KI 'MADE IN GERMANY': WIE OPENGPT-X ZUR DIGITALEN SOUVERÄNITÄT EUROPAS BEITRÄGT

Ausgesprochen digital, der Podcast für digitale Trends.

Intro

[00:00:09.190] - Steffen Wenzel
Wir fangen heute mal etwas anders an, weil ich als erstes Stefanie Liße begrüßen möchte, mit der ich künftig 'Ausgesprochen digital' moderieren werde. Stefanie ist Senior Sales Managerin bei der MMS und ich freue mich darauf, dass sie besonders die Kundenperspektive in diesen Podcast noch stärker mit einbringen wird. Hallo Stefanie.
[00:00:32.560] - Stefanie Liße
Hi Steffen und ich freue mich, heute mit dabei zu sein und auch in nächster Zeit dich mit zu unterstützen. Heute wird es ja wieder einmal um KI gehen und das ist vor allem das Thema, wo mich oftmals unsere Kunden und aber auch andere Unternehmen fragen: Mensch, zum Thema KI, was muss man denn da alles eigentlich berücksichtigen, wenn man KI sinnvoll ins Unternehmen einführen möchte? Gerade so im Hinblick auf Datenschutz oder generell regulatorische Anforderungen, die man erfüllen muss. Und deswegen freue ich mich heute ganz besonders auf diese Folge, denn wir schauen uns heute ein KI-Modell an: „Made in Germany", mit dem Projektnamen OpenGPT-X. Und mit uns dabei sind heute Nicolas Flores-Herr, Teamleiter Conversational AI und Leiter des Standorts Dresden, des Fraunhofer-Instituts für Intelligente Analyse-und Informationssysteme und mein geschätzter Kollege Thomas Wächter, Head of Natural Language Processing bei der Telekom MMS. Hallo zusammen und herzlich willkommen.

Entstehungsgeschichte von OpenGPT-X

[00:01:32.370] - Steffen Wenzel
Ja, Nicolas, wir beginnen mit dir. OpenGPT-X, entwickelt vom Fraunhofer-Institut. Du hast es mit begleitet. Was ist das genau und kannst du uns ein bisschen was zur Entstehungsgeschichte erzählen?
[00:01:44.430] - Nicolas Flores-Herr
Sehr gerne. Erst mal Hallo zusammen. Wir haben im Sommer 2021 einen Forschungsantrag gestellt beim Bundesministerium für Wirtschaft und Klimaschutz. Wir sind ein Konsortium aus Forschungseinrichtungen und Firmen, mit dabei zum Beispiel deutsches Forschungszentrum für künstliche Intelligenz, Forschungszentrum Jülich, aber auch Firmen wie Control Expert waren mit dabei, auch KI Bundesverband. Also wirklich namhafte Organisationen. Wir haben diesen Forschungsantrag eingereicht und haben gesagt, wir wollen ein großes oder mehrere große KISprachmodelle in OpenGPT-X trainieren und wollen diese dann als Open Source veröffentlichen. Das haben wir im Sommer 2021 geschrieben. Im Januar 2022, vor drei Jahren, ging es los und wir haben damals schon gedacht, das wird ein sehr spannendes Thema werden. Hatten noch nicht auf dem Schirm, was dann im Herbst 2022 passiert. Da gab es nämlich diesen sogenannten ChatGPTMoment. ChatGPT wurde zur schnellst-wachsenden App und wirklich jeder hat einen auf das Thema künstliche Intelligenz angesprochen Wochen. Stefanie hat es ja eingangs erwähnt, wie wichtig jetzt dieses Thema auch geworden ist. Aber man muss auch sagen, abgesehen von Firmen, es hat auch wirklich Privatleute, also meine Eltern, Verwandte, Bekannte, alle haben mich angefangen, darauf anzusprechen. Und das war natürlich auch für dieses Projekt ein riesen Push, dass es eben dieses große öffentliche Interesse auch gab.
[00:03:20.680] - Nicolas Flores-Herr
Und jetzt noch mal ein Sprung nach vorne. Dann im November 2024 haben wir tatsächlich Teuken-7B veröffentlicht, was am Forschungszentrum in Jülich trainiert wurde und das wurde dann unter Open

Source veröffentlicht und mit einer Lizenz, die es eben Nutzerinnen und Nutzern ermöglicht, mit diesem Modell alles zu machen, einschließlich eben der kommerziellen Nutzung. Das ist jetzt auch wahrscheinlich so ein bisschen der Brückenschlag in Richtung Unternehmen, was Stefanie eingangs erwähnt hatte, nämlich dass dieses Modell frei von Firmen genutzt werden kann.

Die Vorteile von Open Source

[00:03:55.810] - Steffen Wenzel
Nicolas, du hast gerade erwähnt, dass ihr euch für eine Open Source-Variante entschieden habt. Warum? Was sind die Vorteile davon?
[00:04:02.290] - Nicolas Flores-Herr
Open Source hat mehrere Vorteile aus unserer Sicht. Einmal, es ist natürlich wichtig, wenn man solche Modelle veröffentlicht, dass man eine Community auch schafft von Nutzerinnen und Nutzern, die eben dieses Modell für ihren persönlichen Anwendungszweck eben verwenden, aber auch weiterentwickeln wollen. Und wir hatten bereits nach wenigen Wochen über 60.000 Downloads von Teuken. Ich habe persönlich mit, ich sage mal, 600 bis 700 Downloads gerechnet und wer dann hätte gedacht, das wäre ein Wahnsinnserfolg, mit so einem Interesse haben wir nicht gerechnet und das war schon mal sehr, sehr gut. Und ein Aspekt hiervon ist natürlich, dass das Modell auch mit seinen sieben Milliarden Parametern sich relativ gut noch auf, ich sage mal, mit einer erträglichen Menge an Hardware nutzen lässt. Sprich, als Nutzerin oder Nutzer des Modells lade ich mir das Modell runter, installiere das auf meinem eigenen Server oder in der Cloud-Umgebung meiner Wahl und da nimmt dieses Modell wenig Ressourcen in Anspruch und das macht auch dieses 7B-Format so attraktiv.
Natürlich hat es dann gewisse Einschränkungen auch in der Leistungsfähigkeit, ich sage mal, gegenüber GPT-4 und großen Modellen, aber für die allermeisten – da kommen wir dann auch später noch mal zu – für die allermeisten Anwendungen oder für viele Anwendungszwecke ist das Modell sehr gut geeignet.
[00:05:28.940] - Nicolas Flores-Herr
Jetzt ist es so, jetzt kommen wir noch mal zum Zweiten. Ich habe vorhin gesagt, man kann das in der eigenen Cloud betreiben oder bei einem Cloud-Anbieter seiner Wahl oder auf dem eigenen Server. Unternehmen haben die komplette Kontrolle über ihre eigenen Daten, mit denen sie das Modell abfragen, die sie in das Modell schicken oder Anwendungen, die sie auf Basis dieses Modells eben bauen. Das heißt, sie behalten ihre Datenhoheit. Und das ist der zweite, nicht nur, weil es sozusagen ein Modell umsonst ist, sondern das hat hier sehr, sehr große Perspektiven für die Nutzung.

Telekom Beteiligung am Projekt

[00:06:01.650] - Stefanie Liße
Super. Danke dir, Nicolas. Jetzt interessiert mich ganz persönlich tatsächlich noch, Thomas, wie kam es denn dazu, dass wir als Telekom in dem Projekt hier mitwirken?
[00:06:10.920] - Thomas Wächter
Ja, auch von mir noch mal ein Hallo an die Zuhörer. Hallo Nicolas, hallo Stefanie, hallo Steffen. Für uns ist das Bauen von KI-Anwendungen mit Text, also Textanalyseanwendungen, so Kerngeschäft. Wir bauen immer schon Anwendungen mit Deep-Learning-Modellen und schauen uns natürlich auch Transformer-Architekturen an. Und wir sind mit dem Fraunhofer an der Stelle zusammengekommen im Rahmen eines Industrieprojektes, wo wir uns was ganz anderes über Audiotranscription unterhalten hatten. Und das war gerade in der Phase, als das Training in den Endzügen war und schon überlegt wurde, wie das Release stattfindet, wofür man das verwenden kann und der Abschluss von dem Forschungsprojekt, so vielleicht ein halbes Jahr, bevor das Projekt dann wirklich releast wurde. Und überraschend für uns war, dass die Gruppe, die das wirklich gebaut hat, bei uns hier in Dresden ein paar Kilometer Luftlinie sitzt, mit dem Nicolas und seinen Kollegen, dass auch viele von den Technikern hier sitzen. Das Fraunhofer IAIS hat ja Standorte in verschiedenen Teilen in Deutschland, in Bonn, nahe an unserer Muttergesellschaft und eben hier in Dresden, dort wo auch wir als Textanalyseexperten in der Telekom MMS sitzen. Ja, und ein lokal betreibbares Open-Source-Modell ist für unsere Anwendungslandschaft relevant, um eine souveräne KI-

Anwendung zu bauen und möglich keine Einschränkung zu haben, wo ich das Ganze installieren kann, ist für uns Teil des Baukastens, den wir nutzen, um Anwendung für unsere Kunden zu erstellen.
[00:07:43.740] - Stefanie Liße
Jetzt hast du es gerade schon angesprochen. Wir bauen ja KI-Anwendungen nicht nur zum Eigennutz, sondern natürlich auch für unsere Kunden. Was würdest du denn sagen, für welche Unternehmen oder vielleicht auch Branchen ist dieses Modell denn überhaupt interessant oder gibt es da gar keine Einschränkung?
[00:07:59.900] - Thomas Wächter
Branchenspezifisch gibt es erst mal bei den Large Language Modellen grundsätzlich keine Einschränkung, weil wir hatten viel Text trainiert, können Sachen wie Synonyme, semantische Relationen gut wiedergeben und wir sparen uns das Sammeln von großen Trainingsdatensätzen pro Kunde. Das heißt, einerseits wird es günstiger, solche Anwendungen zu bauen und das zweite ist relevant, dann, wenn wir die Anwendung lokal im Rechenzentrum des Kunden oder auf dedizierten Umgebungen installieren möchten. Zum Beispiel wäre das für öffentliche Auftraggeber so, insbesondere für Finanzen und Gesundheit, wo ich regulatorische Anforderungen habe. Das ist zum Beispiel auch so bei Anwälten und allen Berufsgruppen, die ein Berufsgeheimnis haben, weil dort muss nachgewiesen werden, dass ich möglichst umfangreich Maßnahmen treffe, um auf die Daten, die mir anvertraut wurden, aufzupassen und muss das auch sowohl vertraglich als auch technisch organisatorisch in einem besonderen Maße sicherstellen. Und der einfachste Weg, das zu tun, ist, eine Anwendung bereitzustellen, wo das Modell mit drin ist und die ich dort installieren kann, wo ich sie am Ende auch betreiben darf. Und das gibt uns mit dem 7B-Modellen und mit dem Teuken-Modell erst mal die Flexibilität, diese Modelle, sagen wir, lokal zu installieren. Und das zweite in der Privatwirtschaft ist es eher das Thema Investitionsschutz oder Innovationsschutz in der herstellenden Industrie.
[00:09:28.350] - Thomas Wächter
Wenn ich ein neues Produkt entwickle und dafür Large Language-Modelle, sagen wir, in der Bewertung von Dokumente, von Spezifikationen einsetze, benutze ich das, rufe dieses Modell auf und möchte das auch sehr nah auf meinen Infrastrukturen betreiben und nicht über verschiedenste Cloud Services die Daten verteilen.

Open Source Besonderheit bei OpenGPT-X

[00:09:46.940] - Steffen Wenzel
Wir können das Thema Datenschutz sicherlich später noch mal ein bisschen vertiefen. Das ist sicherlich ein USP, ein Vorteil, von diesem Modell, aber ich möchte noch mal ein bisschen zurück zu diesem Open Source Modell und seinen Vorteilen kommen. Thomas, erst noch mal jetzt auch als dich mit einer technischen Perspektive darauf. Man sagt ja immer, die Community entwickelt mit und es ist frei zugänglich und es wird dadurch besser, dass es natürlich von allen gepflegt und weiterentwickelt wird. Aber es ist natürlich auch immer vielleicht ein größerer Projektmanagement-Effort, den man dann dort zu bewältigen hat. Wie siehst du das? Ist das hier bei diesem Modell gegeben und wie siehst du die Entwicklung?
[00:10:28.280] - Thomas Wächter
Also Open Source ist es nicht immer gleich Open Source. Es kommt drauf an, was wurde Open Source gestellt. In dem Fall die Nutzung des Modells, das Instruct-Modell, das Modell, was ich aufrufe, um sagen wir mal, Text zu generieren, auch die Embedding-Modelle, was ich aufrufe, um Texte miteinander semantisch zu vergleichen. Was auch das Fraunhofer oder letztendlich das Konsortium OpenGPT-X hier anders gemacht hat, ist, dass nicht nur das Modell Open Source gestellt wurde, heißt, welche Quellen wurden verwendet dafür zum Beispiel, sondern auch der komplette Trainingsprozess. Bei den kleinen Modellen ist so, dass man die schon anpassen möchte auf die Wissensdomänen. Wenn ein Anwalt was hat, möchte ich gerne die Paragraphen einlesen. Wenn ich im Gesundheitswesen arbeite, möchte ich gerne vielleicht den Gebührenordnungskatalog mit einlesen und das System damit bekannt machen, dass es Sachen nicht missversteht. Und bei dem Teuken-Modell ist die Trainingspipeline bekannt und mit dem Fraunhofer sogar noch ein Ansprechpartner vorhanden, mit dem wir das im Detail besprechen können und überlegen können, welche Anpassungen für die jeweiligen Anwendungen relevant sind. Und die ist eben auch Open Source gestellt. Das heißt, wir können das auf einer Infrastruktur installieren, können das einfach weitertrainieren. Das ist eine Besonderheit, das kann ich nicht bei allen Modellen und wenn man heute schaut, sagen wir, von der Nutzbarkeit im Projekt, werden die Modelle ja grundsätzlich

auf Hugging Face normalerweise bereitgestellt, also auf einer Infrastruktur, wo die publiziert und verglichen werden.
[00:11:49.210] - Thomas Wächter
Und da kann ich normalerweise so ein Modell einfach runterladen und benutzen. Dann ist für mich erst mal nur relevant, welche Lizenz hat das. In dem Fall haben wir schon gehört, der Apache 2.0 Lizenz, relativ frei nutzbar, einbaubar in Applikationen, bereitstellbar für Kunden, wenn man die Lizenzbedingungen einhält und anpasst, die Anwendung, dass die Lizenzbedingungen so entsprechend berücksichtigt werden. Das heißt, eine relativ freie Nutzung möglich.
[00:12:13.520] - Steffen Wenzel
Wie funktioniert das, wenn ich jetzt so eine Lizenz beantragen will oder wenn ich mich da engagieren will? Thomas, kannst du uns da noch mal kurz helfen?
[00:12:20.280] - Thomas Wächter
Genau, eine Lizenz an sich wird hier nicht gekauft oder bestellt, sagen wir es, ist kein Einkaufsprozess, sondern Open Source wird veröffentlicht und die Lizenz bekanntgegeben. Und damit ist klar, wie ich dieses Stück Software, wie ich das Modell selber zu verwenden habe und was ich dabei berücksichtige. Und bei Apache 2.0 ist es eben so, dass es eine sehr freie Lizenz ist, mit der ich wenige Einschränkungen für die kommerzielle Nutzung habe.

Benchmarks zur Prüfung von KI-Modellen

[00:12:48.070] - Stefanie Liße
Nicolas, wir haben heute schon sehr, sehr viele Begrifflichkeiten gehört und das ist natürlich auch für unsere Kunden oder generell für Unternehmen immer mal schwierig, das einzuordnen. Und es gibt ja auch zahlreiche Open Source Modelle auf dem Markt. Und da wäre meine Frage, ob die vergleichbar sind und nach welchen Kriterien man die vielleicht am Ende auch vergleicht. Kannst du uns da was sagen?
[00:13:07.410] - Nicolas Flores-Herr
Auch Open Source Modelle werden mit sogenannten Benchmarks miteinander verglichen. Diese Benchmarks testen gewisse Eigenschaften von Modellen, beispielsweise ein allgemeines Sprachverständnis, Wissen in verschiedenen Wissensbereichen, die mit Multiple Choice Tests getestet werden und und und. Und diese Benchmarks dienen eben dazu, einen ersten Anhaltspunkt für die Leistungsfähigkeit eines Modells eben zu geben. Das ist natürlich, was man immer auch noch zusätzlich zu diesen Benchmarks untersuchen sollte, ist eben, wie gut sich das Modell in der eigentlichen Anwendung eignet. Jetzt ist es häufig so, dass auf Hugging Face potenzielle Nutzerinnen und Nutzer sich erst anhand dieser Benchmarks orientieren und anhand dieser Benchmarks auch das leistungsfähigste Modell auswählen. Es muss nicht immer das leistungsfähigste Modell jetzt für Ihre spezifische Anwendungen sein. Hier, denke ich, gilt es auch vielleicht, ein bisschen Aufklärungsarbeit zu machen. Gleichzeitig ist natürlich auch Benchmarks für die wissenschaftliche Community wichtig, weil man hier eine Art objektiven Vergleich dieser Modelle hat. Man hat wirklich dann eine Prozentzahl dieser Benchmark. Das Modell hat so und so viel Prozent bei diesem Benchmark und so und so viel Prozent bei dem anderen Benchmark. Das gibt für die Leistungsfähigkeit dieser wirklich komplexen Modelle, da gibt es eine gewisse Griffigkeit.
[00:14:29.080] - Thomas Wächter
Und vielleicht, als Ergänzung, bei den Modellen, bei den Benchmarks werden auch ganz verschiedene Sachen getestet. Also technische Wiederholbarkeit, Faktenlage, gibt sie den wissenschaftlichen Stand der Faktendatenbank genauso wieder oder gibt es alternative Interpretationen, Vollständigkeit der Antwort? Alles Sachen, die einen Hinweis darauf geben, welches Modell für meine Anwendung geeignet ist. Und was man auch noch berücksichtigen muss, ist, dass oft erst mal als Basismodelle bereitgestellt werden und dann es angepasste Modellvarianten gibt, die in dem einen, anderen oder idealerweise in allen Benchmarks besser sind. Und da muss man auch fragen: Womit wurde das trainiert? Welche Daten sind reingegangen? Wird der Prozess so fortgesetzt wie in der Vergangenheit von dem jeweiligen Modellprovider? Ist schon ein bisschen komplizierter, aber als erste Iteration schauen wir natürlich immer, wer steht ganz oben, und gehen von oben nach unten durch und schauen uns die Parameter von den einzelnen Modellen an, ob es generell technisch sauber implementiert wurde.
[00:15:26.860] - Nicolas Flores-Herr
Genau. Wir haben im OpenGPT-X dieses Thema die Benchmarks weiterentwickelt und haben

sozusagen die Benchmarks, die ursprünglich nur die Fähigkeit des Modells auf englischer Sprache getestet haben, haben wir auch in 21 europäischen Sprachen übertragen. Das heißt, man kann jetzt wirklich sehen, wie gut dieser Task X, den zum Beispiel Thomas eben beschrieben hat, auf Griechisch oder auf Italienisch oder Französisch, wie gut der ausgeführt wird. Und wir können wirklich hier europäische Modelle anhand von europäischen Benchmarks beurteilen, aus unserer Sicht die fairste Art und Weise, diese Modelle zu bewerten.

Modellgröße und Trainingsdaten

[00:16:03.590] - Steffen Wenzel
Wir haben jetzt das Modell DeepSeek, was gerade ziemlich eingeschlagen hat, weil es mit weniger Daten auskommt und damit natürlich auch nicht so kostenintensiv ist. Was ist jetzt OpenGPT-X in dem Vergleich? Würdest du das auch als Vorteil betrachten, dass es mit weniger Daten auskommt?
[00:16:20.610] - Thomas Wächter
Also Nicolas hatte ja vorhin schon gesagt, es sind ja nicht generell weniger Daten. Es wird auf einem Korpus von Daten trainiert. In der Regel werden momentan wahrscheinlich alle Modelle auf ähnlich großen, oder den größtmöglichen Korpora trainiert. Die Besonderheit bei dem Teuken-Modell oder bei dem OpenGPT-X-Projekt an sich ist, dass viele europäische Sprachen integriert sind. Das heißt, die Trainingsdaten werden bei der Auswahl so ausgewählt damit, dass sie ein bisschen gleichverteilt über die Nutzung in den verschiedenen Weltregionen verteilt sind und nicht eben nur englischsprachig. Für uns als europäisches Unternehmen mit verschiedenen Unternehmenseinheiten in Europa, in verschiedenen Ländern, sind gerade die europäischen Sprachen auch in kleineren Ländern extrem relevant, dass das funktioniert und dass man eben dort genau eine ähnliche semantische Abbildung, Ähnlichkeitsanalysen über Textkorpora machen kann. Am Ende geht es ja darum, versteht mich das Modell oder missversteht es mich? Halluziniert es, sagt man sozusagen in den News, aber an sich geht es ja darum, wenn ich ein Wort sage und einen Kontext beschreibe, versteht es mich quasi in dem Bereich oder springt es permanent zwischen den verschiedenen semantischen Bedeutungen hin und her?
[00:17:29.300] - Steffen Wenzel
Ja, dann lasst uns sollen wir trotzdem noch mal zu dieser Datengröße kommen, weil das verwirrt, glaube ich, ein bisschen. Auf der einen Seite gibt es aber eine Modellgröße, die dann jetzt davon ausgeht, dass ChatGPT über 100 Milliarden Parameter hat oder umfängt und jetzt Teuken-7B beispielsweise nur 7 Milliarden Parameter. Kannst du das dann noch mal für uns ein bisschen näher erklären?
[00:17:52.960] - Thomas Wächter
Wir trainieren ja die Modelle nicht selber. Das heißt, es sind zwei verschiedene Arten, wie man mit diesen Parametern umgeht. Die großen Modelle werden quasi mit sehr vielen Parametern trainiert. Das heißt, man lässt die einfach viel, viel länger mit den Trainingsdaten sich beschäftigen, unter vielen verschiedenen Aspekten die Trainingsdaten zu bearbeiten. Und so ein Large Language Modell, wenn man es ganz kurz fasst, wird durch Weglassen trainiert. Man lässt ein Stückchen weg und versucht, das wieder zu generieren und das Modell so anzupassen, dass die Gewichte in dem Deep Learning Modell dazu führen, dass das Modell in der Lage ist, ein Stück weggelassenen Text wieder zu rekonstruieren. Jetzt kann man auf viele verschiedene Art und Weisen weglassen. Man kann auch den Text anders codieren und weglassen und das Modell, sagen wir, viel breiter, für einen breiteren Anwendungsfall trainieren. Vorhin hatten wir das LLaMA-Modell besprochen. Das LLaMA 3.1-Modell ist im Ursprung wahrscheinlich 70B-Modell, glaube ich, gewesen. Es gibt auch 7B-Modell-Varianten davon, indem man einfach das Modell, was am Ende erzeugt wurde, noch mal reduziert, entweder durch Quantisierung, glaube ich, oder es gibt verschiedene Verfahren, wie man das reduzieren kann, dass am Ende ein 7B-Modell rauskommt. Und das hat dann den Vorteil, dass ich es auf kleinerer Infrastruktur betreiben kann.
[00:19:07.040] - Thomas Wächter
Die verliert natürlich dabei auch Informationen. Es kommt drauf an, wie ich das reduziere, wie ich das, sagen wir, beim Reduzieren vielleicht auf eine Domäne finetune. Und je größer des Modell, würde man bis vor kurzem gesagt haben, desto allumfassender kann das Fragen beantworten oder kann mit allen Textdokumenten umgehen. Jetzt haben wir gelernt, dass bei DeepSeek ein bisschen anderer Ansatz genommen wurde. Dass man dort anders die Information in das Modell reingebracht hat, indem man sehr, sehr genaue Trainingsdaten und sehr, sehr genaue Referenzdaten erzeugt hat und eben nicht ganz so lange trainiert hat. Ändert aber nichts daran, dass was rauskommt, ist ein ausgewogenes Modell, mit dem ich in jeder Wissensdomäne Sachen miteinander vergleichen kann und Texte vervollständigen kann.

Und durch das Vervollständigen kann ich überprüfen, inwieweit das Modell die Domain wirklich verstanden hat. Und in der Anwendung selber nutze ich diese Funktionalität des Textgenerierens gar nicht so sehr, weil ich möchte eigentlich nicht das Wissen aus dem Modell rausziehen, sondern ich gebe dem Wissen, zum Beispiel ein Textdokument, haben wir vorhin, eines Anwaltes oder eine Bescheinigung in der Klinik und versuche daraus, ein Stück anderen Text, zum Beispiel einen Datensatzeintrag oder eine Entscheidungsvorlage zu generieren. Und dafür muss das
Modell einerseits den Text verstehen und semantisch richtig interpretieren und dann meine Instruktion, also mein Prompt verstehen.
[00:20:28.920] - Thomas Wächter
Dass es weiß, was ich am Ende von dem Modell erwarte. Ich spare mir letztendlich für die Anwendung mit Large Language Modellen das Trainieren des grundlegenden Textverständnisses, wenn ich es aus der Anwendungssicht betrachte.
[00:20:43.870] - Stefanie Liße
Ich würde gerne noch mal zu dem Trainingsprozess ein bisschen tiefer einsteigen, weil genau das ist ja das Thema: Wenn die KI-Modelle trainiert werden mit den Daten, dann geben wir ja ein Stück weit Wissen mit rein. Und da höre ich von unserer Kundenseite sehr oft ein paar Bedenken, dass die ganz klar wissen wollen: „Okay, ist denn sichergestellt, dass jetzt das Modell nicht trainiert wird mit meinen hochsensiblen Kundendaten zum Beispiel? Das ist halt ein Punkt, der natürlich viele umtreibt und sichergestellt werden möchte. Könnt ihr uns da noch mal ein bisschen mehr Transparenz reinbringen, wie das bei eurem Modell am Ende des Tages vielleicht auch abgesichert ist?
[00:21:21.350] - Thomas Wächter
Das ist in der Rolle jetzt ja nicht die Eigenschaft des Modells, sondern wenn ich eine Anwendung aufsetze, was ich im einfachsten Fall, was alle vielleicht kennen, eine ChatGPT-App, dann spreche ich dort rein oder gebe dort Informationen rein. Und die Frage ist, werden diese Informationen gespeichert oder nicht gespeichert? Wären Sie später für Frage-Antwort-Trainingsdaten-Sätze in das Modelltraining übernommen, um das nächste Release des Modells damit zu trainieren. Kann man einerseits sicherstellen, wenn man das lokal ohne Internetverbindung auf seiner Infrastruktur betreut, ist es ganz sicher, dass die Daten nicht abfließen. Kann man aber auch auf europäischen Cloud-Umgebungen, auch auf einer Microsoft Azure Cloud, sicherstellen, dass Fragen oder Anfragen an das Modell nicht genutzt werden, um später in das Modelltraining aufgenommen zu werden. Das ist ein Privatsphäre-Setting in der jeweiligen Cloud-Umgebung.

Anforderungen aus dem EU-AI-Act

[00:22:14.890] - Stefanie Liße
Und im EU-Wirtschaftsraum gibt es ja am Ende des Tages wirklich sehr, sehr viele regulatorische Anforderungen, unter anderem ja auch aktuell den EU-AI-Act, von dem wir einiges gehört haben und, by the way, in einem zukünftigen Podcast noch mal ein bisschen tiefer einsteigen wollen. Erfüllt dann euer Modell diese Anforderungen, die mit dem EU-AI-Act auf uns zukommen, beziehungsweise teilweise ja schon da sind?
[00:22:39.010] - Nicolas Flores-Herr
Also ich bin jetzt auch kein Rechtsexperte hier, aber ich denke mal, das Ansinnen des EU-AI-Acts ist es, dass die Trainingsdaten, die zum Training des Modells verwendet werden, dass die dokumentiert werden können. Und das können wir im Gegensatz zu vielen anderen Herstellen von Open-Source-Modellen eben, das können wir machen. Wir können das dokumentieren und das ist ein Mehrwert, den sozusagen dann auch die Telekom an ihre Kunden dann weitergeben kann.
[00:23:07.740] - Thomas Wächter
Wenn man den AI-Act, so wie er jetzt ist, anschaut, werden die KI-Anwendungen in verschiedene Risikoklassen aufgeteilt und je nach Risikoklassen sind bestimmte Maßnahmen zu ergreifen, um sicherzustellen, dass ich keine KI-Anwendungen baue, die ich in der Form vielleicht nicht bauen sollte, die vielleicht ein Social Profiling macht, die vielleicht bei der Einstellung von Mitarbeitenden unterstützt waren. So die plakativsten Beispiele in den letzten Jahren in der Presse. Und wenn ich eben die Trainingsdaten kenne, wenn ich weiß, wie das Modell trainiert wurde, kann ich auch nachweisen, dass das Modell eben nicht trainiert wurde, um bestimmte Sachen zu übervorteilen und kann auch das Modell jederzeit testen, auch in verschiedenen Entwicklungsstufen testen. Und das sind halt Maßnahmen, die ich zum Beispiel in der Risikoklasse oder der Hochrisikoklasse an der Hand haben muss. Ob das für die

jeweilige Anwendung notwendig ist, kommt drauf an. Der EU-AI-Act sagt ja nicht, dass das Modell quasi an sich eine KI-Applikation ist, sondern die Anwendung, die das Modell nutzt. Und genau wie Datenschutzaspekte, wenn die halt immer aus verwendeten Daten und der Nutzergruppe und dem Verwendungszweck wird der Schutzbedarf letztendlich bewertet. Und das machen wir für unsere Kunden. Der Kunde ist an sich dafür verantwortlich und lässt sich dort unterstützen.

Hostingmöglichkeiten und Nachhaltigkeit

[00:24:24.110] - Steffen Wenzel
Noch mal zurück zum Thema Datenschutz. Das heißt, ein Vorteil ist, dass ihr schon beim Trainieren das Thema Datenschutz berücksichtigt, aber auch im Hosting achtet ihr darauf. Thomas, kannst du das mal erklären, warum?
[00:24:35.770] - Thomas Wächter
Im Idealfall, wenn ein Kunde mit seinen Anforderungen kommt oder wir selber das verwenden wollen, ist das nur von den Daten und von der Nutzergruppe abhängig, wo ich das Modell am Ende betreiben möchte und in welcher Infrastruktur. Das heißt, ich möchte natürlich immer die günstigste, für mich passendste Infrastruktur nehmen und idealerweise ist es auch so, dass ich dann eher die Schnittstellen zu den Daten, wo die herkommen, betrachte, als die Art und Weise, wie ich das Modell betreibe. Daher schauen wir, dass wir jede Anwendung, die wir heute aufsetzen, sowohl auf einer europäischen Datenschutzrichtlinie aufgesetzt haben, in einem europäischen Rechenzentrum mit Datenhaltung und Service in Europa, oder in einer deutschen Infrastruktur. Dort, wo ich regularische Erfordernis habe, ist das dann die einzige Option, das zu betreiben oder lokal in der Infrastruktur des Kunden. Das heißt, dedicated Infrastruktur. Dann muss ich natürlich mir irgendwann einen Rechner kaufen mit einer kräftigen Grafikkarte und in Abhängigkeit, was ich damit tue, auch mehreren. Und die kleineren Open-Source-Modelle kann ich natürlich auf jeder Infrastruktur betreiben. Und die großen Hyperscaler-Modelle, wir hatten GPT-4.0 wäre jetzt das Aktuelle, oder ein google gemini 2.0, die kann ich halt eben nur aus einer europäischen Infrastruktur, also in Europa betrieben, in Europa beserviced, über die Pauschalen des jeweiligen Anbieters in meine Applikation einbinden.
[00:26:00.220] - Thomas Wächter
Eine Besonderheit in der Telekom ist, dass wir jetzt auch einen Service haben, nennt sich AI Foundational Services, wo über eine Schnittstelle sowohl die Hyperscaler-Modelle oder die Modelle aus den Cloud-Infrastrukturen als auch die Open Source Modelle aus der Open Telekom Cloud, wo auch das Teuken betrieben wird, direkt zugreifbar sind. Wir betreiben sozusagen Modelle für Kunden auf Infrastrukturen, die jeder in seine Anwendungen einbinden kann.
[00:26:30.750] - Stefanie Liße
Gerade das Hosting und das Betreiben von solchen Modellen führt tatsächlich auch immer mal zu der Fragestellung, wie die Anwendung von KI-Modellen, du hattest das vorhin schon erwähnt, es gibt ja wirklich sehr, sehr große und sehr performante Modelle, die natürlich wahnsinnig die Rechenkapazität in Anspruch nehmen. Wie das Ganze zusammenpasst zu dem ganzen Thema Nachhaltigkeit? Also da würde mich interessieren, wie ihr das seht, wie KI-Modelle und die Anwendung von KI-Modellen einhergeht mit Energieverbräuchen und Nachhaltigkeitsaspekten und ob vielleicht sogar mit Teuken-7B da eine gute Alternative schon auf dem Markt unterwegs sein könnte?
[00:27:08.240] - Thomas Wächter
Vom Energieverbrauch lässt sich in der Stelle schwer messen, weil es davon abhängt, was man für Daten gibt. Es kann auch sein, dass eine bestimmte Frage schwieriger zu beantworten ist als eine andere, so wie man auch über manche Sachen vielleicht länger nachdenken muss, als über andere, je nachdem, wie das Modell betrieben ist. Was man aber eindeutig sieht, dass die 7B-Modelle natürlich wesentlich günstiger oder auch wir die wesentlich günstiger anbieten als größere Modelle. Das heißt, das sind Teilfaktor vier, fünf, sechs im Preisunterschied pro eine Million Input-Output Tokens. Und das ist ein Indiz dafür, dass sie einfach weniger Rechenkapazität, kleinere Grafikkarten in der Kalkulation im Rechenzentrum dann tatsächlich für die Anwendung verbrauchen.
[00:27:46.500] - Steffen Wenzel
Nico, kannst du dazu noch was ergänzen aus deiner Sicht?
[00:27:48.820] - Nicolas Flores-Herr
Ja, jetzt wird es ein bisschen technisch. Wir haben Teuken-7B, wie ich eingangs erwähnte, haben wir einen sehr großen Prozentsatz an europäischen Sprachen trainiert. Jetzt ist es so, eine zentrale Komponente beim Training der Modelle, aber auch bei der Anwendung, also bei der Inferenz der Modelle,

ist ein sogenannter Tokenizer. Was ist ein Tokenizer? Ein Tokenizer ist im Prinzip ein Algorithmus, ein Verfahren, was große Texte in Token zerstückelt. Idealerweise sind das ganze Worte. Jetzt ist es so, dass es sehr lange auch nur Tokenizer gab, die englische Sprache bevorzugt hat. Wir haben dann gemerkt, dass wenn wir mit vielen europäischen Sprachen trainieren, dass sozusagen diese Tokenizer sehr ineffizient die Worte zerhacken, das heißt, die Worte werden sehr, sehr klein gehackt, was natürlich zu mehr Rechenaufwenden, sowohl beim Training als auch bei der Inferenz führt. Wir haben einen multilingualen Tokenizer entwickelt, der sozusagen bei multilingualen Anfragen, also beispielsweise, wenn es jetzt auf Deutsch und auf Englisch und auf Griechisch funktionieren soll, das Modell, im Schnitt den Energieverbrauch bei der Anwendung auch bis zu 40% senken kann. Das ist sozusagen noch mal auf technischer Seite für multilinguale Anwendungen. Noch mal ein kleiner Add-on zu dem, was Thomas eben gesagt hat.
[00:29:15.110] - Thomas Wächter
Spannend ist schon, dass ihr den Tokenizer wirklich angepasst habt, weil das ist auch das, was wir in den individuell entwickelten Anwendungen tun, den richtigen Tokenizer auswählen, um Sätze, Worte so zu interpretieren, dass ich damit überhaupt erst meine Maschinenlernensprache arbeiten kann. Und der muss halt spezifisch für die jeweilige Landessprache gut funktionieren und nicht einfach nur am Freizeichen zerhacken.
[00:29:39.780] - Stefanie Liße
Thomas, du hattest vorhin schon erwähnt, dass es keine dedizierte Lizenz jetzt in der Art gibt. Und trotzdem könnte ich mir vorstellen, dass wenn Kunden das nutzen wollen, das Modell, dass da trotzdem Kosten anfallen. Kannst du uns dazu was sagen?
[00:29:51.640] - Thomas Wächter
Also ich kann zumindest an den Listenpreisen was sagen. Also das wird ja nach Token abgerechnet und wir rechnen das in der Regel, wenn wir das betreiben der Kunde das nicht auf eigener Infrastruktur betreibt, nach einer Million Input-Output-Token ab. Und die Größenordnungen bei einem 7B-Modell liegen da zwischen zwei bis drei Euro für eine Million Input-Output-Token. Und wenn man sich das überlegt, eine A4-Seite sind ungefähr 400 Worte, weil manchmal sind es dann 300 Token, manchmal sind es 600 Token, je nachdem, was da drauf steht. Aber hat man eine Größenordnung, dass man schon eine relativ große Menge an Text nutzen kann in der Anwendung. Wenn man, sagen wir, chattet, wie ChatGPT, ist das oft vernachlässigbar. Wenn man sogenannte Dokumentsuch-Anwendungen oder semantische Suche oder Kategorisierung von Dokumenten, auch dann große Mengen von Dokumente indexieren möchte, muss natürlich der gesamte Text in den Dokumente interpretiert werden. Dann kann man aber relativ einfach ausrechnen, in welcher Größenordnung man dort Hardware-Ressourcen verwendet.

Anwendungsszenarien

[00:30:59.340] - Steffen Wenzel
Nicolas, lass uns noch mal ein paar Anwendungsbeispiele uns anschauen, damit wir mal ins Konkrete auch kommen. Wo kann Teuken-7B eingesetzt werden?
[00:31:07.100] - Nicolas Flores-Herr
Wir haben in OpenGPT-X, das ist im Prinzip noch mal das Konsortium, wo Teuken-7B entwickelt wurde. Wir hatten auch Anwendungspartner aus der Wirtschaft dort. Ein Beispiel ist der Westdeutsche Rundfunk, der WDR, aber auch die Firma Control Expert aus dem Sicherungsbereich. Und das war sozusagen die Aufgabe, also nicht nur ein Modell zu trainieren und das dann der Forschungscommunity zur Verfügung zu stellen, sondern auch diese trainierten Modelle in einem Anwendungsszenario, prototypisch – so ein Projekt ist natürlich Vorwettbewerblich – dann eben auch einzusetzen. Jetzt ist es so, dass das Teuken-7B ist erst mal eine Art Alleskönner, so out of the box, ja? Um sozusagen, Teuken-7B dann in verschiedenen Anwendungsszenarien, sei es in der Verwaltung, sei es bei den genannten Firmen einzusetzen, ist es häufig nötig, dass sozusagen eine Art Finetuning stattfindet, um eben das Modell auf eine spezielle Domäne zu spezialisieren. Es gibt dann auch noch so sogenannte Retrieval Augmented Generation, also sogenannte RAG-Lösungen, wo eben auch, das sind Anwendungen, die sozusagen um ein Modell herum entwickelt werden, um eben das Modell in spezifischeren Kontexten und in verschiedensten Einsatzbereichen einzusetzen. Es gibt im Prinzip kein, ich sage mal kein Limit, dass man jetzt sagen kann, das ist jetzt nur ein Modell für die Verwaltung, aber nicht für die Logistik, sondern die Limitierungen sind aktuell eben dazu so gesetzt.
[00:32:40.980] - Nicolas Flores-Herr

Es sind genügend Daten für Finetuning vorhanden. Kann man dann den Kunden dann auch gemeinsam Projekte machen. Das sind eher so die Limitierungen. Das ist erst mal so ein technischer Blick, hiermit will ich dann übergeben.
[00:32:52.060] - Thomas Wächter
Genau und bei uns in der Telekom, in dem Einsatz für das OpenGPT-X-Modell, Modell ist es im ersten Schritt hauptsächlich Dokumentanalyse Services, also Dokumente AI Services, weil ich mit diesem Modell Sätze, Paragrafen, Worte semantisch kategorisieren kann, semantisch labeln kann. Ich kann unterscheiden, wenn zwei Dinge ähnlich heißen oder ähnlich aussehen, ob es das eine oder das andere ist, ob es der Gebührenordnungskatalog Code oder der ICD-10 Code in der Gesundheitsanwendung ist oder ob es ein bestimmtes Datum ist, was relevant ist für mich oder personenbezogene Daten zu erkennen. Auch einfache Dinge, wie den Führerschein auszulesen, Fahrzeugpapiere auszulesen, so typische Behördenvorgänge, gehen mit den Open Source Modellen und mit dem Teuken Modell ausreichend gut, dass wir dafür eine Anwendung aufbauen kann, die man lokal betreiben kann. Und für Anwendungsfälle, die Richtung Suche gehen, braucht man eben die sogenannten RAG-Architekturen oder Architekturen, wo ich Dokumente zerlege, indexiere und Teile davon wiederfinde und am Ende das, was relevant ist, wieder dem Modell anbiete, um eine Antwort zu generieren. Und da kommt es, sagen wir, ganz kurz beantwortet auf Vollständigkeit an. Finde ich alle Seiten in einem Dokument, die für die Antwort relevant sind. Wenn ich das vollständig mache, kann ich davon ausgehen, dass ich auch eine vollständige Antwort bekomme.
[00:34:17.820] - Thomas Wächter
Wenn ich bei dem Suchschritt davor einen Fehler mache, zum Beispiel zu wenig die Absätze überlappen lasse und den Kontext verliere oder Sachen auf der ersten Seite stehen und auf der letzten Seite darauf Bezug genommen wird und wenn ich diesen Kontext, diesen Zusammenhang nicht mehr erkenne durch die Art und Weise, wie ich das indexiert habe, dann verliere ich gegebenenfalls Informationen. Und das ist zu prüfen in den Anwendungsfällen. In der Regel, sagen wir in den ersten fünf Tagen schaut man sich das ein bisschen genauer an und kann dann eine Machbarkeitsaussage treffen, ob das auf die Art und Weise in einem Standardweg lösbar ist oder ob man eine sehr individuelle Anwendung bereitstellen müsste.

Projektbeginn

[00:34:52.490] - Stefanie Liße
Wenn ich jetzt als Kundenunternehmen schon eine ganz klare Idee vielleicht habe von einem Use Case, den ich umsetzen wollte, kannst du uns kurz erklären, Thomas, wie es dann bis zum Projekt gehen könnte oder was so die typischen Phasen sind?
[00:35:06.800] - Thomas Wächter
Das kommt drauf an, ob man ein Standardprodukt aufsetzt und grundlegend erst mal davon ausgeht, dass ich hier, zum Beispiel ein Such-und einen Wissensmanagement-Szenario habe, dann nehme ich eben ein Standardprodukt mit einer Standardarchitektur zum Indexieren von Dokumenten. Das funktioniert out of the box. Da muss ich nur die Dokumente hochladen, speichern drücken, fünf Minuten warten und dann kann ich gegen diese Dokumente chatten. Ich kann dann normalerweise noch verschiedene Assistenzmuster anlegen. Das heißt, für verschiedene Nutzergruppen festlegen, wie die Antwort sein soll, weil wenn ein Lehrer eine Frage im Schulkontext stellt, kriegt er vielleicht eine andere Antwort detailliert, als wenn das der Schüler tut oder der Servicemitarbeiter in der Schulbehörde. Das möchte ich beeinflussen. Habe ich also drei verschiedene Assistenten für drei verschiedene Nutzergruppen. So ein bisschen ähnlich wie wenn ich eine Webseite aufbereite und mache eine persönliche Ansprache für die jeweilige Nutzergruppe. Genau das mache ich mit der Regieanweisung, mit dem Prompt pro Assistent. Wenn es jetzt wirklich um einen individuellen Anwendungsfall geht, in dem der Kunde eine gewisse Erwartungshaltung hat, fängt man vom Testset her an, schauen wir die Fragen an, die gestellt werden. Wir schauen uns an, was gute Ergebnisse waren, idealerweise aus der Vergangenheit, als das ein Mensch irgendwann mal prozessiert hat oder manuell ausgefüllt hat.
[00:36:24.760] - Thomas Wächter
Und wir durchlaufen ein klassisches Data-Science-Vorgehen. Heißt, das Verfahren dahinter bedeutet, man schaut sich erst mal an, was ist der Geschäftsnutzen oder generell der Nutzen daran? Dann schaut man sich die Daten an, ob die sich dafür eignen, das umzusetzen. Bereitet die so auf, dass man, zum Beispiel bei einem Large Language Modell, würde man Frage-Antwort-Paare aufbereiten und die richtige Antwort markieren und auch Frage-Antwort-Paare aufbereiten, wo die falsche Antwort drin steht zum

Beispiel, und dann testen, ob die Antworten, die gegeben werden, vollständig sind. Das können jetzt textuelle Antworten sein, wie eine E-Mail, die generiert wird. Es kann aber auch einfach ein Datensatz sein, der aus einem Dokument extrahiert wird, den ich in einer weiteren Anwendung in einem Automatisierungsagenten weiterverwenden kann. Genau, das Modell wird dann modelliert bei einem Large Language Modell, entweder durch verschiedene Prompts und verschiedene Aufrufe, die ich der Reihe nach tätige oder eben, wie der Nicolas schon angesprochen hatte, durch ein Finetuning. Das brauche ich dann, wenn ich in einer Anwendung Inhalte habe, wo ich davon ausgehe, dass das Modell die vielleicht noch nie gesehen hat. Zum Beispiel eine Fertigungsrichtlinie in der Halbleiter-Industrie wird wahrscheinlich in dem Modelltraining-Satz nicht drin gewesen sein. Also weiß dass, kann das... wenn da drei Buchstaben drinstehen kann, kann das Modell das vielleicht falsch interpretieren?
[00:37:42.210] - Thomas Wächter
Da muss ich Beispiele geben. Oder auch in der Fertigungsindustrie eine Sicherheitsrichtlinie für das Anschrauben eines Bauteils. Auch das hat man wahrscheinlich nicht gesehen, weil es eher privatwirtschaftliche Daten sind, die in Internetquellen, in Büchern, in den typischen Quellen für das Training des Modells nicht enthalten wurden sein. Und da kann ich eben mit erweiterten Pre-Training und dann verschiedenen Finetuning-Ansätzen, das Modell darauf vorbereiten, dass es sich auch in dem Thema auskennt. Und das kann ich eben bei denen besonders gut machen, bei den Open Source Modellen und bei den anderen Modellen muss ich davon ausgehen und kann nur testen, ob ich vollständige Informationen habe.

Projektdauer

[00:38:21.120] - Steffen Wenzel
Das hört sich jetzt nach einer umfangreichen Projektvorbereitung an. Kannst du das mal so in Zeit auch einteilen? Ich weiß, es wird wahrscheinlich natürlich wieder nach an Anforderungen oder auf die Anforderungen ankommen, die dir dann gestellt werden. Aber gibt es da ungefähr so eine Formel, dass man sagen kann, so viel brauchen wir für die Projektvorbereitung, bis wir starten können?
[00:38:39.990] - Thomas Wächter
Für diese Pilotphasen gehen wir normalerweise davon aus, haben wir ein Muster uns festgelegt, einfach weil man kann immer beliebig viel Zeit rein investieren, aber man weiß auch, dass es nach einer gewissen Zeit ein Ergebnis ist, wo man erkennen kann, ob es vielversprechend ist. Wir sagen da schon, aber auch seit Jahren schon, Faustregel: 5-10-15. Das bedeutet fünf Tage für die Bewertung der Daten und eine generelle Machbarkeitsaussage zu treffen, zehn Tage, um das technisch nachzuweisen. Das kann manchmal schwieriger sein, weil ich die Daten aufbereiten muss, aber allein der fachliche Teil, das Modell vorzubereiten, sind zehn Tage Aufwand. Und wenn der Kunde dann wirklich ein Pilotsystem braucht, um zu prüfen in der Nutzergruppe zum Beispiel, geht man davon aus, dass man das auch in zusätzlichen 15 Tagen bereitstellen kann. Also in Summe erst mal ein überschaubarer Anteil und was dann natürlich, wie in jedem Individualprojekt dazukommen könnte, sind Live-Schnittstellen an verschiedene Systeme, die aber jetzt keine Neuerung der KI sind, sondern ganz normales Projektgeschäft.

Ausblick

[00:39:39.210] - Stefanie Lße
Apropos Neuerung. Jetzt habt ihr ja schon einiges in eurem Projekt erreichen können und wir haben heute auch schon ganz viel gehört über sehr konkrete Anwendungsfälle, die wir heute nicht nur durchdenken, sondern teilweise eben auch schon in der Praxis überführt haben. Nicolas, was würdest du sagen, wo geht die Reise für euer Projekt noch hin? Beziehungsweise, siehst du Anwendungsfälle, die wir heute noch nicht konkret umgesetzt haben, die aber in Zukunft auf uns dazukommen?
[00:40:05.060] - Nicolas Flores-Herr
Ich versuche mal, die Frage ein bisschen anders zu beantworten, wenn ich darf. Es ist so: Wir sprechen hier jetzt sehr viel über Teuken und ich würde an der Stelle gerne betonen, Teuken ist für uns bei Fraunhofer erst der Anfang. Wir haben jetzt weitere Rechenzeit, steht uns zur Verfügung. Wir werden an den Modellen selber auch weiterentwickeln. Das heißt, wir werden größere Modelle bauen, wir auch kleinere Modelle bauen, werden aber auch Modelle bauen, ich sage mal, DeepSeek inspirierte Reasoning-

Modelle, die eben auch noch mal ganz andere Fähigkeiten haben und die auch noch mal ganz anders in der Lage sind, auf bestimmte Domänen einzugehen und die auch sozusagen die Möglichkeit eröffnen, hier noch mal, ich sage mal, einen deutlichen Mehrwert der KI immer nachgesagt wird, das noch mal deutlich besser zu liefern. Ich sehe es eher in so einer allgemeinen Schiene. Es ist wichtig, dass wir jetzt hier gemeinsam diesen Schulterschluss haben am Forschungsprojekt wie OpenGPT-X, ich sage mal, auf dieser Kooperation mit der Telekom, dass sozusagen Forschungsergebnisse direkt in die Anwendung auch übersetzt werden können. Ich denke, das ist wirklich das Asset, was wir hier haben. Unser Job als Fraunhofer und auch als Konsortium wird es sein, immer bessere Modelle zu liefern und dann kann man Projekte, wie sie Thomas eben beschrieben hat, wirklich auch noch besser, noch cleverer und mit noch größerer Kundenzufriedenheit umsetzen.
[00:41:31.270] - Nicolas Flores-Herr
Und ich denke, das ist so ein bisschen die Richtung, in die wir uns bewegen. An uns ist die Aufgabe gestellt, wirklich wettbewerbsfähige Modelle zu bauen. Und das ist eine Herausforderung, die wir aber auch gerne annehmen.
[00:41:43.480] - Steffen Wenzel
Thomas, wenn du jetzt so ein bisschen in die Zukunft schaust. Du hast jetzt gerade von Nicolas ein paar Szenarien gehört. Was kannst du dazu noch beitragen? Was sind deine Szenarien für die Zukunft, wie sich solche Modelle und auch andere in Konkurrenz dazu Stehende dann weiterentwickeln werden?
[00:41:58.510] - Thomas Wächter
Ich glaube, die Zukunft vorher zu sagen, ist an der Stelle schwierig, weil wir wurden jetzt schon ein paar Mal überrascht. Also alles, was ich im Studium gelernt habe, in der Promotion gelernt habe, zum Thema Textanalyse, hat sich irgendwie auch ein bisschen verändert und teilweise auch überholt, was für uns aber aktuell essentiell ist, ist, dass wir besser mit Bild-und Textdaten in Kombination arbeiten können. Und da ist es so, dass die Modelle, die wir bisher nutzen, in der Regel entweder Bilder oder Texte interpretieren. Und wenn man das alles schön vergleichen möchte, müsste man das Bild und das Text im gleichen Vektorraum, also in der gleichen semantischen Repräsentation, miteinander vergleichen können. Heißt, Audio ebenfalls übrigens. Das heißt, wenn die Stefanie jetzt etwas sagt zum Thema unseres Podcasts und wir schauen uns ein Bild an, was in der Fraunhofer-Publikation ist, und die Publikation selber möchte ich das vergleichen können ohne Transformationsleistung und Informationsverlust dazwischen. Und jetzt gibt es die ersten Modelle, die nicht nur multimodal sind, also nebeneinander stehend quasi das interpretieren, Bilder und Text interpretieren können, sondern die auch Text und Bild gemeinsam interpretieren können, dass ich einfach alle Informationen, die ich irgendwie aufnehme, kommt ein bisschen auch aus der Verhaltenspsychologie, also Sachen, die ich wahrnehme, gemeinsam interpretieren kann, weil manchmal ergibt der Text auch nur Sinn mit dem Bild, was daneben ist und das Bild ist vielleicht nicht noch mal so richtig schön beschrieben.
[00:43:21.780] - Thomas Wächter
Müssen wir heute textuell das Bild auswerten, also den Text generieren, der das Bild beschreibt und können das dann mit dem Text vergleichen, ist aber nicht das Gleiche, wie wenn ich eine technische Zeichnung mit den Hinweisen, die da drin stehen, vergleiche.

Verabschiedung

[00:43:34.420] - Stefanie Liße
Vielen Dank, Thomas. Vielen Dank, Nicolas, dass ihr uns den Einblick heute gewährt habt. Ich bin absolut gespannt, wo die Reise hier noch hingeht und was da für tolle Anwendungsszenarien in Zukunft auf uns noch warten. Und Steffen, dir vielen Dank, dass ich mit dabei sein durfte und jetzt ja demnächst noch öfter mit dabei bin.
[00:43:51.810] - Steffen Wenzel
Ja, das freut mich sehr und ich freue mich auf alle weiteren Podcasts. Und vielen Dank, dass Sie sich auch die Zeit genommen haben, heute sich diesen Podcast anzuhören. In den Shownotes finden Sie noch die Links zu weiteren Infos zu dem heutigen Thema. Und natürlich, wenn Sie keine weiteren Folgen verpassen wollen, dann abonnieren Sie uns doch einfach auf den Ihnen bekannten Kanälen. Bis dahin, alles Gute und ganz liebe Grüße.