FIND THINGS SMARTER

Über den Einsatz von Large Language Models im Enterprise Content Retrieval

Dr. Thomas Kurz
Rupert Westenthaler
Sebastian Baron

Franz-Josef-Straße 15, 5020 Salzburg
Tel.: +43 662 27 66 80 • Mail: office@redlink.at

November 2024

EINLEITUNG

Willkommen zu unserem Whitepaper „Find Things Smarter: Über den Einsatz von Large Language Models im Enterprise Content Retrieval“.

Die digitale Transformation hat die Art und Weise, wie Unternehmen ihre Informationswelten erschließen und nutzen, maßgeblich verändert. In einer zunehmend komplexeren Geschäftswelt ist der Zugang zu präzisen und relevanten Informationen entscheidend. Hierbei spielen LLMs eine zunehmend wichtige Rolle, da sie nicht nur natürlichsprachliche Text verarbeiten, sondern auch kontextbezogene Antworten generieren und so einen natürlichen Zugang zu Informationen bieten können.

Dieses Whitepaper gibt Ihnen einen detaillierten Einblick in die Bedeutung eines effektiven Information Retrievals (IR) im Unternehmenskontext und zeigt, wie moderne Technologien wie Retrieval-Augmented Generation (RAG) den Zugang zu Informationen grundlegend verändern können. Neben den Vorteilen dieser Technologien werden auch die Schwächen und Herausforderungen thematisiert und verschiedene Methoden vorgestellt, diese zu überwinden.

Ziel dieses Whitepapers ist es, IT-Entscheidern, Unternehmensberatern und Innovationsmanagern praktische und theoretische Ansätze an die Hand zu geben, um die Informationsflüsse innerhalb ihres Unternehmens zu optimieren. Dabei wird sowohl auf aktuelle Technologien als auch auf praxisnahe Lösungen eingegangen.

Wir hoffen, dass dieses Whitepaper Ihnen wertvolle Erkenntnisse und Anregungen bietet, die genannten innovativen Technologien auch in ihrem Unternehmen zu etablieren.

Dr. Thomas Kurz

01 NEUE MÖGLICHKEITEN UND HERAUSFORDERUNGEN DURCH LLMs

Large-Language-Modelle (LLMs) existieren seit etwa 2017. Im Laufe des Jahres 2022 wurden sie plötzlich populär, als es mit Modellen wie LLaMA, Bloom oder GPT-3.5 gelang, LLMs in Bezug auf Robustheit, Genauigkeit und Konversationsfähigkeit deutlich zu verbessern. Seitdem haben LLMs viele neue Möglichkeiten gezeigt, um mit Information und Wissensansammlungen zu interagieren.

Insbesondere für Unternehmen mit großen Dokumentsammlungen, kaum genutzten Datensilos und/oder einem verstreuten Wissenskorpus — für die klassische Schlagwortsuche keine ausreichenden Ergebnisse liefert — erscheint eine Integration von LLMs in Unternehmensabläufe daher sehr attraktiv. Laut O'Reilly Tech Trends 2023 gaben 67% der global befragen Unternehmen an, generative AI in Form von LLMs in ihren Unternehmensabläufen zu verwenden.

Schon 2023 gaben 67% der global befragen Unternehmen an, generative AI in Form von LLMs in ihren Unternehmensabläufen zu verwenden. Aber nur 23 % verwenden kommerzielle Modelle, vor allem wegen datenschutzrechtlichen und Sicherheitsbedenken.

Auf der anderen Seite gaben in einer von Predibase durchgeführten Umfrage im August 2023 allerdings nur 23% der Betriebe an, kommerzielle Modelle realisiert zu haben, die Zurückhaltung in dieser Hinsicht resultiere vorwiegend aus datenschutzrechtlichen und Sicherheitsbedenken.

Mittlerweile hat sich auch gezeigt, dass der Einsatz von LLMs nicht immer fehlerfrei funktioniert und faktisch falsche Resultate produzieren kann, auch wenn diese überzeugend formuliert sind. Für derartige Erzeugnisse hat sich der Begriff „Halluzination" etabliert. Sie stellen insbesondere bei einer Verwendung in unternehmensinternem Kontext eine der größten Herausforderungen im Umgang mit LLMs dar.

Um diesem Problem entgegenzuwirken, und um mit einem bestimmten Wissenskorpus arbeiten zu können, hat sich der Einsatz eines als „Retrieval-Augmented-Generation (RAG)“ bezeichneten Verfahrens bewährt. RAG ist ein hybrider Ansatz, bei welchem der zur Behandlung der jeweiligen Anfrage relevante Teil der Dokumentensammlung abgerufen und das Kontextfenster des LLMs mit diesem erweitert wird. Dadurch können Informationen, welche nicht Teil des Trainingsdatensatzes des LLMs waren, zur Behandlung von Anfragen verwendet und mit den Stärken von LLMs in den Bereichen Sprachverständnis und Generierung natürlicher Sprache kombiniert werden.

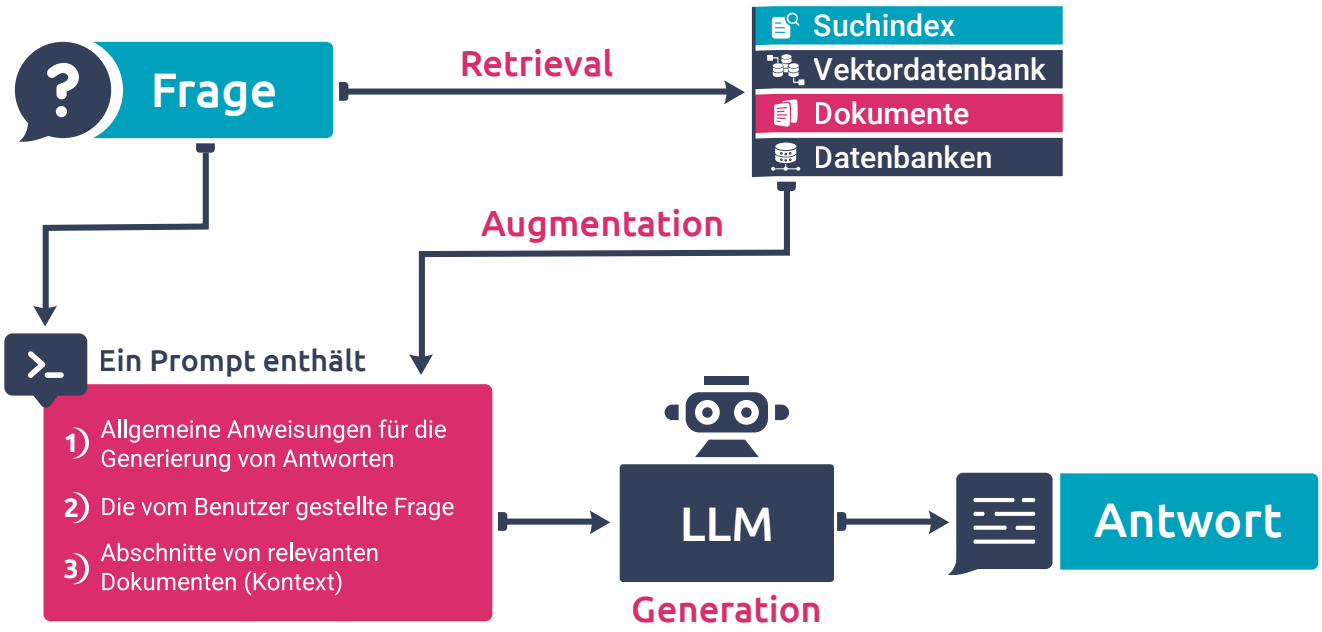
Retrieval-Augmented-Generation (RAG) macht aus einem statischem LLM ein dynamisches System zur Anfrage an spezifisches Unternehmenswissen

02 RETRIEVAL AUGMENTED GENERATION

- Unternehmen können dadurch mit dem eigenen Wissenskorpus interagieren und Informationen aus potenziell riesigen und möglicherweise verteilten Datensammlungen abrufen, ohne dass ein eigenes Modell trainiert werden muss.
- Unternehmenskunden kann ein Chatbot zur Verfügung gestellt werden, um auf potenzielle Kundenanfragen automatisiert Antworten auf Basis der jeweils relevanten Kundendaten aufzubereiten und zurückzumelden.
- Kombiniert mit unternehmensinternen Datenbanken und Websuche kann RAG auch im Bereich Business Analytics und Prognoseerstellung Anwendung finden.

» Da Ausgaben der Sprachmodelle mittels Prompt Engineering umfangreich und detailliert gesteuert werden können, können mit dieser Technik sogar automatisiert entsprechend umfangreiche Berichte erstellt werden.

Die Anforderung an die Präzision eines RAG-Systems ist umso wichtiger, je größer der Datenkorpus ist, aus dem Informationen abgerufen werden müssen, da mit zunehmender Größe für einzelne Informationsabfragen mehr Kandidatendokumente zur Verfügung stehen. Da aber wiederum das Kontextfenster des LLMs, welches die Information aufbereiten und schlussendlich eine Rückmeldung an den Nutzer geben soll, begrenzt ist, ist es entscheidend, dass diese Dokumente zuvor präzise auf jene Kontextabschnitte reduziert werden, welche tatsächlich den größten Beitrag zur Beantwortung der Anfrage leisten können.



Leider funktioniert RAG nicht einfach „out-of-the-box“ also ohne zusätzliche Maßnahmen und Anpassungen. Das liegt unter anderem an den semantischen Differenzen zwischen Nutzeranfragen und den zu deren Beantwortung zur Verfügung stehenden Dokumenten. Einfach ausgedrückt könnte man sagen, dass Fragen und ihre zugehörigen Antworten in Bezug auf die semantische Einbettung oft zu unterschiedlich voneinander sind. Dazu kommt, dass die Vektoreinbettung und die Kosinus-Ähnlichkeit als Abstandsmaß mit ihrer impliziten Annahme, dass alle Dimensionen die gleiche Wichtigkeit besitzen, oft nur grob betrachtet zum Abruf („Retrieval“) korrekter Ergebnisse führen, die Reihung der Relevanz der Dokumente allerdings nicht besonders gut ist. Dies wird insbesondere bei sehr großen Dokumentensammlungen und/oder LLMs mit kleinen Kontextfenstern zum Problem, da man sich in solchen Fällen darauf verlassen muss, dass der Such- und Vergleichsprozess auf Basis der Nutzeranfrage in der Lage ist, präzise Ergebnisse aus einer großen Datenmenge herauszufiltern

RAG funktioniert nicht „out-of-the-box”, sondern erfordert daten- und informationsspezifisches Engineering für das Unternehmen.

03 ENTWICKLUNG DES RAG KONZEPTS

Aus den oben genannten Gründen wurde das RAG-Konzept weiterentwickelt und um zusätzliche Komponenten erweitert, welche anhand ihrer Position innerhalb der RAG-Pipeline in „Pre-Retrieval“ und „Post-Retrieval“ Methoden eingeteilt werden können.

- Pre-Retrieval Methoden befassen sich mit der Überwindung der semantischen Differenzen zwischen Nutzeranfragen und möglicher innerhalb der indizierten Dokumente befindlicher Antworten. Dies geschieht vor allem durch Anpassung bzw. Umformulierung der Nutzeranfrage („Query Rewriting“), so dass eine bessere semantische Vergleichbarkeit mit dem Dokumentenkorpus gegeben ist.

- Post-Retrieval-Methoden versuchen, den abgerufenen relevanten Kontext so zu priorisieren bzw. auszudünnen, dass lediglich jene Information an das LLM weitergegeben wird, die auch tatsächlich zur korrekten Beantwortung der Nutzeranfrage benötigt wird. Re-Ranking, um die abgerufenen Dokumentbestandteile in ihrer Relevanz für die Nutzeranfrage neu zu priorisieren ist dabei die am häufigsten verwendete Methode. Aber auch Ansätze, die den abgerufenen Kontext zusammenzufassen und die Menge an Information, die an das LLM weitergegeben wird, zu reduzieren, kommen hier zum Einsatz.

In jüngster Zeit scheint das RAG-Konzept einen Paradigmenwechsel hin zu einem eher modularen System zu vollziehen, in welchem einzelne Module gewissermaßen zu einem programmartigen RAG-System anstatt einer gerichteten Pipeline angeordnet werden. Dieses kann etwa auch Schleifen und Verzweigungen beinhalten. Die Optimierung der Qualität der Antworten auf die Nutzeranfragen geht allerdings auch zu Lasten der Laufzeit, welche zur Beantwortung einer einzelnen Anfrage benötigt wird, was ihre Einsetzbarkeit für Industrieapplikationen oft fraglich erscheinen lässt.

Anfragen müssen vorverarbeitet werden, damit gute und korrekte Ergebnisse erwartet werden können.

Unabhängig vom Paradigma, unter welchem das RAG-Konzept betrachtet wird, besteht ein zentraler Aspekt darin, die Nutzeranfrage so vorzuverarbeiten, dass der nachfolgende Abgleich mit der Dokumentdatenbank erleichtert wird. Das wird in einer breiteren Sichtweise auch als Anfrageoptimierung („Query-Optimization“) bezeichnet.

04 ANFRAGEOPTIMIERUNG

Es gibt es mehrere Möglichkeiten, wie unterschiedliche Ansätze zur Anfrageoptimierung gegliedert werden können:

- Diese richten sich zum einen nach der Interaktion mit dem Nutzer und unterscheiden dabei zwischen „Ad-hoc“ und „Dialog-Szenarien“, abhängig davon, ob eine Anfrage in einer einzigen Rückmeldung resultieren soll oder eine dialogartige Interaktion möglich ist, wodurch beispielsweise klärende Fragen an den Benutzer im Falle einer ambivalent formulierten Anfrage gestellt werden könnten.

Je nach Interaktion werden „ad hoc“ oder „Dialog“-Szenarien unterschieden, um Anfragen aufzubereiten.

- Eine weitere Unterscheidung betrifft die Frage, ob vor der Optimierung bereits Inhalte aus der Dokumentdatenbank abgerufen wurden. Bei einem modularen Verständnis des RAG-Konzeptes ist es auch möglich, dass eine Nutzeranfrage mehrere Interaktionen mit der Dokumentdatenbank nach sich zieht und beispielsweise die Ergebnisse eines ersten Abrufschrittes dazu verwendet werden, um die Anfrage selbst umzuformulieren und besser an die semantischen Begebenheiten der Dokumente in der Datenbank anzugleichen. Solche Verfahren werden in der Literatur häufig als „Corpus- enhanced“ bezeichnet, während Methoden, welche sich rein an der Nutzeranfrage orientieren, im Folgenden als „LLM-only“ Methoden bezeichnet werden.

Für die Optimierung wird entweder nur ein LLM oder auch der vorhandene Korpus verwendet.

Zuletzt muss auch nach den verwendeten Werkzeugen unterschieden werden: zwischen Prompting-Methoden, also Ansätzen, bei welchen die Erweiterung bzw. Umformulierung der Nutzeranfrage durch zielgerichtete Instruktionen an ein LLM erreicht werden und Trainings-Methoden, bei welchen ein eigenes (LLM-ähnliches aber in Bezug auf die Parameteranzahl häufig kleineres) Modell für diesen Zweck trainiert wird.

Prompting			Training	
Expansion	Rewriting	Routing	Finetuning	Distillation
Erweiterung der ursprünglichen Anfrage	Umformulierung der Anfrage z.B. durch Beispiele	Die Anfrage wird kategorisiert und an ein spezifisches Modell bzw. System geleitet	Das LLM wird trainiert mit neuen, spezifischen Daten	Ein kleineres LLM lernt von einem sehr großen Modell

Da innerhalb der „Ad-hoc-Rewriting“ Situation eine Vielzahl unterschiedlicher Ansätze existieren, lohnt sich hier noch eine feinere Unterteilung der Prompting Methoden in Erweiterungs- („Expansion“) und Umformulierungs- („Rewriting“) Methoden abhängig davon, ob das vordergründige Ziel darin besteht, die Anfrage durch Anreicherung zusätzlicher Komponenten informativer für den Prozess der Dokumentensuche zu machen, oder ob die Anfrage als Ganzes umgeschrieben werden soll. Bei den Trainingsmethoden kann Finetuning und Destillation unterschieden werden, wobei erstere Methode das Ziel verfolgt, ein gegebenes Sprachmodell auf einem vorhandenen Trainingsdatensatz für die Aufgabe der Anfrageoptimierung nachzutrainieren, während man unter dem Konzept der „Knowledge-Distillation“ den Fähigkeitentransfer eines LLMs auf ein kleineres Modell in einem bestimmten Bereich (also z.B. Anfrageoptimierung) versteht. Als letztes steht „Routing“, welches eine Verzweigung der RAG-Pipeline auf Basis bestimmter Merkmale der Nutzeranfrage bedeutet (sodass unterschiedliche Nutzeranfragen

Zur Umsetzung wird unterschieden zwischen „Prompting-“ und „Trainings-Methoden“

mit unterschiedlichen Methoden und Ressourcen bearbeitet werden können), gewissermaßen zwischen den übergeordneten Konzepten des Promptings bzw. Trainings, da es sowohl trainierbare Router gibt als auch solche, die sich durch reines Prompting von LLMs realisieren lassen.

METHODEN FÜR DAS „AD-HOC“ SZENARIO

Bei den Ad-hoc Methoden sind bis auf das Stellen der Anfrage seitens des Nutzers keine weiteren Eingaben bzw. Interaktionen mit dem RAG-System vorgesehen. Das gewährleistet eine rasche Beantwortung der Anfrage des Nutzers, resultiert in einer planbaren Bearbeitungszeit und vermeidet den Medienbruch eines dialogfähigen Assistenten mit einer Suchmaschine.

UMFORMULIERUNG IM DIALOGKONTEXT

Moderne LLMs zeichnen sich nicht nur durch linguistische Fähigkeiten aus, sondern sind auch in der Lage, dialogartige Konversationen zu führen wie ChatGPT, Gemini, etc. eindrucksvoll beweisen. Dementsprechend gibt es auch Bemühungen, RAG-Kontexte dialogartig zu gestalten, um etwa Nutzern zu erlauben, ihre Anfrage zu präzisieren bzw. auf Rückfragen zu antworten. Während dieser Ansatz generell das Potential hat, die Genauigkeit der abgerufenen Inhalte zu verbessern, bringt er auch neue Probleme und Herausforderungen mit sich: Steht ein rascher Zugriff im Vordergrund, ist es eventuell gar nicht wünschenswert, zunächst mehrere Dialogiterationen zu durchlaufen. Des weiteren ist fraglich, ob LLMs ohne weitere Anpassungen in der Lage sind, qualifizierte Rückfragen zu stellen, die die Qualität der abgerufenen Dokumentenabschnitte auch tatsächlich verbessern, da sie hierfür genau genommen den gesamten Kontext der Dokumentdatenbank benötigen würden.

05 EXKURS: PROGRAMMATISCHES PROMPTING

Ein wichtiges Problem in diesem Zusammenhang ist das Design der Prompting-Anweisung an ein LLM, um die gewünschte Aufgabe optimal auszuführen. Prompt Engineering ist zu einem eigenen Forschungsfeld geworden, beispielsweise die optimale Prompting-Methode durch Verwendung der Chain-of-Thought-Prompting-Technik oder das Pseudo-Relevance Feedback, um optimale Anfrageumformulierungen zu erhalten. Prompt Engineering ist dann problematisch, wenn man das verwendete LLM innerhalb einer bestehenden RAG-Pipeline austauschen möchte, weil der optimale Prompt an ein LLM sich sehr stark sowohl zwischen einzelnen LLMs als auch verschiedenen Versionen desselben LLMs unterscheiden kann. Bei DSPy handelt es sich um ein Python-Paket der Universität Stanford in Anlehnung an das „Demonstrate-Search-Predict (DSP)“ Prinzip. Dabei handelt es sich um ein Konzept des modularen RAGs mit dem Ziel, dynamisch LLMs mit Retriever-Modellen zu verbinden. Tatsächlich ist DSPy für den Aufbau modularer RAG-Programme besonders geeignet, da es ermöglicht, mit LLMs in programmatischer Weise zu interagieren, anstatt manuelle Prompting-Methoden zu verwenden.

Demonstrate-Search-Predict (DSPy) ist ein Framework zum programmatischen Einsatz von Prompt-Anweisungen

Da DSPy gute Integrationsmöglichkeiten für unterschiedliche Sprachmodelle (lokal oder via APIs) und Retrieval-Modelle bietet, kann damit ähnlich wie mit LangChain oder LlamaIndex eine komplette RAG-Pipeline aufgebaut werden. Die Vorteile von DSPy gegenüber den erwähnten Frameworks bestehen darin, dass Prompt-Engineering durch programmatische Befehle ersetzt wird und ein bestehendes DSPy-Programm bei Vorhandensein eines entsprechenden Datensatzes in Hinblick auf Prompts und Few-Shot-Beispiele

auch optimiert werden kann. LLM-Programme können im Rahmen von DSPy gewissermaßen ähnlich wie Machine-Learning-Programme angesehen werden: Eine Menge von Modulen kann dynamisch mit einem Programm verbunden werden, um dessen Parameter nach Definition einer Verlustfunktion und eines Optimizers auf einem vorhandenen Trainingsdatensatz zu optimieren. Das Aufkommen von Frameworks wie DSPy und zuvor auch schon LangChain und LlamaIndex machen deutlich, dass LLMs nicht nur isolierte Dienste sind, sondern auch als Komponenten in komplexere Applikationen integriert werden können.

06 ZUSAMMENFASSUNG UND EMPFEHLUNGEN

Die Anwendung von RAG erlaubt es, einer der größten Limitationen im Umgang mit LLMs entgegenzuwirken, nämlich dem Fehlen von domänen-spezifischem Wissen, welches nicht Teil der Trainingsdaten des jeweiligen LLMs war und damit bei Anfragen an „out-of-the-box“ LLMs normalerweise zu Halluzinationen führen würde.

Wir fokussieren hier auf einen wichtigen Aspekt von RAG-Programmen oder -Pipelines, nämlich die Optimierung der Nutzeranfrage im Hinblick auf die Suche nach passenden Passagen, um die Anfrage beantworten zu können. Die Hauptprobleme bestehen dabei darin, dass solche Anfragen oft nur kurz und eventuell uneindeutig sind, vor allem aber, dass sich die Vektoreinbettungen solcher Anfragen nur bedingt für Ähnlichkeitsvergleiche mit auf gleicher Weise eingebetteten Dokumenten eignet, da

sich Anfragen und die Dokumente zu deren Beantwortung semantisch oft deutlich unterscheiden.

Nachfolgend werden unterschiedliche Möglichkeiten aufgezeigt, um mit diesen Problematiken umzugehen:

- Die wichtigste Unterscheidung dieser Ansätze ist dabei jene in Prompting- bzw. Trainings- Methoden. Die Entscheidung, welcher Ansatz geeigneter ist, ist dabei problemspezifisch und hängt davon ab, ob Daten vorhanden sind, mit welchen ein kleineres Modell trainiert werden könnte, oder ein RAG-System neu eingeführt werden soll. Dabei macht die Verwendung von Prompting-Methoden Sinn: nicht nur aufgrund der fehlenden Daten, sondern auch weil Prompting-Ansätze mit weniger Ressourcen implementiert und auch verändert werden können. Das Trainieren oder Transferlernen auch von kleineren Sprachmodellen ist dagegen mit deutlich höherem Rechenaufwand und damit auch höheren Kosten verbunden.

- Die höhere Flexibilität von Prompting- Methoden ist auch in Hinblick auf die Häufigkeit zukünftiger Anpassungen von Vorteil. Müsste das verwendete Modell häufig auf neuen Daten nachtrainiert werden, um Veränderungen gerecht zu werden, wird dieser Ansatz aufgrund der damit verbundenen Kosten eher unattraktiv. Nichtsdestotrotz, falls man schon Erfahrung mit einem bestehenden RAG- oder Query-Optimierungsansatz in Form eines Datensatzes von Nutzeranfragen gemacht hat und sich die Zahl der zukünftigen Veränderungen in Grenzen hält, kann das Trainieren eines spezialisierten Query-Umformulierungsmodells die Ausführungszeit verkürzen, die Anfälligkeit für Halluzinationen reduzieren und auf längere Sicht sogar Kosten einsparen.

- Bei den Prompting-Methoden konnte gezeigt werden, dass die besten Ergebnisse mit Kombinationen aus der Query-to-Document-Methode (Q2D) bzw. der Chain-of-thought-Prompting-Technik im Verbund mit Pseudo-Relevance Feedback erzielt werden konnten. Es wurden allerdings nicht alle erwähnten Methoden verglichen, außerdem sind Performance-Metriken wie Recall oder F1-Score für Unternehmen nicht der einzige zu berücksichtigende

Punkt, sodass dieses Resultat eher als grobe Orientierung angesehen werden kann. In Bezug auf Q2D kann etwa angemerkt werden, dass sich deutliche Laufzeitvorteile dadurch ergeben, dass man den Ansatz umkehrt und bereits bei der Dokumenteinbettung mehrere potenzielle Anfragen erzeugt, welche den späteren Suchprozess unterstützen. Dieser Laufzeit-Performance-Trade-off ergibt sich auch bei den vorgestellten moderneren Corpus-enhanced-Rewriting-Techniken, da hier unter Umständen deutlich mehr als einmal auf die Dokumentdatenbank zugegriffen wird.

- Bei den Trainingsansätzen stellt das Konzept von Knowledge-Distillation eine interessante Möglichkeit dar, ein bereits etabliertes Anfrageoptimierungssystem effizienter zu gestalten, indem die Fähigkeiten des verwendeten LLMs anhand eines Datensatzes auf ein kleineres spezialisiertes Modell übertragen werden. Dabei kann auch ein Modell verwendet werden, welches bereits für die Aufgabe der Anfrageumformulierung trainiert wurde. Da viele der vorgestellten Modelle bzw. deren Modellgewichte öffentlich verfügbar sind, dürfte das die Menge an für den Finetuning-Prozess benötigten Daten deutlich reduzieren. Innerhalb der vorgestellten Knowledge-Distillation-Methoden ist überwachtes Trainieren auf einem annotierten Datensatz mit Abstand die am häufigsten verendete Technik, allerdings erfordert diese die in der Regel manuell durchzuführende Arbeit der Datenannotation. Eine spannende Alternative, um diesen Arbeitsschritt zu vermeiden ist es, Reinforcement-Learning zu verwenden, da hierbei der Reward auch direkt von einem LLM geliefert werden kann, das beispielsweise die Beantwortbarkeit einer Anfrage aufgrund des abgerufenen Kontextes bewertet, was zwar weniger Kontrolle über den Trainingsprozess ermöglicht, allerdings keine manuelle Datenannotation erfordert.

Abschließend ist außerdem zu sagen, dass LLMs für Unternehmen natürlich auch noch andere Anwendungsgebiete bereithalten als solche, die sich durch RAG-Programme realisieren lassen. Gleiches gilt für die Risiken, etwa Sicherheitsrisiken durch den möglichen Zugriff auf unternehmensinterne

Datensammlungen, ethische Fragestellungen, der Black-Box-Charakter solcher Modelle, etc.

Eine Beschäftigung mit diesen allgemeinen Herausforderungen dürfte sich allerdings für viele Unternehmen lohnen, denn auch wenn der allgemeine Hype um LLMs aktuell eher am Abnehmen ist, ist es doch sehr unwahrscheinlich, dass die Verwendung von Künstlicher Intelligenz aus der Unternehmenswelt wieder vollständig verschwinden wird.

07 AUSBLICK

Dieses Whitepaper beleuchtet die Bedeutung eines effektiven Information Retrievals (IR) im Unternehmenskontext und zeigt, wie moderne Technologien wie Retrieval-Augmented Generation (RAG) den Informationszugang revolutionieren können. Neben den Stärken und Schwächen von RAGs wurden praxisorientierte Methoden vorgestellt, um die entstehenden Herausforderungen zu meistern und die Effizienz von Informationsprozessen zu steigern.

Wir hoffen, dass die enthaltenen Erkenntnisse und Erfahrungen für Sie inspirierend und nützlich waren. Sollten sie Unterstützung bei diesem Thema benötigen, stehen wir Ihnen mit unserer jahrelangen Expertise in den Bereichen Information Retrieval und Natural Language Processing gerne zur Seite. Wir würden uns freuen, sie bei ihrer Transformation und beim Erschließen Ihrer Informationswelten zu unterstützen.

Vielen Dank für Ihr Interesse!

Als Basis für dieses Whitepaper dient der Abschlusbericht des von der FFG-geförderten Projektes [„LLM für Unternehmen“, Oktober 2024].