

Realitätscheck bestanden: Wie ein lokal betriebenes GPT-Modell die Kanzleiarbeit verändert

Open Source trifft
Datenschutz & Compliance –
Generative KI für mehr
Sicherheit, Präzision und
Effizienz

*Ein gemeinsames Whitepaper von
Christian Ammer in Zusammenarbeit mit
Fsas Technologies – a Fujitsu company*



Executive Summary

Generative KI: Eine transformative Technologie für den Rechtssektor

Der Rechtssektor steht vor enormen Herausforderungen. Zunehmende regulatorische Anforderungen, Internationalisierung und die Komplexität juristischer Aufgaben treffen auf einen Engpass an hervorragenden Nachwuchsfachkräften. Gleichzeitig bietet die Digitalisierung die Chance, Arbeitsprozesse effizienter und präziser zu gestalten. Generative KI hat sich dabei als Schlüsseltechnologie etabliert, die insbesondere in der textlastigen Arbeit von Kanzleien, Verwaltungen und Rechtsabteilungen ihre Stärken ausspielt.

Mit der Fujitsu Private GPT-Lösung hat Fujitsu eine Plattform entwickelt, die speziell auf die Anforderungen des Rechtssektors zugeschnitten ist. Im Gegensatz zu marktüblichen cloubasierten Systemen bietet Private GPT eine vollständig On-Premise-basierte Lösung mit höchsten Standards in den Bereichen Datenschutz, Datensicherheit und Datensouveränität. Die Plattform kombiniert modernste Large Language Models (LLMs) mit einem Retrieval-Augmented Generation (RAG)-Ansatz, um vertrauliches Wissen gezielt und sicher in juristischen Prozessen einzusetzen

Projektziele und Methodik

Das Ziel des Projekts bestand darin, konkrete Use Cases für den Einsatz generativer KI in Kanzleien und Rechtsabteilungen zu identifizieren, zu evaluieren und deren Umsetzung zu optimieren. Hierbei wurde ein besonderer Fokus auf die Effizienzsteigerung, die Minimierung menschlicher Fehler und die Unterstützung juristischer Fachkräfte durch KI-Assistenten gelegt.

Das Projekt folgte einer strukturierten und agilen Methodik:

Anforderungsanalyse: Definition der zentralen Herausforderungen und Use Cases.

Entwicklung und Anpassung: Konfiguration der Plattform basierend auf Feedback aus der Praxis.

Iterative Tests: Optimierung der Lösung durch wiederholte Tests in Pilotkanzleien.

Rollout und Evaluierung: Implementierung und Bewertung der Ergebnisse in realen Anwendungsszenarien.

Ergebnisse und Erkenntnisse

Die Implementierung von Private GPT hat gezeigt, wie generative KI die juristische Arbeit nachhaltig transformieren kann:

Effizienzsteigerung: Routineaufgaben, wie Vertragsprüfungen oder Dokumentenanalysen, konnten erheblich beschleunigt werden.

Präzision und Sicherheit: Kontextsensitivität und eine geschlossene Datenbasis minimieren Fehler und gewährleisten Vertraulichkeit.

Anwenderfreundlichkeit: Die intuitive Nutzung über natürliche Spracheingabe ermöglicht eine einfache Integration in bestehende Arbeitsprozesse.

Das Projekt offenbarte jedoch auch, dass die Qualität der Ergebnisse maßgeblich von der Datenbasis, den eingesetzten Prompts und der Konfiguration der Plattform abhängt. Daher ist die Schulung von Mitarbeitenden im Bereich des Prompt Engineerings entscheidend für den Erfolg solcher Lösungen. Eine Use Case orientierte Herangehensweise mit klaren Zielen und einem iterativen Vorgehen sind wichtige Voraussetzungen für ein erfolgreiches KI-Projekt.

Potenziale und Ausblick

Private GPT bietet nicht nur unmittelbare Vorteile für die Arbeit im juristischen Umfeld, sondern eröffnet auch Perspektiven für weitere Anwendungsgebiete wie Compliance, Wissensmanagement und Automatisierung. Die Plattform ist skalierbar, flexibel und darauf ausgelegt, mit den steigenden Anforderungen der Branche mitzuwachsen.

Einladung zur Zusammenarbeit

Die Ergebnisse dieses Projekts verdeutlichen, dass generative KI einen klaren Wettbewerbsvorteil schafft. Kanzleien, Justiz und Rechtsabteilungen, die sich frühzeitig mit diesen Möglichkeiten auseinandersetzen, können nicht nur effizienter arbeiten, sondern auch ihren Mandanten einen höheren Mehrwert bieten.

Nutzen Sie die kostenfreie [Testumgebung](#) von Fujitsu, um die Potenziale von Private GPT selbst zu erleben. Das Team von Fujitsu entwickelt gemeinsam mit ihrem Team individuelle Lösungen, die Ihre spezifischen Anforderungen erfüllen und Ihre juristische Arbeit erheblich erleichtert. Die Zukunft der juristischen Arbeit beginnt jetzt.

Inhaltsverzeichnis

Executive Summary	1
Inhaltsverzeichnis	3
1 Einleitung	5
1.1 Hintergrund	5
1.2 Warum generative KI?	5
1.3 Unser Ziel	6
1.4 Zielgruppe	6
2 Projektüberblick	7
2.1 Ziele des Projekts	7
2.2 Projektphasen und Beteiligte	8
2.3 Besondere Anforderungen in diesem Projekt	8
3 Methodik	10
3.1 Die Private GPT Plattform	10
3.2 Iterativer Ansatz	11
3.3 Definition der Use Cases	11
3.4 Vorbereitung und Datenaufbereitung	12
3.5 Tests und Feedback	13
3.6 Anpassungen der Plattform	13
4 Ergebnisse	15
4.1 Verbesserungen bei Präzision und Effizienz	15
4.2 Stärken und Einschränkungen des RAG-Ansatzes	15
4.3 Fazit zu den Ergebnissen	18
5 Anwendungsbeispiele und Business Case	19
5.1 Konkrete Use Cases und Vorteile im Wettbewerb	19
5.2 Wie Private GPT Anwalt*innen besser macht	20
6 Ausblick	22
6.1 Weiterentwicklung der Modelle und Plattform	22
6.2 Weitere Einsatzgebiete der Lösung	23
7 Fazit	24
7.1 Zentrale Erkenntnisse aus dem Projekt	24
7.2 Empfehlungen für die Einführung von generativer KI	25
7.3 Ein Blick nach vorn	25

8	Über die Autoren	26
8.1	Christian Ammer	26
8.2	Florian Zube	26
9	Anhang	27
9.1	Technische Spezifikationen der Plattform	27
9.2	Weiterführende Informationen und Literatur	28
9.3	Konkretes Fallbeispiel: Vertragsanalyse mit Private GPT	29
10	Glossar	33

1 Einleitung

1.1 Hintergrund

Der Rechtssektor sieht sich in den letzten Jahren mit tiefgreifenden Herausforderungen konfrontiert. Digitalisierung, regulatorische Änderungen, eine zunehmende Internationalisierung und die steigende Komplexität insbesondere im Wirtschaftsrecht erschweren die Handhabung juristischer Fälle. Gleichzeitig wird die Situation durch den Mangel an Fachkräften im deutschen Markt verschärft: Die Anzahl der Absolvent*innen im Fachbereich Jura reicht nicht aus, um den steigenden Bedarf an hervorragenden Jurist*innen zu decken. Diese Kombination aus komplexen Anforderungen und begrenzten personellen Ressourcen erfordert innovative Ansätze, um den Anforderungen einer immer anspruchsvolleren Welt gerecht zu werden.

1.2 Warum generative KI?

Generative KI ist spätestens seit ChatGPT das Hype-Thema in der gesamten Rechtsdienstleistungsbranche. Die Technologie basiert auf Modellen, meistens großen Sprachmodellen (Large Language Models, kurz LLMs), die mit gigantischen, textbasierten Datensätzen trainiert wurden, um eine Vielzahl sprachbasierter Aufgaben zu erfüllen. Durch menschliche Interaktion mit dem System (in Form von „Prompts“), können so verschiedenste Anwendungsszenarien realisiert werden: Textgenerierung, Zusammenfassungen erstellen, Informationen extrahieren, Analysen und vieles mehr. Neben ihrer vielfältigen Einsatzmöglichkeiten zeichnet sich diese Technologie gerade dadurch aus, dass die Bedienung im Gegensatz zu klassischen Softwareanwendungen besonders einfach und intuitiv ist. Statt einer komplexen Benutzeroberfläche setzen die allermeisten Anwendungen auf ein Chat-Fenster, wodurch Nutzer*innen ihre Aufgabe in natürlicher Sprache beschreiben können.

Generative KI hat sich in Studien und der Praxis als ein Schlüsselthema für die Transformation des Rechtsmarkts herauskristallisiert, wie unter anderem der aktuelle [Legal Tech Monitor 2025](#) des Legal Tech Verband Deutschland e.V. zeigt. Neben den bereits genannten Herausforderungen, die damit adressiert werden können, bietet die Technologie im Rechtssektor Potenzial für Wettbewerbsvorteile. Diese können z.B. durch eine Steigerung der Effizienz, Reduktion menschlicher Fehler oder verbesserte Genauigkeit realisiert werden. Die Branche profitiert von der textlastig geprägten Kernarbeit, da generative KI für solche Anwendungsszenarien besonders geeignet ist. In Kapitel 5 werden konkrete Beispiele vorgestellt, um diese Potenziale realisieren können.

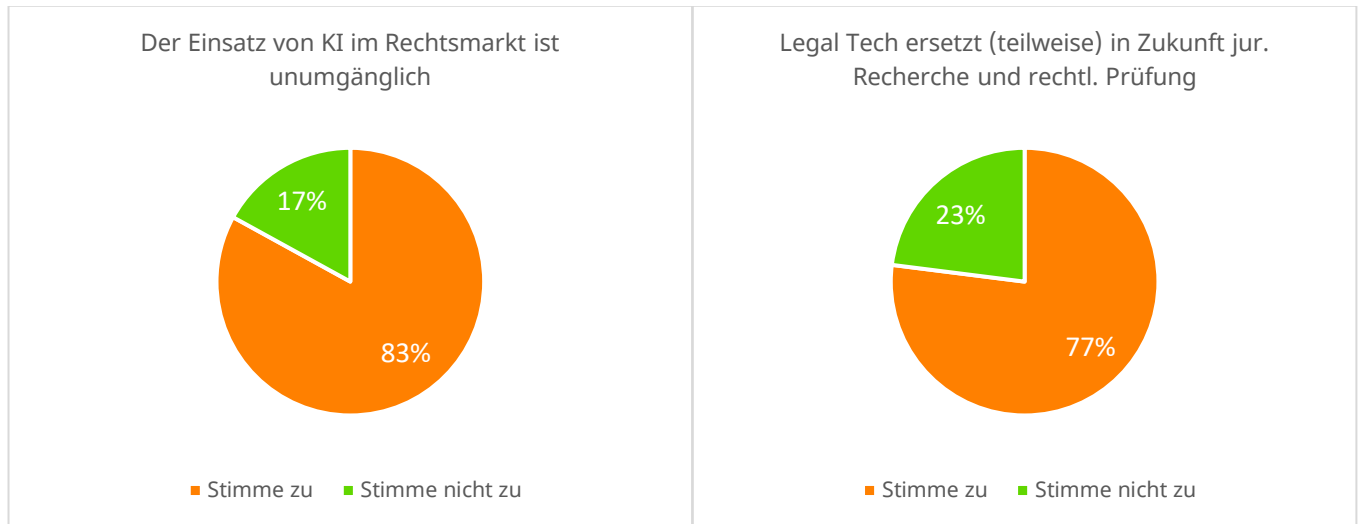


Abbildung 1: Einsatz von KI im Rechtsmarkt.

Abbildung 2: Legal Tech für rechtliche Aufgaben.

1.3 Unser Ziel

Ziel war es, spezifische Use Cases von generativer KI für den Rechtssektor zu entwickeln. In der Untersuchung ging es darum, das Potenzial von generativer KI zu erfassen, gleichzeitig aber auch Limitationen für den Einsatz festzustellen und mögliche Lösungsansätze aufzuzeigen. Das Projekt sollte erforschen, welche Techniken und Faktoren eingesetzt werden können, um die Ergebnisse der Use Cases positiv zu beeinflussen. Aus den Ergebnissen des Projektes ergeben sich eine Vielzahl von Empfehlungen, wie generative KI erfolgreich im Rechtssektor eingesetzt werden kann, was bei der Einführung solcher Lösungen zu beachten ist und welche Grenzen die Technologie heute hat, um so den Nutzen von generativer KI für die Branche maximieren zu können.

1.4 Zielgruppe

Dieses White Paper richtet sich an technologieaffine Interessierte aus Kanzleien und dem juristischen Markt, insbesondere an Rechtsanwält*innen, IT-Verantwortliche großer Kanzleien sowie Anbieter von juristischen Inhalten. Darüber hinaus eignen sich die Erkenntnisse auch hervorragend für den Einsatz in Rechtsabteilungen und der Justiz. Das Paper zielt darauf ab, einen Lösungsweg aufzuzeigen, der sich klar von den bekannten US-Cloud-Angeboten abhebt und stattdessen auf Sicherheit, Vertraulichkeit und ethischen Standards setzt.

Darüber hinaus ist dieses Paper für alle interessant, die mehr über die technische Basis der Private GPT Plattform erfahren wollen, sowie für alle Technologieverantwortlichen, die Anregungen suchen, wie man mit einer generativen KI-Lösung auf OpenSource-Basis zu guten Ergebnissen kommt.

2 Projektüberblick

2.1 Ziele des Projekts

Das Ziel des gemeinsamen Projekts war es, eine möglichst leistungsfähige KI-Lösung für den Rechtssektor zu entwickeln, die die spezifischen Anforderungen von Kanzleien und Rechtsabteilungen berücksichtigt. Mit dieser Lösung soll aufgezeigt werden, wie generative KI als Assistenzsystem die juristische Arbeit bereichern kann, indem sie die Effizienz steigert, Prozesse beschleunigt und die Qualität der Ergebnisse insgesamt verbessert. Gleichzeitig sollten berufsrechtliche, datenschutzrechtliche und ethische Standards konsequent eingehalten werden, um die notwendige Vertraulichkeit für diese Berufsgruppe zu gewähren.

Als Technologie wurde die bestehende Applikation Private GPT von Fujitsu verwendet, die bereits bei verschiedenen Kunden erfolgreich im Einsatz ist. Durch die Entwicklung und Erprobung konkreter Use Cases soll gezeigt werden, wie ein KI-Assistenzsystem die Qualität und Geschwindigkeit juristischer Arbeit erheblich steigern kann. Mit einem Use Case meinen wir eine konkrete Tätigkeit wie eine Vertragsanalyse, die von einer Person ausgeführt wird. Das Ziel lag also darin, diese Use Cases zu identifizieren und mit echten Dokumenten und den passenden Prompts die Person bei der Ausführung des Use Case durch die KI zu unterstützen. Dabei steht nicht die Ersetzung von Fachkräften im Fokus, sondern die Unterstützung durch ein System, das Routineaufgaben übernimmt und Anwältinnen und Anwälte bei Ihrer Arbeit mit Dokumenten entlastet.

Private GPT ist eine Alternative zu den marktbeherrschenden, cloudbasierten und von US-amerikanischen Anbietern geprägten Lösungen darstellt. Private GPT verfolgt einen eigenständigen Ansatz:

Open Source: Die Lösung basiert vollständig auf einem offenen Technologieansatz und gewährleistet dadurch höchste Transparenz.

On-Premise-Betrieb: Kunden können die Plattform lokal oder bei einem Dienstleister ihrer Wahl betreiben, wodurch sie vollständige Kontrolle über ihre Daten behalten.

Europäische Standards: Die Plattform erfüllt höchste Anforderungen an Datensicherheit und Datenschutz nach europäischen Vorgaben.

Ethische Datenquellen: Private GPT verwendet ausschließlich Large Language Modelle (LLMs) der europäischen Firma Mistral. Mistral AI trainiert seine Modelle auf einem proprietären Datensatz und gewährleistet so die Einhaltung der rechtlichen und ethischen Standards der EU.

2.2 Projektphasen und Beteiligte

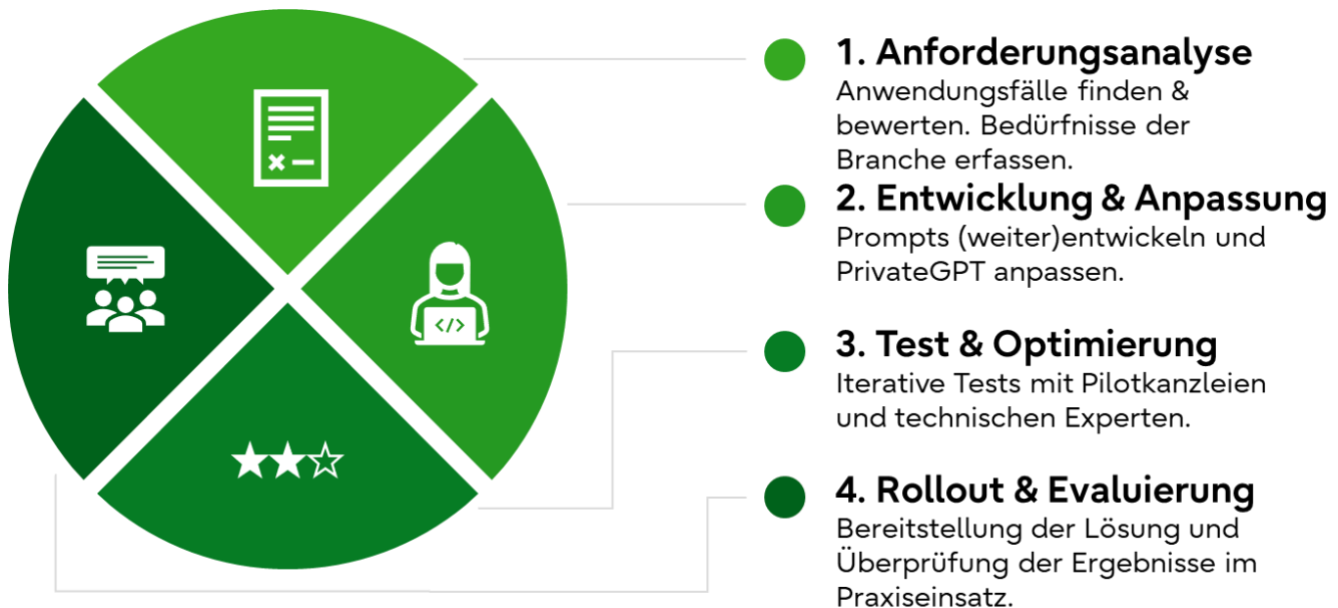


Abbildung 3: Projektphasen und Beteiligte.

- 1. Anforderungsanalyse:** Erhebung der spezifischen Bedürfnisse des Rechtssektors, insbesondere im Hinblick auf Datenschutz, Datensouveränität und Aktualität der Rechtsinformationen. Identifikation und Bewertung der zentralen Anwendungsfälle.
- 2. Entwicklung und Anpassung:** Konfiguration der Private GPT-Lösung, basierend auf den definierten Use Cases und Feedback aus dem Markt.
- 3. Test und Optimierung:** Iterative Tests in enger Zusammenarbeit mit Pilotkanzleien und technischen Expert*innen, um die Lösung weiter zu verfeinern.
- 4. Rollout und Evaluierung:** Bereitstellung der Lösung und Überprüfung der Ergebnisse im Praxiseinsatz.

Das Projektteam bestand dabei aus verschiedenen Expert*innen von Fujitsu, technischen Partnern und Fachleuten mit Expertise in Technologie und der Kanzleiwelt. Durch diesen breit gefächerten Hintergrund der Projektmitglieder konnten technische, fachliche und organisatorische Herausforderungen ganzheitlich adressiert werden.

2.3 Besondere Anforderungen in diesem Projekt

Datenschutz und Sicherheit

Die Einhaltung der europäischen Datenschutzrichtlinien (z. B. DSGVO) war ein zentraler Aspekt. Der Betrieb der Lösung als On-Premise-Installation minimiert das Risiko der Datenverarbeitung durch Dritte und gewährleistet vollständige Kontrolle.

Aktualität der Daten

Juristische Informationen erweitern und ändern sich ständig, wodurch es essenziell ist, die Lösung laufend mit aktuellen und verlässlichen Datenquellen zu versorgen.

EU AI Act

Die Entwicklung erfolgte unter Berücksichtigung der Anforderungen des europäischen KI-Gesetzes, das strenge Vorgaben für den Einsatz von KI in dieser Branche und für deren Anwendungsfälle definiert.

Prozesskonformität

Die Integration von Technologie in bestehende und etablierte dokumenten- und oft noch papierbasierte Arbeitsabläufe.

Ein spezifischer Schwerpunkt lag auf der zeitaufwändigen Sammlung und Aufbereitung von passenden Dokumenten und Daten für die Evaluierung und die Tests der Use Cases. Dieser Prozess ist essenziell für den Erfolg der Lösung, da die Qualität der Ergebnisse maßgeblich von der Qualität und der tatsächlichen Relevanz der Eingangsdaten abhängt. Es braucht also ein breites Fundament an Verträgen, Vorlagen, Musterakten, Konversationen und Gerichtsdokumenten, um hier mit Hilfe der KI auf die Realwelt passende Anwendungsfälle bearbeiten zu können.

3 Methodik

Die Entwicklung der Fujitsu Private GPT-Lösung für den Rechtssektor folgte einer strukturierten und agilen Methodik, um die spezifischen Anforderungen der Zielgruppe effektiv umzusetzen. Dabei wurde besonderer Wert auf iterative Tests, praxisnahes Feedback und die fortlaufende Optimierung der Lösung gelegt.

3.1 Die Private GPT Plattform

Fujitsu Private GPT (PGPT) ist eine vollumfängliche Lösung, die aus der benötigten Hardware, dem Betriebssystem und der eigentlichen Software besteht. Der zentrale Ansatz besteht darin, diese Applikation im eigenen Rechenzentrum oder bei einem vertrauensvollen Hosting-Partner eigener Wahl zu betreiben. Dadurch kann vollständige Datensouveränität garantiert werden, da die Daten das eigene Rechenzentrum niemals verlassen. Falls notwendig, kann Private GPT auch komplett ohne Internetzugriff betrieben werden.

Private GPT setzt ausschließlich auf Open Source Komponenten. Es werden LLMs von der französischen Firma Mistral verwendet, aktuell „NeMo LLM“ und „Mixtral-8x7B“. Diese Modelle sind äußerst leistungsfähig und im Gegensatz zu riesigen Modellen wie GPT-4o durch ihren geringeren Ressourcenbedarf ideal für eine lokale Anwendung geeignet. Eine entscheidende Eigenschaft eines LLMs besteht in der Art der Daten, die für das Training verwendet wurden. Mistral verwendet einen proprietären Datensatz und kann so Konflikte in Bezug auf das Copyright der Trainingsdaten ausschließen.

Neben dem Sprachmodell setzt Private GPT auf die Retrieval Augmented Generation (RAG) Technologie. Dieser Ansatz ist entscheidend, um der KI das unternehmensinterne Wissen zugänglich zu machen. Dazu werden hochgeladene Dokumente von Private GPT in eine Vektordatenbank (Milvus) geladen. Wird jetzt eine Anfrage an das System gestellt, werden die relevanten Teile der bereitgestellten Dokumente basierend auf semantischer Ähnlichkeit identifiziert, und anschließend dem LLM übergeben. Die Antwort basiert dadurch ausschließlich auf den Informationen aus den Dokumenten. Private GPT bietet zusätzlich eine Funktionalität, stattdessen auch das Sprachmodell direkt zu befragen, ohne den Dokumentenkontext zu nutzen.

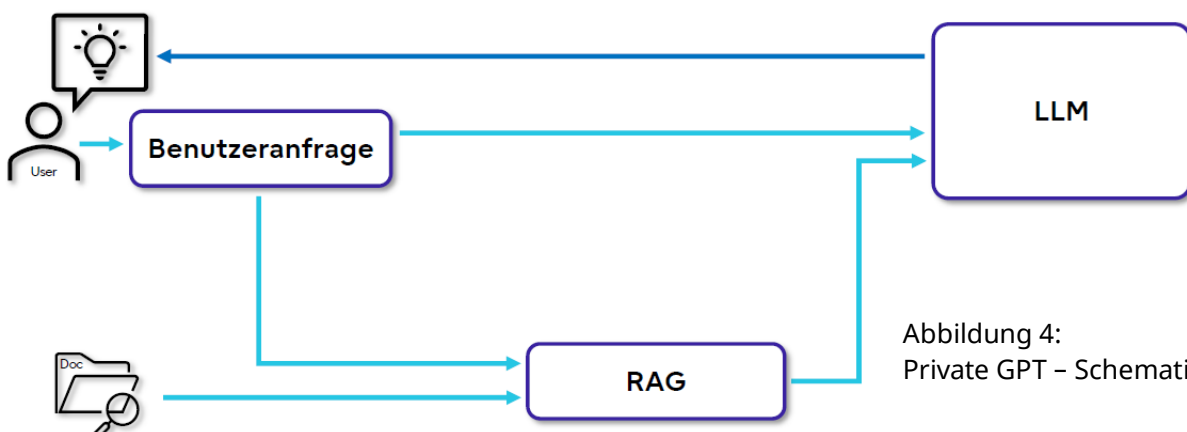


Abbildung 4:
Private GPT – Schematische Darstellung.

Die Lösung läuft auf einem Standard Fujitsu Primergy Server (RX2540 M7), der exakt auf die benötigten Systemvoraussetzungen zugeschnitten ist:

CPU	2x Intel® Xeon® Gold 6542Y 24C 2.9 GHz
Hauptspeicher	16x16GB 1Rx8 DDR5-5600
Festplatte	8x 1.92TB SSD
GPU	NVIDIA L40S (48GB GDDR6)
Betriebssystem	SUSE SLES 15 SP6

Sämtliche Komponenten der Applikation sind containerisiert (Docker), um eine hohe Flexibilität und Zukunftssicherheit zu bieten. Jede Teilkomponente kann somit leicht ausgetauscht werden, wenn beispielsweise ein effizienteres Sprachmodell veröffentlicht wird.

3.2 Iterativer Ansatz

Der gesamte Entwicklungsprozess basierte auf einem iterativen Ansatz, der auf kontinuierliche Verbesserung ausgelegt war. In regelmäßigen Zyklen wurden neue Herangehensweisen, Dokumente, Modelle und Funktionen getestet, Ergebnisse mit Fachexpert*innen aus dem Rechtssektor überprüft, Feedback von den KI-Expert*innen eingeholt und die Lösung entsprechend angepasst. Dieser Ansatz stellte sicher, dass die Lösung nicht nur technisch ausgereift war, sondern auch den praktischen Anforderungen des Rechtssektors entsprach.

Eine starke Projektleitung, die aus einem erfahrenen Projekt-Manager und einem Kanzleiexperten mit einem hohen technologischen Verständnis bestand, koordinierte die Arbeit des interdisziplinär zusammengesetzten Teams. Dieses Team umfasste Data-Science-Expert*innen, technische Entwickler*innen, Consultants und Expert*innen aus dem Rechtssektor mit juristischem Fachwissen. Durch die enge Zusammenarbeit und regelmäßige Abstimmungsrunden konnten Herausforderungen schnell identifiziert und gelöst werden.

3.3 Definition der Use Cases

Der Grundstein für die Entwicklung wurde durch die Identifikation von 26 potenziellen Use Cases gelegt. Diese Use Cases wurden auf Basis von Interviews mit Anwalt*innen und juristisch arbeitenden Expert*innen sowie Assistenzkräften aus unterschiedlichen Kanzleien gesammelt. Die Gespräche deckten eine breite Palette juristischer Arbeitsprozesse ab, um ein möglichst umfassendes Bild der Potenziale durch den Einsatz generativer KI speziell in Wirtschaftskanzleien zu erhalten.

Die 26 identifizierten Fälle wurden anschließend anhand dreier zentraler Kriterien bewertet:

Nutzen: Welcher Mehrwert kann durch den Einsatz von generativer KI in diesem Anwendungsfall erzielt werden?

Machbarkeit: Wie realistisch ist die technische und organisatorische Umsetzung des Use Cases mit hoher Qualität?

Relevanz: Wie wichtig ist der Use Case für die Zielgruppe, insbesondere mit Blick auf die Praxisnähe und die Häufigkeit des Bedarfs?

Auf Basis dieser Bewertung wurde eine Prioritätenliste erstellt, die die Grundlage für die weitere Entwicklung bildete. Der Fokus lag zunächst auf den Use Cases mit dem höchsten Potenzial, um greifbare Ergebnisse zu erzielen. Zu den priorisierten Anwendungsfällen gehörten:



Dokumentenanalyse

Automatische Analyse, Zusammenfassung und Informationsextraktion von offiziellen juristischen Dokumenten (beispielsweise von Behörden und Gerichten).



Knowledge Management

Strukturierte Verwaltung und Abfrage dokumentierten Wissens in unterschiedlichen Dokumenten und Informationspools von Kanzleien.



Verträge

Unterstützung bei der Prüfung und Verbesserung von Verträgen und Vertragsvorlagen.



Fallrecherche und Investigation

Informationen und Zusammenhänge aus Fallakten intelligent recherchieren, zusammenfassen und bewerten lassen.

3.4 Vorbereitung und Datenaufbereitung

Für die priorisierten Use Cases wurden passende Dokumente gesammelt, erstellt und aufbereitet. Die Qualität der Daten spielte eine entscheidende Rolle für den Erfolg und die Relevanz der Testergebnisse, weshalb besonderer Wert auf eine gründliche Datenaufbereitung gelegt wurde. Dabei wurden die Daten sowohl technisch als auch fachlich geprüft, um ihre Eignung für die Tests sicherzustellen.

Zu jedem Use Case wurde der fachliche Hintergrund gemeinsam mit den Fachexpert*innen erarbeitet, um daraus geeignete Prompts abzuleiten. Um die Qualität des KI-Outputs zu bewerten, wurden von Beginn an die konkret erwarteten Ergebnisse dokumentiert. Je Fall wurde eine Musterlösung formuliert, um die Ergebnisse messbar zu machen. Im Beispiel der Vertragsanalyse wurden die Klauseln identifiziert, die das System als riskant bewerten sollte. Je Klausel wurde außerdem die Begründung definiert, warum sie als riskant einzustufen ist. Dieses Vorgehen wurde je Use Case an verschiedenen Szenarien wiederholt (z.B. Geschäftsführervertrag und Vertraulichkeitserklärung).

Dieser Ansatz bietet drei zentrale Mehrwerte:

Verständlichkeit: Für die Prompt Engineers war ein klares Ziel ersichtlich, welches auch ohne einen fachlichen Hintergrund im Rechtsbereich verfolgt werden konnte.

Generalisierung: Durch mehrere Beispiele je Fall konnte sichergestellt werden, dass die Prompts die Antwort nicht bereits vorwegnehmen würden. Der Prompt musste in der Lage sein, unterschiedliche Vertragswerke zu bewerten.

Messbarkeit: Die Ergebnisse konnten klar gemessen werden, um eine Genauigkeit zu ermitteln. Dadurch wurden Prompts untereinander vergleichbar, ebenso wie Systemkonfigurationen, ohne sich auf ein „gefühltes“ Ergebnis zu berufen.

3.5 Tests und Feedback

Die eigentlichen Tests wurden in mehreren Iterationen durchgeführt. Jede Testphase beinhaltete:

1. Die **Anwendung** von Private GPT auf die vorbereiteten Dokumente und Szenarien durch entsprechendes Prompting unter Einsatz von Techniken des Prompt Engineering.
2. Die **Dokumentation** der Testergebnisse und Bewertung der Antwort mit einer fünfstufigen Skala.
3. Die Erhebung von **Feedback** durch fachliche und technische Expert*innen.
4. Die gezielte **Anpassung** der Herangehensweise, Konfiguration und Erwartungshaltung basierend auf den Ergebnissen.

Besonderes Augenmerk wurde dabei auf die Praxisrelevanz und Bedienbarkeit für die Benutzer*innen gelegt, um sicherzustellen, dass die Lösung intuitiv und effizient in der Praxis einsetzbar ist.

Während des gesamten Prozesses fand ein intensiver Austausch zwischen Projektleitung, den technischen Expert*innen und den bei Fujitsu verantwortlichen Führungskräften statt. Dieser Dialog diente dazu, die Stärken und Schwächen der Lösung kontinuierlich zu evaluieren, Ideen für den weiteren Prozess zu entwickeln und alle neuen Erkenntnisse direkt in die Weiterentwicklung einfließen zu lassen.

Durch die iterative Vorgehensweise und die enge Einbindung von Fachleuten aus verschiedenen Disziplinen konnte eine Lösung entwickelt werden, die sowohl technologisch ausgereift als auch praxisorientiert ist.

3.6 Anpassungen der Plattform

Obwohl von vornherein bewusst im Rahmen eines generischen LLMs gearbeitet wurde und somit das Fine-Tuning des Sprachmodells keine Option war, wurde schnell deutlich, dass die Systemparameter die Ergebnisse stark beeinflussen. Folgende Parameter stellten hierbei als die zentralen „Stellschrauben“ heraus:

System Prompt: Insbesondere im Kontext von RAG spielt der System Prompt eine besondere Rolle, da dieser, anders als der Prompt des Users, nicht die Auswahl der Chunks aus den Dokumenten beeinflusst. Somit können hier gezieltere Anweisungen an das Sprachmodell gegeben werden.

Chattiness: Diese Variable beschreibt, wie „gesprächig“ das System eingestellt wird. Dadurch kann beeinflusst werden, ob die Antwort knapp und präzise formuliert werden soll, oder ob das System mehr im Detail erklären soll.

Chunking: Ein Chunk ist ein Teil eines Dokumentes, welcher durch zur Beantwortung des Prompts RAG ausgewählt wurde. Im einfachsten Fall ist ein Chunk einfach eine Seite eines PDFs. Hier waren folgende Anpassungen möglich:

- **Anzahl der Chunks:** Die Anzahl der Chunks, die aus den Quelldokumenten ausgewählt wird. Eine hohe Anzahl bietet mehr Kontext, erhöht aber auch die Chance, dass die gewünschte Information bei zu vielen Chunks verloren geht („Needle Test“).
- **Größe der Chunks:** Die Größe spielt gemeinsam mit der Anzahl der Chunks eine zentrale Rolle. Größere Chunks ermöglichen mehr Kontext, liefern aber möglicherweise unwichtige Informationen und „verschwenden“ damit Platz im Context-Window
- **Chunk-Grenze:** Chunks werden nach ihrer Relevanz bewertet. Liegt sie unterhalb dieser Grenze, wird der Chunk aussortiert.

Temperatur: Beeinflusst die Kreativität des Systems. Geringe Werte reduzieren die Chance auf Halluzinationen, schränken aber auch die Kreativität ein.

Output-Länge: Eine längere Ausgabe kann z.B. bei der Vertragserstellung eine Voraussetzung sein, erhöht aber auch die Antwortzeit sowie die Chance auf Halluzinationen.

Die initialen Tests basierten auf der Basiskonfiguration von Private GPT. Nachdem die grundlegenden Prompts verfasst und bewertet waren, wurde die Konfiguration im weiteren Testverlauf wöchentlich angepasst. Die Anpassung basierte auf der gemeinsamen Evaluation der Ergebnisse, um Potenziale ausschöpfen zu können. Die Tests wurden anschließend auf der neuen Konfiguration wiederholt und ausgewertet, um beide miteinander zu vergleichen. So konnten schrittweise die Systemparameter optimiert werden. Hierbei ist zu erwähnen, dass es keine „perfekte“ Einstellung für den gesamten Rechtssektor gibt. Zwar sind bestimmte Konfigurationen in jeder Hinsicht besser als andere, aber die besten Ergebnisse setzen in der Regel Konfigurationen für bestimmte Use Cases voraus, die wiederum die Ergebnisse anderer Use Cases möglicherweise beeinflussen. Konkrete Empfehlungen hierzu folgen in den nächsten Kapiteln und im Anhang.

4 Ergebnisse

Die Implementierung und der Test der Private GPT-Lösung haben wertvolle Einblicke geliefert, wie generative KI die juristische Arbeit in der Praxis verändern kann. Durch den gezielten Einsatz können bestehende Herausforderungen adressiert und zudem auch neue Möglichkeiten für Effizienz und Präzision erzielt werden. Gleichzeitig hat das Projekt klar gemacht, welche Rahmenbedingungen erforderlich sind, um das volle Potenzial dieser Technologie auszuschöpfen. In diesem Kapitel werden die wichtigsten Ergebnisse zusammengefasst, die sowohl die Stärken der Lösung als auch ihre aktuellen Einschränkungen beleuchten.

Im Anhang finden Sie einige der erzielten Ergebnisse am Beispiel der Vertragsanalyse.

4.1 Verbesserungen bei Präzision und Effizienz

Die Implementierung von Private GPT hat messbare Verbesserungen in der Präzision und Effizienz der juristischen Arbeit ermöglicht. Tester*innen berichten von einer deutlichen Reduktion des Aufwands für Routineaufgaben, wie etwa die Analyse umfangreicher Dokumente oder die Suche nach relevanten Präzedenzfällen. Gleichzeitig konnte die Genauigkeit der Ergebnisse durch den gezielten Einsatz generativer KI gesteigert werden, da die Lösung kontextsensitiv relevante Informationen liefert.

Richtig eingesetzt führen diese Verbesserungen nicht nur zu einer deutlichen Entlastung der Fachkräfte, sondern auch zu einer höheren Zufriedenheit der User. Besonders in Situationen mit hohem Arbeitsaufkommen, wie etwa bei der Bearbeitung langwieriger Rechtsstreitigkeiten oder der Erstellung umfassender Vertragswerke, bietet Private GPT einen echten Mehrwert für die User. In Kapitel 5 werden konkrete Anwendungsszenarien beispielhaft vorgestellt.

4.2 Stärken und Einschränkungen des RAG-Ansatzes

Der Einsatz von Retrieval-Augmented Generation (RAG) als technologische Grundlage hat sich in vielen Bereichen bewährt, bringt jedoch auch Herausforderungen mit sich. Durch bestimmte Techniken oder angepasste Workflows, können einige Einschränkungen kompensiert werden.

Die hier dargestellten Aspekte beziehen sich speziell auf einen klassisch generalistischen Ansatz.

4.2.1 Stärken

Aktualität der Daten

Der RAG-Ansatz erlaubt es, aktuelle und spezifische Informationen aus einer definierten Wissensbasis abzurufen, wodurch die Lösung immer auf dem neuesten Stand der Rechtsprechung gehalten werden kann. Die Daten können mit einfachen Mitteln in die Vektordatenbank eingefügt, entfernt oder aktualisiert werden, insbesondere durch eine direkte Anbindung an bestehende Dokumentenverzeichnisse.

Relevanz der Daten

Der Zugriff auf das eigene Wissen stellt sicher, dass Ergebnisse relevant für das Unternehmen sind. Somit werden z.B. Richtlinien und Falldaten zur Beantwortung der Frage verwendet, statt allgemeines Wissen zu verwenden.

Keine Halluzination

Ein RAG-Ansatz reduziert die Chance auf Halluzinationen stark, da nur Informationen aus der Wissensbasis verwendet werden.

Skalierbarkeit

Private GPT kann aktuell vertikal skaliert werden. Durch zusätzliche Systemhardware und GPUs, können z.B. die Menge der verarbeitbaren Dokumente oder die Anzahl paralleler Nutzer ohne großen Aufwand erhöht werden. Perspektivisch wird via Loadbalancer auch eine horizontale Skalierung möglich sein.

Datensicherheit

Durch die Verwendung einer geschlossenen Datenbasis wird sichergestellt, dass vertrauliche Informationen geschützt bleiben und Daten niemals außerhalb der eigenen Organisation verarbeitet werden.

Vielseitiger Einsatz

RAG ermöglicht es, die Lösung auf eine Vielzahl von Anwendungs-fällen anzupassen, da die zugrundeliegende Wissensbasis jederzeit und einfach erweitert oder angepasst werden kann.

4.2.2 Einschränkungen

Abhängigkeit von der Datenqualität: Die Präzision der Ergebnisse hängt stark von der Qualität und Struktur der zugrundeliegenden Dokumente ab. In Fällen, in denen die darin enthaltenen Informationen unvollständig, schlecht aufbereitet oder inkonsistent sind, können die Ergebnisse beeinträchtigt sein.

Verständnis komplexer Kontexte: Trotz der generativen Fähigkeiten stößt der RAG-Ansatz bei komplizierten juristischen Fragestellungen an Grenzen, insbesondere wenn die Antworten fachliche Interpretationen erfordern.

Abhängigkeiten im Kontext: Bei Quellen, dessen Inhalte logisch aufeinander aufbauen und über den Text weit verteilt sind, kann RAG diesen Zusammenhang oft nicht richtig identifizieren, da diese internen Abhängigkeiten eine eigene Verarbeitung jedes Chunks durch das LLM erfordern würden.

Größe der Quelldaten: Abhängig vom Kontextfenster des Sprachmodells kann nur eine beschränkte Anzahl Chunks ausgewählt werden. Bei einer höheren Anzahl steigt außerdem das Risiko auf schlechte Ergebnisse. Anwendungsfälle wie eine Zusammenfassung langer Texte oder alle Vorkommnisse einer Information zu finden sind dadurch eingeschränkt.

Rechenaufwand: Der Ansatz erfordert signifikante Rechenressourcen, insbesondere bei der Verarbeitung großer Datenmengen.

Textbasierte Daten: Nicht-textbasierte Dateiformate, wie z.B. Tabellen oder Grafiken, sind derzeit ungeeignet. Durch zusätzliche Module, die z.B. Bilder textuell beschreiben, kann diese Einschränkung perspektivisch ausgeglichen werden.

4.2.3 Beispiele für Stärken und Einschränkungen

Stärke – Aktualität: Im Falle der Fallrecherche konnte der RAG-Ansatz effektiv genutzt werden, um auch langjährige und komplexe Prozesse effektiv handzuhaben. In den Tests konnten z.B. bei einer Korrespondenz hunderter E-Mails die Rollen bestimmter Personen und dessen Mitwirkung analysiert, fachliche Zusammenhänge erklärt und wichtige Daten identifiziert werden. Die Informationen konnten in wenigen Sekunden erfasst werden und haben auch Sachverhalte berücksichtigt, die bereits Jahre in der Vergangenheit lagen.

Einschränkung – Größe der Quelldaten: Im Beispiel der Vertragsanalyse wurden bei der Verwendung von RAG oft nicht alle Risiken korrekt identifiziert. Die Ursache liegt daran, dass die maximale Anzahl der Chunks begrenzt ist und zum Zeitpunkt des Retrievals noch nicht klar ist, welche Klauseln risikobehaftet sein könnten, da diese Analyse erst nachgelagert durch das Sprachmodell erfolgt. Häufig wurden daher bestimmte Seiten des Vertrages nicht übermittelt, wodurch die entsprechenden Risiken nicht gefunden werden können.

4.3 Fazit zu den Ergebnissen

Die Ergebnisse des Projekts zeigen, dass Private GPT in der Lage ist, als Assistenzsystem Fachkräfte mit ganz unterschiedlichen Schwerpunkten und Verantwortlichkeiten erheblich in ihrer Arbeit zu unterstützen. Die Lösung punktet durch Effizienzsteigerungen und eine hohe Präzision, während gleichzeitig Vertraulichkeit, Datenschutz und Datensouveränität gewährleistet bleiben. Dennoch verdeutlichen die identifizierten Einschränkungen, dass der Erfolg solcher Technologien maßgeblich von der Qualität der Datenbasis und einer sorgfältigen Integration in bestehende Prozesse abhängt. Prompt Engineering wurde als wichtiger Erfolgsfaktor identifiziert, was die Bedeutung von Schulungen der Fachkräfte in diesem Bereich unterstreicht.

5 Anwendungsbeispiele und Business Case

Die Stärke von der Private GPT-Lösung zeigt sich besonders in konkreten Anwendungsfällen, die den juristischen Alltag vereinfachen können. Über mehrere exemplarische Use Cases, die während des Projekts entwickelt und getestet wurden, war es das wichtigste Ziel, zu untersuchen, wie Private GPT die Arbeitsweise von Anwälten verbessert und diese in die Lage versetzt, schneller und präziser zu handeln. Damit konnte aufgezeigt werden, welchen geschäftlichen Mehrwert diese Technologie bietet, sowohl qualitativ als auch in Bezug auf Effizienzsteigerungen.

5.1 Konkrete Use Cases und Vorteile im Wettbewerb

Private GPT bietet Anwalt*innen und Kanzleien Werkzeuge, die in spezifischen Arbeitsbereichen erhebliche Vorteile schaffen. Zu den wichtigsten Anwendungsfällen gehören:

1. Dokumentenanalyse

Private GPT unterstützt juristisch arbeitende Expert*innen bei der Analyse umfangreicher juristischer Dokumente, indem es automatisch relevante Informationen identifiziert, strukturiert und zusammenfasst. Dies reduziert die Bearbeitungszeit signifikant und minimiert das Risiko, kritische Details zu übersehen.

Beispiel: In einem laufenden Rechtsstreit werden eingehende gerichtliche Dokumente durch Private GPT zusammengefasst und in ihren Fallkontext eingeordnet. Das System erkennt in den Schreiben genannte Fristen und berechnet die Daten anhand des Eingangsdatums.

2. Interne Fallrecherche und Investigation

Private GPT unterstützt die Bearbeitung komplexer und langwieriger Fälle und Projekte, indem bestimmte Informationen oder Sachverhalte recherchiert und Zusammenhänge erläutert werden können, und auch ältere Informationen in die Auswertung einfließen. Somit kann die Bearbeitung der Fälle beschleunigt und die Argumentation stichhaltiger werden, indem sämtliche Informationen bestmöglich eingesetzt werden.

Beispiel: Aus einer Fallakte mit über 150 E-Mails konnte die Historie und der Zusammenhang fünf involvierter Unternehmen durch einen Prompt in wenigen Minuten erfasst werden (Zeit für die Ausführung des Prompts + Lesen der Antwort).

3. Recherche und Knowledge Management

Durch den Zugriff auf eine strukturierte Wissensbasis ermöglicht Private GPT eine effiziente Recherche nach Fachliteratur, Präzedenzfällen, Urteilen oder Gesetzestexten. Die Ergebnisse sind nicht nur schneller verfügbar, sondern auch präziser und relevanter. Die Möglichkeiten gehen weit über eine klassische Suche hinaus, da Informationen analysiert, verglichen oder zusammengefasst werden können.

Beispiel: Ein Anwalt recherchierte mit auf der Basis verschiedener Literaturquellen Präzedenzfälle aus drei Ländern für einen internationalen Vertragsstreit. Statt stundenlanger Arbeit lieferte die KI in 10 Minuten eine gute Übersicht relevanter Urteile, Gesetzestexte und wesentlicher Unterschiede.

4. Vertragsprüfung

Private GPT kann bestehende Verträge prüfen, indem es kritische Klauseln hervorhebt und auf potenzielle Risiken hinweist. Es gibt dadurch einen sofortigen ersten Überblick, dient als zweite Meinung und reduziert so den Zeitaufwand, während gleichzeitig die Präzision gesteigert werden kann.

Beispiel: In einem Geschäftsführervertrag konnten sechs zentrale Risiken in Bereichen wie Haftung und Compliance durch einen Prompt identifiziert und korrekt begründet werden.

5. Fallmanagement / Case Management

Der effektive Zugriff auf die Fallhistorie der Kanzlei kann entscheidende Vorteile in der Bearbeitung neuer Anfragen oder Fälle bringen. Private GPT kann ältere Fälle auswerten, um zum Beispiel ähnliche Fälle aufzuzeigen oder wichtige Erfolgsfaktoren darzustellen. Auch in die andere Richtung kann Private GPT genutzt werden, um Zusammenfassungen von Fällen zu erstellen und sicher abzulegen.

Beispiel: Bei der Recherche nach alten Fällen findet Private GPT alle Fälle im Themenbereich Datenschutz in der Region Deutschland, und stellt diese mit einem Kurzsteckbrief dar.

5.2 Wie Private GPT Anwalt*innen besser macht


Private GPT wurde als Assistenzsystem entwickelt, um Anwalt*innen in ihrer täglichen Arbeit zu unterstützen, nicht um diese zu ersetzen. Die Lösung befähigt sie, sich stärker auf die strategischen und fachlich anspruchsvollen Aspekte ihrer Fälle zu konzentrieren, während einfache, repetitive und unverhältnismäßig zeitaufwändige Aufgaben durch die KI übernommen werden. Dies führt zu einer höheren Qualität der Arbeit und einer besseren Entscheidungsfindung, da Anwalt*innen und ihre Assistenzkräfte damit mehr Zeit und Ressourcen für die inhaltliche Auseinandersetzung mit ihren Fällen haben.

Beispiel für verbesserte Entscheidungsfindung

In einem komplexen Vertragsstreit sind mehrere rechtliche Optionen zu bewerten, darunter potenzielle Klagewege und Vergleichsmöglichkeiten. Private GPT wird eingesetzt, um eine umfassende Analyse der relevanten Vertragsklauseln, Gerichtsentscheidungen und rechtlichen Präzedenzfälle durchzuführen. Die KI liefert innerhalb kürzester Zeit eine strukturierte Übersicht über die verschiedenen Optionen sowie Hinweise auf potenzielle Risiken.

Anhand dieser Analyse kann eine fundierte Strategie abgeleitet werden, die sowohl die Interessen des Mandanten optimal berücksichtigt als auch mögliche Fallstricke minimiert. Die Lösung ermöglicht es Benutzer*innen, sich auf die strategischen Überlegungen zu konzentrieren, anstatt wertvolle Zeit mit der Recherche nach Informationen zu verbringen.

Diese exemplarischen Anwendungsfälle zeigen, dass Private GPT nicht nur Effizienzsteigerungen ermöglicht, sondern auch qualitative Wettbewerbsvorteile schafft. Kanzleien, die die Technologie



nutzen, können schneller auf die Bedürfnisse ihrer Mandant*innen reagieren, komplexe Fälle effizienter bearbeiten und eine höhere Präzision in der Arbeit gewährleisten. Dadurch entsteht ein erkennbarer Mehrwert für Mandanten, der sich sehr schnell in der Wettbewerbsfähigkeit der Kanzlei niederschlagen kann.

6 Ausblick

6.1 Weiterentwicklung der Modelle und Plattform

Angesichts der schnellen Entwicklung von Technologie und im KI-Markt ist die kontinuierliche Weiterentwicklung der Private GPT-Plattform essenziell, um den weiter steigenden Anforderungen des Rechtssektors gerecht zu werden. Im Fokus stehen dabei insbesondere folgende Aspekte:

Erweiterung der Wissensbasis

Um die Aktualität und Relevanz der Ergebnisse weiter zu verbessern, könnten zusätzliche Datenquellen, wie internationale Rechtsprechung, branchenspezifische Fallstudien oder interne Wissensdatenbanken von Kanzleien, Informationsdiensten oder externen Wissensdatenbanken, integriert werden. Diese Daten können in vektorisierter Form bereitgestellt und so gemeinsam mit dem internen Wissen der Kanzlei eingesetzt werden, um Fragestellungen noch umfangreicher zu beantworten.

Verbesserung der Benutzerfreundlichkeit

Die Integration von intuitiven Dashboards und Visualisierungstools könnte die Interaktion mit der Plattform noch einfacher gestalten und den Usern helfen, komplexe Sachverhalte schneller zu erfassen. Insbesondere die bereits in Entwicklung befindlichen „Szenarien“ werden es möglich machen, Prompts abzuspeichern und gemeinsam mit den passenden Systemparametern zu einem Anwendungsfall zu gruppieren. Dadurch können häufig auftretende und komplexe Use Cases mit einer dedizierten Konfiguration besonders einfach ausgeführt werden und dabei die bestmöglichen Ergebnisse liefern, ohne andere Use Cases negativ zu beeinträchtigen.

Automatisierung

Eine Erweiterung der Kernfunktionalitäten durch z.B. Agenten würde das Automatisierungspotenzial deutlich steigern. Das System wäre in der Lage, nicht nur Fragen zu beantworten, sondern eigenständig Aktionen auszulösen. Dieser Ansatz erfordert die Anbindung an Schnittstellen und sollte für solche Szenarien eingesetzt werden, wo die Geschwindigkeit interner Prozesse eine wichtige Rolle spielt und eventuelle Ungenauigkeiten der KI toleriert werden können. Für wichtige Tätigkeiten, wie der Mandantenkommunikation, sollte die KI zwar Vorschläge liefern, die letztliche Entscheidung muss jedoch stets bei dem/der Anwalt/Anwältin liegen.

6.2 Weitere Einsatzgebiete der Lösung

Private GPT ist nicht auf den Rechtssektor beschränkt. Die zugrunde liegenden Technologien bieten auch Potenzial für den Einsatz in anderen Bereichen:

Compliance und Regulatorik

Unterstützung bei der Einhaltung von regulatorischen Anforderungen in stark regulierten Branchen wie dem Finanz- oder Gesundheitswesen

Wissensmanagement in Unternehmen

Strukturierte Verwaltung und Abfrage von Informationen in großen Organisationen.

Trainings & Schulungen

Erstellung von Lernmaterialien und Simulationen für unterschiedliche Weiterbildungen, Trainings und Zielgruppen.

Intelligente Chatbots

Aufbau eines Chatbots auf Grundlage unternehmensinterner Daten, z.B. für Kommunen in der Bürgerkommunikation.

Code-Generator

Technische Unterstützung in der Programmierung durch z.B. Debugging von Code oder Formulierung von Funktionen. Speziell für proprietäre Entwicklungen von Relevanz.

Übersetzung

Texte in andere Sprachen oder einfache Sprache übersetzen lassen. Die Formatierung bleibt erhalten.

Das hier vorgestellte Projekt hat nicht nur eine leistungsstarke KI-Lösung für den Einsatz bei Kanzleien hervorgebracht, sondern zeigt vor allem, wie wichtig der Dialog und die Zusammenarbeit mit verschiedenen Stakeholdern ist. Denkt man das Thema weiter, bieten sich weitere Kooperationsmöglichkeiten für Unternehmen unterschiedlicher Branchen an:

Partnerschaften mit Rechtsabteilungen und Justiz: Gemeinsame Entwicklung neuer Use Cases und maßgeschneiderter Lösungen für spezifische Anforderungen von Inhouse-Jurist*innen und der Justiz.

Technologiepartnerschaften: Zusammenarbeit mit Anbietern juristischer Datenbanken oder Plattformen, um die Wissensbasis weiter auszubauen.

Pilotprojekte in weiteren Industrien: Entwicklung und Test neuer Einsatzmöglichkeiten in regulierten Branchen mit ähnlich strengen Anforderungen an Datenschutz, Vertraulichkeit und Präzision.

Aus den 26 von uns identifizierten Use Cases wurden im Rahmen dieses Projektes nur ein Ausschnitt bewertet. Hier besteht großes Potenzial, aufbauend auf den gesammelten Erkenntnissen, die weiteren Use Cases in einem gemeinsamen Projekt mit einer Kanzlei zu entwickeln. Neben dem perspektivischen Einsatz dieser Anwendungsfälle in der Praxis, könnte eine solche Zusammenarbeit die kanzleiinterne KI-Strategie schärfen und weitere Einblicke geben, was KI-Lösungen speziell für den Rechtssektor leisten müssen.

7 Fazit

Der Prozess für die Weiterentwicklung der Private GPT-Lösung hat deutlich gemacht, welches Potenzial generative KI insbesondere im Rechtssektor entfalten kann. Die Lösung adressiert zentrale Herausforderungen wie den Mangel an Personal, die wachsende Komplexität juristischer Fragestellungen, den zunehmenden Zeitdruck und den hohen Zeitaufwand für Routineaufgaben. Sie bietet einen klaren Mehrwert, indem sie Anwälte bei der Analyse, Recherche und Erstellung von Dokumenten unterstützt und ihnen mehr Zeit für die wirklich strategischen und komplexen Aspekte ihrer Arbeit verschafft.

7.1 Zentrale Erkenntnisse aus dem Projekt

Effizienz und Präzision als Wettbewerbsvorteil

Private GPT zeigt, dass generative KI nicht nur Effizienzgewinne ermöglicht, sondern richtig angewandt auch die Qualität juristischer Arbeit steigert. Anwalt*innen profitieren von präziseren Analysen, einer weiteren Instanz der Qualitätskontrolle und schnelleren Prozessen, was ihnen hilft, ihre Mandant*innen besser und effektiver zu unterstützen.

Technologische Stärken und praktische Grenzen

Der Einsatz des RAG-Ansatzes hat bewiesen, dass aktuelle und relevante Daten entscheidend für den Erfolg beim Einsatz von KI sind. Gleichzeitig hat das Projekt verdeutlicht, dass die Qualität der Ergebnisse stark von der Dokumenten- und Datenbasis abhängt und fachlich komplexe Kontexte weiterhin menschliches Expertenwissen erfordern.

Anpassbarkeit der Lösung

Im Verlauf des Projektes hat sich gezeigt, dass Systemparameter, wie die Chunking-Strategie oder der Systemprompt, einen entscheidenden Unterschied bei den Ergebnissen machen. Je nach Fachgebiet und Use Case, liefern bestimmte Konfigurationen deutlich bessere Resultate. Die Anpassbarkeit von Private GPT ist daher ein entscheidender Erfolgsfaktor.

Potenzial für Weiterentwicklungen

Die Plattform bietet eine solide Grundlage für künftige Innovationen, wie die Erweiterung der Wissensbasis, Unterstützung grafischer bzw. tabellarischer Datentypen oder den Einsatz in anderen Themenfeldern wie Compliance und Wissensmanagement.

Zusammenarbeit als Schlüssel zum Erfolg

Die enge Zusammenarbeit mit juristischen und technischen Expert*innen war entscheidend, um eine praxisnahe und bedarfsgerechte Lösung zu entwickeln. Diese kompetenzübergreifende Herangehensweise wird für alle Akteure in Zukunft eine zentrale Rolle spielen.

7.2 Empfehlungen für die Einführung von generativer KI

Starten Sie mit einer klaren Zielformulierung: Was wollen Sie mit KI erreichen? Welche Prozesse sollen unterstützt werden? Ist generative KI wirklich die beste Technologie dafür? Haben Sie die passenden Daten und Informationen dazu verfügbar?

Konzentrieren Sie sich auf spezifische Use Cases: Führen Sie nicht einfach irgendeine Lösung ein und geben dem Team freien Zugriff. Legen Sie stattdessen Use Cases fest, definieren Sie gemeinsam die erwarteten Ergebnisse und sammeln Sie die besten Prompts in einer Library.

Wählen Sie eine geeignete Lösung aus: Vergleichen Sie die Funktionalitäten, Betriebskosten, Lizenzmodelle und Sicherheitsaspekte diverser Lösungen. Wenn Ihnen Datensouveränität, Datensicherheit und die Arbeit mit internem Wissen besonders wichtig sind, dann kann die Fujitsu Private GPT die beste Wahl sein. Große Kanzleien können auch von mehreren Lösungen für verschiedene Fachbereiche oder Anwendungsfälle profitieren.

Fangen Sie mit einfachen Fällen an: Nachdem die Lösung eingeführt ist, sollten Sie mit einer kleinen Anzahl einfacher Fälle starten. So steigern Sie die Akzeptanz der Lösung, bauen Expertise auf und können die Anwendungsbereiche kontinuierlich erweitern.

Seien Sie so genau wie möglich: Die besten Prompts beschreiben die tatsächliche Aufgabe so genau wie möglich. Je exakter Sie beispielsweise die logischen Arbeitsschritte einer Vertragsanalyse und die potenziell zu berücksichtigenden Risiken beschreiben, desto besser und brauchbarer wird Ihr Ergebnis sein.

Trainieren Sie regelmäßig: Prompt Engineering wird als Kompetenz immer wichtiger. Bieten Sie regelmäßige Trainings für möglichst alle Mitarbeiter*innen an und betreuen Sie das Team gerade in der Einführungsphase intensiv. Schlechte Ergebnisse können schnell demotivieren, aber fast immer durch ein zielgerichtetes Prompting vermieden werden.

7.3 Ein Blick nach vorn

Private GPT steht nicht nur für technologische Innovation, sondern auch für ein Umdenken im Umgang mit juristischer Arbeit. Kanzleien und Rechtsabteilungen, die sich frühzeitig mit den Möglichkeiten generativer KI auseinandersetzen, können einen klaren Wettbewerbsvorteil erzielen. Gleichzeitig eröffnet sich die Chance, Arbeitsprozesse so zu gestalten, dass sie effizienter und gleichzeitig erfüllender für Anwält*innen werden.

Wir laden Sie ein, sich näher mit den Möglichkeiten von Private GPT auseinanderzusetzen. Ob durch die Implementierung in Ihrer Kanzlei, die Entwicklung spezifischer Use Cases oder die Zusammenarbeit an neuen Ideen – die Zukunft der juristischen Arbeit beginnt jetzt. Unsere kostenfreie [Testumgebung](#) („Test-Drive“) bietet Ihnen die Möglichkeit, sich selbst von den Potenzialen zu überzeugen. Sprechen Sie mit Fujitsu, um herauszufinden, wie auch Sie von dieser Technologie profitieren können.

8 Über die Autoren

8.1 Christian Ammer

Experte für digitale Transformation und Fractional Leadership

<https://www.linkedin.com/in/cha23x> // ca@f-ectives.com

Christian Ammer ist Gründer und Geschäftsführer der F.ECTIVES GmbH, dem führenden Anbieter für Fractional Leadership im deutschen Mittelstand. Mit über 20 Jahren Erfahrung in der Digitalisierung von Geschäftsprozessen und der Leitung von Transformationsprojekten - unter anderem bei Siemens und als Chief Information Officer (CIO) bei der internationalen Wirtschaftskanzlei Noerr - ist er ein ausgewiesener Experte für den erfolgreichen Einsatz innovativer Technologien in Unternehmen.

Seine berufliche Laufbahn zeichnet sich durch eine langjährige Expertise im Bereich der Künstlichen Intelligenz aus: Bereits 2001 beschäftigte er sich im Rahmen einer wissenschaftlichen Arbeit mit Neuronalen Netzen und Data Mining im E-Commerce. Seitdem arbeitet er kontinuierlich an der Schnittstelle zwischen Technologie und Wirtschaft, unter anderem mit der Entwicklung von Scoring-Modellen und KI-gestützten Risikoanalysen für Banken. Heute berät er Anwaltskanzleien, Technologiedienstleister und Unternehmen zu den Potenzialen von KI und begleitet sie bei der Einführung praxisorientierter Lösungen.

In diesem Projekt brachte er seine Erfahrung im Bereich Technologie, Prozessmanagement und juristische Digitalisierung ein. Sein Ziel ist es, Kanzleien und Unternehmen im Rechtsmarkt durch den Einsatz von generativer KI und datenbasierter Automatisierung konsequent digital und damit zukunftsfähig aufzustellen. Neben der konzeptionellen Arbeit legt er besonderen Wert auf die Verbindung von technologischer Exzellenz mit konkretem Geschäftsnutzen.

Er ist gefragter Sprecher auf Veranstaltungen zu digitalen Themen und gibt regelmäßig über LinkedIn und andere Plattformen Einblicke in die Welt der KI und der Unternehmensführung. Mit seinem Fractional Leadership Modell bietet er Unternehmen flexible, fokussierte und effektive Führung für die Herausforderungen der digitalen Transformation.



8.2 Florian Zube

Berater und Projektleiter für KI und IT-gestützte Transformation

Florian Zube ist IT-Consultant bei Fujitsu und Experte für künstliche Intelligenz, IT Service Management (ITSM) und Projektmanagement. Mit seinem tiefen Verständnis für innovative Technologien und deren praktischer Anwendung in Unternehmen hat er zahlreiche Projekte erfolgreich begleitet. Sein Schwerpunkt liegt in der Entwicklung und Implementierung zukunftsorientierter IT-Lösungen, die Unternehmen helfen, ihre Prozesse effizienter und skalierbarer zu gestalten.

In seiner Rolle als Berater hat Florian Zube umfassende Erfahrung in der Koordination interdisziplinärer Teams und der Gestaltung und Implementierung von IT-Service-Prozessen. Besonders hervorzuheben ist seine Fähigkeit, komplexe technische Konzepte in praxisnahe Lösungen zu übersetzen, die einen direkten Mehrwert für Unternehmen schaffen.

Als Leiter des Private GPT-Projekts steuerte Florian Zube seine Expertise im Projektmanagement bei und führte die Entwicklung der Use Cases voran. Seine fundierten Kenntnisse in der Konfiguration und Optimierung von KI-Lösungen sowie sein ausgeprägtes Verständnis für die Bedürfnisse des Rechtssektors trugen maßgeblich zum Erfolg des Projekts bei.



9 Anhang

Der Anhang dient als Ergänzung zum Hauptteil des White Papers und bietet technische Details zur Private GPT-Plattform sowie weiterführende Informationen für interessierte Leserinnen und Leser. Dies gibt technikaffinen Nutzern und Entscheidungsträgern die Möglichkeit, sich tiefergehend mit den Grundlagen und Potenzialen der Lösung auseinanderzusetzen.

9.1 Technische Spezifikationen der Plattform

Überblick der Systemarchitektur

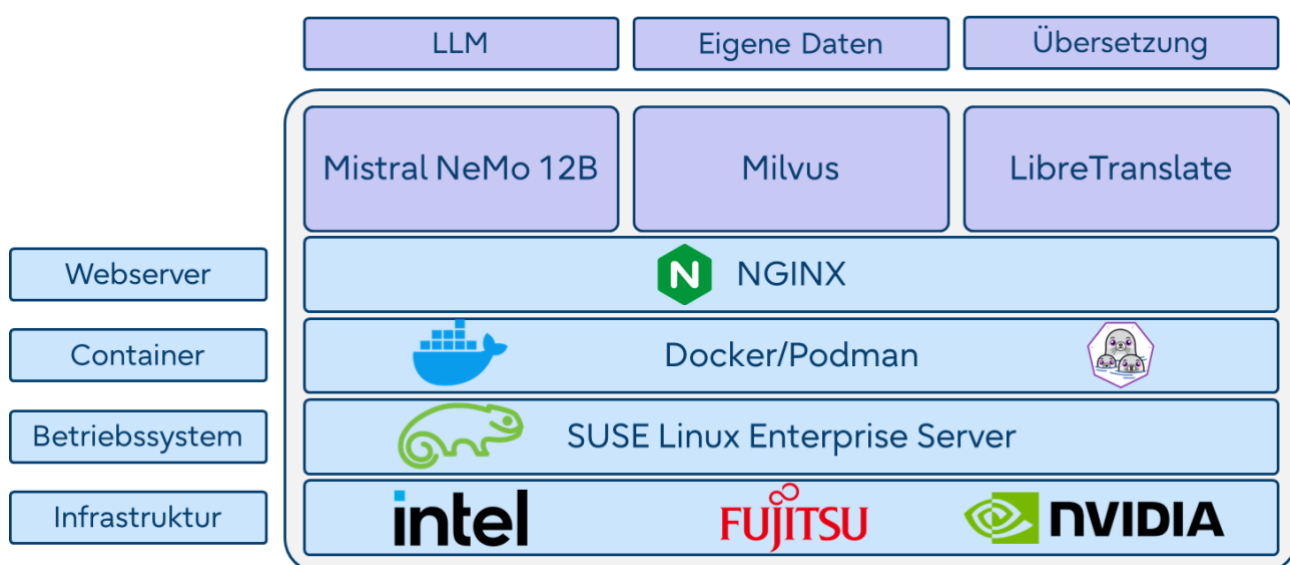


Abbildung 5: Private GPT Systemarchitektur

Technische und organisatorische Anforderungen

Um das System erfolgreich implementieren, betreiben und Nutzen draus permanent weiterzuentwickeln, sind folgende Anforderungen zu beachten:

Aus rein technischer Sicht wird ein 19" Rack Slot mit 2 Höheneinheiten zur Installation des Servers benötigt. Eine Ethernet Verbindung regelt den Zugriff auf das Webportal und den sFTP Upload. Die Netzwerkkonfiguration muss Benutzern den HTTPS Zugriff (Port 443) auf die Webapplikation und auf sFTP (Port 2222) ermöglichen. Für Admins wird SSH (Port 4422) Zugriff benötigt. Auf Benutzerebene ist ein Browser nötig.

Aus organisatorischer Sicht sind folgende Kenntnisse und Fähigkeiten empfehlenswert (Fujitsu bietet in allen Bereichen Unterstützung und Beratung an):

- Data Science und Prompt Engineering Expertise
- Product Owner für die Private GPT Plattform
- Knowledge Management
- Schnittstellenentwicklung

Integration und Anpassungsmöglichkeiten

Private GPT bietet ein REST-API zur Anbindung an Bestandssysteme. Es stehen 3 APIs zur Verfügung: Chat API, Dokumentenmanagement API und Benutzermanagement API. Schnittstelle zu einigen Systemen stehen als Base-Connector zur Verfügung (z.B. Confluence). Weitere Schnittstellen können selbst oder im Rahmen von Consultingleistungen entwickelt werden, oder Teil der Roadmap werden. Darüber hinaus sind aktuell 2 MCP (Model Context Protocol) Server für Private GPT Verfügbar. Als neuer, universeller Standard kann MCP die Integration in bestehende Systeme deutlich vereinfachen.

Private GPT kann z.B. durch Konfiguration der Systemparameter auf spezifischen Branchen oder Anwendungsfälle angepasst werden.

Leistungsumfang

Fujitsu Private GPT wird als vollständige Appliance angeboten. Die Lösung enthält:

- **Software**
 - Standard Perpetual License
 - Standard Maintenance
 - Standard Implementation
 - SUSE Linux Enterprise (OS) License & Support
- **Hardware**
 - PRIMERGY Server Hardware
 - Product Related Support (PRS)
- **Support**
 - Single point of contact L1 Service Desk (hardware & software)

Inkl. Implementierung bieten wir das vollständige Paket (Software, Hardware, Support) für ca. 200.000€ an (Lizenz: 5 Jahre). Für kleinere und mittlere Unternehmen sind flexible Optionen verfügbar. Wir bieten auch eine Leasing-Option auf monatlicher Basis via Fujitsu uScale an. Die Leistung beinhaltet sämtliche Updates auf der Roadmap während des Lizenzzeitraums.

9.2 Weiterführende Informationen und Literatur

1. Kontaktinformationen



ca@f-ectives.com (Christian Ammer)



[LinkedIn Profil \(Christian Ammer\)](#)



www.fujitsu.com/emeia/private-gpt



AI.Team@Fujitsu.com

2. Berichte und White Papers von Fujitsu

- „Private GPT – a Fujitsu AI Solution“, 2024, veröffentlicht von Fujitsu
- “Fujitsu Private GPT – Einsatzszenarien für echte Vorteile“, 2024, veröffentlicht von Fujitsu

3. Fachpublikationen zur generativen KI im juristischen Umfeld

- Fachbroschüre "Künstliche Intelligenz in der Kanzleipraxis - Fünf Einsatzfelder von Sekretariat bis Übersetzung" veröffentlicht 2024 unter Beck-Shop.de
- Studie: „Future Ready Lawyer Studie 2024“ von Wolters Kluwer, veröffentlicht auf wolterskluwer.com
- Diskussionsschrift: „Generative KI - jenseits von Euphorie und einfachen Lösungen" veröffentlicht 2024 von der Nationalen Akademie der Wissenschaften Leopoldina
- Magazin: „ChatGPT (noch) effizienter nutzen“ veröffentlicht 2024 unter ki-in-kanzleien.de

9.3 Konkretes Fallbeispiel: Vertragsanalyse mit Private GPT

Eine der Hauptaufgaben eines Anwalts ist die Kontrolle und Überprüfung von Verträgen. Da es hier oft um sprachliche Details geht, ist dies eine Aufgabe, die Genauigkeit erfordert und somit zeitintensiv ist. Die Idee ist, das System Verträge nach bestimmten Inhalten scannen zu lassen, die als riskant, ungewöhnlich, möglicherweise rechtswidrig oder anderweitig interessant identifiziert werden, damit der fachkundige Anwalt einen zweiten Blick darauf werfen kann. Der Einsatzzweck der KI-Lösung liegt also in der Unterstützung des Anwalts, indem diese Aufgabe beschleunigt werden kann, und das System sozusagen als zweites Augenpaar und Qualitätssicherung dienen kann. Somit wird das Risiko minimiert, dass wichtige Klauseln außer Acht gelassen werden.

9.3.1 Vorbereitung

Als Grundlage wurde ein existierender und geprüfter Arbeitsvertrag eines Geschäftsführers verwendet. Der Vertrag wurde anonymisiert und anschließend modifiziert, um einige Risiken und ungewöhnliche Klauseln einzubauen. Diese Klauseln waren z.B. eine ungewöhnlich hohe Abfindung, ein fehlender Anspruch auf einen Jahresurlaub und eine beschränkte Haftung des Geschäftsführers auf einen ungewöhnlich geringen Betrag. Somit waren sowohl Risiken für die Gesellschaft als auch für den Geschäftsführer vorhanden, die das System identifizieren sollte. Die zu identifizierenden Klauseln wurden als erwartetes Ergebnis festgehalten, um die Testergebnisse messbar zu machen. Zur Dokumentation wurde eine Tabelle vorbereitet und eine quantitative Bewertungsskala festgelegt, die eine Genauigkeit von „unbrauchbar“ (0) bis „exzellent“ (1) zuließ.

9.3.2 Tests und Feedback

Die ersten Tests basierten auf einfachen Prompts, z.B. *„Analysiere den gesamten Vertrag und identifiziere alle ungewöhnlichen oder risikoreichen Klauseln, die potenzielle Probleme für die Gesellschaft oder den Geschäftsführer darstellen könnten.“* Als Ergebnis lieferte das System einige der

modifizierten Klauseln, ließ andere jedoch außer Acht. Zusätzlich wurden andere Klauseln fälschlicherweise als riskant eingestuft. In der Analyse wurden folgende Probleme abgeleitet:

1. Das System unterscheidet nicht, für welche Partei das Risiko gilt.
2. Es ist nicht klar, auf welche Arten von Risiken geprüft werden soll.
3. Es fehlt der Kontext, im Vergleich zu was eine Klausel als ungewöhnlich einzustufen ist
4. „Normale“ Risiken eines Arbeitsvertrages werden aufgeführt, obwohl sie in jedem Arbeitsvertrag zu erwarten sind (z.B. der Fakt, dass eine Kündigung grundsätzlich möglich ist).

In den nächsten Iterationen wurden diese Probleme durch Optimierungen des Prompts kontinuierlich eliminiert. Der finale Prompt war: *„Analysiere den gesamten Vertrag und identifiziere alle ungewöhnlichen oder risikoreichen Klauseln. Unterscheide klar zwischen den Risiken, die potenzielle Probleme für die Gesellschaft darstellen, und denen, die potenzielle Probleme für den Geschäftsführer darstellen könnten. Konzentriere dich dabei auf relevante Risiken in Bezug auf:*

- *Haftung*
- *Finanzielle Verpflichtungen*
- *Kündigungsbedingungen*
- *Compliance-Anforderungen*

Betrachte Klauseln als ungewöhnlich, wenn sie:

- *Von Standardverträgen abweichen*
- *Zusätzliche Verpflichtungen auferlegen.“*

Dieser Prompts weist auf die explizite Unterscheidung der Risiken für die jeweilige Partei hin, stellt sicher, dass nur relevante Risiken genannt werden, schafft durch die Arten von Risiken den nötigen Kontext und erklärt, wann Klauseln als ungewöhnlich betrachtet werden sollen. Durch die optimierte Formulierung erreicht dieses Beispiel eine hohe Genauigkeit und liefert alle erwarteten Risiken mit korrekter Begründung zurück, ohne weitere Klauseln als fälschlich riskant zu bewerten.

Die Tests wurden jeweils mit- und ohne Dokumentenkontext ausgeführt. Grundsätzlich wurden drei verschiedene Ansätze evaluiert:

- **Vertrag ohne Dokumentenkontext in das Chatfenster kopieren** → Ergebnis basiert auf den Trainingsdaten und dem Skill des Sprachmodells
- **Vertrag in den Dokumentenkontext laden** → Solange der vollständige Vertrag durch das Retrieval selektiert wird, ist das Ergebnis nahezu identisch zu Variante 1, aber in der Handhabung komfortabler, da das Kopieren entfällt
- **Vertrag in das Chatfenster kopieren und relevante Gesetze im Dokumentenkontext verwenden** → Ungeeignet, da die Reihenfolge des Workflows nicht funktioniert. Das System wählt erst basierend auf der Useranfrage die Teile der Dokumente aus, die dann für die Beantwortung genutzt werden. Um die relevanten Gesetzestexte zu identifizieren, müsste aber ERST die eigentliche Bewertung des Vertrages durch das LLM erfolgen, um risikobehaftete Kandidaten zu identifizieren, und DANN passende Stellen der Gesetze gezogen werden.

Hier haben die Tests ergeben, dass Variante 1 die besten Ergebnisse liefert. Variante 2 ist fast identisch, allerdings nur, wenn der vollständige Vertrag selektiert wird. Besonders lange Verträge sind in Abhängigkeit der Anzahl der Chunks, die selektiert werden, dafür ungeeignet. Variante 3 ist in jedem Fall ungeeignet, da hier ein deutlich komplexerer Workflow gebaut werden müsste. Um Gesetzestexte effektiv für eine Vertragsanalyse zu verwenden, sollten sie in das Sprachmodell trainiert werden, etwa via Fine-Tuning. Um die Handhabung der Variante 1 zu verbessern, bietet PGPT mittlerweile einen „Single Document Chat“, wo einzelne Dokumente direkt hochgeladen werden können, ohne vektorisiert zu werden.

9.3.3 Anpassungen der Plattform

Aus den Ergebnissen wurden einige Optimierungspotenziale deutlich, die durch angepasste Systemparameter realisiert werden konnten. Es wurden folgende Bereiche identifiziert:

- **Struktur der Antwort:** Die Vertragsanalyse sollte übersichtlich strukturiert sein.
- **Argumentation:** Die Auswahl der Klauseln sollte umfangreich und treffend begründet werden.
- **Vollständigkeit:** Das System soll alle risikobehafteten Klauseln identifizieren.
- **Formulierung:** Die Antwort soll formal sein und rechtliche Fachbegriffe beinhalten.

Im ersten Schritt wurden die Parameter „Chattiness“ und die Anzahl der Chunks im Kontext des RAG angepasst. Die Verdoppelung der Chunks von 5 auf 10 führte dazu, dass der Vertrag nun vollständig durch RAG selektiert wurde, was die Tests vereinfachte. Diese Anpassung hatte bei diesem Fall keine Auswirkung auf die Ergebnisse, bei anderen Fällen aber eine sehr wesentliche. Deutlich stärker war der Effekt der angepassten Chattiness. Durch eine dedizierte „Legal Chattiness“, wurde die Formulierung stichhaltiger und formaler. Die Argumentation wurde besser, indem das System nun aus einer rechtlichen Perspektive argumentierte (im Vergleich zu einer persönlicheren Perspektive, wie es vorher der Fall war).

Im nächsten Schritt wurde der Systemprompt angepasst. Anfänglich war im System ein generischer Systemprompt hinterlegt, der dann auf den rechtlichen Kontext zugeschnitten wurde. Der neue Systemprompt lautete: *„Du bist ein hilfreicher KI-Assistent, der sich auf rechtliche Anfragen spezialisiert hat. Du antwortest immer freundlich, wahrheitsgemäß und professionell. Deine Antworten sollen präzise und klar sein und, falls notwendig, rechtliche Begründungen enthalten. Wenn du eine Frage nicht beantworten kannst, sage dies offen und erfinde keine Antwort.“* Dadurch wurde die Argumentation noch etwas besser, die Anpassung hatte aber keine Auswirkung auf die Struktur und Vollständigkeit der Antwort.

In unserer letzten Iteration haben wir daher den Systemprompt weiter angepasst und ihn auf den konkreten Anwendungsfall der Vertragsanalyse fokussiert. Die erwartete Struktur wurde klar vorgegeben und die Aufgabe, relevante Risiken zu finden, war ebenfalls Teil des Prompts. Auf diesem Wege konnten wir unsere insgesamt besten Ergebnisse erzielen, da der Output nun die passende Struktur hatte und sämtliche Risiken korrekt identifiziert wurde (siehe Abbildung x). Der angepasste Systemprompt lies auch eine höhere Toleranz beim User Prompt zu, der jetzt auch etwas weniger komplex sein konnte. Auf der anderen Seite wurden andere Anwendungsfälle durch diesen spezifischen Ansatz sehr negativ beeinflusst, was die Bedeutung spezifischer Szenarien mit dedizierten Systemprompts unterstreicht.

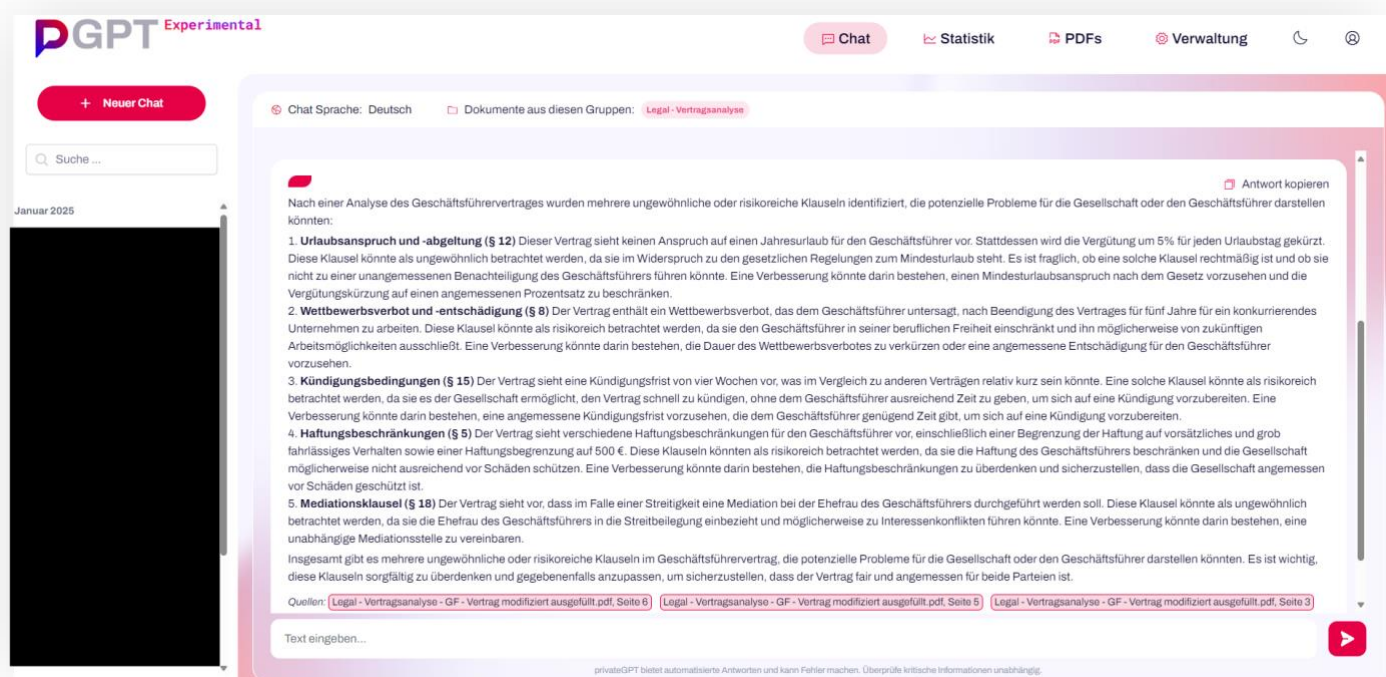


Abbildung 6: Prompt Ergebnis der Vertragsanalyse.

10 Glossar

Begriff	Erläuterung
Chunk	Damit RAG große Datenmengen effektiv verarbeiten kann, werden diese in kleinere, handhabbare Teile aufgeteilt. Das sind Chunks. Im einfachsten Fall entspricht ein Chunk einer Seite eines PDFs.
Context Window	Jedes Sprachmodell hat ein begrenztes Context Window, was die Anzahl der vom Modell verarbeitbaren Zeichen (genauer: Tokens) beschreibt. Praktisch gesehen ist das die maximale Länge der Eingabe, die vom System für einen Prompt akzeptiert wird. Mixtral NeMo kann mit dem Context Window von 128.000 Tokens z.B. ca. 213 Din A4 Seiten verarbeiten.
Finetuning	Finetuning ist das vertiefte Trainieren eines Sprachmodells auf einen spezifischen Anwendungsfall oder eine Branche.
Prompt	Ein Prompt ist eine textbasierte Eingabe, die dem Sprachmodell sagt, was es tun soll. Ein Prompt kann eine Aufgabe, Frage, Aussage oder ähnliches sein.
Prompt-Engineering	Prompt-Engineering ist der Ansatz, Eingaben an die KI, textlich, strukturell und inhaltlich so aufzubereiten, dass bestmögliche Ergebnisse geliefert werden.
Retrieval Augmented Generation (RAG)	Retrieval Augmented Generation (RAG) ist eine Technologie, die einem LLM gezielt zusätzliche Informationen zur Beantwortung einer Aufgabe liefert, ohne das LLM selbst zu verändern. Im Fall von Private GPT werden relevante Inhalte aus den Daten ausgewählt und dem LLM zur Verfügung gestellt, um Antworten auf das Wissen des Unternehmens zu stützen und die Qualität zu steigern.
System-Prompt	Der System-Prompt wird dem Sprachmodell zusätzlich übergeben, in Ergänzung zum User-Prompt. Der System-Prompt ist somit eher statisch und instruiert das System, <i>wie</i> es antworten soll. Hier wird der Schreibstil, die Struktur der Antwort und die Rollenbeschreibung der KI festgelegt.
Use Case	Ein Use Case beschreibt hier eine konkrete Tätigkeit oder Aufgabe, die von der KI durchgeführt werden soll (und normalerweise von einem Menschen durchgeführt wird). Darunter fällt zum Beispiel die Vertragsanalyse.
User-Prompt	Der User-Prompt wird durch den Benutzer eingegeben. Dieser Prompt ist individuell und beschreibt, <i>was</i> das System tun soll.
Vektor-Datenbank	Eine Vektordatenbank stellt Informationen als mathematische Vektoren dar. Die Vektoren können z.B. Texte, Bilder oder Videos repräsentieren. Die Datenbank ist darauf optimiert, ähnliche Informationen zu finden, was besonders für RAG entscheidend ist.