Heart-a-Tech Podcast by MaibornWolff

Folge 01:

Personal Assistant für alle!
Wie LLM aus Daten Mehrwert für Teams und Kunden schafft

Mit Experte Konrad Schreiber

Intro Konrad Schreiber: Trotzdem ist das ein Transformationsprozess, in dem wir uns da gerade befinden, für jeden einzelnen Mitarbeiter, für jede einzelne Mitarbeiterin. Alle müssen schauen: Was mache ich jetzt eigentlich mit diesem Werkzeug, das ich da habe?

Brigitte Streibich: Herzlich willkommen bei Heart-a-Tech, dem Podcast rund um alles, was du wissen musst, um neue IT-Trends und Technologien erfolgreich in deinem Unternehmen zu implementieren. Dieser Podcast ist happily hosted by MaibornWolff. Ich bin Brigitte, und wir sprechen heute in Folge 1 über das Thema "Personal Assistent für alle! Wie LLM aus Daten Mehrwert für Teams und Kunden schaffen". Ich habe mir einen ganz besonderen Gast eingeladen. Es geht um Konrad Schreiber. Er ist Experte für die Anwendung von Machine Learning im Unternehmenskontext. Hallo, Konrad!

Konrad Schreiber: Hallo.

Brigitte Streibich: Wir sind ja hier im Podcast Heart-a-Tech. Wenn du an deine Arbeit denkst, was lässt dein Herz höherschlagen?

Konrad Schreiber: Was mein Herz höherschlagen lässt? Ich bin absolut technikbegeistert, tatsächlich schon immer gewesen. Deswegen habe ich auch irgendwann mal angefangen, Informatik zu studieren und habe davor schon in den Neunzigern rumprogrammiert. Ich finde es faszinierend, was die Technologie uns mittlerweile für Möglichkeiten bietet - insbesondere was im Bereich Data Science, AI schon immer möglich war. Schon in den Nullerjahren, in den Zehnerjahren kam das Deep Learning auf, so eine Art neuer Advent der neuronalen Netzwerke. 2012 war es tatsächlich das erste Mal möglich, sinnvoll Bilderkennung zu machen und das hat sich irrsinnig weiterentwickelt. Das finde ich unglaublich faszinierend. Und dann ist es natürlich toll zu schauen, welche Herausforderungen wir draußen in den Organisationen dieser Welt - insbesondere in Deutschland - jetzt haben und welche der AI-Methoden helfen können. Wenn ich irgendwo ein Werkzeug finde, das ich kenne und das auf ein Problem einer unserer Geschäftspartner passt, das lässt mein Herz dann höherschlagen.

Brigitte Streibich: Man hört oft in der Öffentlichkeit, in den Diskussionen in den Medien bei Künstliche Intelligenz, das Thema ChatGPT. Jetzt setzen wir das in den Unternehmenskontext. Ich stell mir das ein bisschen so vor: Der Produktentwickler kommt morgens zur Arbeit, klappt seinen Laptop auf, gibt einen Prompt ein, holt sich einen Kaffee und schaut dann einfach nur zu, wie die KI arbeitet. Die KI beantwortet E-Mails, weiß genau, was gestern in den Meetings besprochen wurde, erledigt schon mal To-dos, schickt Status-Updates an Team-Mitglieder - und das über unterschiedliche Tools hinweg. Ist das die Zukunft, oder sind wir da heute schon?

Konrad Schreiber: Sehr gute Frage. Was bring die Zukunft? Da sind wir gerade an einem Punkt: Das kann wahrscheinlich niemand so ganz genau sagen. Gerade mit der Technologie Foundation Models, Large Language Models oder auch Generative AI haben wir ganz neue Werkzeuge an die Hand bekommen, die jetzt tatsächlich funktionieren. Nach und nach finden wir jetzt erst raus, was diese Werkzeuge tatsächlich leisten können.

Grundsätzlich bin ich immer ein Freund davon zu schauen, was jetzt technisch schon möglich ist und nicht so sehr den Blick in die Zukunft zu werfen. Denn: Wir können es nicht wirklich vorhersehen. Es gibt aber jetzt schon viele Anwendungsfälle. Vieles von dem, was du gesagt hast, ist jetzt bereits umsetzbar mit den verfügbaren Technologien und wird sich natürlich noch weiterentwickeln. Diese Überlegung von einem persönlichen Assistenten - gerade auch im Berufskontext - wird kommen. Wir sind tatsächlich schon dabei, so etwas mit Kunden umzusetzen. Es gibt ein paar Challenges, über die wir jetzt wahrscheinlich ein bisschen sprechen werden.

Brigitte Streibich: Da du gerade schon die Herausforderungen ansprichst. Wenn du jetzt zum Kunden gehst, wo stehen die, wenn man sich das auf einer Skala von null bis zehn anschaut? Wie weit sind sie bei den Themen LLM und auch GPT?

Konrad Schreiber: Vielleicht ist es sinnvoll zu gucken, welche Reifegrade haben wir und wie kann man sie einordnen. Stehen wir gerade am Anfang, um so was zu bewerten? Ich würde mal sagen, wir haben da eine breit gefächerte Skala. Es geht damit los, dass man Tools verwendet, bei denen ein Large Language Model irgendwo unter der Haube läuft. Zum Beispiel ChatGPT. Man geht hin, hat ein Problem, stellt eine Frage, kriegt die Antwort. Man benutzt eine Bing Search oder jetzt auch eine Bing Enterprise Search für eine Suche, das Large Language Model interagiert dann schon mit dem Internet, mit Informationsquellen und synthetisiert daraus eine Antwort. Bei dem, was da unter der Haube passiert, sind wir von dem Reifegrad her ein Stück weiter. Dass der Anwender diese Tools verwendet, wäre Reifegrad eins.

Der nächste Schritt ist, solche Tools zu entwickeln. Das ist für viele unserer Kunden sehr interessant. Der erste Schritt, versteh mich nicht falsch, ist auch wichtig: jeder Mitarbeiter, jede Mitarbeiterin in dem Unternehmen sollte LLM-getriebene Tools für die tägliche Arbeit verwenden und auch selber herausfinden, wo sie mir helfen und wo nicht. Die Challenges drum herum, das Juristische muss die Organisation klären. Das ist der Reifegrad eins.

Beim Reifegrad zwei bauen wir solche Applikationen selbst Inhouse. Das geht auf verschiedene Arten. Über eine API kannst du zum Beispiel auf ein Large Language Model zugreifen. Da gibts gerade bei Azure einen sehr guten Zugang. Über die Azure Open AI Services lässt sich über eine API direkt auf GPT 3.5, GPT 4 zugreifen und mit dem Modell interagieren. So kann man beliebig komplexe Applikationen bauen. Man ist allerdings immer von dem Modell noch ein Stück weit entfernt. Das Nächste, was man machen kann, ist dieses Modell, was dort von einem Hyperscaler gehostet wird, nehmen und seinen eigenen Bedürfnissen ein Stück weit anpassen. Das heißt dann Fine-Tuning oder auch Transfer Learning. Und auch das geht über eine API. Man kann beispielsweise so einem Modell Vorlagen schicken und sagen: Verhalte dich so, wie es in diesen Vorlagen beschrieben steht, und das Modell tut es. Dann ist es fine-getuned auf einem bestimmten Anwendungsfall.

Brigitte Streibich: Das wäre jetzt Reifegrad drei, oder?

Konrad Schreiber: Genau. Wir sind jetzt bei drei angekommen, beziehungsweise drei A. Drei B ist dann tatsächlich ein großer Schritt. Es ist sehr spannend, was aktuell passiert ist. Es wurde jetzt das Llama 2 Modell von Meta Open Source released. Das ist in der Open Source Community wirklich eingeschlagen wie eine Granate. Das ist ein 70 Milliarden Parameter Modell, von den Parametern ungefähr halb so groß wie GPT 3. Allerdings ist es ungefähr so leistungsfähig und es steht mit allen internen Gewichten, also mit allem, was in diesem Modell so drinsteckt, einer Open-Source-Gemeinde zur Verfügung. Der Clou: 150 GB, die man sich runterladen, damit arbeiten und eben auch weiterentwickeln kann. Da ließe sich jetzt ein Transfer Learning wirklich selber machen. Dafür braucht es allerdings eine ganz neue Latte an Experten, Expertinnen, die Data Science und Deep Learning verstehen und damit umgehen können.

Den Reifegrad vier erreichen ganz wenige und würde ich wenigen Kunden wirklich empfehlen: Ganz eigene Foundation Models, also ein eigenes Sprachmodell zu entwickeln, beispielsweise auf einem eigenen Datensatz oder auf einer ganz eigenen Daten Domäne und es dann zu von Scratch trainieren, sozusagen Bottom-up. Das wäre dann der höchste Reifegrad, den wir gerade mit dieser Technologie haben. Das schaffen derzeit nur einige Big Shots und ein paar Unternehmen gönnen sich das.

Man kennt das Modell von Open AI. Mit ChatGPT sind sie damit im letzten November, Dezember an den Markt gegangen. Natürlich haben die anderen, etwa Google, Meta und Amazon auch ihre eigenen Modelle. Und es gibt ein paar Companies, die sich mit draufgesetzt haben, zum Beispiel Bloomberg. Es gibt ganz viele andere Organisationen, die eigene Modelle bauen, natürlich immer verbunden mit sehr hohem Ressourcenaufwand.

Brigitte Streibich: Es gibt also diese Skala von vier Reifegraden. Was stellst du fest? Wo steht die Mehrheit der Kunden aktuell in Deutschland?

Konrad Schreiber: Viele ganz am Anfang muss man ganz ehrlich sagen. Aber dort sehr gut. Ich war jetzt Anfang des Jahres auf einem Workshop eingeladen bei einem Kunden in Norddeutschland, einem Energieversorger. Die haben einen Innovationstag gemacht und mich gebeten, einen Vortrag zu halten zum Thema kognitive Suche. Das ist eine semantische Suche nach Inhalten und keine Standardsuche in Dokumenten, die auf Schlagworten basiert. Dafür stellt man dem System eine Frage und bekommt eine Antwort. ChatGPT wurde da gerade veröffentlicht, für eine breite Masse zugänglich gemacht und fing gerade an einzuschlagen. In dem Vortrag saßen ungefähr 150, 200 Leute. Den Vortrag habe ich mir damals bereits in Richtung Large Language Models gebogen, denn das passt wie die Faust aufs Auge zu diesem Thema Knowledge Retrieval, Cognitive Search. Ich habe die Frage gestellt, wer hat eigentlich schon – jetzt im Januar - mit ChatGPT gearbeitet hat. Und da ging die Hälfte der Hände hoch. Über die Weihnachtsferien haben sich die Leute sich dieses Tool wirklich angeschaut.

Das ist dieser Reifegrad eins: Ich verwende das Tool, um eine kleine oder große Aufgabe zu erledigen. Gleichzeitig erkennen die Unternehmen das Potenzial, eigene Anwendungen zu entwickeln. Sie sehen: GPT kann mir Fragen beantworten, basierend auf einem Wissensfundus bis 2021, bei manchen Modellen auch 2022. Es ist möglich, dass die Modelle mittlerweile öffentliche Datenquellen anbinden und mit einem Bing Chat lässt sich das Internet durchsuchen. Da wäre es doch total praktisch, wenn ich das auf meinen eigenen Unternehmensdaten machen könnte.

Genau da setzen wir an: „Chat with your Data, Chat with Organisational Data." Wie schaffe ich es, die Power, die hinter so einem Language Model steckt, auf die Unternehmensdaten anzuwenden? Als erstes stellt sich die Frage, ob ich ein eigenes Modell auf meine eigenen Daten trainieren? Glücklicherweise nicht. Diese Sprachmodelle haben gelernt, mit Datenquellen zu interagieren. Diese Fähigkeit der Sprachmodelle können wir nutzen. Das ist der Reifegrad zwei.

Wir können über einen API Zugang Systeme designen, die in der Organisation helfen, ein Language Model, zum Beispiel GPT, zu nehmen und das auf die eigenen Datenquellen zusetzen. Dann kann ich mit meinen eigenen Datenquellen oder mit eigenen Dokumenten chatten. Da bewegt sich aktuell das größte Interesse unserer meisten Kunden.

Brigitte Streibich: Würdest du empfehlen, schon in diesen zweiten Schritt einzusteigen? Oder sagt ihr, wenn ihr in so Projekte reingeht: Fangt erst mal mit der Basis an, vielleicht mit Schritt eins, probiert euch ein bisschen aus und dann gucken wir mal, wie wir das mit euren Daten umsetzen können?

Konrad Schreiber: Beides ist wichtig. Die Mitarbeiterinnen und Mitarbeiter in einem Unternehmen sollten erst mal alle durch die Bank verstehen: Was ist die Technologie, was kann die Technologie? Die Reise wird ja irgendwo hingehen. Am Anfang hast du die Frage gestellt, wie das mit diesen persönlichen Assistenten aussieht. Man muss natürlich ein Stück

weit vorbereitet sein, wenn man jetzt plötzlich so einen Assistenten an der Seite hat. Idealerweise läuft vieles intuitiv und gut und automatisch. Trotzdem ist es ein Transformationsprozess, in dem wir uns da gerade befinden, für jeden einzelnen Mitarbeiter, für jede einzelne Mitarbeiterin. Alle müssen schauen: Was mache ich jetzt eigentlich mit diesem Werkzeug, das ich da habe?

Brigitte Streibich: Da leistet ihr ja tatsächlich mit MaibornWolff eine Grundlagenarbeit. Es gibt auch das Whitepaper, was du verfasst hast. Das verlinken wir euch gerne noch mal in den Shownotes. Sehr empfehlenswert, um zu verstehen, wie funktioniert GPT und wie lässt es sich im Unternehmenskontext einsetzen. Ihr geht ja tatsächlich zum Kunden und klärt auch auf. Wie würden dann die nächsten Schritte aussehen?

Konrad Schreiber: Dazu haben wir auch hier im Podcast noch eine ganz interessante Folge. Da geht es darum, wie ich solche ersten Prototypen bauen kann und was da wichtig ist. Das wäre tatsächlich der nächste Schritt. Entscheidet sich ein Kunde: Wir möchten was tun und möchten die Technologie besser kennenlernen als nur diesen Reifegrad eins. Tools gibts mittlerweile wie Sand am Meer, da ist es eher schwierig, einen Überblick zu behalten. Welche kann ich benutzen, auch wegen der regulatorischen Implikationen. Darauf kommen wir später in einer anderen Folge: Was darf ich überhaupt als Unternehmen, worauf muss ich achten? Aber die Tools gibts da draußen und wenn man ein bisschen vorsichtig ist, kann man sie auch verwenden, um damit seine Probleme zu lösen.

Das nächste ist dann üblicherweise ein PoC, ein Proof of Concept. Ich habe einen Use Case, eigene Daten, eine eigene Datenquelle oder wahrscheinlich sogar mehrere Datenquellen und die möchte ich integrieren. Zu diesen Daten möchte ich eine Conversational UI beziehungsweise eine Sprach-User-Schnittstelle haben. Das bereitet die Bahn für ein größeres System, wo mehr und mehr Datenquellen dran sind. Wir hatten es am Anfang, technisch ist es bereits möglich, deine E-Mails, deine Meeting Notes von letzter Woche oder sogar von vor einem halben Jahr einzubinden. Tatsächlich diskutieren wir gerade auch Teams Chat-Verläufe mit reinzunehmen. Wir müssen uns auch Gedanken machen, was da an regulatorischen und Datenschutzthemen mit dranhängt, technisch können wir aber all diese Datenquellen hernehmen. Qualitätsdatenbanken oder Jira-Ticketsysteme mit bestimmte Cases aus etwa dem Customer Support - all diese Datenquellen können wir zusammenführen und ein Sprachmodell drauf loslassen, das sich durch die Daten pflügt und ein Assistentensystem bietet. Technisch ist es machbar, man muss es allerdings bauen. Der Clou liegt dann in der Integration dieser ganzen Datenquellen.

Brigitte Streibich: Wenn du dir jetzt KI, GPT in Unternehmen heute und in Zukunft anschaust, wo geht die Reise hin? Was wäre dein Herzenswunsch, wo KI vielleicht in fünf Jahren, in zehn Jahren bei deinen Kunden steht?

Konrad Schreiber: Die Large Language Models haben sich über einen Zeitraum von fünf Jahren eigentlich von Scratch entwickelt. Die Transformer Architektur, das T in GPT steht ja für

Transformer, eine bestimmte neuronale Netzwerk Architektur, wurde in 2017 entwickelt oder zu mindestens publiziert. Damit wurde damals das Übersetzungsproblem Maschinenübersetzung eigentlich gelöst. Das Modell konnte noch nicht viel mehr, aber da wurde der Attention Mechanismus das erste Mal publiziert und kam zur Transform Architektur. Dann hat es fünf Jahre gedauert - und das ist eigentlich eine verdammt kurze Zeit – bis hin zu Sprachmodellen, die in unserer Sprache mit uns über beliebige Themen reden und die auf Datenquellen zugreifen können.

In den nächsten fünf Jahren wird es sicherlich mehr als eine Verdopplung dieser Fähigkeiten geben. Wir sind durch diese Sprachmodelle jetzt in der Lage, die weitere Entwicklung zu beschleunigen. Man kann sich damit Source Code erstellen lassen, man kann damit Ideation-Prozesse unterstützen. Überall, wo wir vorher viel Kapazität, auch geistige und Zeit gebraucht haben, um einen Prototypen hinzustellen, kann uns jetzt ein Language Model helfen.

Die Firmen, die diese Modelle entwickeln, kennen sie selbst natürlich auch sehr gut. Da hat ein Kunde schön formuliert: "We drink our own champagne". Wir trinken unseren eigenen Champagner natürlich auch. Diese Firmen nutzen ihre eigene Technologie, um ihre eigene Technologie weiter nach vorne zu bringen. Da werden wir ein exponentielles Wachstum sehen, was die Fähigkeiten dieser Modelle angeht. Das Stichwort sind da Scaling Laws, also wie entwickeln sich die Modelle weiter und welche emergenten Fähigkeiten treten auf. Konkret festzunageln, was in fünf Jahren geht, ist schwierig.

Auf der anderen Seite muss man aber auch beachten, wie schnell wir darin sind, so was zu adaptieren. Sam Altman, das ist der CEO von Open AI, war neulich in München und hat hier an der TU München einen Vortrag gehalten und gesagt: "It´s a lot of inertia in the system". Es gibt viel Trägheit, gerade in den Organisationen die Technologie zu adaptieren und das auch richtig zu machen. Er rechnet damit, dass die Veränderungsprozesse mit dem, was jetzt schon möglich ist, was wir jetzt schon haben könnten, fünf Jahre dauern, bis wir es überhaupt in verschiedenen Firmen und Organisationen realisieren können.

Brigitte Streibich: Das heißt, es ist der Mensch, der eigentlich zu langsam ist oder der nicht ganz hinterherkommt mit diesen Entwicklungen?

Konrad Schreiber: Das ist, glaube ich, so ein grundsätzliches Thema in der Digitalisierung. Unsere Technologie rennt uns da massiv voran oder treibt uns auch ein Stück weit. So ein bisschen sind wir da wie auf hoher See, es tobt ein Sturm, es kommen von überall neue Wellen, neue Tools, neue Nachrichten. Da einfach ein bisschen ruhig zu bleiben und zu sagen: Schauen wir doch mal, was jetzt geht. Wir gehen es Schritt für Schritt an und gucken, dass wir uns vielleicht nicht den großen übergreifenden Assistenten in die Firma denken oder bauen möchten.

Das kann ein Ziel sein. Dieser visionäre Gedanke ist super, den unterstütze ich. Aber lass uns erst mal mit ganz konkreten Use Cases starten. Das ist auch das, was viele Kunden von sich

aus schon tun. Sie suchen sich ein bestimmtes Feld aus, oftmals irgendwo im Bereich Knowledge Retrieval, Wissensmanagement und bitten uns da eine PoC zu bauen und nennen das schon teilweise virtueller Assistent oder Assistenzsystem. Das sind so die internen Bezeichnungen bei unseren Kunden für diese Systeme, die wir da bauen. Und einige gehen damit auch an die Presse raus und verkünden stolz, was für Leuchtturmprojekte sie gerade schon bauen.

Brigitte Streibich: Also ist dein Credo: Einfach machen, einfach mal loslegen und dann schauen, wie man das nach und nach aufbauen kann.

Konrad Schreiber: Das auf jeden Fall.

Brigitte Streibich: Super, ich danke dir, Konrad, dass du bei mir warst. Ich glaube, wir sehen uns in einer der nächsten Folgen wieder und sprechen dann über das Thema Sicherheit und Datenschutz im Zusammenhang mit Künstlicher Intelligenz. Ich freue mich drauf.

Konrad Schreiber: Genau!

Brigitte Streibich: Bis bald!

Konrad Schreiber: Jetzt haben wir geguckt, was geht. Und in der nächsten oder übernächsten Folge schauen wir, was man eigentlich darf. Ich freue mich auch drauf.

Brigitte Streibich: Ja, ich freue mich, danke dir.