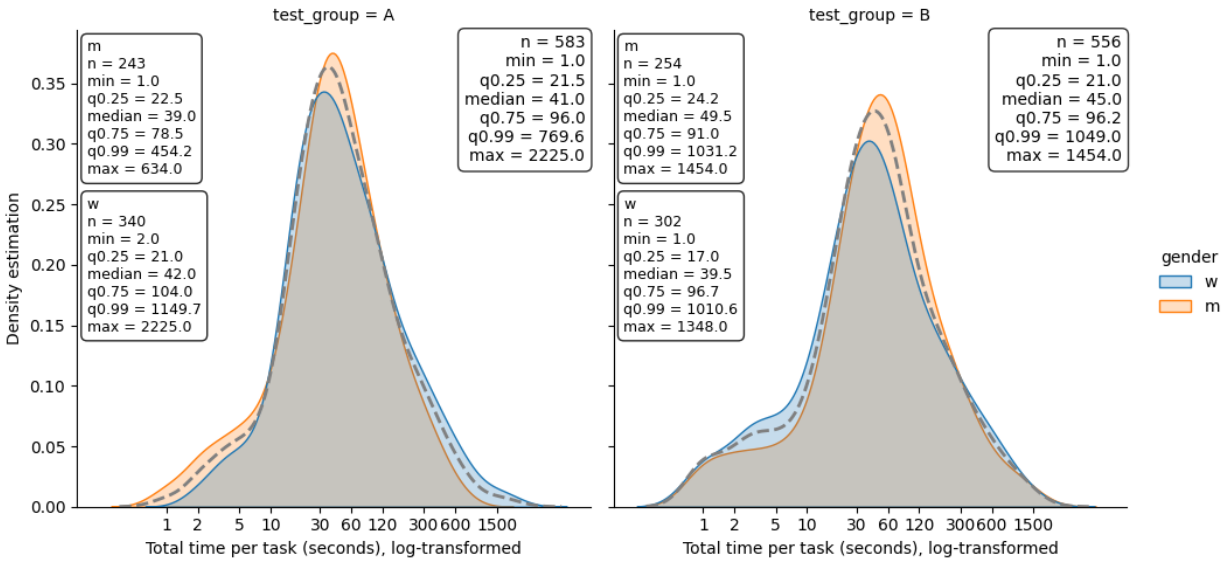

Abbildung 25 - Kerndichtediagramm nach Test-Gruppe und Gender (ohne right-censored)

9.2.3 Test-Gruppe und Dienstalter

Zum Vergleich nach Dienstalter wurden drei Dienstalter-Intervalle gebildet: 0 bis 5 Jahre, ab 5 bis 20 Jahre, 20 und mehr Jahre Diensterfahrung. Dienstalter-Gruppen zeigen einen ähnlichen Einfluss auf die Zeitmessungen in den beiden Test-Gruppen (Tabelle 20 - Deskriptive Kennzahlen in Test-Gruppen nach Dienstalter, alle Zeitmessungen Tabelle 21). Die zweite Dienstalter-Gruppe (Dienstalter 5 bis 20 Jahre) hat die niedrigsten Medianwerte im Dienstalter-Vergleich (A: 35s, B: 41s sowohl über alle Zeitmessungen hinweg als auch ohne ungelöste), das erste und das dritte Quartil liegen ebenfalls unter den Werten der dienstjüngsten (0 bis 5 Jahre) und der dienstältesten Gruppen (20 und mehr Jahre).

Die dienstjüngsten Teilnehmerinnen und Teilnehmer (0-5 Jahre) haben Medianwerte von 41s bzw. 39s ohne right-censored in der Gruppe A und 47s bzw. 44s ohne right-censored in der Gruppe B. Die höchsten Parameterzahlen zeigen die dienstältesten Teilnehmerinnen und Teilnehmer in beiden Test-Gruppen, hier liegt etwa der Medianwert bei 52,5s bzw. 50s in der Gruppe A, bei 56s bzw. 52s in der Gruppe B.

Im Vorfeld der Studie wurden folgende mögliche Dienstalter- und Gruppen-Zusammenhänge angenommen:

- Die Personen mit weniger Diensterfahrung würden besonders von der KI-gestützten Suche profitieren;

- Die erfahrensten Teilnehmerinnen und Teilnehmer würden durch ihre Erfahrung im Umgang mit dem BMEIA-Intranet einen Vorteil haben.

Auf Basis der explorativen Betrachtung können die beiden Annahmen nicht bestätigt werden.

Tabelle 20 - Deskriptive Kennzahlen in Test-Gruppen nach Dienstalter, alle Zeitmessungen

Test-Gruppe	Dienst-alter (Jahre)	count	mean	std	min	25%	50%	75%	max
A	0-5	217	90.34	183.15	1	23	41	86	1692
	5-20	115	64.03	96.65	1	18	35	64.5	759
	>=20	290	122.88	204.25	2	25	52.5	138.75	2225
B	0-5	232	117.67	209.68	1	22.75	47	117.5	1348
	5-20	195	83.92	140.51	1	20	41	71	1096
	>=20	175	120.43	186.15	1	25.5	56	128	1454

Tabelle 21- Deskriptive Kennzahlen in Test-Gruppen nach Dienstalter, ohne right-censored

Test-Gruppe	Dienst-alter (Jahre)	count	mean	std	min	25%	50%	75%	max
A	0-5	204	85.98	181.34	1	21.75	39	79.25	1692
	5-20	111	63.68	97.39	1	17.5	35	64.5	759
	>=20	268	119.54	207.96	2	25	50	125.25	2225
B	0-5	215	116.97	215.49	1	21	44	88	1348
	5-20	175	82.94	141.92	1	19	41	72	1096
	>=20	166	118.49	188.65	1	24.25	52	125.5	1454

9.2.4 Test-Gruppe, Dienstalter und Gender

Die Untergruppen aufgeteilt nach Test-Gruppe, Dienstalter und Gender bestehen aus wenigen Studienteilnehmerinnen und -Teilnehmern (min. 2, max. 7) und haben daher geringe Aussagekraft. Tabelle 22, Tabelle 23 und Abbildung 26-Abbildung 28 bieten einen Überblick über die Zeitmessungen in den Untergruppen.

Tabelle 22 - Deskriptive Kennzahlen in Test-Gruppen nach Dienstalter und Gender, alle Zeitmessungen

Test-Gruppe - Gender	Anzahl Personen	Dienstalter	count	mean	std	min	25 %	median	75 %	max
A - m	4	0-5	74	77.16	124.04	1	20.75	40	75.5	748
	3	5-20	67	59.64	82.69	1	8	27	63	381
	6	>=20	126	97.05	112.77	3	29.25	53	115.5	634
A - w	6	0-5	143	97.15	207.31	2	23	41	89	1692
	2	5-20	48	70.17	113.98	4	21	39.5	64.25	759
	7	>=20	164	142.72	251.6	2	20	52.5	154.75	2225
B - m	4	0-5	93	99.32	172.1	1	22	46	87	1130
	4	5-20	88	80.3	146.11	2	22.75	42.5	65.25	1096
	5	>=20	100	136.42	214.55	1	29	71	165.75	1454
B - w	6	0-5	139	129.95	231.24	1	23	48	139.5	1348

	5	5-20	107	86.9	136.35	1	14.5	36	93.5	715
	4	>=20	75	99.12	138.17	2	21	43	94	747

Tabelle 23 - Deskriptive Kennzahlen in Test-Gruppen nach Dienstalter und Gender, ohne right-censored

Test-Gruppe - Gender	**Anzahl Personen**	**Dienstalter**	**count**	**mean**	**std**	**min**	**25 %**	**median**	**75 %**	**max**
A - m	4	0-5	71	67.73	97.57	1	20.00	38.0	67.50	595
	3	5-20	63	58.73	83.19	1	7.50	27.0	63.00	381
	6	>=20	109	82.70	97.09	3	27.00	48.0	103.00	634
A - w	6	0-5	133	95.71	212.71	2	22.00	40.0	87.00	1692
	2	5-20	48	70.17	113.98	4	21.00	39.5	64.25	759
	7	>=20	159	144.79	255.06	2	20.00	52.0	157.50	2225
B - m	4	0-5	83	96.46	175.05	1	20.50	45.0	82.50	1130
	4	5-20	73	80.08	151.00	2	23.00	44.0	65.00	1096
	5	>=20	98	135.87	216.31	1	29.00	71.0	164.50	1454
B - w	6	0-5	132	129.87	237.15	1	21.00	43.5	117.00	1348

	5	5-20	102	84.98	135.78	1	13.25	35.5	87.50	715
	4	>=20	68	93.44	137.19	2	17.00	38.0	92.00	747

In der dienstjüngsten Gruppe überlappen sich die Gender-Verteilungen jeweils stark (Abbildung 26).

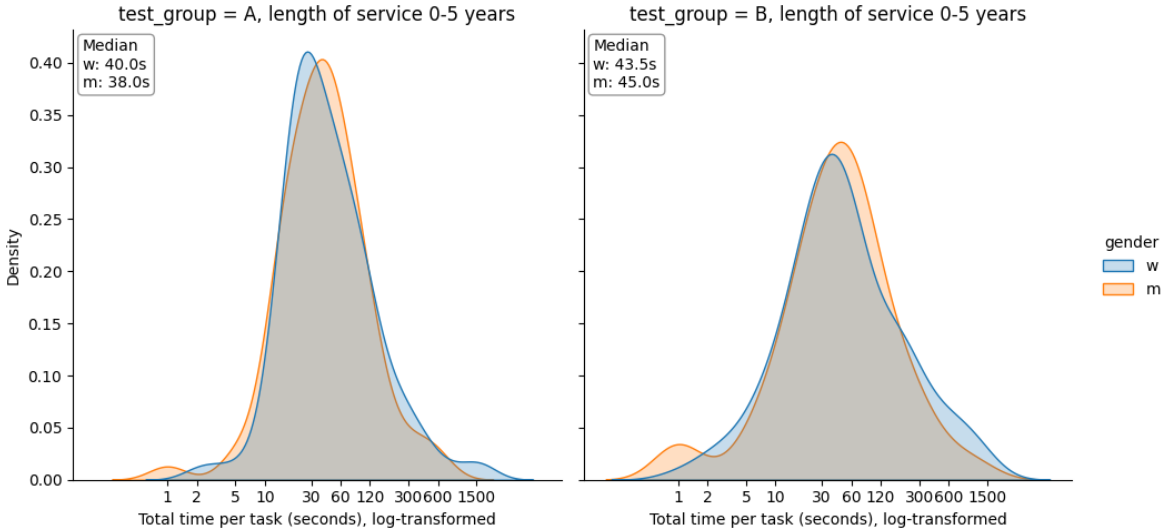

Abbildung 26 - Kerndichtediagramm nach Test-Gruppe und Gender, Dienstalter 0-5 (ohne right-censored)

In der zweiten Dienstalter-Gruppe sind Unterschiede sowohl innerhalb der Test-Gruppen nach Gender als auch zwischen den Test-Gruppen bemerkbar (Abbildung 27): Während in der Gruppe A bei Männern höhere Dichte im unteren Zeitbereich beobachtet wird, ist es in der Gruppe B bei Frauen der Fall. Der Medianwert der Frauen in der Gruppe A ist mit 39,5s um 12,5s höher als jener der Männer (27s), in der Gruppe B liegt der Medianwert der Frauen mit 35,5s um 8,5s unterhalb dem Median der Männer (44s).

In der dienstältesten Gruppe (20 Jahre und mehr, Abbildung 28) können ebenfalls Unterschiede zwischen den Gender-Gruppen erkannt werden. In der Gruppe A liegt der Median-

wert der Frauen (48s) um 4s unter dem der Männer (52s). In der Gruppe B ist der Medianwert der Frauen mit 38s um 33s niedriger als jener der Männer (71s), das Kerndichtediagramm zeigt im Vergleich zu Männern eine Linksverschiebung.

Die beobachteten Unterschiede können jedoch aufgrund der begrenzten Anzahl Personen in jeweiligen Gruppen maßgeblich von einzelnen Teilnehmerinnen und Teilnehmern beeinflusst werden.

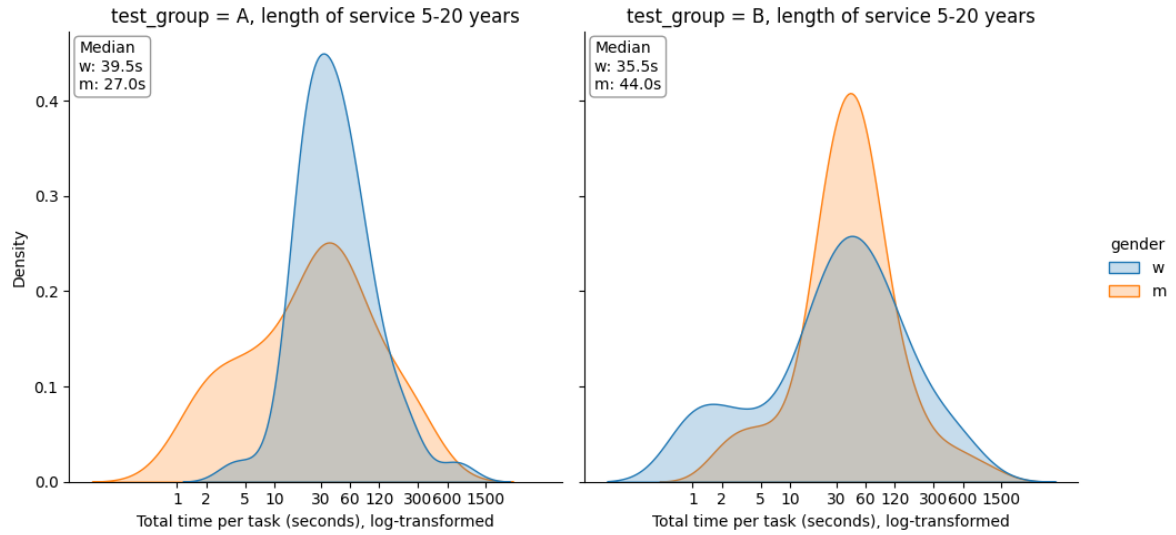

Abbildung 27 - Kerndichtediagramm nach Test-Gruppe und Gender, Dienstalter 5-20 (ohne right-censored)

Abbildung 28 - Kerndichtediagramm nach Test-Gruppe und Gender, Dienstalter 20+ (ohne right-censored)

9.2.5 Aufgabentyp

Der Aufgabentyp bestimmt, welche Funktionen eines Wissensmanagement-Systems für die erfolgreiche Lösung erforderlich sind. Während für die Dokument-Suche insbesondere eine intelligente Suche und die Relevanz und Übersichtlichkeit der Suchergebnisse wichtig sind, erfordert die Inhalts-Suche ein einfaches Herunterladen bzw. Öffnen eines Dokumentes oder im Falle des RAG-Systems eine Auswahl der relevanten Textabschnitte für die Antwort-Generation. Die Text-Erstellung-Aufgaben erfordern außerdem eine Kombination des Wissens aus unterschiedlichen Quellen (Dokument und BMEIA-Webseite) und müssen auch formelle Kriterien erfüllen.

Aufgrund der unterschiedlichen Anforderungen sind gruppenspezifische Unterschiede zu erwarten. Die Kennzahlen und Verteilungen der Zeitmessungen nach dem Aufgabentyp sind in Tabelle 24, Tabelle 25 und Abbildung 29, Abbildung 30 dargestellt. Für beide Gruppen sind die Dokument-Suche Aufgaben (beurteilt nach dem Zeitaufwand) am einfachsten, an der zweiten Stelle ist die Inhalts-Suche, und besonders anspruchsvoll sind die Text-

Erstellung-Aufgaben. Die meisten Zeitmessungen sind aufgrund der hohen Anzahl Aufgaben (15 Aufgaben[5]) für die Dokument-Suche verfügbar, die wenigsten für die Text-Erstellung (nur 2 Aufgaben).

Die Gruppe B profitiert bei den Dokument-Suche Aufgaben: Median, erstes und drittes Quartil liegen jeweils unter den Werten der Gruppe A, sowohl unter Betrachtung aller Zeitmessungen als auch ohne unvollständige Beobachtungen. Der Unterschied in den Medianwerten ist jedoch mit 2s gering.

Bei der Inhalts-Suche schneidet die Gruppe A besser ab: Median, erstes und drittes Quartil liegen deutlich unter den Werten der B-Gruppe. Bei der Text-Erstellung liegt das erste Quartil der Gruppe B unter jenem der Gruppe A, der Median und das dritte Quartil sind jedoch deutlich höher, die Werte der Gruppe B sind weiter gestreut.

Tabelle 24 - Deskriptive Kennzahlen in Test-Gruppen nach Aufgabentyp, alle Zeitmessungen

Test-Gruppe	task_type	count	mean	std	min	25%	median	75%	max
A	**document search**	392	67.34	103.29	1	20	35	69	758
	content search	183	92.73	176.26	2	28.5	56	104	2225
	generation	47	409.26	359.33	11	198.5	309	465.5	1692
B	**document search**	383	49.98	67.31	1	15	33	58.5	747
	content search	180	155.58	196.92	2	38	84.5	188.25	1454
	generation	39	451.13	370.53	5	150	376	639	1348

[5] Ohne 12.1.

Tabelle 25 - Deskriptive Kennzahlen in Test-Gruppen nach Aufgabentyp, ohne right-censored

Test-Gruppe	task_type	count	mean	std	min	25%	median	75%	max
A	**document search**	374	64.07	99.26	1	19.25	34	65	758
	content search	168	91.83	182.38	2	28	53	102.5	2225
	generation	41	420.78	374.54	11	211	309	479	1692
B	**document search**	363	47.03	63.27	1	14	32	56.5	747
	content search	158	157.56	203.22	2	37.25	87.5	192.75	1454
	generation	35	496.17	364.4	53	189	426	681.5	1348

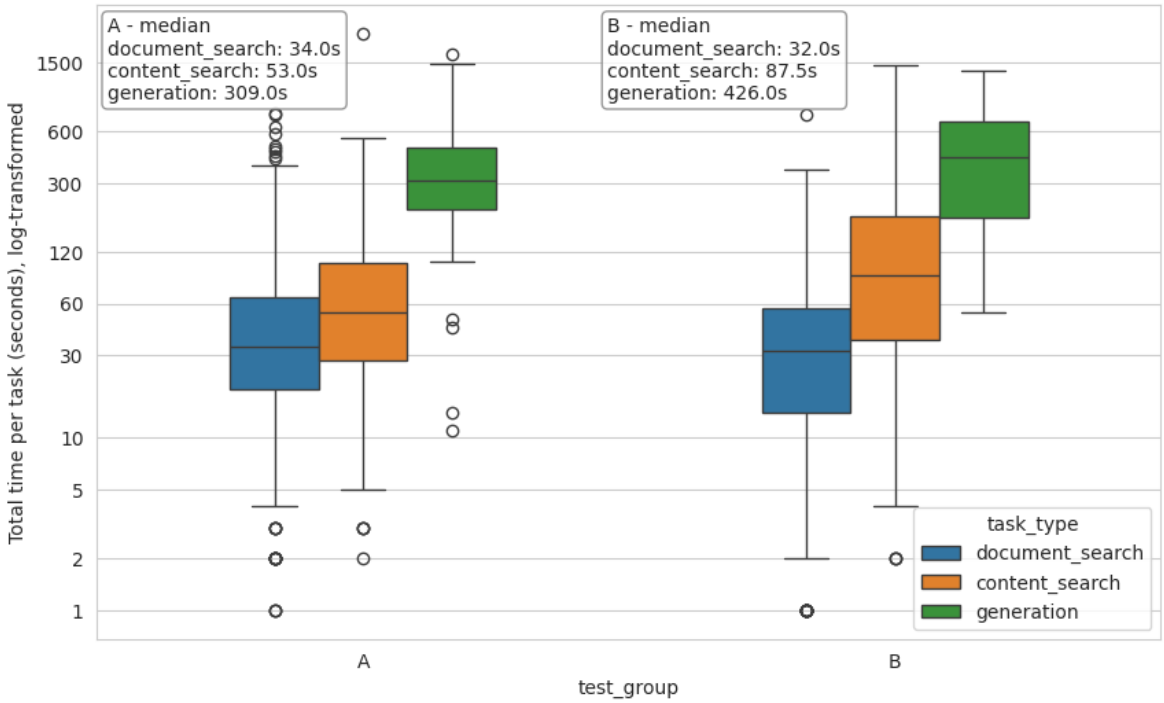

Abbildung 29 - Box-plot Zeitmessungen der Test-Gruppen nach Aufgabentyp, ohne right-censored

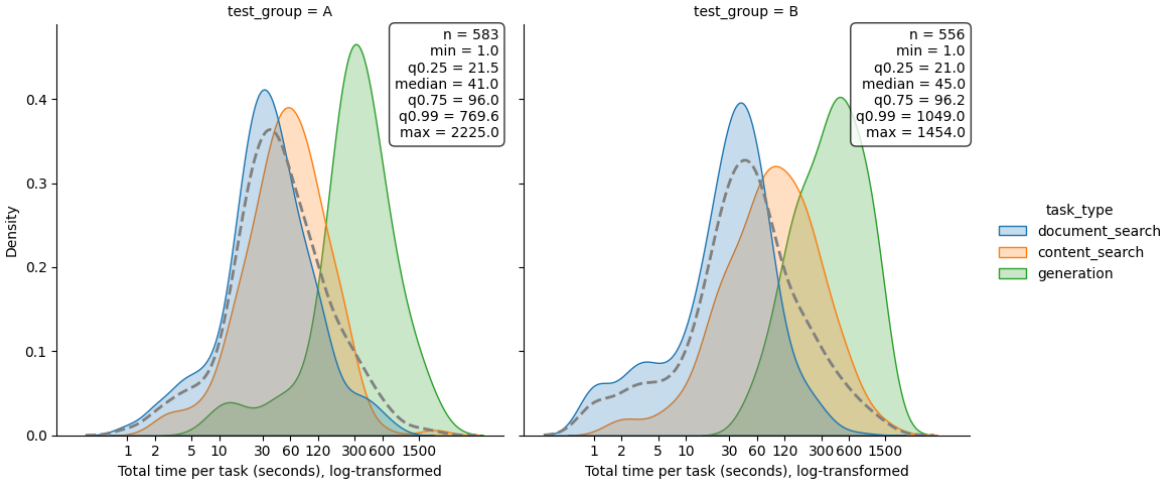

Abbildung 30 - Kerndichtediagramm nach Test-Gruppe und Aufgabentyp, ohne right-censored

Die Abbildung 31 - Abbildung 33 zeigen Gender-Verteilungen in Gruppen je nach Aufgabentyp. In der Gruppe B fallen Zeitmessungen der Frauen bei der Dokument-Suche niedriger aus als jene der Männer sowohl in derselben Gruppe als auch in der Kontrollgruppe: Der Median der Frauen (29s) liegt um 4s bzw. 5,5s unter dem Median der Männer (A: 33s, B: 34,5s), das erste Quartil (11s) um 6s bzw. 7,7s (A: 17s, B: 18,7s), das dritte Quartil (55s) um 4s bzw. 5s (A: 59s, B: 60s). Die Kennzahlen der Frauen in der Gruppe A liegen über jenen der Frauen in der Gruppe B, Männer in der Gruppe A, und im Medianwert ähnlich mit Männern der Gruppe B (Frauen-A: 34s vs. Männer-B: 34,5s). Dies kann darauf hindeuten, dass Frauen von der intelligenten Suche profitieren würden.

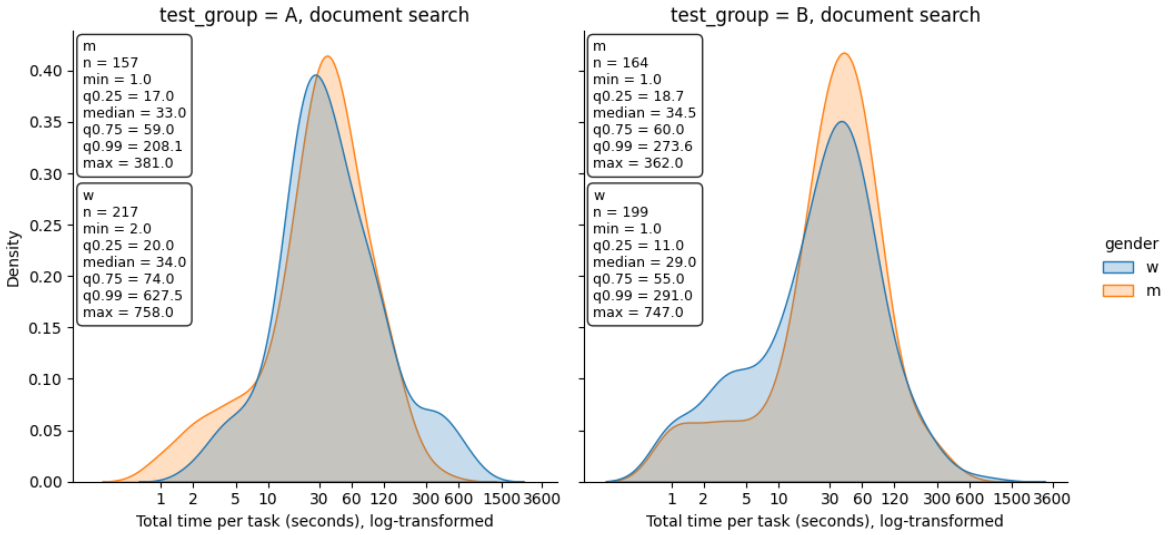

Abbildung 31 - Kerndichtediagramm Dokument-Suche nach Test-Gruppe und Gender, ohne right-censored

Niedrigere Zeitmessungen bei Frauen im Vergleich zu Männern in der Gruppe B können auch bei der Inhalts-Suche beobachtet werden (Abbildung 32). Die Kennzahlen der Gruppe B liegen hier bei beiden Gender-Gruppen deutlich über jenen der Gruppe A; ungeachtet des Geschlechts sind Inhalt-Suche Aufgaben eher schwieriger für die Gruppe B verglichen mit Kontrollgruppe.

Bei der Text-Erstellung (Abbildung 33) liegen deutlich weniger Zeitmessungen pro Gender-Gruppe vor im Vergleich zu den anderen Aufgabentypen. Hier liegt der Medianwert der Frauen in der Gruppe B (426s) über jenem der Männer (400,7s), erstes Quartil fällt deutlich höher aus (Frauen-B: 234,9s, Männer-B: 152s), drittes Quartil liegt darunter (Frauen-B:

680,7s, Männer-B: 693,3s). In der Gruppe A fällt der Median der Frauen niedriger aus (Frauen-A: 288,3s, Männer-A: 334s), das erste und das dritte Quartil jedoch höher.

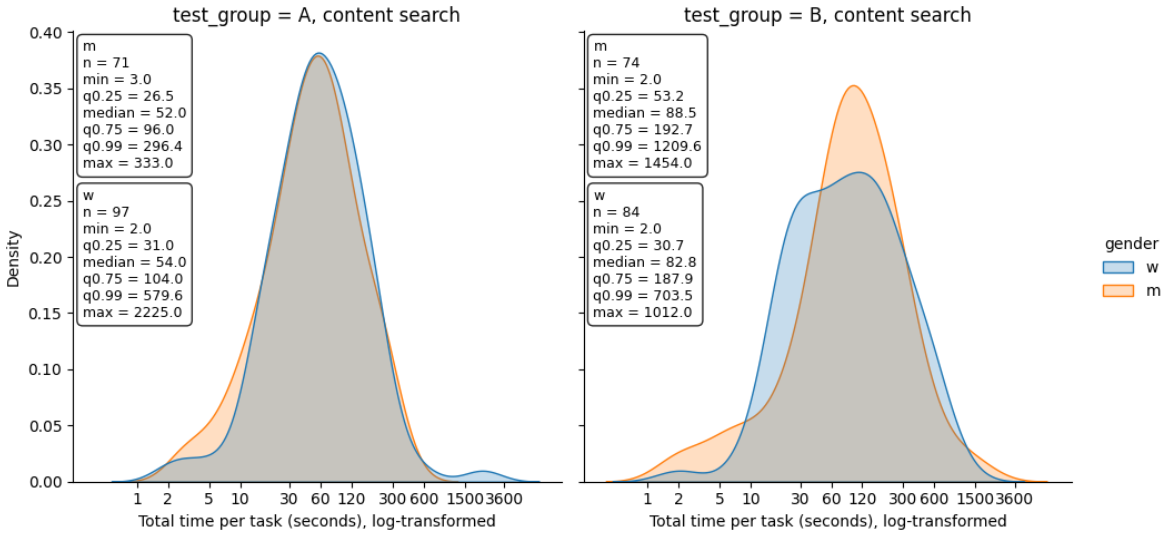

Abbildung 32 - Kerndichtediagramm Inhalts-Suche nach Test-Gruppe und Gender, ohne right-censored

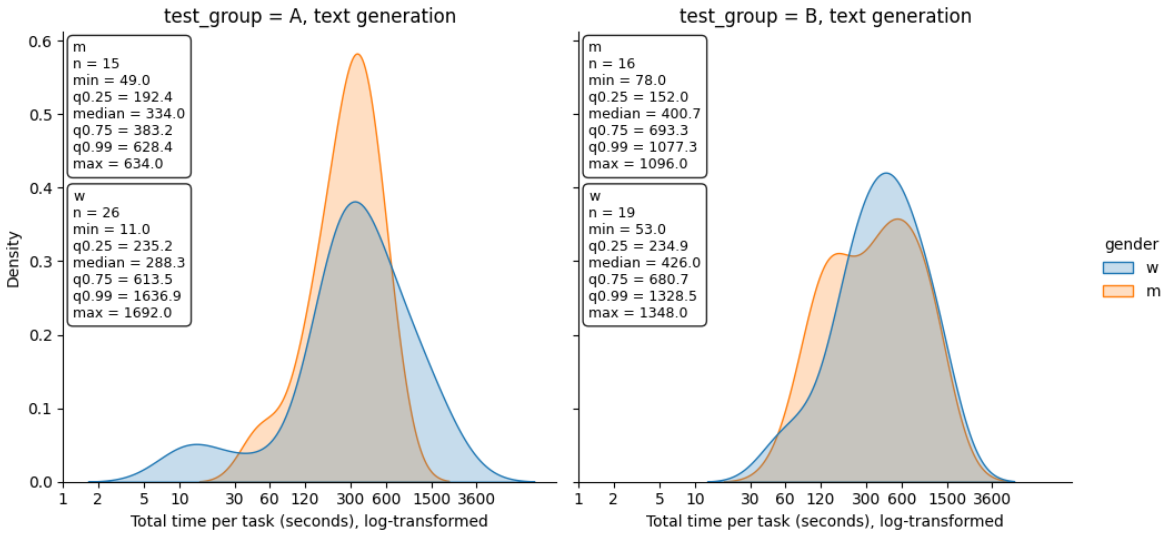

Abbildung 33 - Kerndichtediagramm Text-Generation nach Test-Gruppe und Gender, ohne right-censored

9.2.6 Erfolgsrate und Erfolg beim ersten Versuch

Ergänzend zu Zeitmessungen ist die Erfolgsrate in den Gruppen von Bedeutung: Der Anteil der Aufgaben, bei denen Userinnen und User Erfolg beim ersten Versuch erzielen ("first try success" - Erfolg beim ersten Versuch) und der Anteil der gelösten Aufgaben, ungeachtet der erforderlichen Anzahl Versuche, zu insgesamt versuchten Aufgaben (Erfolgsrate allgemein).

Tabelle 26 zeigt die Erfolg beim ersten Versuch-Rate in den Gruppen nach Gender und gesamt, ohne weitere Aggregation. Abbildung 34 zeigt die entsprechenden Erfolgsraten der Userinnen und User, berechnet als der Anteil der "first-try-success" in den insgesamt von Userin oder User versuchten Aufgaben. Die Erfolgsrate beim ersten Versuch, eine Aufgabe zu lösen, fällt in der Gruppe A höher aus. Frauen in beiden Gruppen fanden seltener korrekte Lösungen beim ersten Versuch als Männer.

Tabelle 26 - Erfolg beim ersten Versuch in den Test-Gruppen nach Gender

test_group	gender	n_attemp-ted	n_first_try	first_try_rate
A	**m**	267	202	0.76
	w	355	242	0.68
	total	622	444	0.71
B	**m**	281	198	0.70
	w	321	205	0.64
	total	602	403	0.67

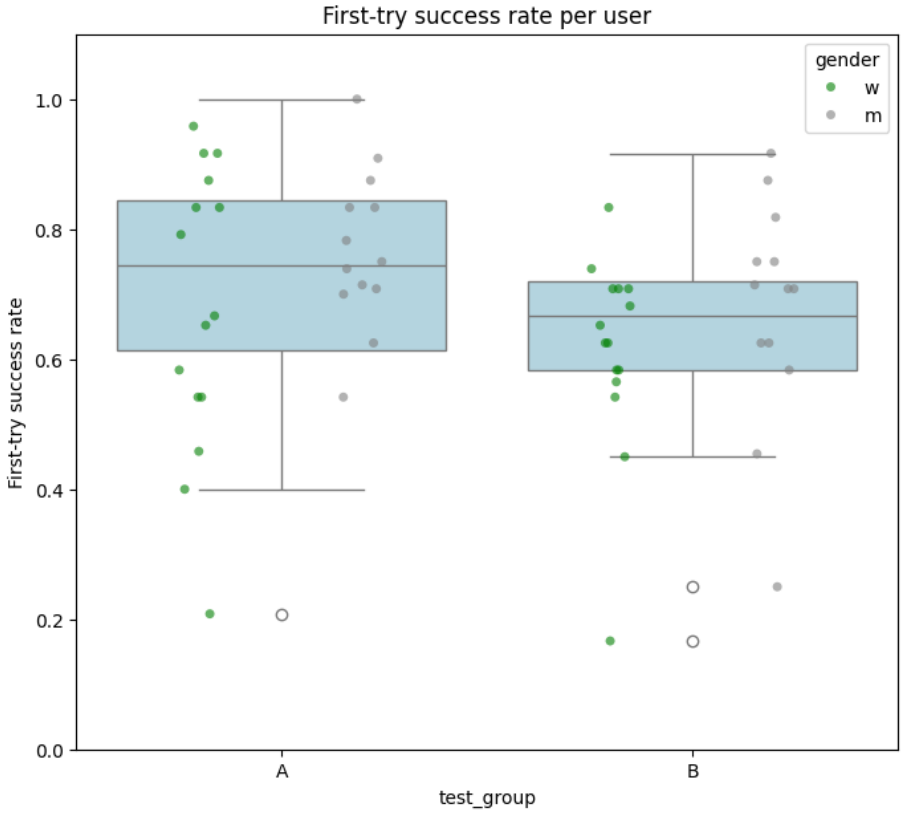

Abbildung 34 - Box-Plot: Erfolg beim ersten Versuch in Gruppen, aggregiert nach User

Tabelle 27 zeigt die allgemeine Erfolgsraten (ungeachtet der Anzahl erforderlicher Versuche) berechnet nach Gruppe und Gender, Abbildung 35 zeigt die Erfolgsraten der Userinnen und User. Die allgemeine Erfolgsrate in der Gruppe B ist etwas niedriger; in beiden

Gruppen ist sie höher bei Frauen. Die Erfolgsraten der Userinnen und User in Gruppen sich ähnlich verteilt.

Tabelle 27 - Allgemeine Erfolgsrate in Test-Gruppen nach Gender

test_group	gender	n_attemp-ted	n_success	success_rate
A	**m**	267	243	0.91
	w	355	340	0.96
	total	622	583	0.94
B	**m**	281	254	0.90
	w	321	302	0.94
	total	602	556	0.92

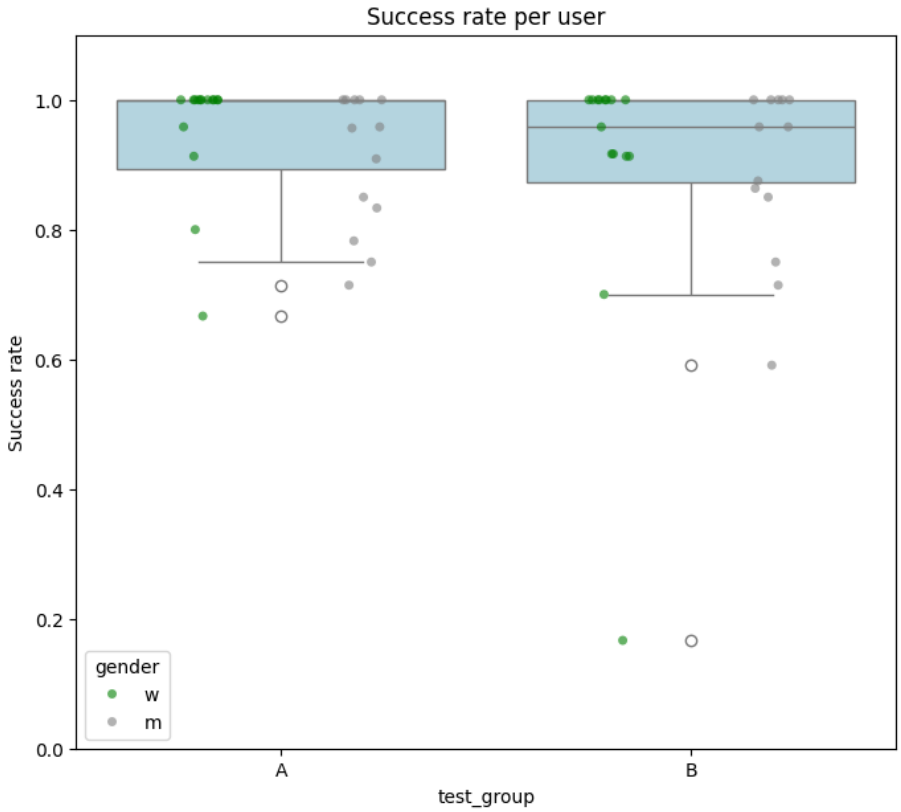

Abbildung 35 - Box-Plot: Erfolgsraten der Userinnen und User in den Test-Gruppen

9.2.7 Veränderungen im Zeitverlauf (Lerneffekte)

Aufgrund der steigenden Erfahrung im Umgang mit dem neuen Tool wurden mögliche Lerneffekte mit Zeitverlauf in der Gruppe B vermutet: Userinnen und User würden mit mehr versuchten Aufgaben an der Erfahrung gewinnen und ggf. schneller werden. Außerdem könnte steigende Erfahrung im Umgang mit dem Evaluierungsframework beide Gruppen beeinflussen.

Teilnehmerinnen und Teilnehmer konnten die Aufgaben in beliebiger Reihenfolge versuchen zu lösen. Außerdem variieren die Aufgaben nach Schwierigkeit (benötigter Zeit) sowohl abhängig von dem Aufgabentyp als auch innerhalb des Aufgabentyps (Abbildung 36).

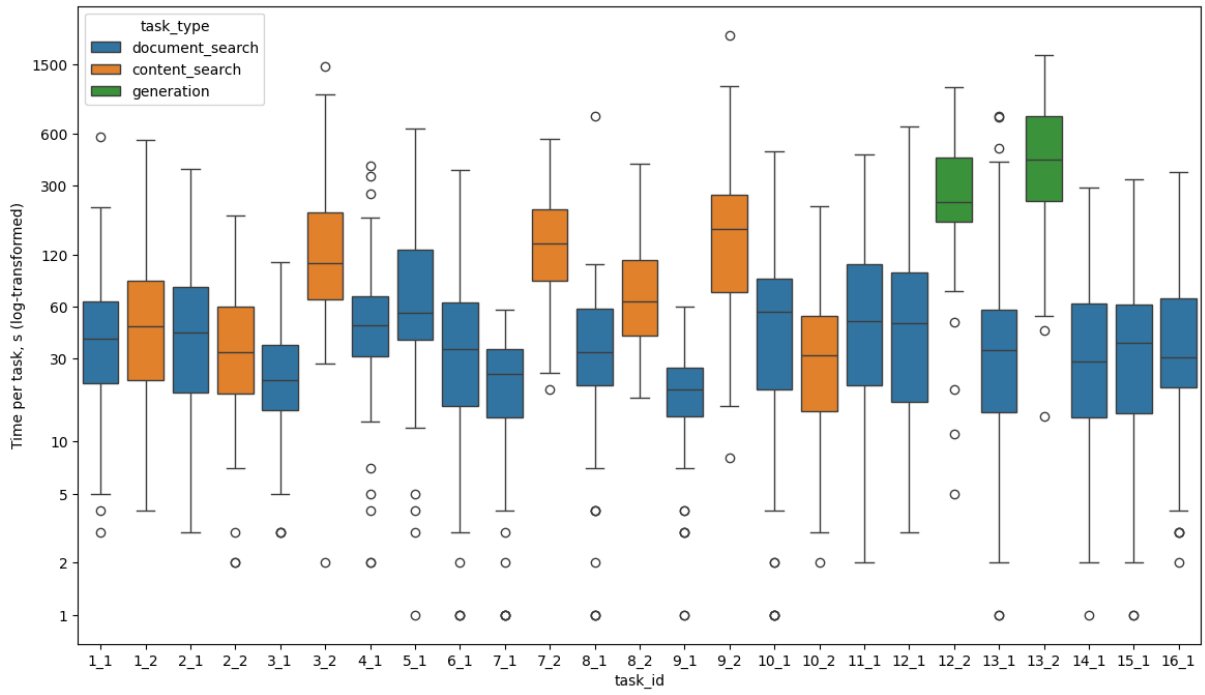

Abbildung 36 - Box-Plot: Bearbeitungszeiten nach Test-Cases

Um die Abweichungen in den Zeiten vergleichbar zu machen, wurden die log-transformierten Zeitmessungen für jede Aufgabe Test-Gruppen-übergreifend standardisiert (Z-transformiert, sodass für jede Aufgaben gilt: mean=0, std=1).

Userinnen und User konnten Aufgaben im Rahmen der Studie zu jedem Zeitpunkt bearbeiten. Um eine gemeinsame Zeitachse für alle Teilnehmerinnen und Teilnehmer herzustellen, wurden die Aufgaben für jede/n Userin oder User chronologisch nach dem Zeitstempel des ersten Versuchs gereiht und ihnen die Reihenfolge des Versuchs (1-25) zugewiesen. Die Beobachtungen mit derselben Reihenfolge-Nummer innerhalb der Gruppen wurden aggregiert (Mittelwert).

Abbildung 37 zeigt die Abweichungen von der normalisierten log-transformierten durchschnittlichen Aufgaben-Bearbeitungszeit (0) nach der Reihenfolge der Versuche. Negative Werte sind als schneller als Durchschnitt zu interpretieren, positive Werte als langsamer als Durchschnitt. Abbildung 38 zeigt die Abweichungen im gleitenden Zeitfenster (n=5). Ein möglicher Lerneffekt würde sich als ein Abwärtstrend manifestieren, dieser ist jedoch nicht erkennbar. Lerneffekte können jedoch auch nicht ausgeschlossen werden, diese sind ggf. schwächer als der Einfluss durch den Aufgaben-Mix und andere Effekte.

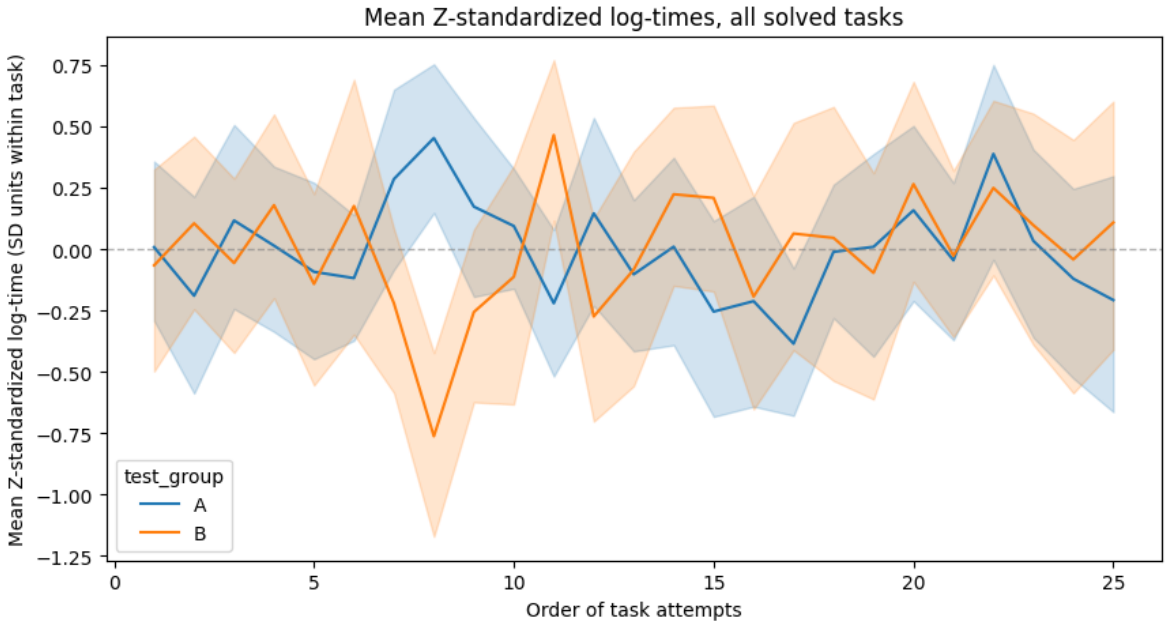

Abbildung 37 - Mittlere Abweichungen der Z-standardisierten log-transformierten Aufgabenzeit in Test-Gruppen nach Reihenfolge der Aufgaben-Bearbeitung, mit 95% CI

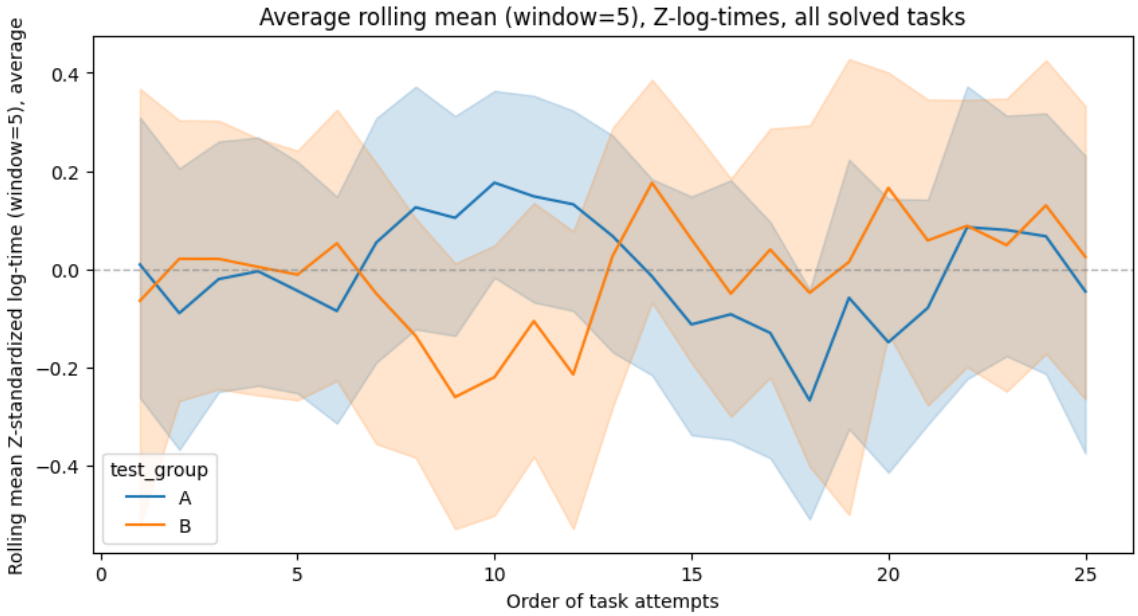

Abbildung 38 - Mittlere Abweichungen der Z-standardisierten log-transformierten Aufgabenzeit in Test-Gruppen nach Reihenfolge der Aufgaben-Bearbeitung, in rolling window (5), mit 95% CI

Abbildung 39 zeigt die Gruppen-Trends auf Dokument-Suche Aufgaben, die Abbildung 40 zeigt sie im gleitenden Fenster. Während hier ebenfalls kein anhaltender Abwärtstrend erkannt werden kann, kann bei der Gruppe B eine Lernphase am Anfang vermutet werden. Die Abweichungen der Gruppe B liegen außerdem vorwiegend im negativen Bereich, was mit der vorherigen Beobachtung der niedrigeren Zeitmessungen bei der Dokument-Suche übereinstimmt.

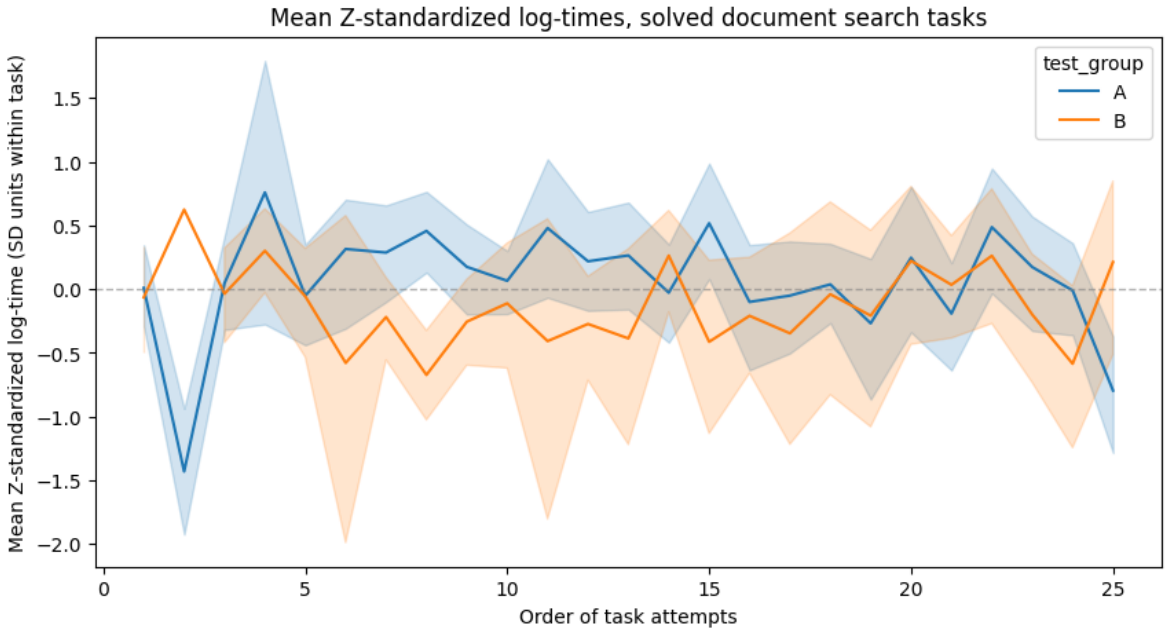

Abbildung 39 - Mittlere Abweichungen der Z-standardisierten log-transformierten Dokument-Suche Aufgabenzeit in Test-Gruppen nach Reihenfolge der Aufgaben-Bearbeitung, mit 95% CI

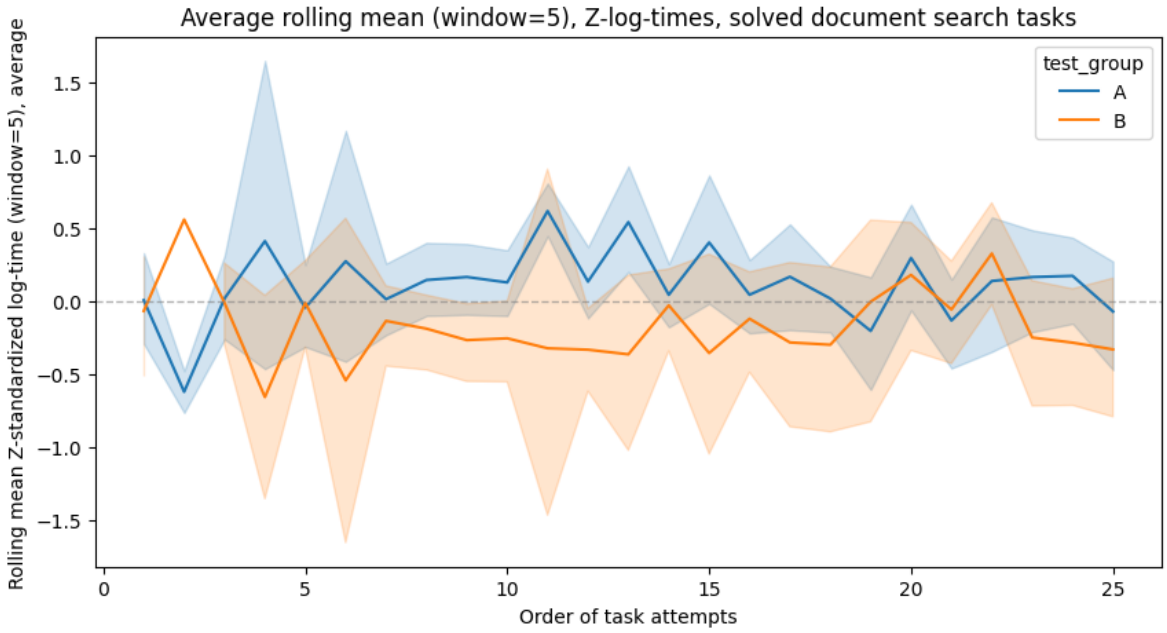

Abbildung 40 - Mittlere Abweichungen der Z-standardisierten log-transformierten Dokument-Suche Aufgabenzeit in Test-Gruppen nach Reihenfolge der Aufgaben-Bearbeitung, in rolling window (5), mit 95% CI

9.3 Inferenzstatistik

Zur Evaluierung der möglichen systematischen Unterschiede und deren statistischen Signifikanz in den Test-Gruppen wurden Mann-Whitney U-Test (Rangsummentest) und Welch T-Test (auf log-transformierten Daten) durchgeführt. Die aggregierten Zeitmessungen pro Userin oder User und Aufgabe bilden User-Cluster und erfüllen daher die Voraussetzung der Unabhängigkeit der Datenpunkte für die Tests nicht. Es werden Medianwerte pro Userin oder User (je 28 Datenpunkte pro Gruppe) berechnet (Tabelle 28, Abbildung 41), die unabhängig sind.

Die unvollständigen right-censored Beobachtungen konnten in den Tests nicht korrekt behandelt und mussten daher ausgeschlossen werden. Das Miteinbeziehen solcher Datenpunkte würde zur impliziten Annahme führen, dass Userinnen und User in der angegebenen Zeit die Aufgaben gelöst hätten, dies würde zu Verzerrungen führen.

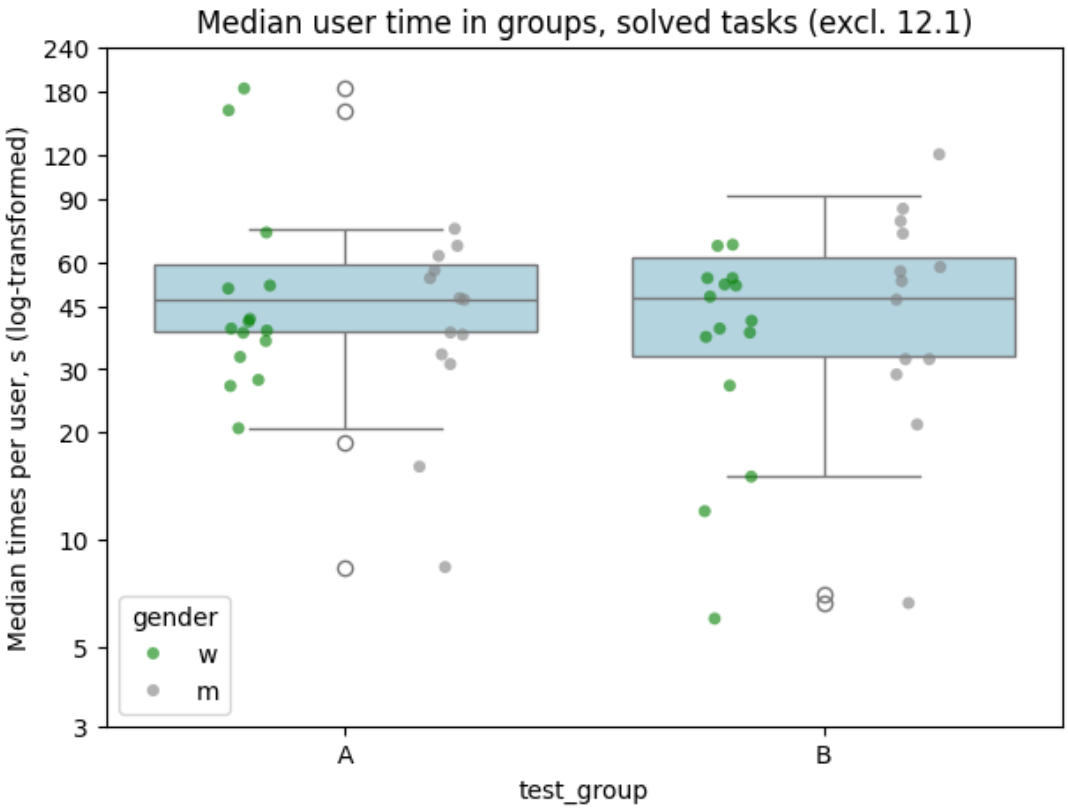

Abbildung 41 - Box-Plot: Medianzeiten der Teilnehmerinnen und Teilnehmer in Test-Gruppen

Tabelle 28 - Deskriptive Kennzahlen: Medianzeiten der Teilnehmerinnen und Teilnehmer

Test-Gruppe	gender	count	mean	std	min	25%	median	75%	max
A	**m**	13	44.38	19.87	8.5	33.00	47.0	57.0	78.5
	w	15	57.37	48.34	20.5	34.25	39.0	51.0	184.5
	total	28	51.34	37.82	8.5	32.88	40.0	54.75	184.5
B	**m**	13	53.12	30.40	7.5	32.00	53.0	72.0	120.0
	w	15	40.57	18.74	6.0	32.00	41.0	53.0	67.0

	total	28	46.39	25.17	6.0	31.25	47.75	56.88	120.0

9.3.1 U-Test (Rangsummen-Test)

Nullhypothese: Unter der Voraussetzung des Erfolgs (Aufgabe gelöst) ist es gleich wahrscheinlich, dass ein zufällig aus der einen Population ausgewählter Wert größer oder kleiner ist als ein zufällig ausgewählter Wert aus der anderen Population.

Alternativhypothese: Unter der Voraussetzung des Erfolgs ist es nicht gleich wahrscheinlich, dass ein zufällig aus der einen Population ausgewählter Wert größer oder kleiner ist als ein zufällig ausgewählter Wert aus der anderen Population. Die Medianzeiten der Teilnehmerinnen und Teilnehmer sind unterschiedlich verteilt.

Es wurde beidseitiger Test durchgeführt (Tabelle 29). Die Nullhypothese kann nicht verworfen werden (p-value > 0,05). Common language effect size (CLES) beträgt 0,49, was eine Wahrscheinlichkeit von 49% bedeutet, dass ein zufällig ausgewählter Wert aus der Stichprobe A höher ist als ein zufällig ausgewählter Wert aus der Stichprobe B (Gruppen zeigen keinen systematischen Unterschied in eine Richtung). Biseriale Rangkorrelation r_{rb} gibt die Effektstärke an und liegt bei etwa 0 (vernachlässigbarer Effekt), sie zeigt außerdem die Richtung, Gruppe A ist vernachlässigbar schneller.

Dem Rangsummentest zufolge besteht kein statistisch signifikanter Unterschied in den Medianzeiten der Userinnen und User in den Test-Gruppen.

Tabelle 29 - Ergebnisse U-Test

Mann-Whitney U-statistic	**p-value**	**CLES**	**r_{rb}**
384.00	0.90	0.49	-0.02

9.3.2 Welch T-Test

Welch T-Test vergleicht Mittelwerte zwei unabhängigen Stichproben ohne Voraussetzung der gleichen Varianzen und wurde auf log-transformierten Daten durchgeführt (beidseitiger Test). Der arithmetische Mittelwert der log-transformierten Daten entspricht nach dem Exponenzieren dem geometrischen Mittel der nicht-transformierten Zeiten.

Nullhypothese: Unter der Voraussetzung des Erfolgs sind die Mittelwerte der log-transformierten Zeiten der zwei unabhängigen Stichproben A, B gleich.

Alternativhypothese: Unter der Voraussetzung des Erfolgs sind die Mittelwerte der log-transformierten Medianzeiten der zwei unabhängigen Stichproben A, B unterschiedlich.

Die Nullhypothese kann nicht verworfen werden, s. Tabelle 30 ($p > 0,05$). Die Effektstärke ist gering (<0.2). Der Mittelwert der log-transformierten Medianzeiten in der Gruppe A ist vernachlässigbar höher als in der Gruppe B. Es besteht kein statistisch signifikanter Unterschied in den Mittelwerten der Gruppen.

Tabelle 30 - Welch T-Test Ergebnisse

Welch T-Test-statistic	p-value	Cohen's d	Hedge's g
0.59	0.55	0.16	0.16

Die Ergebnisse des Welch T-Tests können als eine Kontrolle und Bestätigung der Ergebnisse des U-Rangsummentests gesehen werden. Eine Korrektur der p-Werte ist entfallen, da p-Werte weit oberhalb der Signifikanzgrenze ausfallen.

9.3.3 Auswertung ab dem Aktualisierungsprozess-Abschluss des KnowHow-Tools

Wie in 9.1.1 unter "Besondere Ereignisse" erwähnt, wurde am 21.01.2026 der im Laufe der Studie noch anhaltende Aktualisierungsprozess abgeschlossen. Die Updates hatten u.a. das Ziel, bekannte Probleme zu beheben und User Experience (UX) zu verbessern. Infolge dieser Updates hätten Unterschiede zwischen den zwei Systemen bemerkbar wer-

den können. Stichprobengrößen ab 22.01.2026: 41 Teilnehmerinnen und Teilnehmer, davon 19 aus der Gruppe A, 22 aus der Gruppe B, insgesamt 733 aggregierte Aufgabenzeiten, davon 661 gelöst und 72 ungelöst.

Tabelle 31, Abbildung 42 bieten einen Überblick über den verfügbaren Datenbestand und die Verteilung der Zeitmessungen, aggregiert als Teilnehmer-Median, ab dem 22.01.2026 nach Test-Gruppen und Gender.

Die zu testenden Hypothesen sind durch die tatsächlichen Veränderungen im Datenbestand bestimmt und unterscheiden sich daher von den in 9.3.1, 9.3.2 formulierten Hypothesen.

Die zu testendenden Hypothesen können wie folgt formuliert werden:

U-Test:

Nullhypothese: Nach den vorgenommenen Verbesserungen des KnowHow-Tools, unter der Voraussetzung des Erfolgs ist es gleich wahrscheinlich, dass ein zufällig aus der einen Population ausgewählter Wert größer oder kleiner ist als ein zufällig ausgewählter Wert aus der anderen Population.

Alternativhypothese: Nach den vorgenommenen Verbesserungen des KnowHow-Tools, unter der Voraussetzung des erzielten Erfolgs sind Median-Zeiten der Userinnen und User aus Gruppen A und B unterschiedlich verteilt – die Werte einer der Gruppen tendieren dazu, systematisch größer oder kleiner zu sein als Werte der anderen Gruppe.

Welch T-Test:

Nullhypothese: Nach den vorgenommenen Verbesserungen des KnowHow-Tools, unter der Voraussetzung des erzielten Erfolgs sind Mittelwerte der Log-transformierten Median-Zeiten der Userinnen und User in Gruppen A und B gleich.

Alternativhypothese: Nach den vorgenommenen Verbesserungen des KnowHow-Tools, unter der Voraussetzung des erzielten Erfolgs, sind Mittelwerte der Log-transformierten Median-Zeiten der Userinnen und User in Gruppen A und B unterschiedlich.

Tabelle 31 - Deskriptive Kennzahlen: Medianzeiten der Teilnehmerinnen und Teilnehmer ab 22.01.2026

Test-Gruppe	gender	count	mean	std	min	25%	median	75%	max
A	**m**	11	45.18	21.44	8.5	34.25	47.50	59.75	78.5
	w	8	120.00	155.67	28.0	36.62	56.75	105.75	484.0
	total	19	76.68	105.45	8.5	35.00	47.50	69.50	484.0
B	**m**	10	68.15	58.46	7.5	29.75	57.50	73.50	206.5
	w	12	70.75	86.90	12.0	36.12	46.25	66.12	339.0
	total	22	69.57	73.64	7.5	29.75	51.75	66.38	339.0

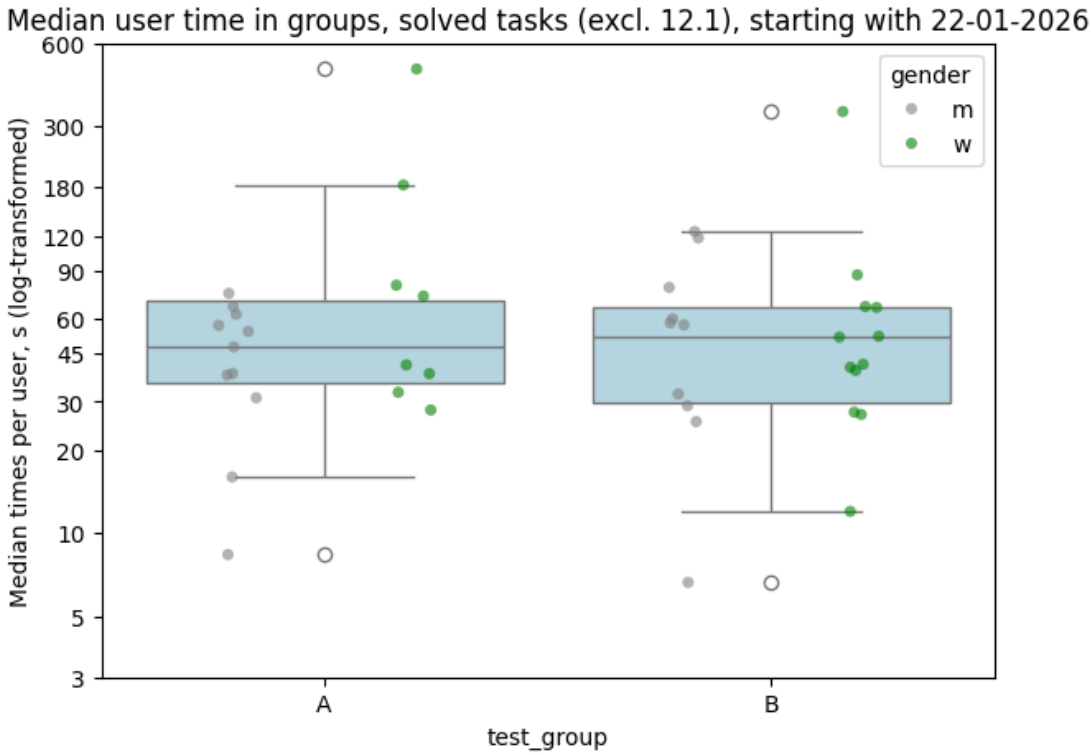

Abbildung 42 - Box-Plot: Medianzeiten der Teilnehmerinnen und Teilnehmer in Test-Gruppen ab 22.01.2026

Sowohl der U-Test als auch der Welch T-Test zeigen keine signifikanten Unterschiede in den Gruppen (Tabelle 32, Tabelle 33) nach den vorgenommenen Aktualisierungen des KnowHow-Tools.

Tabelle 32 - Ergebnisse U-Test ab 22.01.2026

Mann-Whitney U-statistic	**p-value**	**CLES**	**r_{rb}**
211.50	0.96	0.51	0.012

Tabelle 33 - Ergebnisse Welch T-Test ab 22.01.2026

Welch T-Test-statistic	p-value	Cohen's d	Hedge's g
0.22	0.83	0.07	0.07

9.4 Regressionsmodell

Die Zeit-bis-zum-Ereignis (Zeit-bis-zum-Erfolg: Userin oder User hat Aufgabe gelöst) wurde mit einem Cox's proportional hazard-Regressionsmodell (python lifelines[6]-Implementierung) modelliert. Dieses Modell kann die right-censored Beobachtungen berücksichtigen und modelliert die Zeit-bis-zum-Ereignis in Abhängigkeit von einer oder mehreren Variablen (Kovariaten) und deren Interaktionen. Das Verfahren kann außerdem ein Clustering nach einer Variable enthalten, sodass die User-bezogene Beobachtungen-Cluster korrekt behandelt werden. Das Cox's Regressionsmodell macht keine Annahmen über die Verteilung der sogenannten Risiken (Hazards). Das Verhältnis des Risikos (in Subgruppen nach Kovariaten) wird als ein konstanter Wert modelliert. Zwar können sich die Risiken im Zeitverlauf ändern, jedoch wird angenommen, dass dieses Verhältnis konstant bleibt. Diese Annahme kann geprüft werden.

Durch eine Stratifizierung nach dem Aufgabentyp können unterschiedliche Aufgabenschwierigkeiten berücksichtigt werden, ohne dass der Aufgabentyp als Prädiktor im Modell fungiert. In diesem Fall wird für jeden Wert der Stratifizierungsvariable ein eigenes Baseline-Risiko angenommen. Die Stratifizierung nach dem Aufgabentyp erlaubt einen Vergleich der Effekte über alle Aufgabentypen hinweg, allerdings wird angenommen, dass der Effekt gleich für alle Aufgabentypen ist. Miteinbeziehen des Aufgabentyps als Kovariate erlaubt es hingegen, den Effekt in Gruppen abhängig von dem Aufgabentyp zu modellieren.

[6] https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html#cox-s-proportional-hazard-model

Dienstalter zeigte in der explorativen Analyse einen nicht-linearen Effekt: Die Bearbeitungszeit sinkt zunächst mit dem zunehmenden Dienstalter und steigt später wieder an. Dieser Effekt kann mit einem Basis-Spline modelliert werden.

9.4.1 Aufgabentyp als Strata

Stratifizierung nach dem Aufgabentyp modelliert eine eigene Risikofunktion für jeden Aufgabentyp und erlaubt einen Vergleich der Effekte anderer Variablen über alle Aufgabentypen hinweg. Benutzer-ID wurde als Clustering-Variable verwendet.

Test-Gruppe (test_group), Dienstalter (length of service, los) bzw. Basis-Spline (bs(los)) mit drei Koeffizienten, Geschlecht (gender) und Interaktionen wurden im Regressionsmodell als Kovariaten verwendet. Nach dem Informationskriterium (AIC) wurde das AIC-optimale Modell ausgewählt. Das volle Modell (Maximalmodell) inkludiert alle potenziellen Prädiktoren und Interaktionen.

Das AIC-optimale Modell kann mit einer folgenden Formel beschrieben werden:
time_to_event ~ test_group + bs(los, df=3).

Tabelle 34 zeigt zentrale Modellgütemaße zur Bewertung der Anpassung des AIC-optimalen Regressionsmodells. Die Konkordanz beträgt 0,5487, was einem schwachen Modell entspricht (leicht über dem Zufallsniveau von 0,5). Der Likelihood Ratio Test ist statistisch signifikant ($p < 0,005$), was darauf hindeutet, dass das Modell insgesamt eine signifikante Verbesserung gegenüber einem Nullmodell bietet. Die Proportional-Hazards-Annahme wurde nicht verletzt.

Tabelle 34 - Modellgütemaße (AIC-optimales Modell mit Strata)

Concordance	0.5487
AIC	12100.45
log-likelihood ratio test	33.16 on 4 df, p<0.005

Tabelle 35 enthält Regressionsergebnisse für das AIC-optimale Modell. Die Referenz-Test-Gruppe ist A. Das exponenzierte Hazard Ratio für die Test-Gruppe B beträgt 0,88, sodass die Gruppe B im Verhältnis zu Gruppe A langsamer geschätzt wird; dieser Effekt ist jedoch nicht signifikant (p=0,21). Einer der Dienstalter-Basis-Spline Koeffizienten ist signifikant (p=0,02), dieser kann nicht direkt interpretiert werden.

Tabelle 35 - Koeffizientenschätzungen des AIC-optimalen Modells mit Strata

covariate	coef	exp (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	p
test_group[B]	-0.13	0.88	-0.34	0.08	0.71	1.08	0.21
bs(los)[1]	1.36	3.90	0.22	2.50	1.25	12.21	**0.02**
bs(los)[2]	-0.75	0.47	-1.65	0.14	0.19	1.15	0.10
bs(los)[3]	0.01	1.01	-0.58	0.60	0.56	1.82	0.97

Die Abbildung 43 zeigt den modellierten Dienstalter-Effekt als Hazard Ratio im Verhältnis zum Median-Dienstalter (16,33 Jahre, Hazard=1.0). Werte über 1.0 bedeuten höheres Risiko (schnellere Lösung) im Vergleich zum Median-Dienstalter, unter 1.0 niedrigeres Risiko (langsamere Lösung). Die Spline-Funktion zeigt den nicht-linearen Zusammenhang. Dieser spiegelt die früheren Beobachtungen wider: Die jüngeren und die älteren Dienstalter-Gruppen sind langsamer (im Verhältnis zum Median-Dienstalter), die mittleren hingegen schneller.

Beim Anstieg im oberen Wertebereich (Dienstalter über 30 Jahre) handelt es sich um einen Randeffekt aufgrund begrenzter Daten in diesem Bereich.

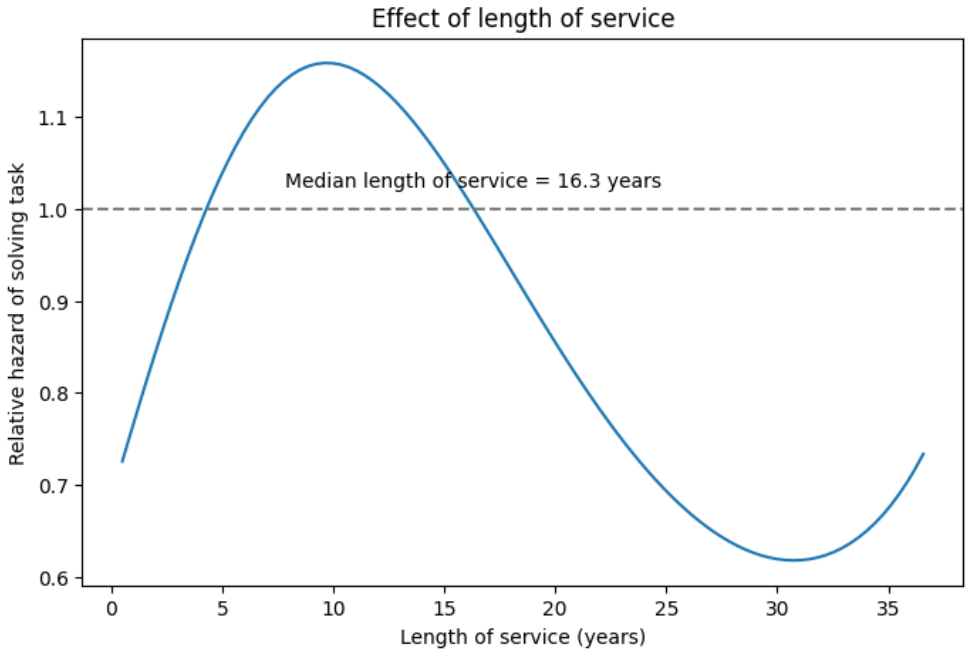

Abbildung 43 - Dienstalter-Effekt: Relatives Hazard Ratio der Teilnehmerinnen und Teilnehmer im Vergleich zum Median-Dienstalter (16,33 Jahre)

Das volle Modell kann mit einer folgenden Formel beschrieben werden:
*time_to_event ~ test_group * gender * bs(los, df=3)*.

Tabelle 36 zeigt zentrale Modellgütemaße zur Bewertung der Anpassung des vollen Regressionsmodells. Die Konkordanz beträgt 0,5589, was eine leichte Verbesserung der Vorhersagekraft gegenüber dem AIC-optimalen Modell bedeutet. Der Likelihood Ratio Test ist statistisch signifikant ($p < 0{,}005$).

Tabelle 36 - Modellgütemaße (volles Modell mit Strata)

Concordance	0.5589
AIC	12109.53
log-likelihood ratio test	46.07 on 15 df, p<0.005

Tabelle 37 enthält Regressionsergebnisse des vollen Modells (Referenz: Test-Gruppe A, Gender m). Während die Prädikatoren die Vorhersagekraft leicht verbessert haben, sind die Koeffizienten der Test-Gruppe, Gender, Zweifach- und Dreifach-Interaktionen nicht signifikant. Einer der Koeffizienten des Dienstalter-Splines zeigt deutliche Signifikanz ($p<0.005$).

Tabelle 37 - Koeffizientenschätzungen des vollen Modells mit Strata

covariate	coef	exp (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	p
test_group[B]	-0.13	0.87	-0.49	0.22	0.61	1.25	0.46
gender[w]	-0.16	0.86	-0.60	0.29	0.55	1.33	0.49
bs(los)[1]	1.68	5.37	-1.05	4.41	0.35	82.32	0.23
bs(los)[2]	-1.66	0.19	-2.71	-0.62	0.07	0.54	**<0.005**
bs(los)[3]	0.43	1.54	-0.46	1.32	0.63	3.76	0.34
test_group[B]: gender[w]	-0.03	0.97	-0.50	0.44	0.60	1.55	0.89
test_group[B]: bs(los)[1]	-0.94	0.39	-3.19	1.30	0.04	3.68	0.41
test_group[B]: bs(los)[2]	1.24	3.47	-0.60	3.08	0.55	21.85	0.19
test_group[B]: bs(los)[3]	-0.58	0.56	-1.42	0.27	0.24	1.31	0.18
gender[w]: bs(los)[1]	0.14	1.15	-2.36	2.64	0.09	14.01	0.91
gender[w]: bs(los)[2]	0.97	2.65	-0.51	2.46	0.60	11.73	0.20
gender[w]: bs(los)[3]	-0.51	0.60	-1.46	0.44	0.23	1.56	0.29
test_group[B]: gender[w]: bs(los)[1]	0.77	2.17	-1.56	3.11	0.21	22.34	0.52

test_group[B]: gender[w]: bs(los)[2]	-1.45	0.24	-3.79	0.90	0.02	2.46	0.23
test_group[B]: gender[w]: bs(los)[3]	1.26	3.54	-0.24	2.77	0.78	15.99	0.10

Die Dreifach-Interaktionen zeigten mögliche Verletzung der Proportional-Hazard-Annahme (Tabelle 38). Dieser war jedoch nicht konsistent in den Test-Statistiken. Nachdem die Interaktionen keinen signifikanten Effekt in den Regressionsergebnissen zeigten, wurde auf eine weitere Modellierung der Zeit-Abhängigkeit bzw. Stratifizierung nach Altersgruppen verzichtet.

Tabelle 38 - Mögliche Verletzungen der Proportion-Hazard-Annahme

covariate	test_statistic (KM)	p (KM)	test_statistic (rank)	p (rank)
test_group[B]: gender[w]: bs(los)[1]	6.92	0.01	0.30	0.59
test_group[B]: gender[w]: bs(los)[2]	5.89	0.02	0.72	0.40
test_group[B]: gender[w]: bs(los)[3]	6.35	0.01	0.17	0.68

9.4.2 Aufgabentyp als Kovariate

Der Aufgabentyp als Kovariate ermöglicht es, die Abhängigkeit der Zeit-bis-zum-Erfolg von dem Aufgabentyp bzw. seine Interaktionen mit anderen Kovariaten zu modellieren. Die Benutzer-ID wurde als Clustering-Variable verwendet.

Das AIC-optimale Modell (Tabelle 39) kann mit der folgenden Formel beschrieben werden: *time_to_event ~ test_group*task_type + bs(los, df=3).*

Das Modell hat höhere Konkordanz≈0,66 im Vergleich zu Modellen mit Strata, der Aufgabentyp verbessert die Vorhersagekraft des Modells. AIC kann nicht direkt mit Modellen mit Strata verglichen werden. Der Likelihood Ratio Test ist statistisch signifikant (p < 0,005).

Tabelle 39 - Modellgütemaße (AIC-optimales Modell mit Aufgabentyp als Kovariate)

Concordance	0.6568
AIC	13665.92
log-likelihood ratio test	334.73 on 8 df, p<0.005

Dieses Modell zeigt eine klare Verletzung der Proportional-Hazard-Annahme für den Aufgabentyp Dokument-Suche als Kovariate (KM-Statistik=11.85, p<0.005; Rank-Statistik=12.28, p<0.005). Die Tests für sonstige Variablen sind nicht signifikant. Dokument-Suche steht in Zusammenhang mit höherem Hazard im niedrigeren Zeitwert-Bereich und sinkt später (Abbildung 44). Dieses Muster deutet darauf hin, dass Dokument-Suche-Aufgaben in vielen Fällen schnell abgeschlossen werden, während sich länger andauernde Aufgaben im Zeitverlauf zunehmend anderen Aufgabentypen annähern.

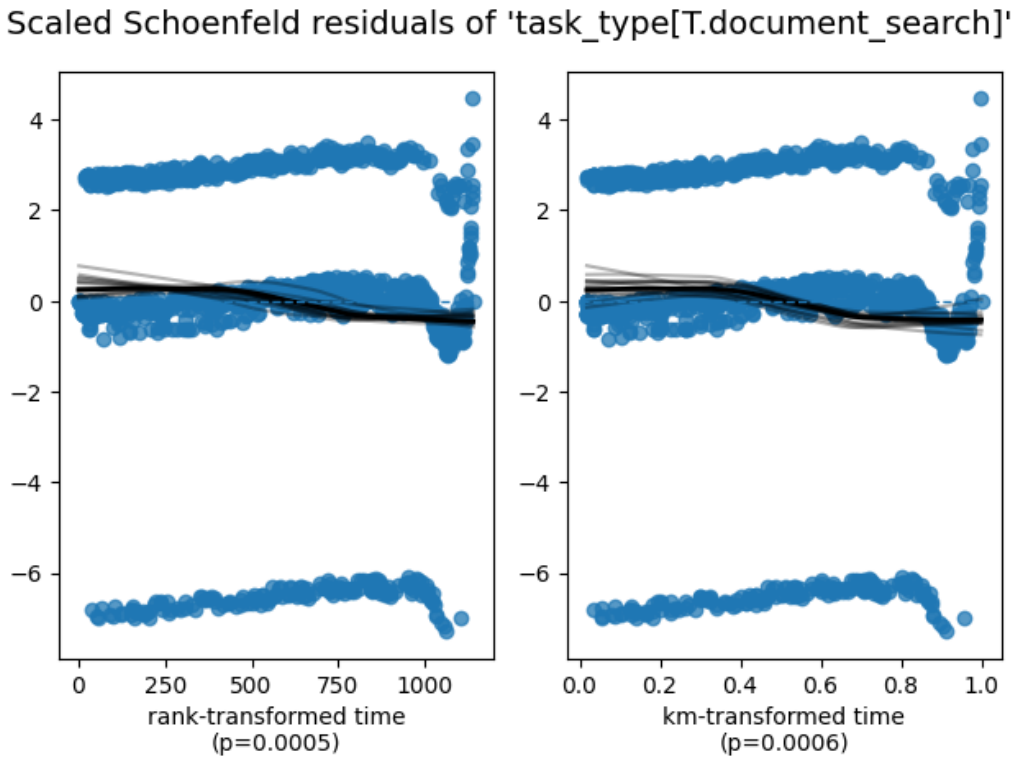

Abbildung 44 - Schoefeld-Residuals Dokument-Suche

Tabelle 40 stellt die Regressionsergebnisse dar (Referenz: Test-Gruppe A, Aufgabentyp Inhalts-Suche). Die Koeffizienten der Test-Gruppe B, des Aufgabentyps Dokument-Suche, Text-Generation und der Interaktion zwischen der Test-Gruppe B und dem Aufgabentyp Dokument-Suche zeigen eine deutliche Signifikanz (p<0,005). Die Interaktion zwischen der Test-Gruppe B und dem Aufgabentyp Text-Generation ist an der Signifikanzgrenze (p=0,05), einer der Koeffizienten des Dienstalter-Splines ist signifikant (p=0,01).

Tabelle 40 - Koeffizientenschätzungen des AIC-optimalen Modells mit Aufgabentyp als Kovariate

covariate	coef	exp (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	p
test_group[B]	-0.56	0.57	-0.80	-0.32	0.45	0.73	**<0.005**
task_type[document_search]	0.38	1.46	0.22	0.53	1.25	1.70	**<0.005**

task_type[generation]	-1.28	0.28	-1.60	-0.97	0.20	0.38	**<0.005**
bs(los)[1]	1.49	4.43	0.33	2.65	1.38	14.15	**0.01**
bs(los)[2]	-0.89	0.41	-1.80	0.03	0.17	1.03	0.06
bs(los)[3]	-0.02	0.98	-0.63	0.58	0.53	1.79	0.94
test_group[B]: task_type[document_search]	0.65	1.92	0.39	0.91	1.48	2.50	**<0.005**
test_group[B]: task_type[generation]	0.40	1.48	0.00	0.79	1.00	2.20	**0.05**

Haupteffekte: In dem Referenzaufgabentyp Inhalts-Suche hat Gruppe B im Vergleich zu Gruppe A ein signifikant niedrigeres Risiko (Hazard) (HR = 0,57), was auf eine deutlich langsamere Bearbeitung durch Gruppe B hinweist. Die Dokument-Suche Aufgaben wurden von der Referenzgruppe A schneller gelöst als Inhalts-Suche (HR = 1,46), die Text-Generations-Aufgaben deutlich langsamer (HR = 0,28).

Da die Proportional-Hazard-Annahme verletzt ist, ist der geschätzte Hazard-Ratio-Koeffizient als zeitgewichtetes Mittel zu interpretieren[7]. Der Effekt ist früh stärker und nimmt im Zeitverlauf ab: Der Aufgabentyp Dokument-Suche verkürzt anfangs die Bearbeitungszeit, später sinkt dieser Effekt im Vergleich zu anderen Aufgabentypen (s. Residuen Abbildung 44).

Die **Interaktionseffekte** zwischen Gruppe B und den Aufgaben zeigen, wie sich der Effekt der Aufgaben im Vergleich zum Referenzaufgabentyp Inhalts-Suche für Gruppe B im Vergleich zu Gruppe A verändert. exp(coef) > 1 bedeutet, dass der Unterschied zwischen dem Aufgabentyp und der Inhalts-Suche für Gruppe B stärker ist als für Gruppe A, während exp(coef) < 1 einen schwächeren Effekt für Gruppe B bedeutet. Die Dokument-Suche im Vergleich zu Inhalts-Suche wurde durch die Interaktion beschleunigt (HR = 1,92), auch die Text-Generation (HR = 1,48), wobei hier der p-Wert an der Signifikanzgrenze liegt.

[7] https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Do-I-need-to-care-about-the-proportional-hazard-assumption?

Hazards für Gruppe B in Aufgabentypen im Vergleich zu Kontroll-Gruppe können durch Kombination der Koeffizienten ermittelt werden. Bei der Dokument-Suche war Gruppe B im Vergleich zu Gruppe A etwas schneller (HR = exp(-0,56 + 0,65) = 1,09), bei der Text-Erstellung hingegen etwas langsamer (HR = exp(-0,56 + 0,40) = 0,85).

9.5 Feedback-Auswertung

Studienteilnehmerinnen und Studienteilnehmer konnten ihr Feedback nach jedem Bearbeitungsvorgang und auf diesen bezogen abgeben, sodass mehrere Einträge pro Teilnehmerin oder Teilnehmer gesammelt wurden, die jedoch unterschiedliche Bewertungen und Rückmeldungen enthalten können. Teilnehmerinnen und Teilnehmer konnten Fragen unbeantwortet lassen. Die Feedback-Fragen bestehen aus Auswahlfragen und Freitext-Feedback. Insgesamt haben 50 von 56 Teilnehmerinnen und Teilnehmer das Feedback abgegeben: 26 aus der Gruppe A (13 Frauen, 13 Männer) und 24 aus der Gruppe B (12 Frauen, 12 Männer), Abbildung 45.

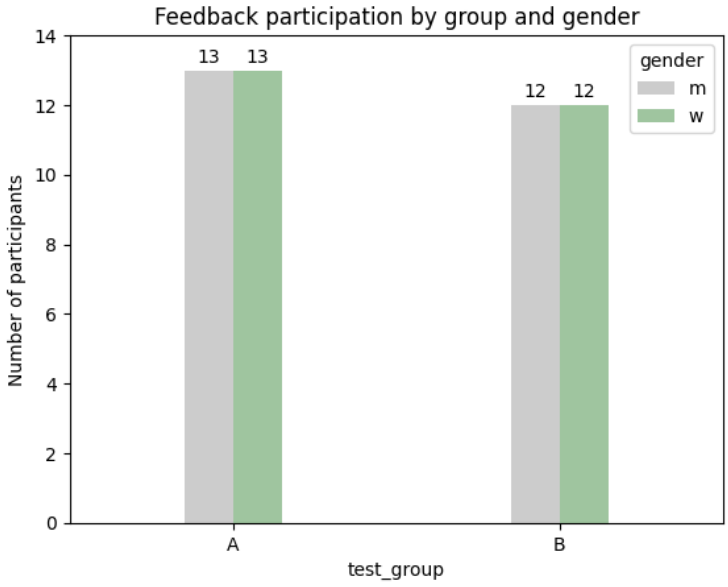

Abbildung 45 - Feedback-Beteiligung

Die Antworten auf die Auswahlfragen werden sowohl nach Test-Gruppen als auch nach Test-Gruppe und Gender aufgeschlüsselt.

9.5.1 Auswahlfragen

Bei den Auswahlfragen konnten Userinnen und User eine von mehreren Antwort-Möglichkeiten auswählen. Einige Fragen wurden nur einer der Test-Gruppen gestellt bzw. wurden unterschiedlich formuliert.

Hinweis: Zum Zeitpunkt des Ausfüllens des Fragebogens wussten Teilnehmenden nicht über die Korrektheit der gefundenen Lösung Bescheid. Diese war erst später, nach dem Absenden des Fragebogens ersichtlich. Somit handelt es sich um eine subjektive Einschätzung, was bspw. die Relevanz der Ergebnisse oder die Schwierigkeit der Lösungsfindung betrifft.

Anschließend sind Ergebnisse je nach Frage dargestellt.

1. "Wie einfach war es für Sie, die benötigten Informationen zu finden?" (A, B)

Die Mehrheit der Teilnehmenden aus beiden Gruppen beantwortete diese Frage mit den Top-2 Kategorien "sehr einfach" oder "eher einfach", wobei Gruppe B über der Gruppe A liegt (A: 75,4%, B: 81,6%), insbesondere in der Kategorie "sehr einfach" (A: 39%, B: 53,5%). Die Gruppen sind ähnlich bei der Einschätzung "sehr schwer" (A: 5,7%, B: 5,8%); Gruppe B bewertete die Suchvorgänge seltener als "eher schwer" (A: 18,9%, B: 12,5%). Insgesamt fand somit die Gruppe B (subjektiv) die benötigten Informationen etwas leichter als Gruppe A. S. Abbildung 46.

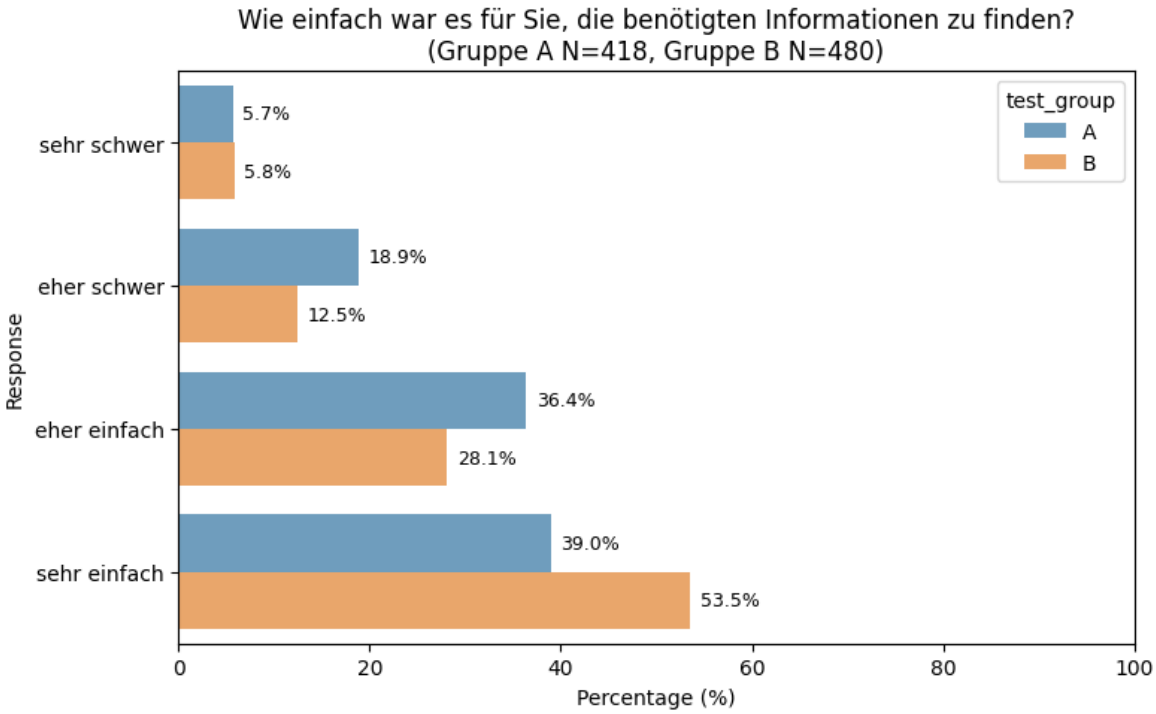

Abbildung 46 - Feedback nach Test-Gruppen: "Wie einfach war es für Sie, die benötigten Informationen zu finden?"

Betrachtet nach Gender-Gruppen (Abbildung 47), lassen sich Unterschiede innerhalb der Gruppen erkennen. Während Frauen und Männer in der Gruppe B ähnlich oft die Einfachheit der Lösungsfindung mit "sehr einfach" bewerteten (Männer: 54,7%, Frauen: 52,3%), ist der Unterschied in der Gruppe A höher (Männer: 46,5%, Frauen: 30,8%). In der Gruppe A fanden Frauen hingegen die Lösungen "eher einfach" öfter als Männer, unter Betrachtung der Top-2 Kategorien ist der Gender-Unterschied geringer (Männer: 77,8%, Frauen: 72,6%). In beiden Gruppen fanden Frauen die gefragten Informationen öfter "eher schwer" als Männer und etwas seltener "sehr schwer".

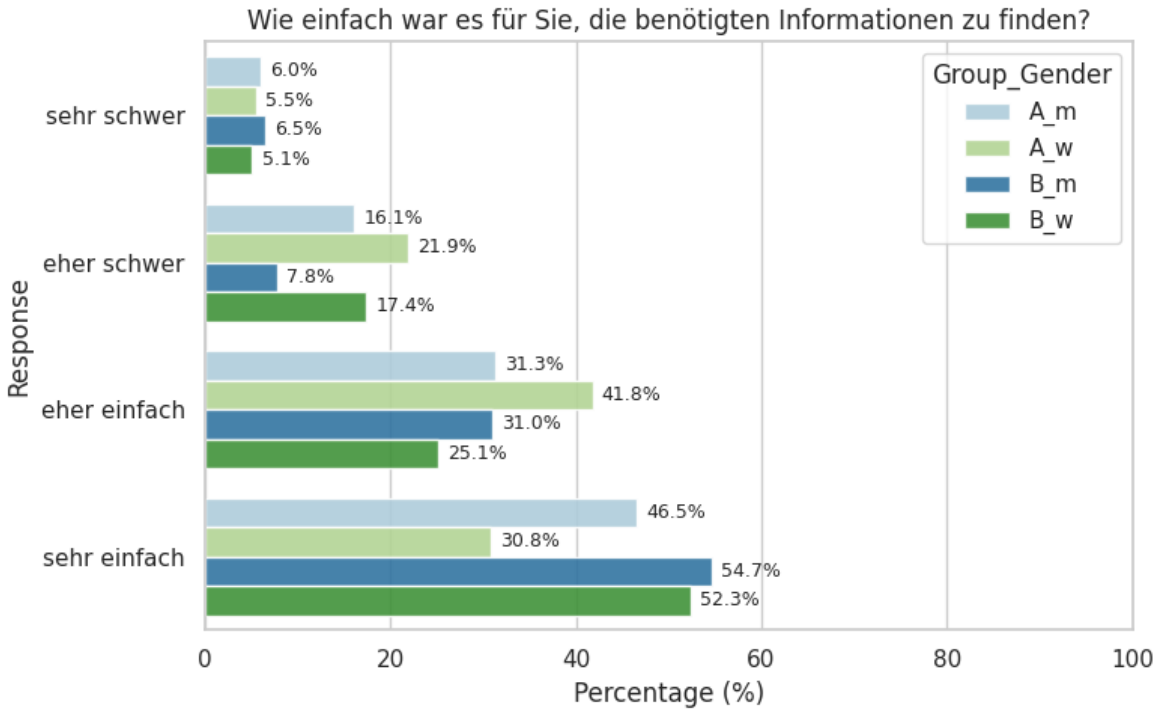

Abbildung 47 - Feedback nach Test-Gruppen und Gender: "Wie einfach war es für Sie, die benötigten Informationen zu finden?"

2. "Wie zufrieden sind Sie mit der Übersichtlichkeit der Oberfläche?" (A, B)

Die Mehrheit der Teilnehmerinnen und Teilnehmer war mit der Oberfläche des jeweiligen Tool zufrieden ("sehr zufrieden" oder "eher zufrieden", A: 65,5%, B: 80,1%), wobei Gruppe B höhere Zufriedenheit zeigt, insbesondere in der Kategorie "sehr zufrieden" mit 44% gegenüber nur 23,8% in der Gruppe A. Teilnehmerinnen und Teilnehmer aus der Gruppe A waren entsprechend öfter "eher unzufrieden" oder "sehr unzufrieden". Insgesamt lässt sich die Zufriedenheit mit der Oberfläche in der Gruppe B als tendenziell höher im Vergleich zu A bewerten (Abbildung 48).

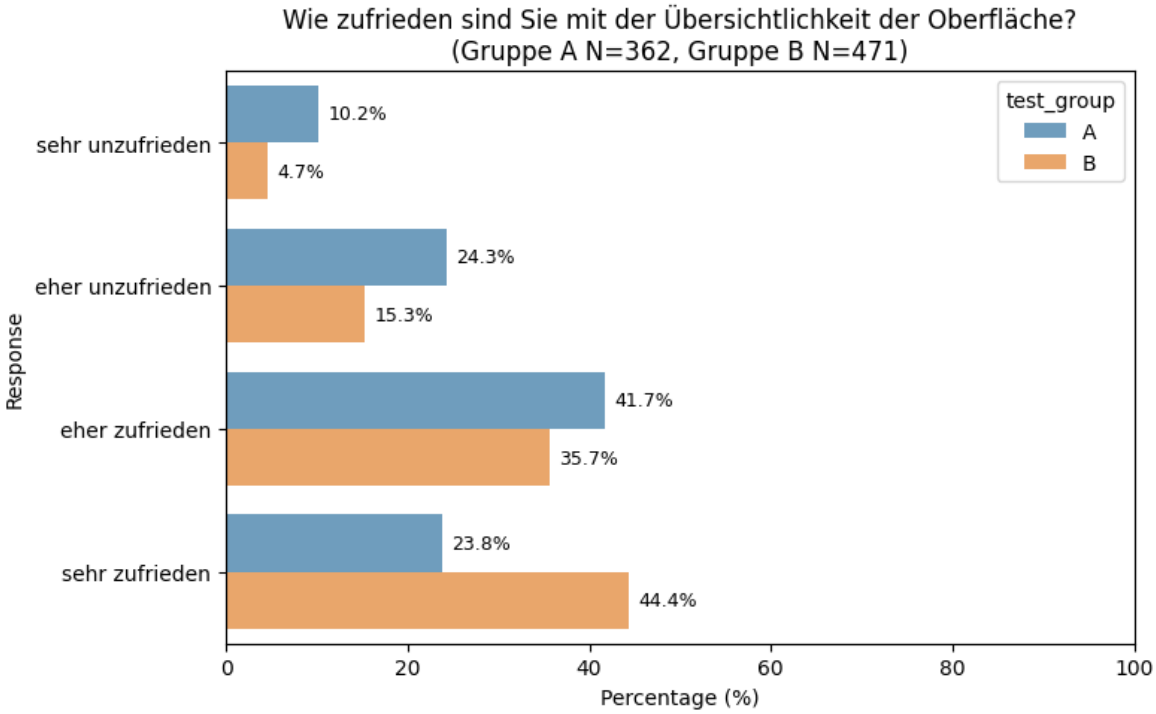

Abbildung 48 - Feedback nach Test-Gruppen: "Wie zufrieden sind Sie mit der Übersichtlichkeit der Oberfläche?"

In den Gender-Gruppen (Abbildung 49) zeigt sich unterschiedliches Bild nach Test-Gruppen: In der Gruppe A waren Frauen besonders selten mit der Oberfläche "sehr zufrieden" (13,5%) im Vergleich zu Männern (31,4%) und deutlich öfter "eher unzufrieden" (31,6%; Männer: 18,8%). Nach den Top-2 Kategorien waren Frauen in beiden Test-Gruppen jeweils weniger zufrieden als Männer, dies insbesondere in der Gruppe A (Frauen: 56,7%, Männer: 72%), Gruppe B (Frauen: 75,4%, Männer: 84,5%).

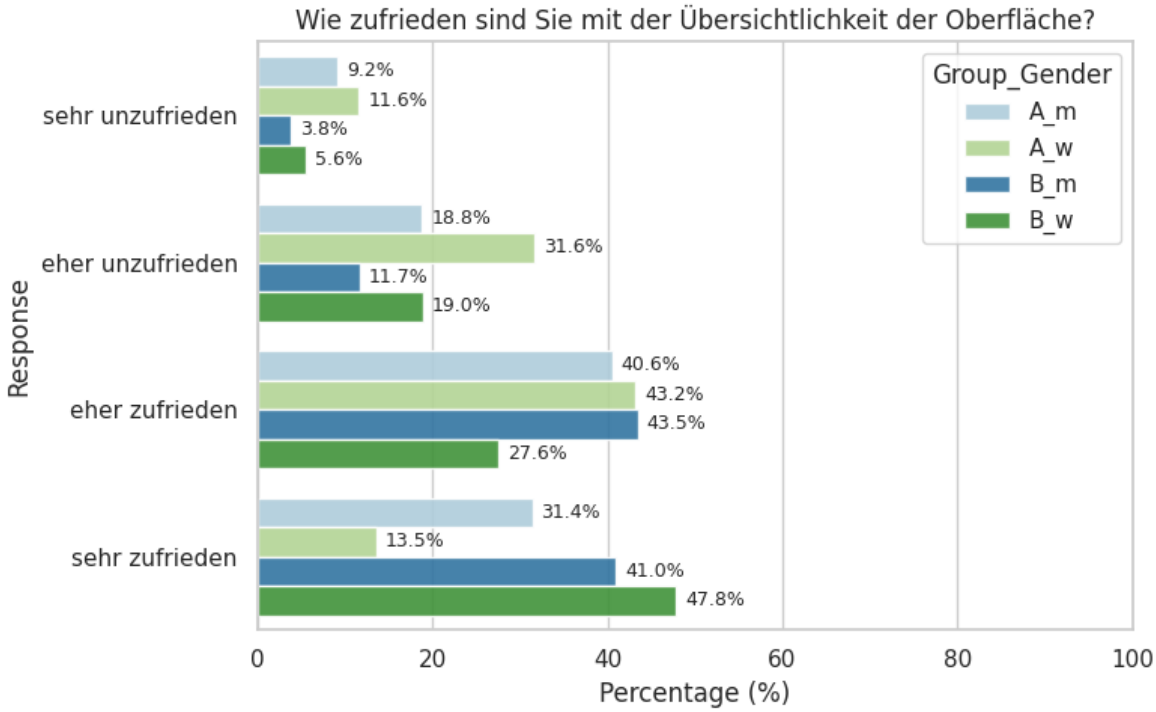

Abbildung 49 - Feedback nach Test-Gruppen und Gender: "Wie zufrieden sind Sie mit der Übersichtlichkeit der Oberfläche?"

3. "Mussten Sie Umwege gehen (…), um ans Ziel zu kommen?" (A, B)

Gruppe B hat etwas seltener als Gruppe A Umwege gebraucht, um Lösungen zu finden (Abbildung 50): Teilnehmende aus der Gruppe B haben diese Frage in 76,8% der Fälle mit "Nein" beantwortet, Gruppe B in 74,3%.

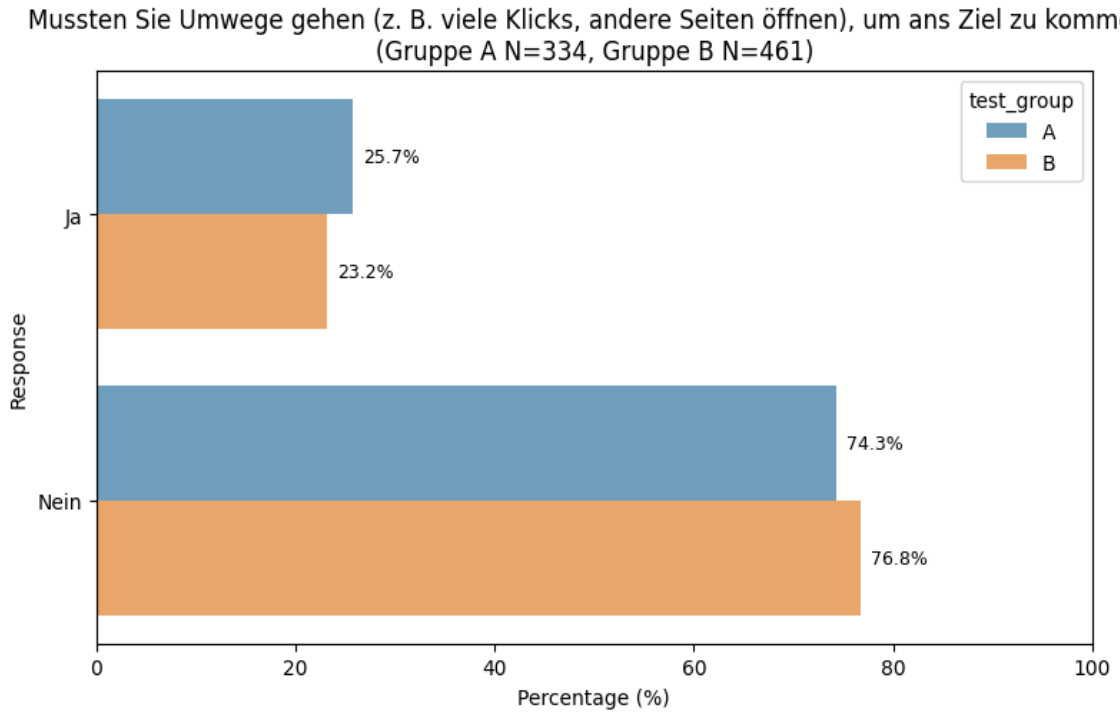

Abbildung 50 - Feedback nach Test-Gruppen: "Mussten Sie Umwege gehen, um ans Ziel zu kommen?"

Nach Gender-Gruppen haben insbesondere Frauen aus der Gruppe A öfter Umwege gebraucht mit 30% "Ja" (Abbildung 51). In beiden Test-Gruppen haben Frauen öfter als Männer Umwege gehen müssen, wobei der Unterschied in der Gruppe B geringer ausfällt (3,3%) im Vergleich zu A (7,7%).

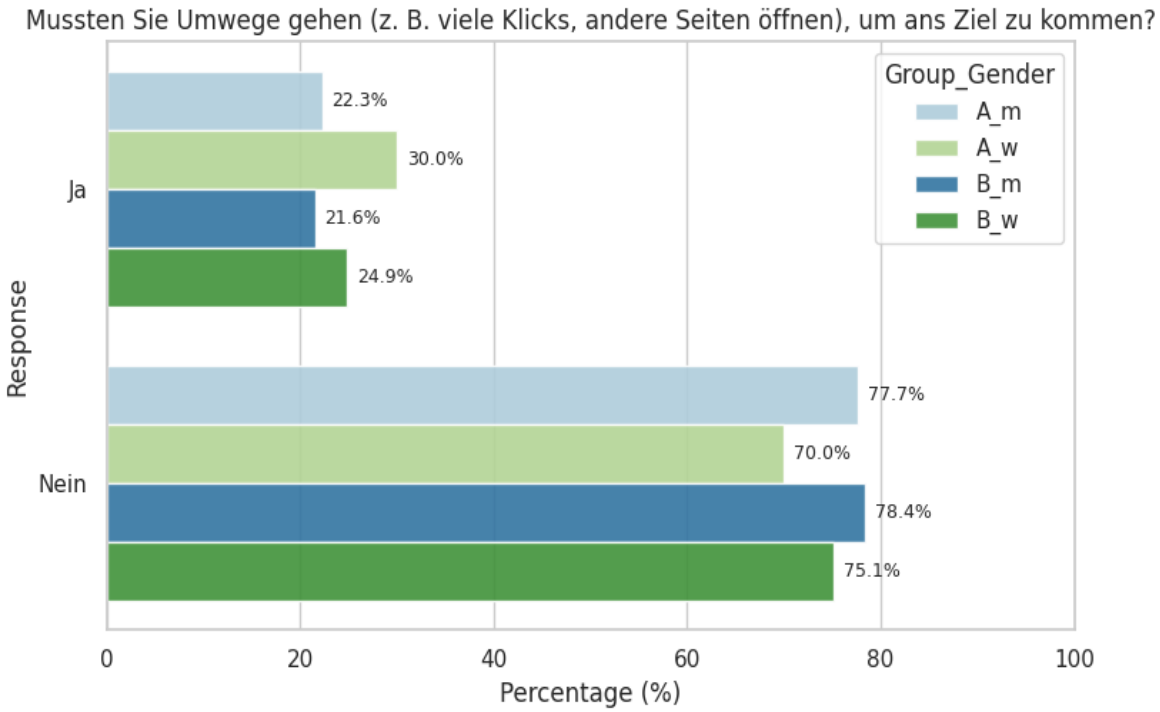

Abbildung 51- Feedback nach Test-Gruppen und Gender: "Mussten Sie Umwege gehen, um ans Ziel zu kommen?"

4. "Hat das KI-Tool Ihre Frage inhaltlich richtig verstanden?" (B)

Die meisten Teilnehmerinnen und Teilnehmer in der Gruppe B beantworteten diese Frage positiv (57,1%), in den beiden Top-2 Kategorien ("eher ja", "ja") waren es 83,5% (Abbildung 52).

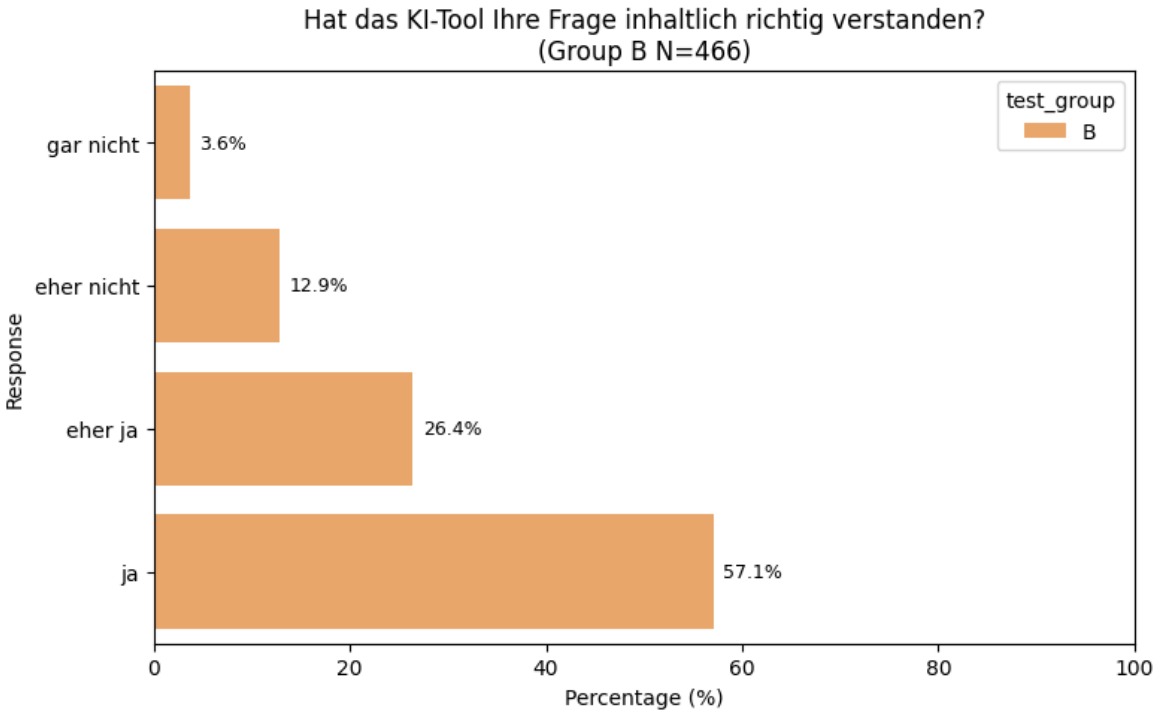

Abbildung 52 - Feedback Gruppe B: "Hat das KI-Tool Ihre Frage inhaltlich richtig verstanden?"

Frauen haben seltener als Männer die Frage mit "ja" beantwortet (50,4% vs. 63,8%). Hingegen fanden sie in 19,2% der Fälle, dass das KI-Tool die Frage "eher nicht" richtig verstanden hat (Männer: 6,5%).

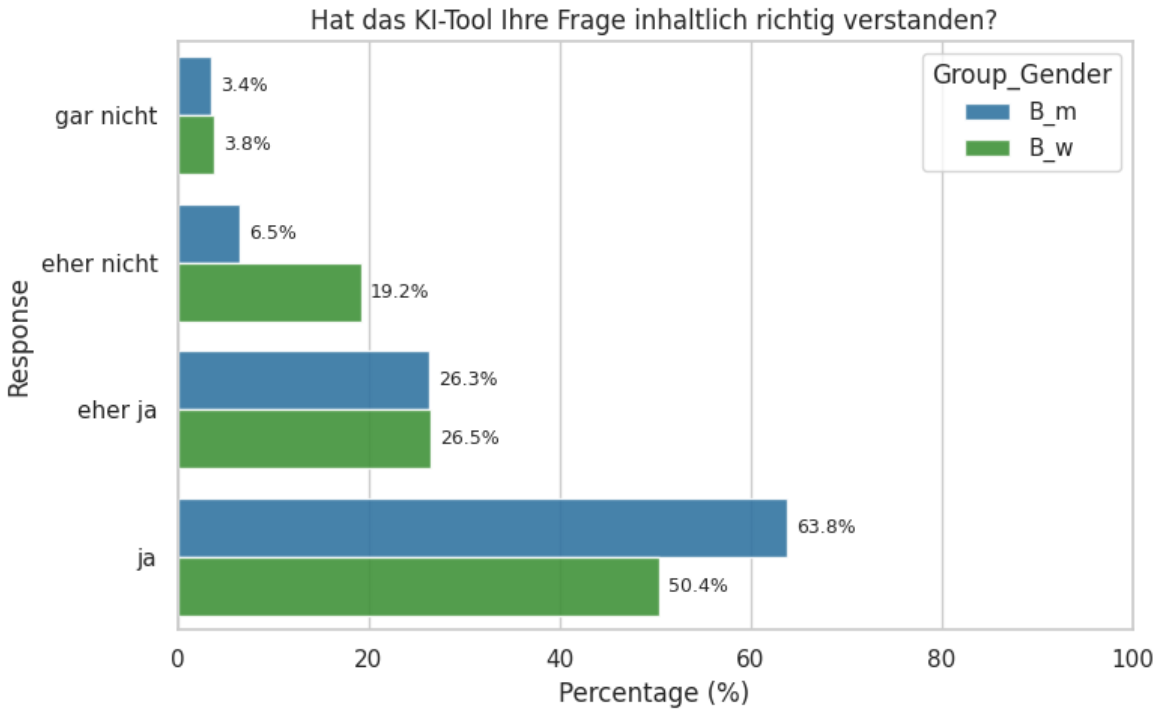

Abbildung 53- Feedback Gruppe B nach Gender: "Hat das KI-Tool Ihre Frage inhaltlich richtig verstanden?"

5. "Wie hilfreich waren die angezeigten Ergebnisse?" (B)

In 46,9% der Fälle fanden Teilnehmerinnen und Teilnehmer in der Gruppe B die Ergebnisse "sehr hilfreich". In den Top-2 Kategorien sind es 81,5%. In 5,2% der Fälle wurden Ergebnisse als "gar nicht hilfreich" empfunden (Abbildung 54).

Frauen fanden die angezeigten Ergebnisse seltener "sehr hilfreich" als Männer (42,4% vs. 51,3%) und etwas seltener "eher hilfreich" (33,3% vs. 35,8%). Um 13% öfter (19,9%) als Männer fanden sie die Ergebnisse "eher nicht hilfreich" (Abbildung 55).

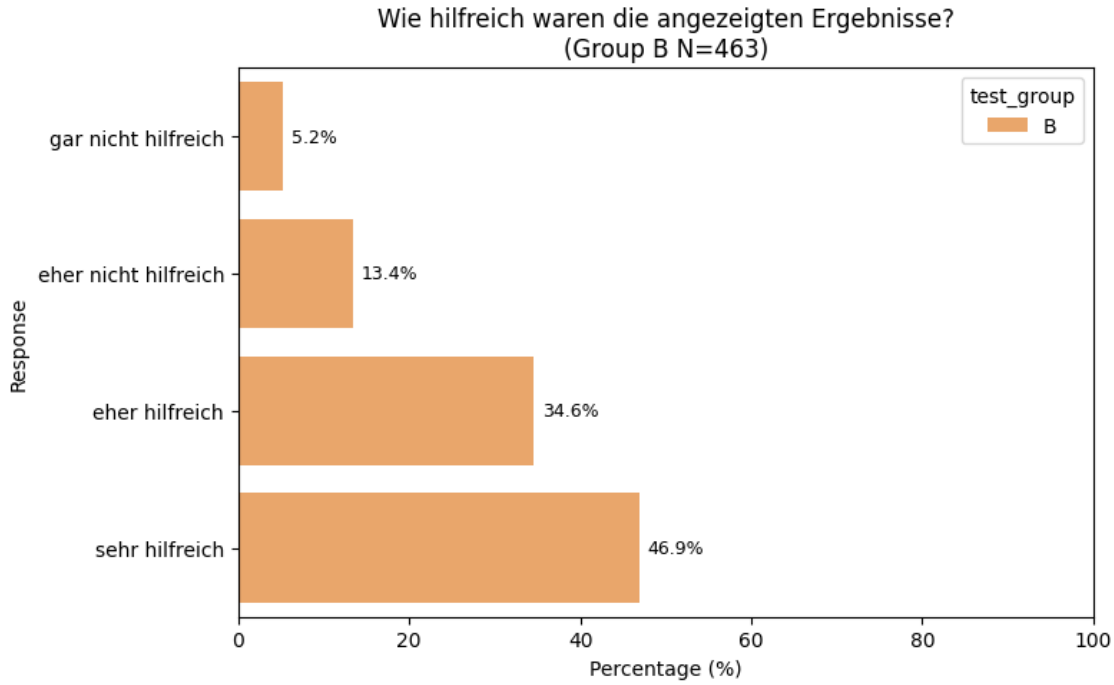

Abbildung 54 - Feedback Gruppe B: "Wie hilfreich waren die angezeigten Ergebnisse?"

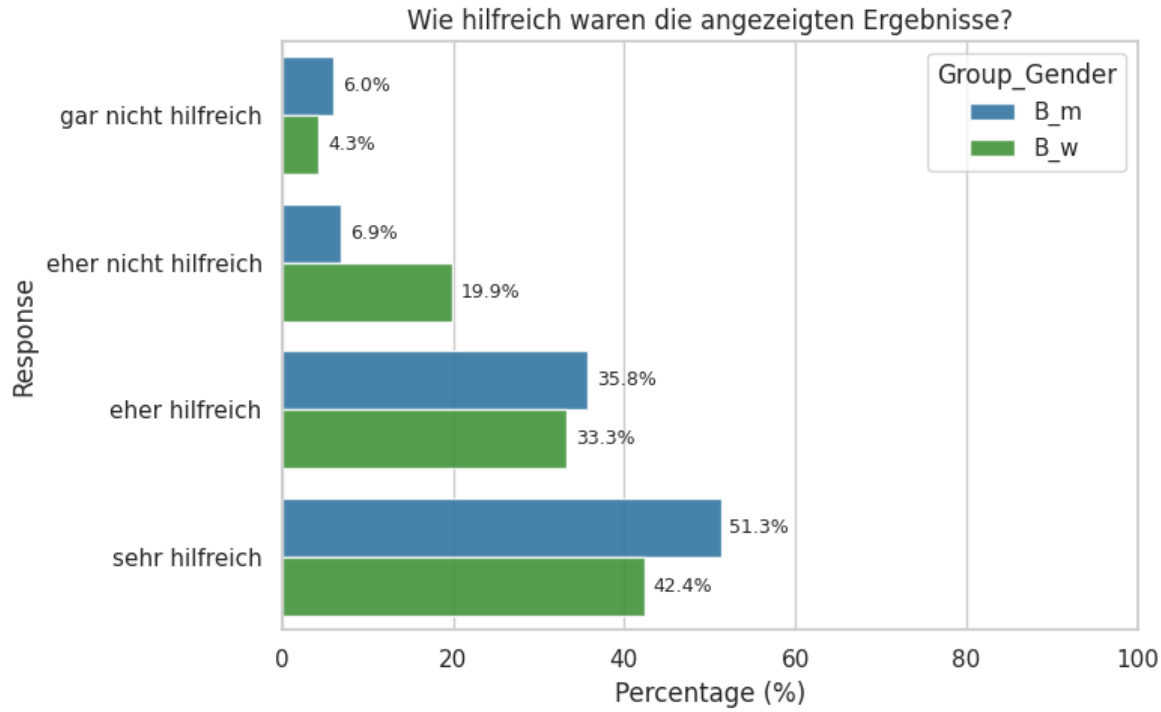

Abbildung 55 - Feedback Gruppe B nach Gender: "Wie hilfreich waren die angezeigten Ergebnisse?"

6. "War die Relevanz der angezeigten Dokumente hoch?" (A)

Diese Frage kann mit der Frage 5 (Gruppe B) verglichen werden. Jedoch ist zu beachten, dass Gruppe A bzgl. der Relevanz der Suchergebnisse gefragt wurde, währenddessen Gruppe B auch die Generations-Ergebnisse des RAG-Systems mitbewerten konnte. Gruppe A fand die angezeigten Dokumente in 47,8% der Fälle hochrelevant und in 32,7% der Fälle eher relevant (Abbildung 56). Die Zustimmung ist ähnlich der in der Gruppe B.

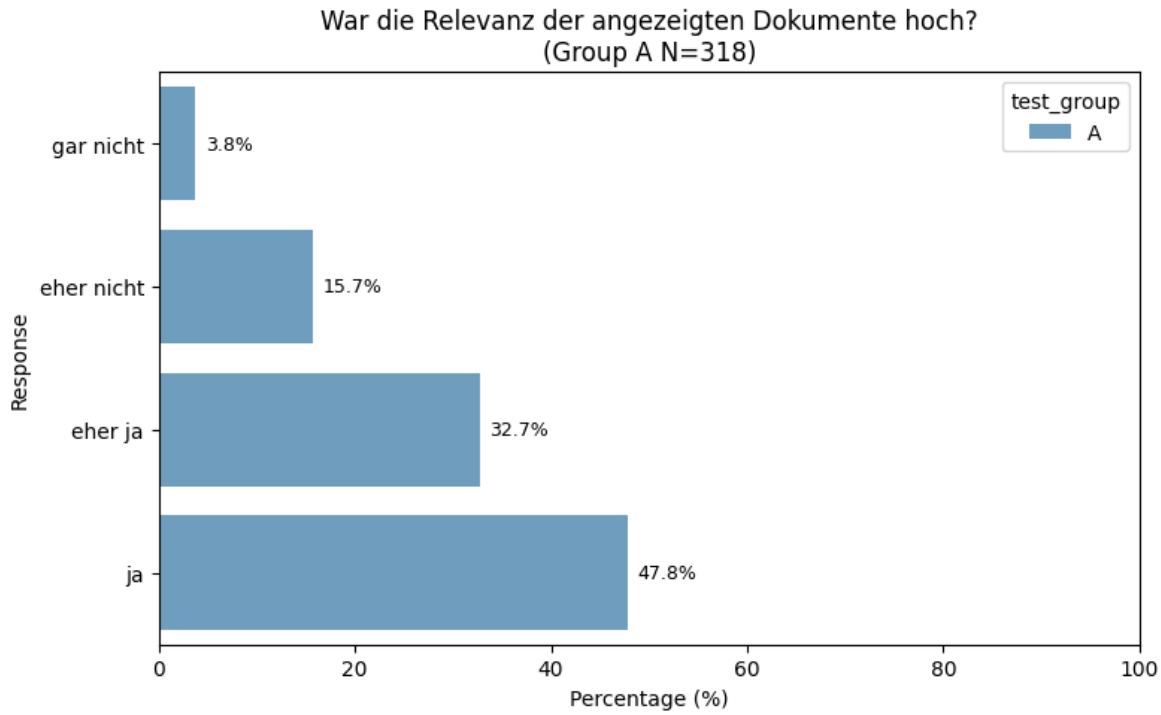

Abbildung 56 - Feedback Gruppe A: "War die Relevanz der angezeigten Dokumente hoch?"

In den Gender-Gruppen sind deutlich höhere Unterschiede im Vergleich zu Gruppe B bemerkbar (Abbildung 57): Frauen fanden nur in 32,4% der Fälle die Ergebnisse hochrelevant ("ja"), Männer hingegen in 60,2%. Frauen bewerteten die Ergebnisse um 8,3% öfter als eher relevant im Vergleich zu Männern, jedoch auffällig oft als "eher nicht" relevant (25,4% vs. 8%). Diese Bewertungen deuten darauf hin, dass in der Gruppe A ein gender-spezifischer Unterschied betreffend der wahrgenommenen Relevanz der Suchergebnisse besteht.

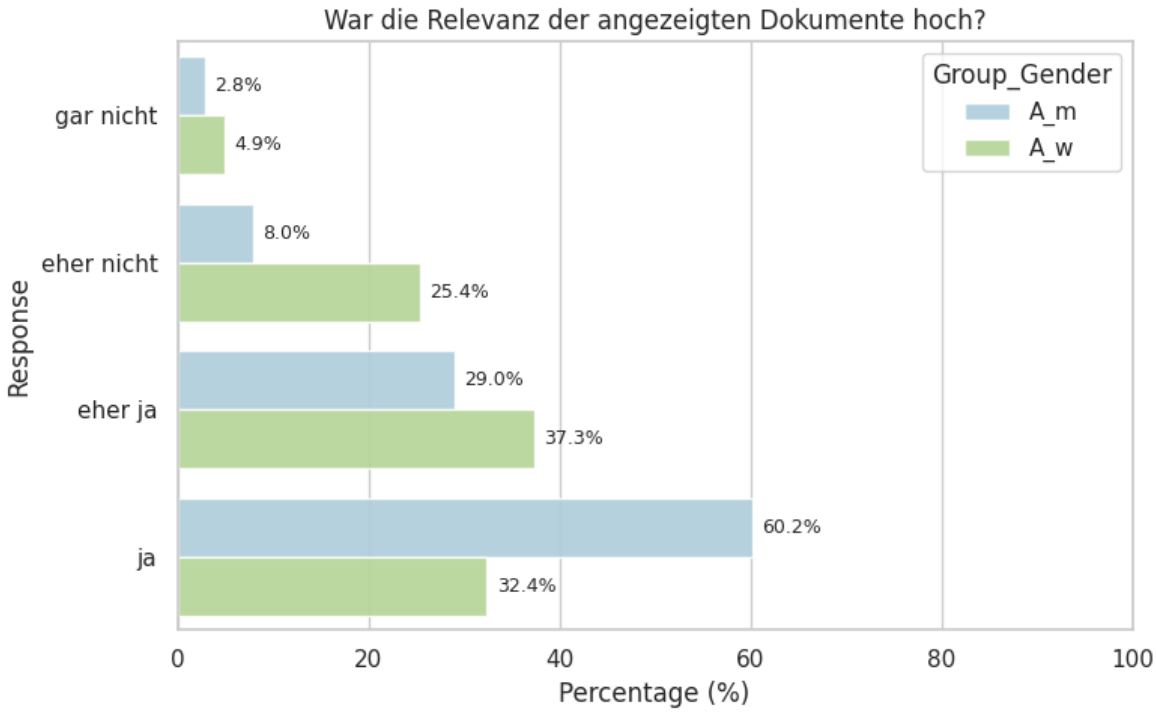

Abbildung 57 - Feedback Gruppe A nach Gender: "War die Relevanz der angezeigten Dokumente hoch?"

7. "Haben Sie die gesuchte Information direkt gefunden oder mussten Sie die Frage neu formulieren?" (B)

In 69,3% der Fälle haben Teilnehmende die Information direkt gefunden und in 25% der Fälle mussten sie die Frage neu formulieren (Abbildung 58). In 5,7% der Fälle wurde die Information nicht gefunden.

Frauen haben die gefragten Informationen seltener als Männer direkt gefunden (62,7% vs. 76,2%) und haben öfter nachjustieren müssen (32,2% vs. 17,6%), etwas seltener (5,2% vs. 6,2%) haben sie die Informationen gar nicht gefunden, s. Abbildung 59.

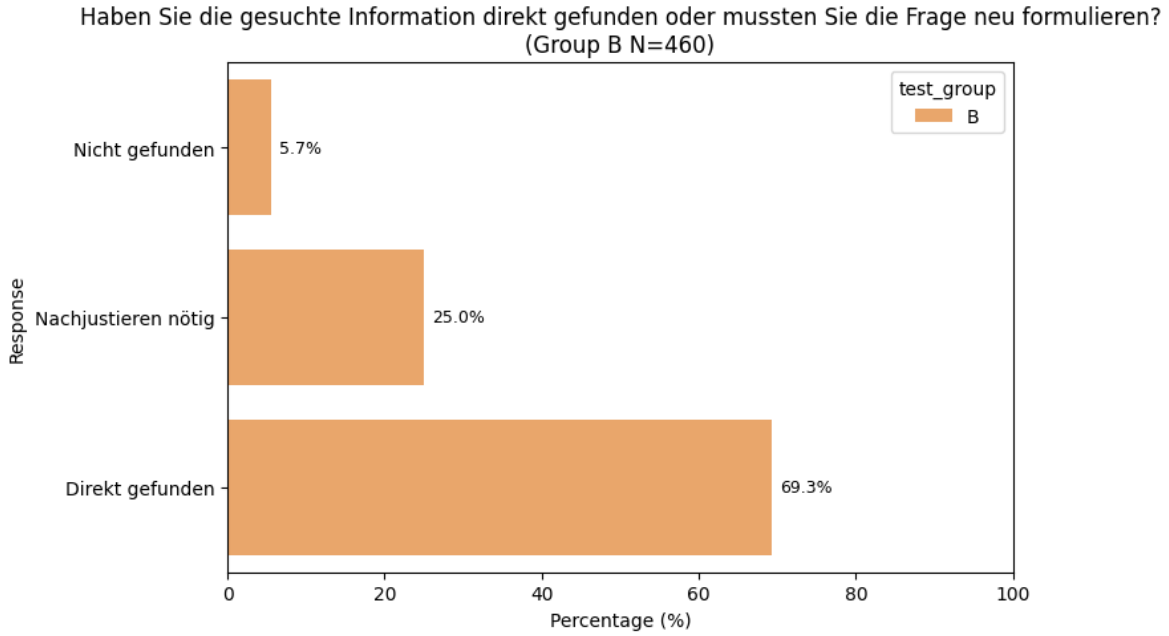

Abbildung 58 - Feedback Gruppe B: "Haben Sie die gesuchte Information direkt gefunden oder mussten Sie die Frage neu formulieren?"

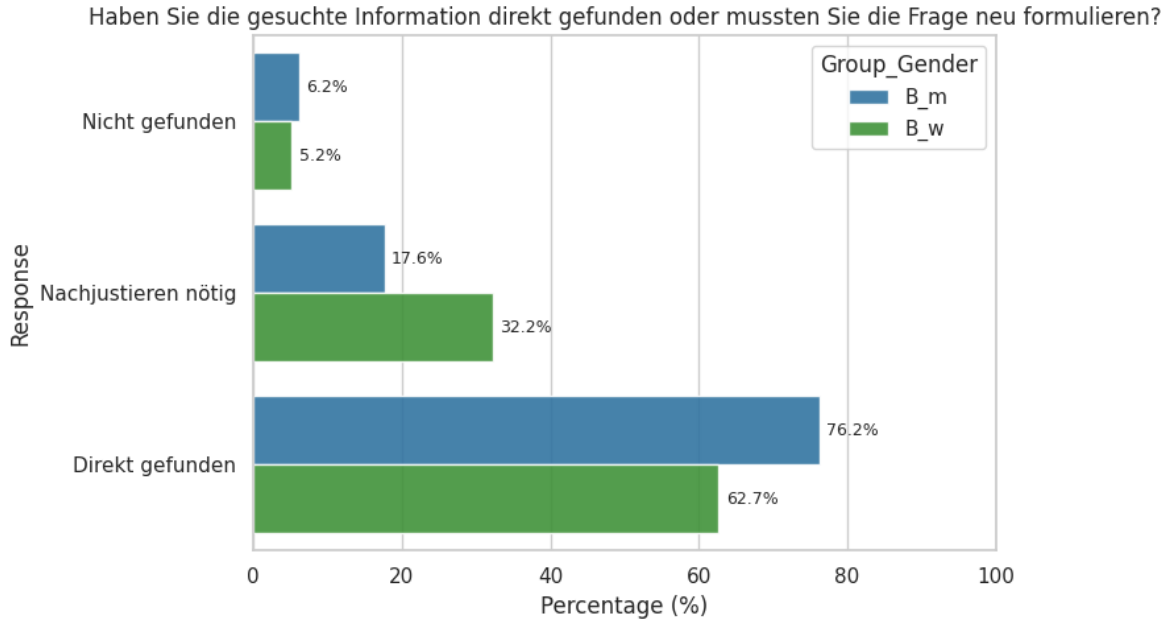

Abbildung 59- Feedback Gruppe B nach Gender: "Haben Sie die gesuchte Information direkt gefunden oder mussten Sie die Frage neu formulieren?"

8. "Wie lange haben Sie (gefühlt) gebraucht, um die gesuchte Antwort zu finden?" (A, B)

Gruppe B hat in 56,5% der Fälle das Gefühl gehabt, weniger als eine Minute gebraucht zu haben, Gruppe A mit 51% etwas seltener. Weniger als 3 Minuten (Top-2 Kategorien) wurde von der Gruppe A in 85,6% der Fälle, Gruppe B in 82,6% ausgewählt (Abbildung 60). Nach der subjektiven Einschätzung war Gruppe B etwas öfter sehr schnell, wenn jedoch eine Aufgabe länger als eine Minute in Anspruch genommen hat, haben die Teilnehmerinnen und Teilnehmer gefühlt öfter als Gruppe A zwischen 3 und 10 Minuten gebraucht.

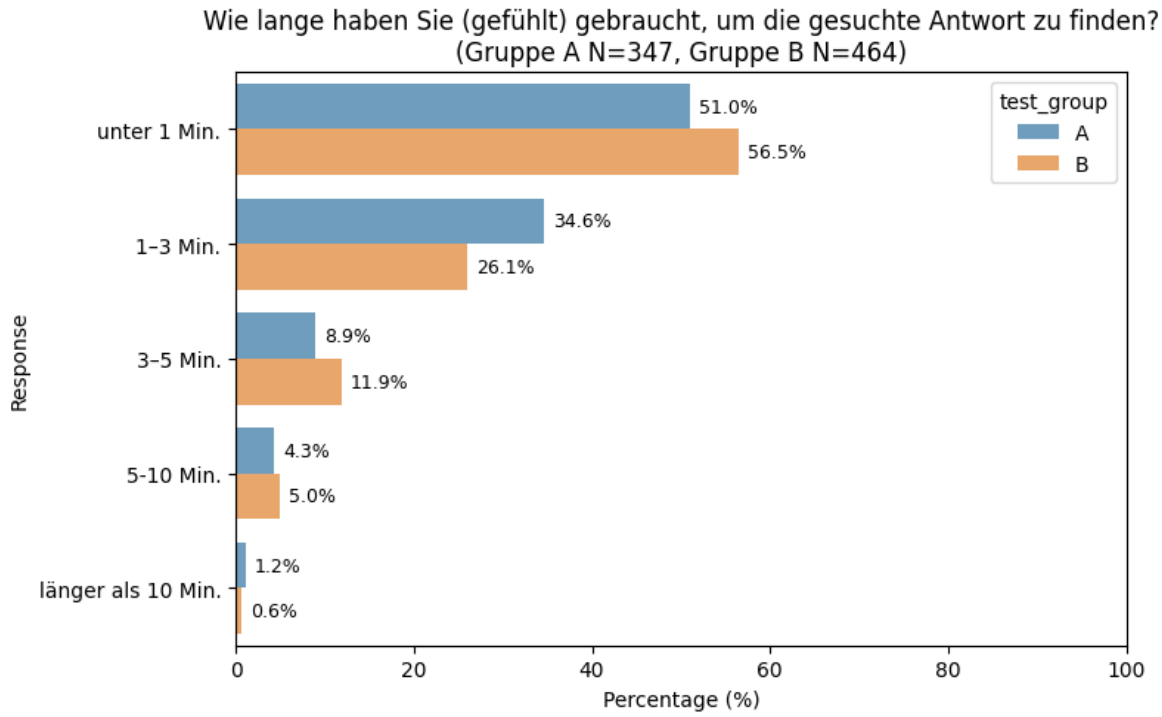

Abbildung 60 - Feedback nach Test-Gruppen: "Wie lange haben Sie (gefühlt) gebraucht, um die gesuchte Antwort zu finden?"

Der größte Unterschied in den Gender-Gruppen (Abbildung 61) ist in der Kategorie "unter 1 Minute" in der Gruppe A bemerkbar: In fast 10% der Fälle haben Frauen in der Gruppe A seltener gefühlt unter 1 Minute für die Aufgabe gebraucht als Männer. Die Werte der Gruppe B sind hingegen ähnlich (Frauen: 56,9%, Männer: 56%). Frauen in der Gruppe A haben gefühlt öfter mehr als eine und weniger als 3 Minuten gebraucht als Männer (37% vs. 32,6%), hingegen Frauen in der Gruppe B seltener (22% vs. 30,2%). In beiden Gruppen haben Frauen öfter als Männer gefühlt 3 bis 5 Minuten gebraucht. Mit 6% haben Frauen

als der Gruppe B nach subjektiver Einschätzung öfter 5 bis 10 Minuten gebraucht. Männer aus der Gruppe A haben nie länger als 10 Minuten gebraucht, hingegen Frauen in 2,6% der Fälle.

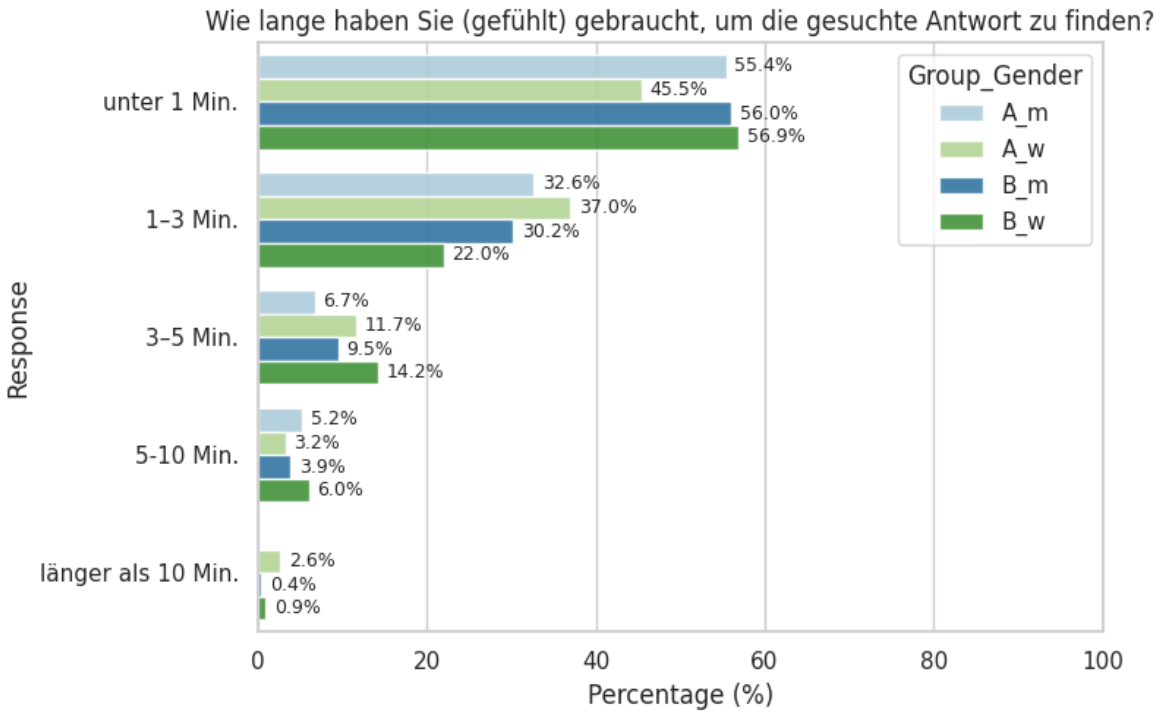

Abbildung 61 - Feedback nach Test-Gruppen und Gender: "Wie lange haben Sie (gefühlt) gebraucht, um die gesuchte Antwort zu finden?"

9. "Hatten Sie das Gefühl, schneller ans Ziel zu kommen, als Sie es sonst gewohnt sind?" (A, B)

Gruppe B hatte in 65,5% der Fälle das Gefühl, schneller ans Ziel zu kommen als gewohnt, während Gruppe A die Frage nur in 12,9% der Fälle mit "ja" beantwortete (Abbildung 62). Während die niedrige Zustimmung in der Gruppe A aufgrund des gewöhnlichen Suchsystems erwartbar ist, deutet die hohe Zustimmung in der Gruppe B auf die wahrgenommenen Vorteile des neuen Tools hin.

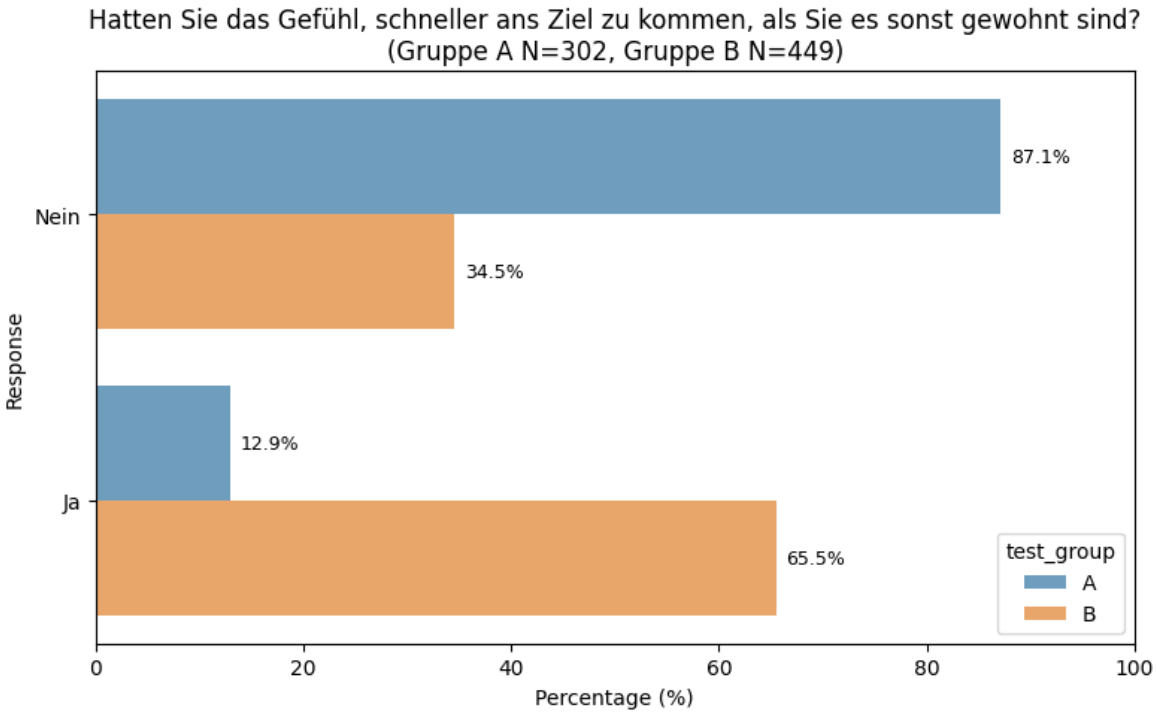

Abbildung 62 - "Feedback nach Test-Gruppen: Hatten Sie das Gefühl, schneller ans Ziel zu kommen, als Sie es sonst gewohnt sind?"

Männer in der Gruppe B hatten besonders oft gefühlt schneller als gewohnt die Informationen gefunden (68,6%), Frauen etwas weniger oft (62,4%). Besonders selten (6,4%) hatten Männer aus der Gruppe A die Frage mit "ja" beantwortet, Frauen aus der Gruppe A mit 19,9% um 13,5% öfter (Abbildung 63).

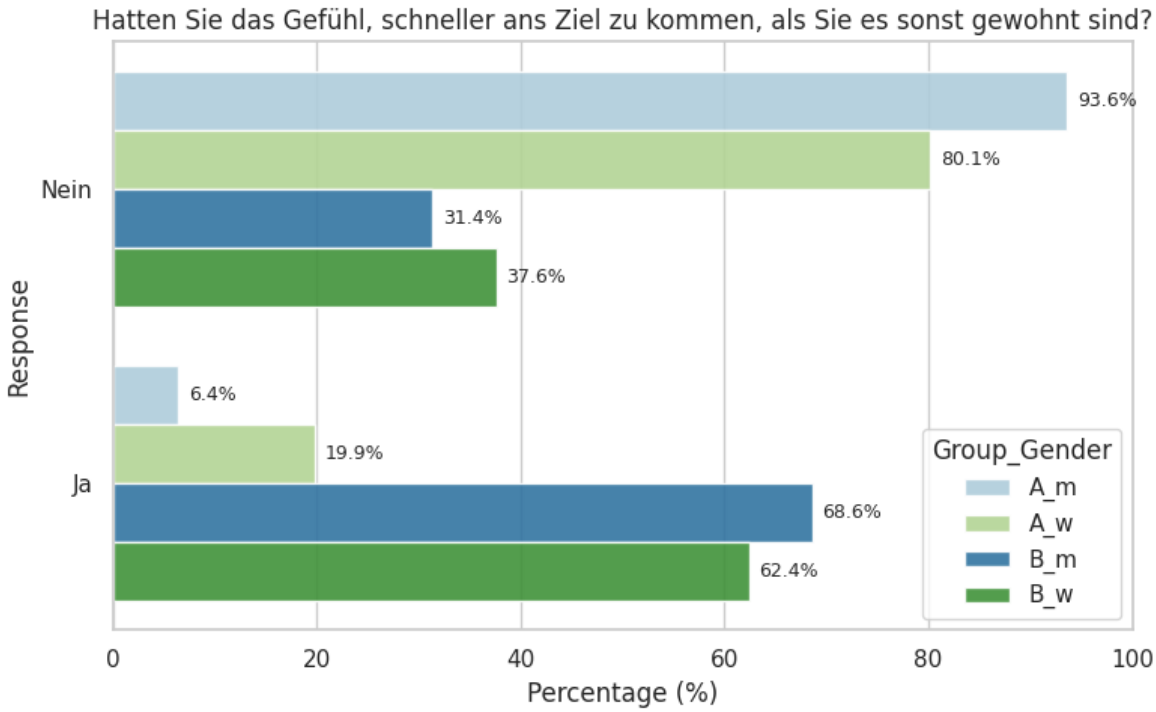

Abbildung 63 - "Feedback nach Test-Gruppen und Gender: Hatten Sie das Gefühl, schneller ans Ziel zu kommen, als Sie es sonst gewohnt sind?"

10. "Wie sehr vertrauen Sie den angezeigten Informationen?" (A, B)

Das Vertrauen in die angezeigten Informationen in den Gruppen (Abbildung 64) ist grundsätzlich hoch nach Top-2 Kategorien: Gruppe A - 89,6%, Gruppe B - 88,1%. Der höchste Unterschied zwischen den Gruppen beträgt 3,4% in der Kategorie "ja, sehr", wobei Gruppe B etwas seltener den angezeigten Informationen sehr vertraut hatte. Insgesamt halten sich die Vertrauensbewertungen der Gruppen die Waage, mit geringen Unterschieden.

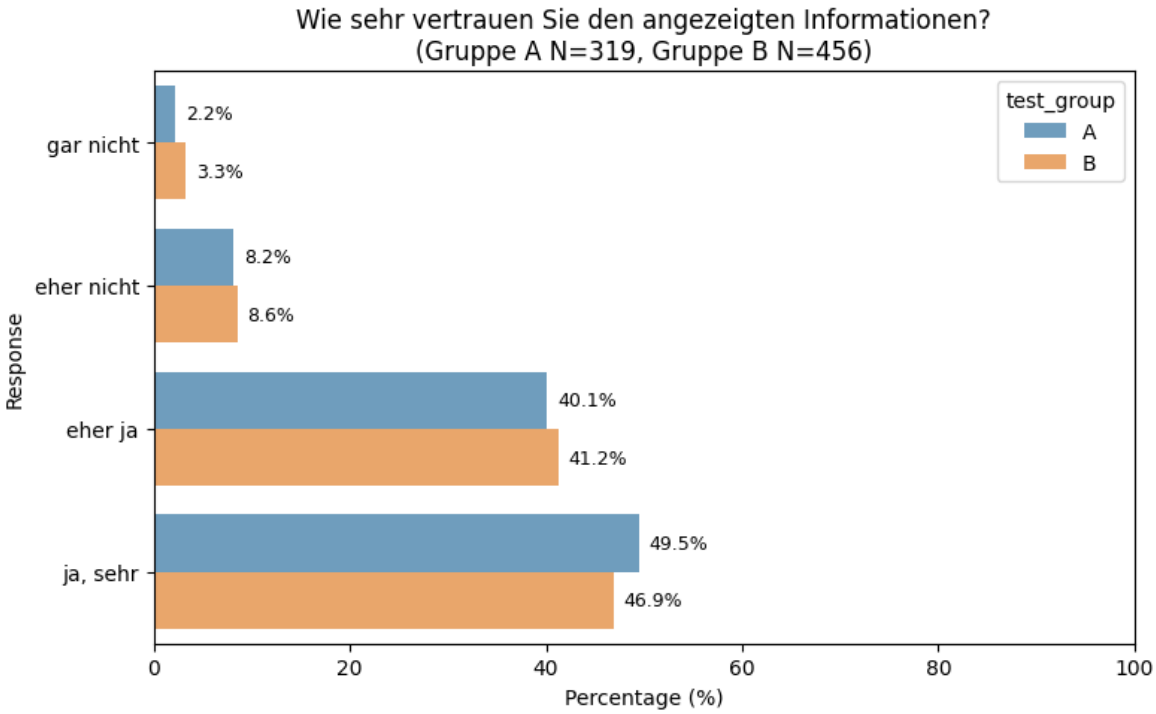

Abbildung 64 - Feedback nach Test-Gruppen: "Wie sehr vertrauen Sie den angezeigten Informationen?"

Bemerkbare Unterschiede bestehen in den Gender-Gruppen der Test-Gruppe A: Besonders selten (28,6%) vertrauten Frauen hier den angezeigten Informationen sehr, Männer hingegen in 67,4% und somit etwa 2/3 der Fälle. In 51,7% haben Frauen in der Gruppe A die Frage mit "eher ja" beantwortet, Männer in 30,2%. Insgesamt nach den Top-2 Kategorien vertrauten Männer 97,6% der Fälle den angezeigten Ergebnissen, während Frauen nur in 80,3% der Fälle.

In der Gruppe B waren die Unterschiede in den Top-2 Kategorien weniger ausgeprägt: Frauen 85,3%, Männer 91,1%. In beiden Gruppen vertrauten Frauen den angezeigten Ergebnissen weniger als Männer.

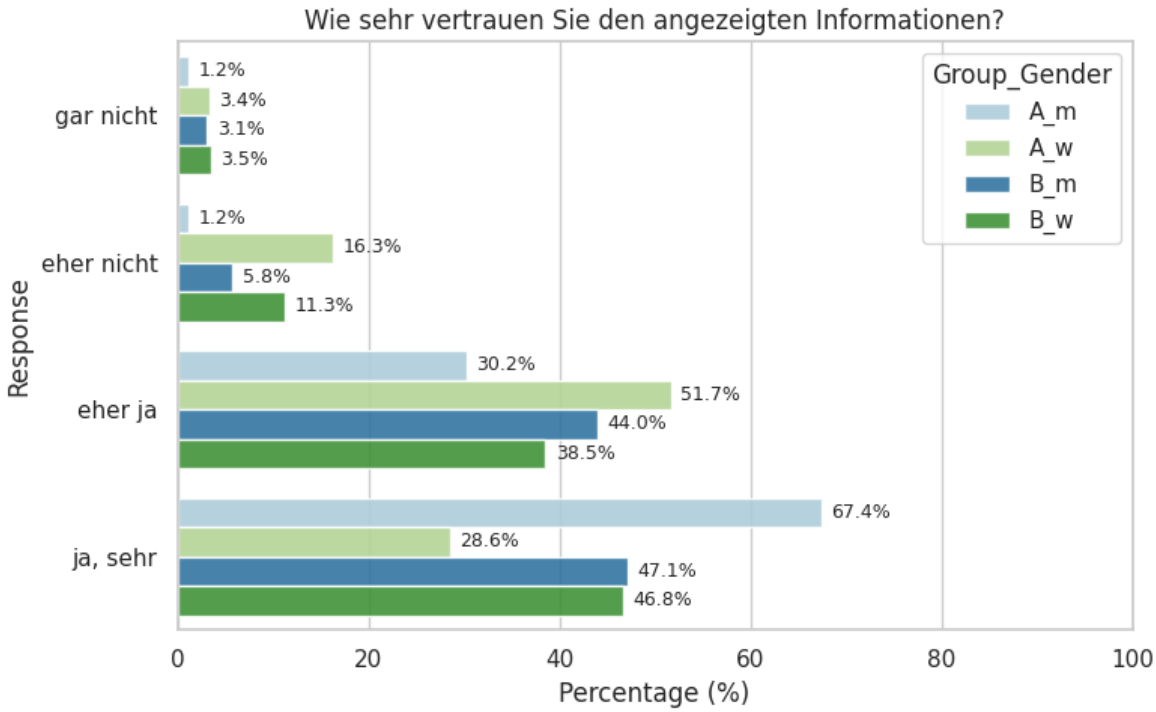

Abbildung 65- Feedback nach Test-Gruppen und Gender: "Wie sehr vertrauen Sie den angezeigten Informationen?"

11. " Haben Sie das Gefühl, die Ergebnisse nochmals gegenprüfen zu müssen?" (A, B)

Gruppe B hatte öfter das Gefühl, die Ergebnisse überprüfen zu müssen (33,2%) im Vergleich zur Gruppe A (20,8%), Abbildung 66. Dabei hatten Frauen in beiden Gruppen dieser Aussage öfter zugestimmt als Männer: Gruppe A um 11,1%, Gruppe B um 14,1%. Dies kann im Zusammenhang mit der Frage "Wie sehr vertrauen Sie den angezeigten Informationen?" als eine Bestätigung des tendenziell geringeren Vertrauen bei Frauen gesehen werden.

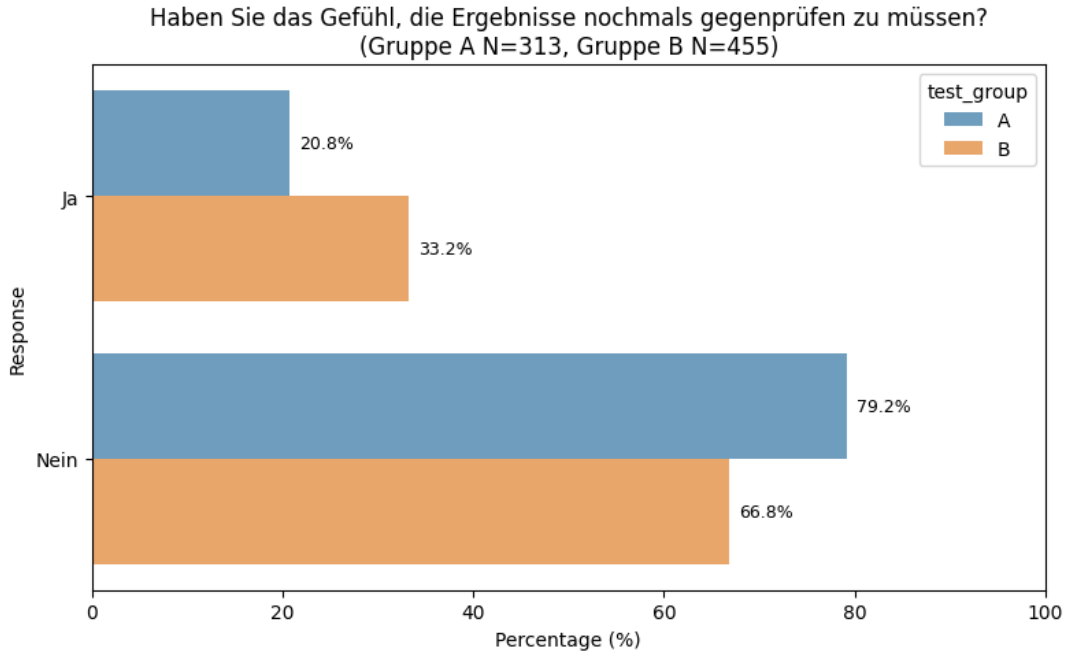

Abbildung 66 - Feedback nach Test-Gruppen: "Haben Sie das Gefühl, die Ergebnisse nochmals gegenprüfen zu müssen?"

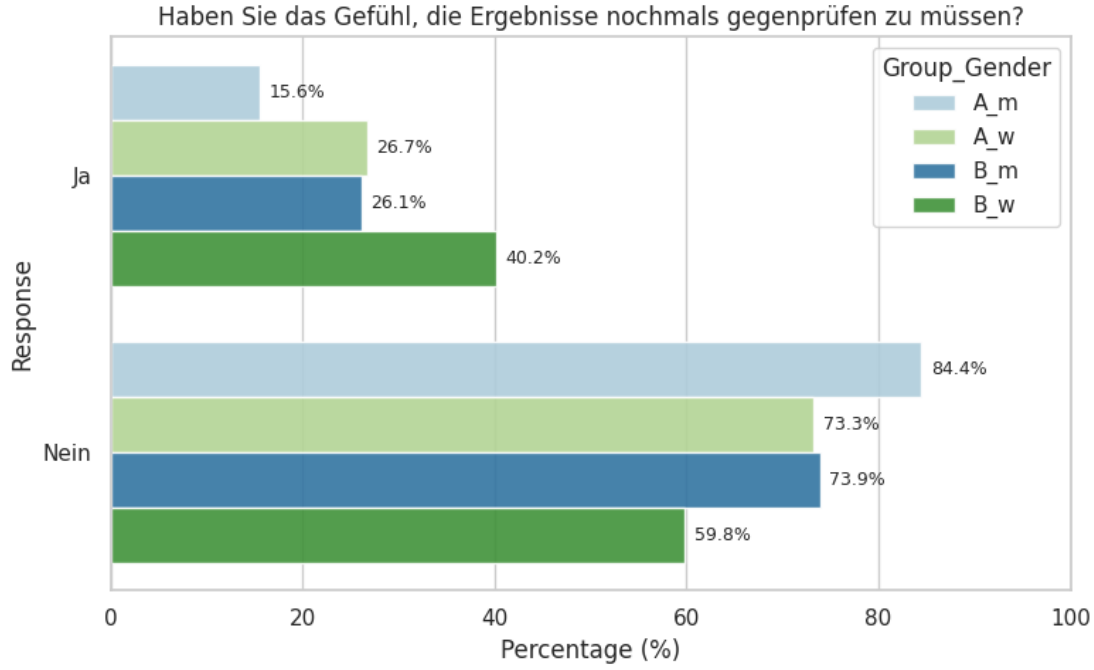

Abbildung 67- Feedback nach Test-Gruppen und Gender: "Haben Sie das Gefühl, die Ergebnisse nochmals gegenprüfen zu müssen?"

12. "Würden Sie dieses System (klassisches Intranet/ KnowHow-Tool) regelmäßig nutzen, wenn es dauerhaft verfügbar wäre?" (A, B)

Diese Frage wurde leicht unterschiedlich je nach Test-Gruppe formuliert und bezog sich jeweils auf das in der Test-Gruppe verwendete System. Gruppe B hatte eine regelmäßige Nutzung der KI-unterstützten Suche in 77,7% der Fälle positiv beantwortet. Gruppe A hatte der regelmäßigen Nutzung des Intranets in 64,8% der Fälle und somit seltener zugestimmt. In 29,3% der Fälle haben Teilnehmerinnen und Teilnehmer der Gruppe A nicht gewusst, ob sie das Intranet regelmäßig nutzen würden, Gruppe B lag mit 16,2% darunter. In ähnlich vielen Fällen wurde die regelmäßige Nutzung des jeweiligen Systems abgelehnt (A: 5,9%, B: 6,1%). Insgesamt lässt sich eine hohe Zustimmung der regelmäßigen Nutzung der KI-unterstützten Suche erkennen, wobei das klassische Intranet in fast 2/3 der Fälle auch gerne regelmäßig genutzt werden würde.

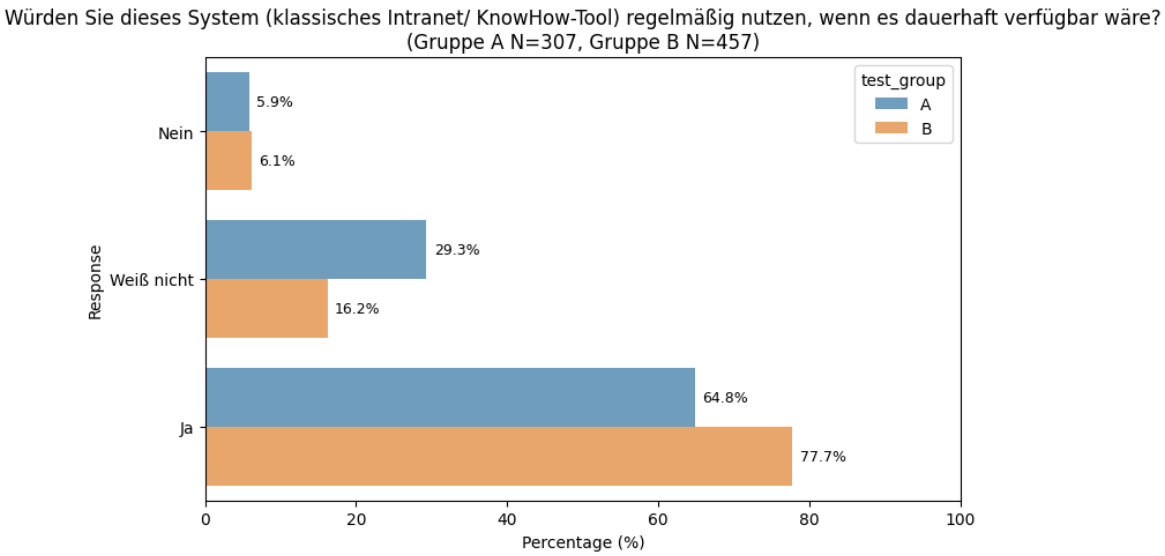

Abbildung 68 - Feedback nach Test-Gruppen: "Würden Sie dieses System (klassisches Intranet/ KnowHow-Tool) regelmäßig nutzen, wenn es dauerhaft verfügbar wäre?"

In den Gender-Gruppen sind auch in der Frage der regelmäßigen Nutzung deutliche Unterschiede bemerkbar: Insbesondere Frauen in der Gruppe A wussten eher nicht, ob sie das Intranet regelmäßig nutzen würden (54,6%) und stimmten der regelmäßigen Nutzung nur in 39,7% der Fälle zu, während Männer mit 86,1% eine hohe Zustimmung der regelmäßigen Intranet-Nutzung erteilten. Ähnlich hoch ist die Zustimmung bei Männern in der Gruppe B, hier wurde in 86,5% der Fälle der regelmäßigen Nutzung des KnowHow-Tools

zugestimmt. Frauen mit 68,9% sind zurückhaltender, jedoch ist die Zustimmung dem KnowHow-Tool deutlich höher als jene dem Intranet bei Frauen aus der Gruppe A.

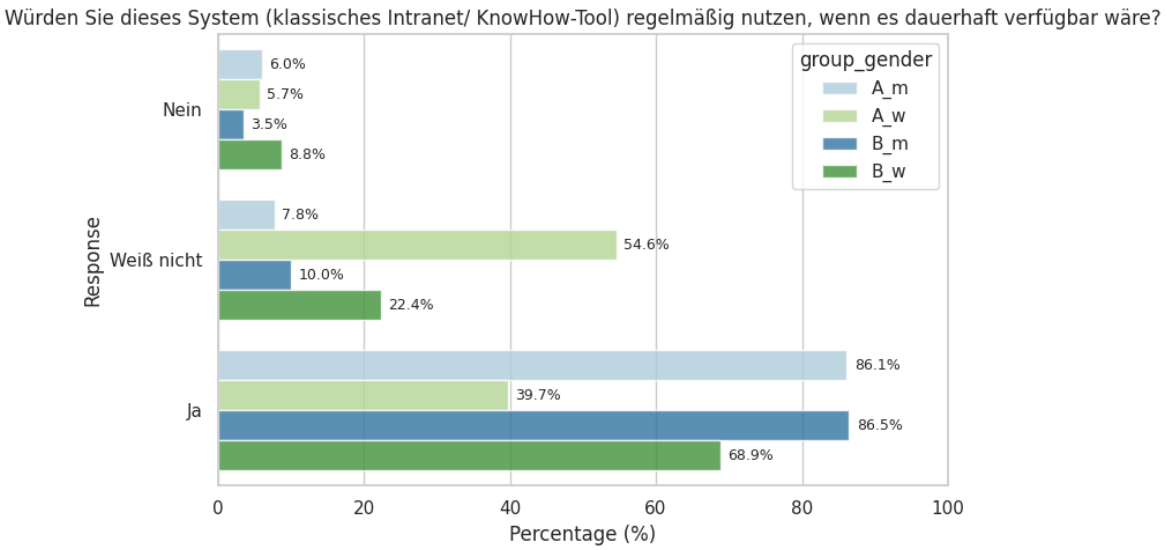

Abbildung 69- Feedback nach Test-Gruppen und Gender: "Würden Sie dieses System (klassisches Intranet/ KnowHow-Tool) regelmäßig nutzen, wenn es dauerhaft verfügbar wäre?"

13. "Was würden Sie bevorzugen, wenn Sie nur eines der beiden Systeme nutzen dürften (klassisches Intranet oder KI-gestützte Suche)?" (B)

Gefragt nach der Tool-Präferenz, hat Gruppe B in 57,8% der Fälle die KI-gestützte Suche bevorzugt. In 35,4% der Fälle wurde die Frage als "unentschieden" beantwortet. Nur in 6,8% der Fälle wurde das Intranet bevorzugt (Abbildung 70).

Gender-bezogen zeigt sich ein differenziertes Bild (Abbildung 71): Männer bevorzugten mit 83,1% klar das KnowHow-Tool während Frauen nur in 33% der Fälle dieses ausgewählt haben. In der Mehrheit der Fälle waren Frauen unentschieden (57,8%) und erteilten eine etwas höhere Zustimmung dem Intranet als Männer (9,1% vs. 4,4%).

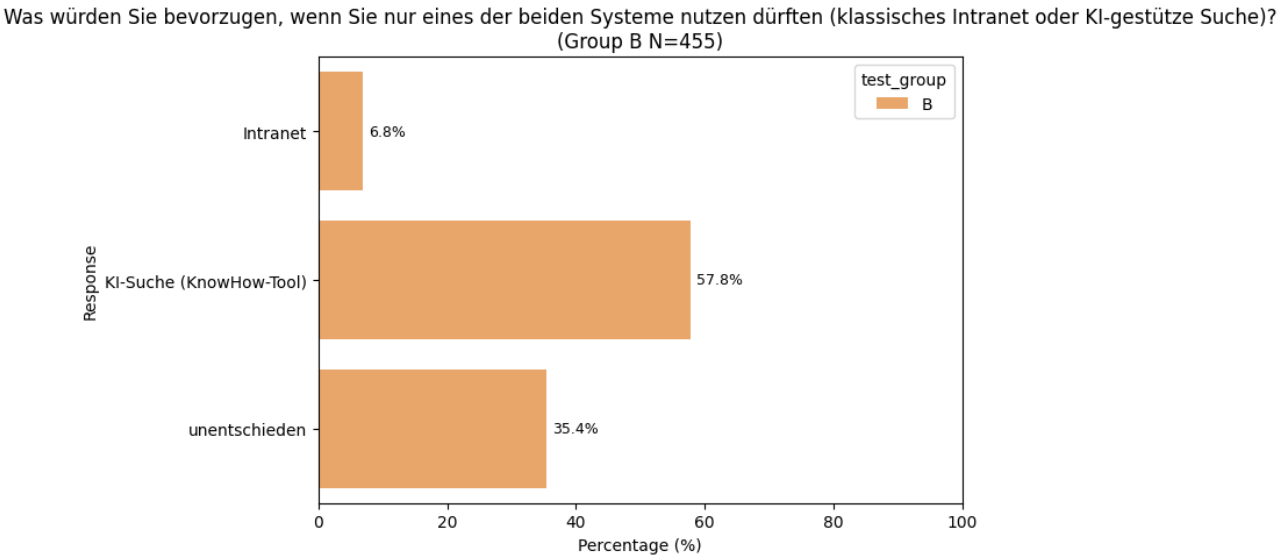

Abbildung 70 - Feedback Gruppe B: "Was würden Sie bevorzugen, wenn Sie nur eines der beiden Systeme nutzen dürften (klassisches Intranet oder KI-gestützte Suche)?"

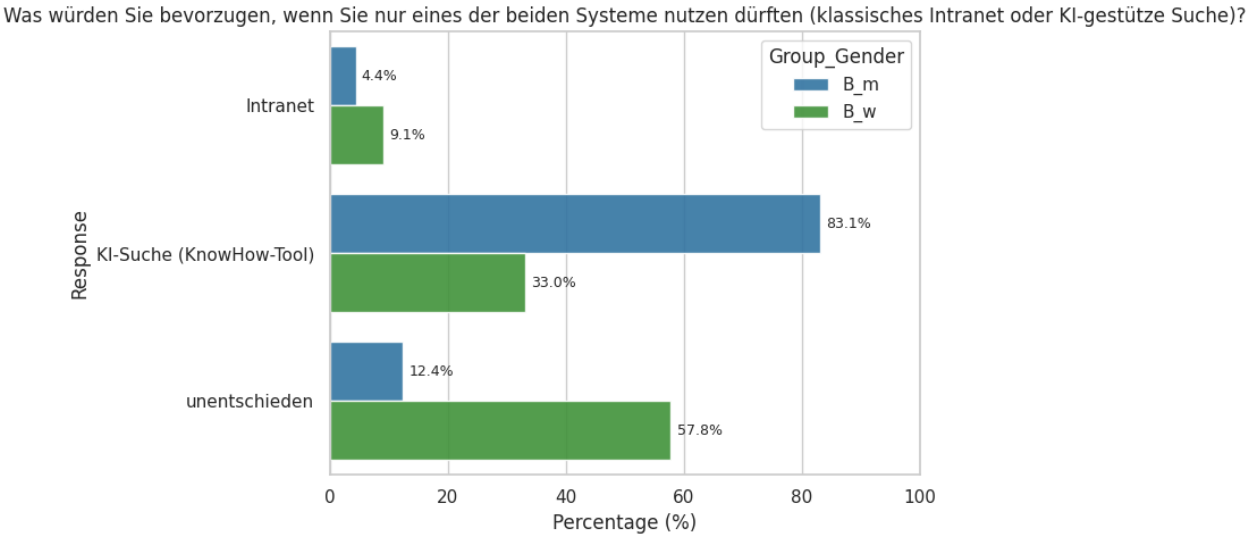

Abbildung 71 - Feedback Gruppe B nach Gender: "Was würden Sie bevorzugen, wenn Sie nur eines der beiden Systeme nutzen dürften (klassisches Intranet oder KI-gestützte Suche)?"

14. **"Würden Sie sich eine KI-gestützte Suche wünschen?" (A)**

Gruppe A würde sich in 88,3% der Fälle eine KI-gestützte Suche wünschen (Abbildung 72). Die Zustimmung ist dabei bei Frauen um 11,5% höher als bei Männern (Abbildung 73).

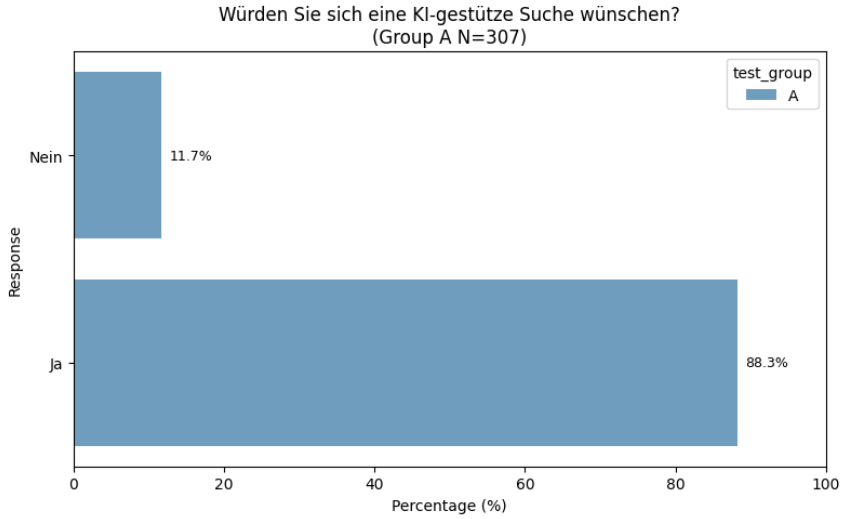

Abbildung 72 - Feedback Gruppe A: "Würden Sie sich eine KI-gestützte Suche wünschen?"

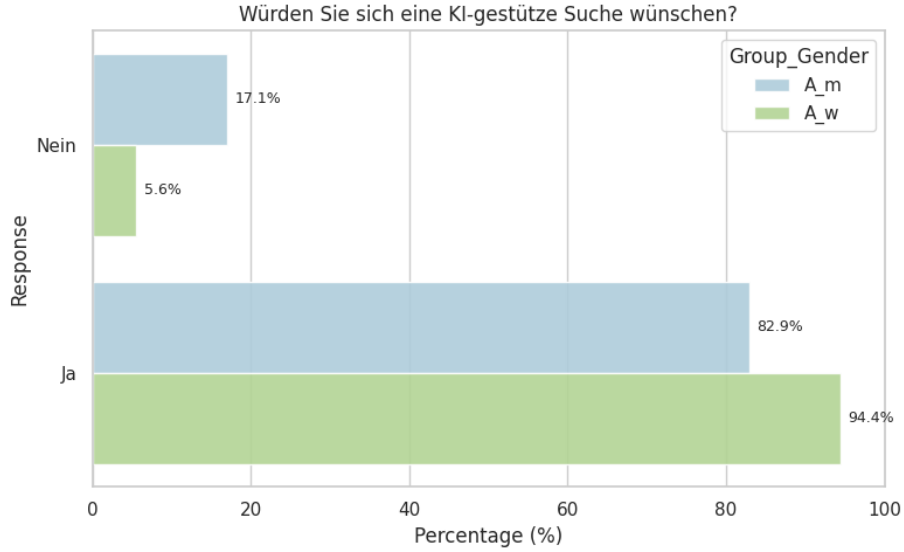

Abbildung 73 - Feedback Gruppe A nach Gender: "Würden Sie sich eine KI-gestützte Suche wünschen?"

9.5.2 Inferenzstatistik Auswahlfragen

Zur Detektion der statistisch signifikanten Unterschiede zwischen den Test-Gruppen in den Rückmeldungen wurde Feedback pro Frage und Teilnehmerin oder Teilnehmer aggregiert, um Unabhängigkeit der Datenpunkte zu gewährleisten. Für Fragen mit abgestufter Skala (bspw. "sehr zufrieden", … "gar nicht zufrieden") wurde Median ermittelt. Für dichotome Fragen ("Ja" / "Nein") und kategoriale Fragen (bspw. "Intranet"/ "KnowHow-Tool"/ "Weiß nicht") wurde Modus der Teilnehmerin oder Teilnehmer ermittelt. Für dichotome und kategoriale Fragen wurden der Chi-Quadrat-Test bzw. bei kleinen Stichproben der Fisher-Exact-Test verwendet. Für Fragen mit Skala wurde Mann-Whitney U-Test verwendet.

Mehrfache Tests erfordern eine Korrektur der p-Werte, was dazu führen kann, dass ursprünglich signifikante Ergebnisse ihre Signifikanz verlieren. Um die Anzahl der Tests zu reduzieren, werden nur Fragen mit Unterschieden von mindestens 20 % (im Median bzw. in den Antwortraten) weiter analysiert. Nach der Auswahl wurden Unterschiede in der Rückmeldungen zu 3 Fragen getestet (Tabelle 41). Die Korrektur erfolgte mit Benjamini-Hochberg False Discovery Rate (FDR) Methode.

Tabelle 41 - Inferenzstatistiken Auswahlfragen

question	type	test	n_A	n_B	abs. diff.	statistic	p_value	p_adj
Hatten Sie das Gefühl, schneller ans Ziel zu kommen, als Sie es sonst gewohnt sind?	boolean	Chi-squared	24	23	0.40	6.14	0.01	**0.04**
Wie einfach war es für Sie, die benötigten Informationen zu finden?	scale	Mann-Whitney	26	24	0.25	233.00	0.10	0.15
Wie sehr vertrauen Sie den angezeigten Informationen?	scale	Mann-Whitney	25	24	0.25	337.50	0.40	0.40

In den Rückmeldungen zur Frage " Hatten Sie das Gefühl, schneller ans Ziel zu kommen, als Sie es sonst gewohnt sind?" zeigt sich statistisch signifikanter Unterschied in den Gruppen (p-value korrigiert = 0,04). Teilnehmerinnen und Teilnehmer in der Gruppe B haben öfter das Gefühl gehabt, schneller als gewohnt zu sein (Tabelle 42).

Tabelle 42 - Aggregierte Rückmeldungen zur gefühlt schnelleren Lösungen als gewohnt

test_group/ response	A	B
Nein	18 (0.75)	8 (0.35)
Ja	6 (0.25)	15 (0.65)

9.5.3 Freitext-Feedback

Die Freitextantworten geben einen Einblick in die praktische Nutzung des Know-How Tools. Sie zeigen, wie die Nutzerinnen und Nutzer das Tool erlebt haben, wo es bereits gut funktioniert und an welchen Stellen es noch Verbesserungsbedarf gibt. Insgesamt ergibt sich ein gemischtes, aber überwiegend konstruktives Bild. Positive und negative Rückmeldungen halten sich in etwa die Waage.

Userinnen und User wurden nach der Absolvierung ihrer Aufgaben um Feedback auf die folgenden Fragen gebeten. Die Antworten sind als Freitext formuliert.

- Sonstige Kommentare zur Effizienz
- Sonstige Kommentare zur Such-Qualität
- Sonstige Kommentare zur Tool-Bedienung
- Sonstige Kommentare zur Vertrauenswürdigkeit und Sicherheit der Informationen
- Was hat Ihnen am meisten geholfen?
- Wo sehen Sie Verbesserungsbedarf?

Hier sei nochmal in Erinnerung gerufen, dass Gruppe A das Intranet und Gruppe B das KnowHow tool zur Absolvierung der Studie verwendet hat.

9.5.3.1 Sonstige Kommentare zur Effizienz

Beim Thema Effizienz zeigt sich, dass der Nutzen des Tools stark vom jeweiligen Anwendungsfall abhängt. In Gruppe A werden vor allem Struktur, Übersichtlichkeit und Suchlogik s

angesprochen. Die Rückmeldungen bleiben hier eher neutral und deuten darauf hin, dass Effizienz eng mit einer klaren und gut nachvollziehbaren Suchführung verbunden ist. In

	Gruppe A	Gruppe B
Positiv	1	3
Neutral	5	4
Negativ	1	6
Gesamt	7	13

Tabelle 43: Sonstige Kommentare zur Effizienz

Gruppe B wird häufiger hinterfragt, ob das Tool bei einfachen Suchanfragen tatsächlich einen spürbaren Vorteil gegenüber dem Intranet bietet. Gleichzeitig gibt es aber auch positive Rückmeldungen, in denen die Nutzung als schnell und unkompliziert beschrieben wird. Insgesamt spricht das dafür, dass das Tool vor allem dann als effizient wahrgenommen wird, wenn die Suchanfrage klar ist und rasch ein passendes Ergebnis gefunden wird.

Beispiele:

> “Bessere Strukturierung einzelner Themen [wäre gut]” (A)
> “Bei einfachen Suchanfragen besteht kaum ein Vorteil zum Intranet.“ (B)
> „top! easy und schnell“ (B)

9.5.3.2 Sonstige Kommentare zur Such-Qualität

Die Suchqualität ist eines der zentralen Themen in den Freitextantworten. In Gruppe A wird vor allem beschrieben, dass relevante Informationen nicht immer leicht auffindbar sind. Gruppe B berichtet zusätzlich, dass Ergebnisse teilweise falsch, unvollständig oder nur schwer überprüfbar sind. Zwar gibt es auch einzelne positive Rückmeldungen, in denen Dokumente direkt gefunden wurden, insgesamt wird die Suchqualität aber als klarer Verbesserungsbereich sichtbar. Die Qualität der Suche hat damit großen Einfluss darauf, wie hilfreich und verlässlich das Tool insgesamt erlebt wird.

	Gruppe A	Gruppe B
Positiv	0	10
Neutral	9	13
Negativ	0	9
Gesamt	9	32

Tabelle 44: Sonstige Kommentare zur Such-Qualität

Beispiele:

„Die Suche war eher schwierig.“ (A)
„Suche hat leider zu lange gedauert.“ (A)
„Das Dokument habe ich direkt gefunden“ (B)

9.5.3.3 Sonstige Kommentare zur Tool-Bedienung

Die Rückmeldungen zur Bedienung fallen gemischt aus. In Gruppe A überwiegen eher neutrale Kommentare. In Gruppe B zeigt sich ein differenzierteres Bild: Einerseits wird erwähnt, dass das Tool nach der ersten Anwendung verständlicher wird und dann auch als sinnvoll und zeitsparend wahrgenommen werden kann. Andererseits werden konkrete Schwierigkeiten bei der Nutzung genannt. Vor allem die Suchoberfläche wird teilweise als unübersichtlich oder verwirrend beschrieben, etwa wegen der vielen optionalen Eingabefelder. Insgesamt lässt sich daraus ableiten, dass das Tool grundsätzlich erlernbar ist, der Einstieg aber noch einfacher gestaltet werden sollte.

	Gruppe A	Gruppe B
Positiv	0	4
Neutral	17	53
Negativ	1	16
Gesamt	18	73

Tabelle 45: Sonstige Kommentare zur Tool-Bedienung

Beispiele:

"Anfangs verwirrend, nach der ersten Anwendung der KI aber sehr sinnvoll und zeitsparend“ (B)
„Tool Bedienung okay (klappt auch ohne die Anleitung zu lesen)“ (B)
„Die "Suche"-Seite des Tools scheint mit den vielen Feldern, die man optional ausfüllen kann, etwas verwirrend.“ (B)

9.5.3.4 Vertrauenswürdigkeit und Sicherheit der Informationen

Ein zentrales Ergebnis betrifft das Vertrauen in die bereitgestellten Informationen. In beiden Gruppen zeigt sich, dass Nutzerinnen und Nutzer die Ergebnisse häufig zusätzlich im Intranet oder in anderen Quellen überprüfen. Auch wenn die Antworten des Tools oft als plausibel wahrgenommen werden, besteht offenbar ein klares Bedürfnis nach zusätzlicher Absicherung. Das deutet darauf hin, dass die Ergebnisse derzeit noch nicht vollständig als eigenständig verlässlich eingeschätzt werden. Für die weitere Entwicklung ist es daher wichtig, die Nachvollziehbarkeit und Transparenz der Antworten weiter zu verbessern.

	Gruppe A	Gruppe B
Positiv	0	1
Neutral	1	5
Negativ	1	2
Gesamt	2	8

Tabelle 46: Vertrauenswürdigkeit und Sicherheit der Informationen

Beispiele:

“Ich prüfe alle Antworten immer noch einmal nach, das es im Intranet sehr viel tote Akten gibt und ich ja die letzte Information zu jedem Thema benötige“ (A)
„Ergebnis klingt plausibel“ (B)
„Hätte zur Verifizierung kurz das Dok durchgeblättert“ (B)

9.5.3.5 Was hat Ihnen am meisten geholfen?

	Gruppe A	Gruppe B
Positiv	7	1
Neutral	1	2
Negativ	3	1
Gesamt	11	4

Tabelle 47: Was hat Ihnen am meisten geholfen?

Gruppe A zeigt sich vor allem, dass klare oder vorgegebene Suchbegriffe zu schnellen und passenden Ergebnissen führen können. In Gruppe B werden ebenfalls hilfreiche Aspekte genannt, gleichzeitig wird dort aber häufiger auf andere Informationsquellen verwiesen, zum Beispiel auf das Intranet oder externe Websites. Das zeigt, dass das Tool in bestimmten Situationen einen spürbaren Nutzen bietet, dieser Nutzen aber nicht in allen Fällen deutlich über alternativen Quellen liegt.

Beispiele:

„Vorgegebener Suchbegriff führt zu raschem und passendem Ergebnis.“ (A)
„Sofern das KI-Tool einen Mehrwert gegenüber dem Intranet bietet, bevorzuge ich natürlich die KI-Suche.“ (B)
„Diese Antwort wurde nicht mit Know How Tool entworfen, sondern mit google/website BMEIA.“ (B)
„Am meisten hat mir das Intranet … geholfen.“ (B)

9.5.3.6 Wo sehen Sie Verbesserungsbedarf?

Auch beim Verbesserungsbedarf zeigen sich klare Schwerpunkte. Gruppe A nennt vor allem Verbesserungen bei der Suche, etwa eine bessere Kennzeichnung der aktuellen oder letztgültigen Version eines Dokuments. Gruppe B spricht darüber hinaus auch technische und funktionale Punkte an, zum Beispiel die Erweiterung der Informationsquellen oder eine bessere Darstellung und Zusammenfassung der Suchergebnisse. Insgesamt wird damit deutlich, dass es nicht nur um die Suchqualität selbst geht, sondern auch um die Aufbereitung der Ergebnisse und die Einbindung relevanter Inhalte.

	Gruppe A	Gruppe B
Positiv	0	0
Neutral	22	9
Negativ	0	9
Gesamt	22	18

Tabelle 48: Wo sehen Sie Verbesserungsbedarf?

Beispiele:

„Kennzeichnung der letztgültigen (aktuellen) Version“ (A)
„Erweiterung der Informationsquellen. Einbindung der BMEIA-Homepages in die Suchmaschine.“ (B)
„Darstellung / Zusammenfassung in Suche” (B)

9.6 Interpretation und Zusammenfassung

Im Rahmen der explorativen Datenanalyse wurden Unterschiede in den Zeitmessungen in den Test-Gruppen, aufgeteilt nach Gender- und Dienstalter-Subgruppen, Aufgaben-Typen betrachtet. Gesamt gesehen sind die Zeitmessungen in den Gruppen ähnlich verteilt. Dies wurde durch die Inferenzstatistiken (U-Test und Welch T-Test) auf den aggregierten Median-Zeitwerten der Teilnehmerinnen und Teilnehmer bestätigt, es wurden keine statistisch signifikanten Unterschiede zwischen den Test-Gruppen festgestellt.

Aufgrund einer möglichen natürlichen Veränderung in den Daten, bedingt durch den abgeschlossenen Aktualisierungsprozess des KnowHow-Tools wurden außerdem Tests auf dem verringerten Datensatz ab dem Zeitpunkt des Updates-Ende durchgeführt. Hier konnten ebenfalls keine signifikanten Unterschiede in den Zeitmessungen unter Voraussetzung, dass Aufgaben gelöst wurden, in den Test-Gruppen festgestellt werden.

Ergänzend zu den Bearbeitungszeiten wurden die Erfolgsraten in den Test-Gruppen analysiert. Während Gruppe B etwas niedrigere Erfolg-beim-ersten-Versuch Rate hat, was womöglich auf die mangelnde Erfahrung im Umgang mit dem neuen Tool zurückzuführen ist, sind die allgemeinen Erfolgsraten (ungeachtet der Anzahl Versuche) in den Test-Gruppen vergleichbar.

Explorativ wurden mögliche Lerneffekte, die insbesondere bei der Gruppe B ggf. zu erwarten waren, betrachtet, dies auf allen Aufgaben und auf den Dokument-Suche Aufgaben. Während kein stetiger Abwärtstrend erkannt werden konnte, können Lerneffekte nicht ausgeschlossen werden, da die aggregierten Beobachtungen trotz der Z-Standardisierung vom Aufgabenmix beeinflusst werden. Bezogen auf die Dokument-Suche Aufgaben kann anfangs ein Lerneffekt in der Gruppe B vermutet werden.

In der explorativen Datenanalyse zeigten sich Unterschiede in den Verteilungen betrachtet nach Gender- und Dienstalter-Subgruppen, diese konnten jedoch aufgrund der geringen Stichprobengrößen nicht interpretiert werden. Mögliche Effekte wurden mit Cox's Proportional Hazard-Regressionsmodellen analysiert. Dieses Verfahren modelliert Zeit-bis-zum-Ereignis (Zeit bis zur Lösung der Aufgabe) unter Berücksichtigung der zensierten (right-censored) Beobachtungen, die in U-Test und T-Test nicht miteinbezogen werden können. Es konnten dabei keine signifikanten Effekte in Bezug auf das Geschlecht der Studienteilnehmerinnen und Studienteilnehmer bzw. Interaktionen mit anderen Kovariaten festgestellt werden.

Das Dienstalter zeigte in der explorativen Analyse einen nicht-linearen Zusammenhang mit den Bearbeitungszeiten, wobei die Bearbeitungszeiten in beiden Test-Gruppen anfangs sinken mit dem zunehmenden Dienstalter und steigen später wieder an. Dienstalter als Kovariate des Regressionsmodell wurde daher mit einem Basis-Spline modelliert, um die nicht-linearen Effekte abbilden zu können. Koeffizienten des Dienstalter-Splines zeigten statistische Signifikanz der Kovariate.

Die Annahmen, dass die dienst-jüngeren Teilnehmerinnen und Teilnehmer von der neuen Technologie und die dienst-älteren hingegen vom klassischen Intranet besonders profitieren würden, fanden keine Bestätigung in Rahmen der Datenexploration. Auch die Interaktionen des Dienstalter-Splines mit der Test-Gruppe zeigten keine signifikanten Effekte in den Regressionsmodellen.

In Bezug auf die Aufgabentypen zeigten sich deutliche Unterschiede in den Bearbeitungszeiten. Wobei die Dokument-Suche in beiden Gruppe tendenziell am wenigsten Zeit in Anspruch genommen hat, waren die Inhalts-Suche Aufgaben schwieriger, dies insbesondere in der Gruppe B, und die Text-Erstellung für beide Gruppen deutlich schwieriger als andere Aufgabentypen. Während die Regressionsanalyse mit der Stratifizierung nach dem Aufgabentyp die Ergebnisse des U-Tests und Welch T-Tests bestätigte, indem hier kein signifikanter Effekt der Test-Gruppe festgestellt wurde, wurden unter Einbeziehung des Aufgabentyps als Kovariate Unterschiede in den Gruppen bemerkbar.

In Bezug auf die Haupteffekte konnte festgestellt werden, dass Dokument-Suche im Vergleich zur Referenz, Inhalts-Suche, in der Referenz-Gruppe A die Zeit-bis-zur-Lösung signifikant beschleunigt. Diese Kovariate verletzt jedoch die Proportional-Hazard-Annahme. Dokument-Suche steht in Zusammenhang mit höherem Hazard im niedrigeren Zeitwert-Bereich und sinkt später, was darauf hindeutet, dass Dokument-Suche-Aufgaben in vielen Fällen schnell abgeschlossen werden konnten, während sich länger andauernde Aufgaben im Zeitverlauf zunehmend der Referenz annähern. Die Text-Erstellung als Aufgabentyp verlängert hingegen die Zeit-bis-zur-Lösung signifikant im Verhältnis zur Inhalts-Suche in der Referenz-Gruppe A. Die beiden Effekte sind aufgrund des Aufgaben-Designs erwartbar und entsprechen den Beobachtungen während der Datenexploration.

In der Gruppe B führt der Referenz-Aufgabentyp Inhalts-Suche zu längerer Zeit-bis-zur-Lösung im Verhältnis zur Gruppe A. Bei der Dokument-Suche hat Gruppe B hingegen den Vorteil gegenüber der Referenzgruppe A, die Kombination der Koeffizienten zeigt eine verkürzte Zeit-bis-zur-Lösung. Diese Effekte sind signifikant. Bei der Text-Generation verlängert sich die Zeit-bis-zur-Lösung in der Gruppe B im Verhältnis zu A, wobei hier der Interaktions-Effekt ein p-value an der Signifikanzgrenze aufweist.

Zusammenfassend lässt sich sagen, dass die Test-Gruppen über alle Aufgabentypen hinweg betrachtet keine statistisch signifikanten Unterschiede in den Bearbeitungszeiten aufweisen. Es konnten keine Gender-Effekte festgestellt werden. Dienstalter zeigte in der Regressionsanalyse einen signifikanten nicht-linearen Effekt, dieser war unabhängig von der

Test-Gruppe. Unter Berücksichtigung des Aufgabentyp in der Modellierung als Kovariate waren Unterschiede in den Test-Gruppen signifikant: Gruppe B profitierte vom Aufgabentyp Dokument-Suche, war jedoch langsamer bei der Inhalts-Suche als die Referenz-Gruppe A.

9.6.1 Zusammenfassung Feedback

In den Auswahlfragen zeigt sich eine positive Wahrnehmung in Bezug auf das KnowHow-Tool. So haben bspw. Teilnehmerinnen und Teilnehmer aus der Gruppe B die Einfachheit der Suche (82%) und die Übersichtlichkeit der Oberfläche (80%) besser als Gruppe A mit je 75%, 66% in Top-2 Kategorien bewertet. Gruppe B hat wesentlich öfter als A das Gefühl gehabt, schneller ans Ziel zu kommen als gewohnt (66% vs. 13%), der Unterschied in dieser Frage war nach dem Chi-Quadrat-Test auf aggregierten Daten statistisch signifikant. Beider Gruppen haben ähnlich oft Umwege gebraucht, um Informationen zu finden, Gruppe B jedoch etwas seltener.

Gruppe B war meist der Meinung, dass das KI-Tool die Frage richtig verstanden hat (84%). Gruppen haben ähnliche Bewertungen abgegeben in Bezug darauf, wir relevant oder hilfreich die Ergebnisse waren, so fand Gruppe B in über 80% der Fälle die Ergebnisse hilfreich. Zu fast 70% haben Teilnehmerinnen und Teilnehmer der Gruppe B die Informationen direkt gefunden, in 25% der Fälle mussten sie nachjustieren. Gruppen haben keine starken Unterschiede in den subjektiv wahrgenommenen Bearbeitungszeiten gezeigt, wobei Gruppe B etwas öfter das Gefühl hatte, weniger als eine Minute gebraucht zu haben (+5%), dafür seltener zwischen einer und 3 Minuten und öfter zwischen 3 und 5 Minuten benötigt zu haben.

Beide Gruppen vertrauten den angezeigten Ergebnissen ähnlich stark, das Vertrauen in der Gruppe B war leicht niedriger (-1,5% in den Top-2 Kategorien). Gruppe B hatte außerdem öfter das Gefühl, die Ergebnisse nochmals gegenprüfen zu müssen (33% zu 21%).

In Bezug auf die Tool-Präferenz bevorzugte Gruppe B mit 58% mehrheitlich das KnowHow-Tool (nur 7% Intranet, 35% "unentschieden") und würde dieses auch eher regelmäßig nutzen als Gruppe A das Intranet (78% vs. 65%). Gruppe A würde sich zu 88% eine KI-gestützte Suche wünschen.

Betrachtet nach Gender sind in beiden Test-Gruppen, insbesondere aber in der Gruppe A deutliche Unterschiede erkennbar. Frauen bewerten ihre Erfahrungen mit dem Intranet

tendenziell weniger positiv als Männer. Frauen in der Gruppe A bewerteten die Einfachheit der Suche seltener als Männer mit "sehr einfach", nach Top-2 Kategorien waren Frauen auch weniger zufrieden mit der Einfachheit der Suche in beiden Gruppen. Frauen waren weniger zufrieden mit der Übersichtlichkeit der Oberfläche, insbesondere deutlich in der Gruppe A. Frauen haben jeweils etwas öfter Umwege gehen müssen, dies mehr in der Gruppe A im Vergleich zu Männern.

In der Gruppe B waren Frauen seltener überzeugt, dass das KI-Tool ihre Frage richtig verstanden hat, und sie fanden die Ergebnisse auch etwas seltener hilfreich als Männer. In der Gruppe A, gefragt nach der Relevanz der Ergebnisse, waren Frauen deutlich weniger zufrieden mit der Relevanz der Ergebnisse als Männer, insbesondere was die Bewertung "sehr relevant" betrifft, was auf einen möglichen gender-spezifischen Unterschied betreffend der wahrgenommenen Relevanz der Suchergebnisse besteht.

Frauen in der Gruppe B haben die Informationen etwas seltener als Männer direkt gefunden und haben öfter nachjustieren müssen. In Bezug auf die subjektive Einschätzung der erforderlichen Bearbeitungszeit waren Frauen in der Gruppe A weniger optimistisch als Männer und haben seltener gefühlt unter 1 Minute für die Aufgabe gebraucht, während Frauen in der Gruppe B leicht über den Männer-Bewertungen liegen. Männer aus der Gruppe A haben nie länger als 10 Minuten gebraucht, hingegen Frauen in 2,6% der Fälle. In der Gruppe A haben Frauen mit fast 20% etwa 3-mal so oft wie Männer das Gefühl gehabt, Lösungen schneller als gewohnt zu finden, in der Gruppe B hingegen etwas weniger oft als Männer.

In beiden Gruppen vertrauten Frauen den angezeigten Ergebnissen weniger als Männer, der Unterschied ist jedoch bei der Gruppe A größer, insbesondere auffällig selten im Vergleich zu Männern vertrauten Frauen den angezeigten Informationen "sehr". Ebenso haben Frauen jeweils eher das Gefühl gehabt, die Ergebnisse nochmals gegenprüfen zu müssen, dies etwas mehr in der Gruppe B im Vergleich zu Männern als in der Gruppe A.

Starke Unterschiede sind in der Tool-Präferenz bemerkbar: In Mehrheit der Fälle (55%) wussten Frauen in der Gruppe A wussten nicht, ob sie das Intranet regelmäßig nutzen würden und stimmten der regelmäßigen Nutzung nur in etwa 40% der Fälle zu (Männer 86%). Während die Zustimmung bei Männern in der Gruppe B der regelmäßigen Nutzung des KnowHow-Tools ähnlich hoch war, war der Unterschied zu Frauen (69% Zustimmung) deutlich geringer. In der Gruppe B, gefragt nach der exklusiven Präferenz (entweder das KnowHow-Tool oder das Intranet), bevorzugten Männer klar das KnowHow-Tool (83%),

Frauen hingegen mi 33% viel seltener. In der Mehrheit der Fälle waren Frauen unentschieden (58%). Anders ist das Bild in der Gruppe A: Hier würden sich Frauen mit 94% öfter als Männer (83%) eine KI-Suche wünschen.

Insgesamt lässt sich sagen, dass das KnowHow-Tool von Userinnen und Usern in allen befragten Aspekten positiv wahrgenommen wurde und nach subjektiver Einschätzung schnellere Lösungen ermöglichte. Die Vertrauensaspekte werden etwas weniger hoch bewertet im Vergleich zum Intranet. Gruppe A würde sich klar eine KI-gestützte Suche wünschen. Frauen bewerten beide Systeme tendenziell weniger positiv als Männer, jedoch sind diese Unterschiede höher in der Gruppe A: Frauen sind in manchen Aspekten deutlich weniger zufrieden mit dem Intranet als Männer. Frauen (Gruppe A) würden sich öfter als Männer KI-Suche wünschen, als einziges Tool (Gruppe B) es jedoch wesentlich seltener als Männer wählen.

Die Freitextantworten zeigen, dass das KnowHow-Tool vor allem dann als hilfreich erlebt wird, wenn mit klaren Suchbegriffen gearbeitet wird und komplexere Fragen bearbeitet werden. Der wahrgenommene Mehrwert hängt jedoch stark davon ab, ob die Ergebnisse nachvollziehbar, überprüfbar und ohne größere Hürden zugänglich sind. Insgesamt überwiegen neutrale und konstruktive Rückmeldungen. Gleichzeitig wird deutlich, dass in den Bereichen Suchqualität, Bedienbarkeit und Vertrauenswürdigkeit noch Entwicklungspotenzial besteht.

10 Fazit

Die KnowHow-Studie untersuchte, wie die Anwendung von modernen KI-basierten Techniken, insbesondere semantischer Suche, LLMs und RAG-Systeme, Knowledge Graphs ein effizientes Wissensmanagementsystem für das BMEIA unterstützen könnte und die Effizienz des KI-unterstützten prototypischen Systems gegenüber dem klassischen BMEIA Wissensmanagementsystem (BMEIA-Intranet) zu quantifizieren und zu evaluieren. Zu diesem Zweck wurde ein A/B-Testing Design aufgesetzt, wobei Gruppe A (Kontroll-Gruppe) im Verlauf der Evaluierungsphase das BMEIA-Intranet verwendete und Gruppe B den im Projekt entwickelten Demonstrator.

Im Rahmen der Literatur- und Marktrecherche wurden verfügbare Techniken und Ansätze und deren Anwendbarkeit und Mehrwert in Bezug auf die Problemstellung des Bedarfsträgers analysiert. In einem iterativen Prozess unter Mitwirkung der Expertinnen und Experten des BMEIA, AIT und CIB wurden Anforderungen an den Demonstrator und das Evaluierungsframework formuliert, welche eine Grundlage für die Spezifikation des KnowHow-Tools und des Evaluierungsdesigns bildeten.

Das KnowHow-Tool wurde entsprechend des Spezifikationen als eine Docker-Compose-Applikation umgesetzt und integrierte die am AIT entwickelte Knowledge Graph-Komponente. Die Evaluierungsumgebung wurde als eine separate Docker-Compose-Applikation aufgesetzt. Im Rahmen mehrerer Workshops wurden 25 Test-Cases entwickelt, deren Design als Dokument-, Inhalts-Suche und Text-Erstellung das Testen und Evaluieren verschiedener Funktionen erlaubte.

Beide Applikationen wurden in der BMEIA-Ziel-Umgebung in Absprache und Zusammenarbeit mit BMEIA-Expertinnen und - Experten deployt und getestet. Vor dem Beginn der Evaluierungsstudie wurde Daten-Ingest durchgeführt. Studienteilnehmerinnen und Studienteilnehmer wurden im Umgang mit Tools eingeschult und mit Schulungsunterlagen sowie laufendem Support während der Evaluierung unterstützt.

Im Verlauf der Evaluierungsphase (12.01.2026 - 22.02.2026) wurden anonymisierte Daten über die Bearbeitungsvorgänge gesammelt. Ergebnisse aus der Evaluierungsphase, inkl.

Feedback der Teilnehmerinnen und Teilnehmer, wurden bereinigt, transformiert und sowohl explorativ als auch mit Inferenzstatistiken und mittels geeigneter Regressionsmodelle analysiert.

Es konnten keine statistisch signifikanten Unterschiede unter der Berücksichtigung aller Aufgabentypen zwischen den Test-Gruppen festgestellt werden. Dieses Ergebnis war konsistent über verschiedene Analyse-Verfahren. Im Rahmen der Regressionsanalyse konnten signifikante Effekte der Test-Gruppen in verschiedenen Aufgabentypen festgestellt werden, wobei Gruppe B (KnowHow-Tool) einen Vorteil in der Dokument-Suche hatte und einen Nachteil in der Inhalts-Suche im Vergleich zur Kontrollgruppe. Statistisch signifikante Gender-Effekte konnten nicht nachgewiesen werden.

Die Rückmeldungen der Teilnehmerinnen und Teilnehmer in Bezug auf das KnowHow-Tool waren überwiegend positiv, in mehreren Aspekten wurde der Technologie-Demonstrator positiver bewertet als Intranet. Eine KI-gestützte Suche wurde in den Fragebögen präferiert (Gruppe B) bzw. gewünscht (Gruppe A). In den Gender-Subgruppen fielen die Rückmeldungen der Frauen in beiden Test-Gruppen und über mehrere Aspekte hinweg weniger positiv aus als jene der Männer, dies jedoch insbesondere in der Gruppe A. Während Frauen in vielen Aspekten wesentlich unzufriedener mit dem BMEIA-Intranet waren als Männer, waren die Unterschiede in der Gruppe B deutlich geringer.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Y. Gao u. a., „Retrieval-Augmented Generation for Large Language Models: A Survey“, arXiv, arXiv:2312.10997, März 2024. Zugegriffen: 26. April 2024. [Online]. Verfügbar unter: http://arxiv.org/abs/2312.10997

[2] Y. Wang, N. Lipka, R. A. Rossi, A. Siu, R. Zhang, und T. Derr, „Knowledge Graph Prompting for Multi-Document Question Answering“, 25. Dezember 2023, arXiv: arXiv:2308.11730. Zugegriffen: 28. November 2024. [Online]. Verfügbar unter: http://arxiv.org/abs/2308.11730

[3] D. Edge u. a., „From Local to Global: A Graph RAG Approach to Query-Focused Summarization“, 24. April 2024, arXiv: arXiv:2404.16130. Zugegriffen: 28. November 2024. [Online]. Verfügbar unter: http://arxiv.org/abs/2404.16130

[4] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, und C. D. Manning, „RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval“, 31. Januar 2024, arXiv: arXiv:2401.18059. doi: 10.48550/arXiv.2401.18059.

[5] M. Glass, G. Rossiello, M. F. M. Chowdhury, A. R. Naik, P. Cai, und A. Gliozzo, „Re2G: Retrieve, Rerank, Generate“, 13. Juli 2022, arXiv: arXiv:2207.06300. Zugegriffen: 22. Oktober 2024. [Online]. Verfügbar unter: http://arxiv.org/abs/2207.06300

[6] J. Hsia, A. Shaikh, Z. Wang, und G. Neubig, „RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems“, 12. August 2024, arXiv: arXiv:2403.09040. Zugegriffen: 14. Oktober 2024. [Online]. Verfügbar unter: http://arxiv.org/abs/2403.09040

[7] D. Rau u. a., „BERGEN: A Benchmarking Library for Retrieval-Augmented Generation“, 1. Juli 2024, arXiv: arXiv:2407.01102. Zugegriffen: 14. Oktober 2024. [Online]. Verfügbar unter: http://arxiv.org/abs/2407.01102

[8] J. J. Pan, J. Wang, und G. Li, „Survey of Vector Database Management Systems“, 21. Oktober 2023, arXiv: arXiv:2310.14021. doi: 10.48550/arXiv.2310.14021.

[9] M. B. Chouikha Zouari und S. B. Dhaou Dakhli, „A Multi-Faceted Analysis of Knowledge Management Systems“, Procedia Comput. Sci., Bd. 138, S. 646–654, 2018, doi: 10.1016/j.procs.2018.10.086.

[10] M. H. Jarrahi, D. Askay, A. Eshraghi, und P. Smith, „Artificial intelligence and knowledge management: A partnership between human and AI“, Bus. Horiz., Bd. 66, Nr. 1, S. 87–99, Jan. 2023, doi: 10.1016/j.bushor.2022.03.002.

[11] S. Packowski, I. Halilovic, J. Schlotfeldt, und T. Smith, „Optimizing and Evaluating Enterprise Retrieval-Augmented Generation (RAG): A Content Design Perspective“, 1. Oktober 2024, arXiv: arXiv:2410.12812. doi: 10.48550/arXiv.2410.12812.

[12] P. Jiang, W. Niu, Q. Wang, R. Yuan, und K. Chen, „Understanding Users' Acceptance of Artificial Intelligence Applications: A Literature Review“, Behav. Sci., Bd. 14, Nr. 8, S. 671, Aug. 2024, doi: 10.3390/bs14080671.

[13] K. Tomanek, S. Cai, und S. Venugopalan, „Parameter Efficient Tuning Allows Scalable Personalization of LLMs for Text Entry: A Case Study on Abbreviation Expansion“, 21. Dezember 2023, arXiv: arXiv:2312.14327. doi: 10.48550/arXiv.2312.14327.

[14] S. Weiss, „Enhancing Diversity in RAG Document Retrieval Using Projection-Based Techniques“, Medium. Zugegriffen: 4. Dezember 2024. [Online]. Verfügbar unter: https://medium.com/@samcarlos_14058/enhancing-diversity-in-rag-document-retrieval-using-projection-based-techniques-9fef5422e043

[15] CHAT doc, 2 2025. [Online]. Available: https://chatdoc.com/.

[16] 2 2025. [Online]. Available: https://sharly.ai/ai-summarizer.

[17] 2 2025. [Online]. Available: https://askyourpdf.com/de.

[18] pdf.ai, 2 2025. [Online]. Available: https://pdf.ai/.

[19] MATHEMA GmbH, 2 2025. [Online].

[20] textcortex, 2 2025. [Online]. Available: https://textcortex.com/de/chatdoc-alternative.

[21] AI for Verticals, 2 2025. [Online]. Available: https://docanalyzer.ai/.

[22] acto, 2 2025. [Online]. Available: https://www.heyacto.com/blog-posts/chat-with-your-data.

[23] inovex, 2 2025. [Online]. Available: https://www.inovex.de/de/chat-with-your-data/.

[24] LlamaIndex, „LlamaDeploy Documentation,“ [Online]. Available: https://docs.llamaindex.ai/en/v0.12.15/module_guides/llama_deploy/.

[25] LlamaIndex, „Documentation,“ [Online]. Available: https://developers.llamaindex.ai/python/framework/.

[26] OpenSearch, „Approximate k-NN search,“ [Online]. Available: https://docs.opensearch.org/latest/vector-search/vector-search-techniques/approximate-knn/.

[27] E. S. Labs, „Diversifying search results with Maximum Marginal Relevance (MMR),“ [Online]. Available: https://www.elastic.co/search-labs/blog/maximum-marginal-relevance-diversify-results.

Abkürzungen

KI	Künstliche Intelligenz
RAG	Retrieval Augmented Generation
LLM	Large Language Model
KMS	Knowledge Management System
KM	Knowledge Management
QA	Question Answering
OCR	Optical Character Recognition
OE	Organisationseinheit
RE	Runderlass
UX	User Experience
NER	Named Entity Recognition
US	User Story
KG	Knowledge Graph
MMR	Maximum Marginal Relevance

AIT Austrian Institute of Technology GmbH
Giefinggasse 4, 1210 Wien
+43 50 550-0
office@ait.ac.at
https://www.ait.ac.at/

CIB solutions GmbH
Hadikgasse 64, 1040 Wien
+43 361 5539-0
https://www.cib.de/