Felina Wellner

Experten-Interview mit Jan Fischer

TYP DES DOKUMENTS | TYPE OF THE DOCUMENT
Interview | Interview

Experten-Interview mit Jan Fischer

Vom 17.06.2024 und 01.07.2024

Jan Fischer ist Wissenschaftlicher Mitarbeiter am Business Innovation Lab der HAW Hamburg
https://www.haw-hamburg.de/hochschule/beschaeftigte/detail/person/person/show/jan-fischer/

Felina Wellner: Sie haben bereits Chatbots entwickelt. Welche spezifischen Funktionen und Merkmale müssen Chatbots haben, um sicherzustellen, dass die bereitgestellten Informationen korrekt und nützlich sind?

Jan Fischer: RAG – Retrival Augmented Generation. Die Nutzung von eigenen „Wissensdatenbanken“ anstatt dem trainierten Wissen des LLM. Und besser noch (read-only) API-Zugängen zu „Hard-Facts“ wie Terminkalendern oder Lagerbeständen.

Felina Wellner: Welche Herausforderungen sind Ihnen bei der Entwicklung von Chatbots begegnet, insbesondere hinsichtlich der Vermeidung von Halluzinationen und Desinformationen, und wie haben Sie diese gelöst?

Jan Fischer: Halluzinationen sind nie zu 100% auszuschließen, aber durch RAG und eine niedrige „Temperatur“ (Die Temperatur ist ein Parameter, der in Modellen zur Verarbeitung natürlicher Sprache verwendet wird, um das "Vertrauen" eines Modells in seine wahrscheinlichste Antwort zu erhöhen oder zu verringern) erhält man präzisiere, aber weniger kreative Antworten bzw. weniger Varianz in den Antworten.

Felina Wellner: Welche langfristige Rolle sehen Sie für Chatbots und generative KI-Systeme in der digitalen Transformation von Unternehmen?

Jan Fischer: GenAI im speziellen LLM basierte Chatbots, werden nach und nach in die verschiedensten Produkte integriert sein, der Microsoft Office 365 Copilot oder Apple Intelligence sind da die besten Beispiele. Die Nutzung wird so alltäglich und fließen sein wie heute die Rechtschreibprüfung. Außerdem werden Chatbots (bzw. Agenten) immer mehr Routineaufgaben im Kundensupport oder in hochautomatisierten Workflows übernehmen können.

Felina Wellner: Wie berücksichtigen Sie die Tonalität von Chatbots wie ChatGPT in Ihrer Arbeit?

Jan Fischer: Durch entsprechendes Prompting kann man die Tonalität der Ausgaben anpassen. Hervorzuheben ist hierbei die Arbeit auf einer Meta-Ebene, anstatt bspw. bei einer Stellenanzeige, einzelne Formulierungen zu ändern, passe ich den Prompt an um eine Anzeige „enthusiastischer“ oder „weniger starr“ klingen zu lassen.

Felina Wellner: Welche generativen KI-Systeme kommen typischerweise in Unternehmen zum Einsatz? Handelt es sich oft um Tools von Anbietern wie OpenAI?

Jan Fischer: Microsoft Copilot, ChatGPT und GitHub Copilot warden vermutlich den größten Anteil ausmachen, und die basieren alle auf OpenAI. Ich beobachte aber auch zunehmend das Interesse an „lokalen“ Lösungen, also ohne externe Clouds / APIs, wenn auch vergleichsweise gering.

Felina Wellner: Wie verläuft der Integrationsprozess generativer KI-Systeme in Unternehmen? Wird manuelles oder individuelles Training durchgeführt? Wenn ja, wie?

Jan Fischer: Das kommt komplett auf den Anwendungsfall an. Häufig reicht ein allgemeiner Chatbot mit RAG, gehostet durch einen Fremdanbieter und integriert über ein Java-Script-Snippet auf die Webseite des Unternehmens. Wenn es in die Richtung von automatisierten Agenten geht, ist eine sorgfältiger Integration in die Unternehmenssystem notwendig.

Felina Wellner: Welche Maßnahmen sind bei der Datenzufuhr erforderlich, um die Qualität und Zuverlässigkeit der generierten Inhalte sicherzustellen?

Jan Fischer: Wenn es um das Fine-Tuning geht, dann gilt immer ‚garbage in, garbage out‘. Es reicht nicht wahllos Daten zu nutzen, sondern für den Anwendungsfall relevante und manuell geprüfte Daten. Wenn es um RAG geht, sollten die Daten so aufbereitet sein, dass sie auf die relevanten Inhalte reduziert werden und ggf. strukturiert sind.

Felina Wellner: Wie kann man KI-Systemen die spezifische Tonalität und den Stil im Sinne eines Unternehmens vermitteln?

Jan Fischer: Entweder durch Analyse der eigenen Texte und Ableitung der Tonalität und dann Anwendung via Prompt oder vermutlich besser, durch das Finetuning auf die Texte des Unternehmens.

Felina Wellner: Sie erwähnten, dass der Anwendungsfall entscheidend sei – könnten Sie ein konkretes Beispiel nennen, bei dem die Integration eines einfachen Chatbots erfolgreich in den Unternehmensprozess umgesetzt wurde?

Jan Fischer: Man hat einen Chatbot, der auf die eigene Website, auf einen Produktkatalog oder Ähnliches trainiert wird. Man kennt sie vielleicht selbst von einigen Webseiten: Dann kommt so eine kleine Chatblase und dann kann der Kunde Fragen zu irgendwelchen Produkten, Terminverfügbarkeiten oder Ähnliches, stellen. Es gibt entweder die Möglichkeit, dass man einem kommerziellen Chatbot von OpenAI oder von irgendeinem anderen Anbieter ganze Produktdatenblätter zur Verfügung stellt. Die Texte werden dann einmalig in Tokens gewandelt und jedes Mal, wenn der Kunde eine Anfrage stellt, diese ebenfalls in Tokens umgewandelt. Dann wird ein Ähnlichkeitsmaß angewendet, also ein Abstand, wie weit die Tokens voneinander entfernt sind, um passend zu der Frage des Kunden den Textblock zu finden, wo die Antwort drinstehen könnte.

Und dieser Textblock und die Anfrage des Kunden wird dann an den kommerziellen Chatbot gesendet. Der nimmt dann diese Kontextinformation und gibt dann dem Kunden die Antwort. Das hat den großen Vorteil, dass man nur einmal die Daten digitalisieren oder in Tokens umwandeln muss und beim zweiten Mal die Kosten sparen kann, weil man nicht hunderte von Seiten PDF jedes Mal in die Anfrage mit einstellt, die jedes Mal aufs Neue tokenbasiert behandelt werden müssen. Das ist die einzige Möglichkeit, den Chatbot

kosteneffizient mit eigenen Informationen zu füttern. Ein Fine-Tuning macht für gewöhnliche Unternehmen wenig Sinn, da sich die Daten zu schnell ändern. Zwar könnten Konzerne einen eigenen Chatbot trainieren, doch für klassische Unternehmen ist das nicht praktikabel. Der RAG-Prozess, wie eben beschrieben, ermöglicht zudem die Angabe von Quellen, also aus welchem Textblock oder Dokument die Antwort stammt, was eine zusätzliche Quellensicherheit bietet – eine Funktion, die beim regulären Fine-Tuning fehlt.

Felina Wellner: Welche Schritte sind dann für ein Unternehmen fällig?

Jan Fischer: Es macht natürlich Sinn, im Vorfeld zu wissen, wonach die Kunden häufig fragen. Das heißt, man guckt entweder in seine Mailprotokolle oder Ähnliches und stellt Informationen bereit, die Kunden auch haben wollen. Wenn ich Produktdatenblätter nutze, müssen diese zudem aufbereitet werden. Da sind manchmal viele Tabellen oder Schemazeichnungen oder Ähnliches dabei. Das sind Informationen, die ein Chatbot in der Regel im ersten Schritt nicht verarbeiten kann. Das heißt, man kann die unnötigen Informationen entfernen und letztendlich macht man dann eine Sammlung davon. Und dann kann man die Daten selbst gehostet, also auf einer Software, die auf den eigenen Geräten verfügbar ist, oder über kommerzielle Anbieter hochladen und dem Chatbot zur Verfügung stellen.

Felina Wellner: Gibt es dann noch weitere Schritte, die im laufenden Betrieb getätigt werden müssen, was Wartung oder Optimierung betrifft?

Jan Fischer: Auch das Prompting kann angepasst werden. Wenn ich zum Beispiel sehr schnell ein Upselling erzeugen möchte, dann könnte ich den Chatbot so anpassen, dass er gezielt auf bestimmte Produkte verweist oder Empfehlungen gibt. Ansonsten werden Anpassungen immer dann nötig, wenn sich Daten ändern. Das heißt, wenn sich meine Produkte aktualisieren, muss ich die Daten anpassen. Und ansonsten würde ich immer empfehlen, dass man sich regelmäßig anguckt, was häufig gestellte Fragen sind und dann gegebenenfalls nachzuschärfen. Ich kann mir angucken, welche Fragen gestellt werden. Ich kann mir angucken, welche Antworten er gibt. Ich kann den Chatbot auch so bauen, dass er stets auf Kontaktinformationen oder Ähnliches verweist, damit ich sehe, ob ich damit wirklich aktiv Leads generiere oder nicht.