[00:00:02.030] - Sprecher 1
Ausgesprochen digital. Der Podcast für digitale Trends.

[00:00:07.100] - Sprecher 2
...dass etwa 55 % Code Productivity mehr entsteht. Und das ist natürlich ein Versprechen. Das kann ein Softwarehaus nicht ungetestet lassen und unversucht lassen. Es muss dieses Potenzial für sich erkennen. Es muss dieses Potenzial heben.

[00:00:29.090] - Sprecher 3
Am 30. November 2022 wurde mit ChatGPT ein textbasiertes KI Sprachmodell weltweit für die Öffentlichkeit freigegeben. Ein Tag, der vielleicht in die Geschichte der Menschheit eingehen wird, aber zumindest schon heute einen Meilenstein darstellt. Vier Monate später prognostizierte der Microsoft Gründer Bill Gates, dass eine soziale Schockwelle auf die Menschheit zurollen könnte, weil die Veränderungen in unserem Leben durch KI enorm weitreichend für uns alle sein werden. Andererseits glaubt Gates aber auch, dass mithilfe von künstlicher Intelligenz zum Beispiel im Gesundheitssektor enorme Fortschritte erzielt werden könnten oder die Bekämpfung des Klimawandels uns besser gelingen sollte. Wo stehen wir also heute? Mein Name ist Steffen Wenzel. Herzlich willkommen bei ausgesprochen digital. In dieser Folge wollen wir uns genau mit diesem Spannungsfeld zwischen Chancen und Risiken von KI Modellen auseinandersetzen. Wir blicken jetzt kurz vor Weihnachten auf das Jahr und seine Entwicklung zurück, aber auch nach vorne. Und das machen wir zusammen mit Frank Schönefeld, dem ehemaligen CTO der Telekom MMMS und heute Berater für Künstliche Intelligenz, mit dem wir bereits Anfang des Jahres über das Thema gesprochen hatten. Herzlich willkommen, Frank.

[00:01:42.710] - Sprecher 2
Ja, hallo in die Runde.

[00:01:44.240] - Sprecher 3
Frank, Du kommst gerade aus den USA von einer Technologiekonferenz. Welche Vorträge hast du da gehört und welche Rolle spielt das Thema KI?

[00:01:51.860] - Sprecher 2
Ja, es war eine Konferenz, die etwas breiter angelegt war, also die technologische Fortschritte über vielen Gebieten untersucht hat. Ein Schwerpunkt war Biomaterialien und wie ich die über den 3D Drucker erzeugen kann. Es ist faszinierend, welche Fortschritte es dort gibt. Wir können also inzwischen Organe drucken und wir können Extremitäten drucken und die dann auch. Das bleiben lebendige Materialien, und die können dann auch transplantiert werden. Also faszinierende Fortschritte. Obwohl das so breit angelegt war, hat KI in sehr vielen Vorträgen eine Rolle gespielt, weil es eben zur dominierenden Datenverarbeitungstechnologie geworden ist oder auch zur dominierenden Anwendung von generativen KI. Ich denke, ich habe in jedem dritten Vortrag einen Verweis auf generative KI gehört und das verbindet mich mit einem Eröffnungsstatement, dass das wirklich ein großer Tag für die Menschheit war. Dieses Bitte loslassen auf die Menschheit, Damit ist wirklich was Großes in Gang gesetzt worden. Die Konferenz hat das einmal mehr bestätigt und hat dann solche Themen. Wie kann ich mit generativer KI mehr Testdaten generieren? Welche Regulierungsansätze gibt es schon? Sind die alle gleichermaßen relevant?

[00:03:15.650] - Sprecher 2
Sind sie berechtigt und hat das also auch mit adressiert? Insofern sehen wir also, dass dieses KI Thema und generative KI Thema schlechthin ein Motor für Technologieentwicklung geworden ist.

[00:03:30.080] - Sprecher 3
Wir haben damals glaube ich, auch im ersten Podcast, den wir zu dem Thema gemacht haben, gesagt, dass wir mit Chats ja nur die Spitze des Eisbergs sehen. Was hat sich denn darunter so entwickelt?

[00:03:41.210] - Sprecher 2
Ja, beim normalen Eisberg spricht man glaube ich von diesem 6/7 unter der Wasseroberfläche und

1/7 drüber. Wenn wir mal sagen die Spitze des Eisbergs ist immer noch gibt. Vier dann würde ich aber eher sagen, wir haben es mit so einer 9/10 oder vielleicht auch 19/20 Eisberg zu tun. Denn in der Tat, es gibt immer noch von es ein Spitzenprodukt ist eine Spitzentechnologie. Aber was alles in diesem Jahr 2023 zusätzlich passiert ist, unter der Oberfläche oder scheinbar unter der Oberfläche rechtfertigt wirklich von einer dramatischen Entwicklung zu sprechen. In meinen Vorträgen verwende ich häufig das Wort Cambridge Explosion, in dem ich sage, mit diesem Startschuss durch BT ist was in Gang gesetzt worden, was in rasender Geschwindigkeit eine völlig neue Welt erzeugen wird. So wie damals die Cambridge Explosion des Lebens vor 600 Millionen Jahren. Ohne jetzt tiefer darauf eingehen zu wollen, den Begriff hat da auch schon mal der CEO von Nvidia erwähnt oder verwendet, um auf die Entwicklung der KI in Gänze einzugehen. Und wir können das vielleicht in einzelnen Strängen dann auch noch untersuchen. Aber nur um mal zu sagen, was sind denn die anderen 90 % oder auch 19/20, die da unter der Oberfläche laufen oder schwimmen?

[00:05:10.610] - Sprecher 2
Wir haben eine wahnsinnige Entwicklung im Open Source Bereich gesehen. Das kann ich noch mal mit Zahlen unterlegen. Also das ist wichtig, sich anzuschauen und zu verstehen, was dort passiert. Wir haben das Rennen der Großen gesehen, während am Anfang scheinbar nur GPT. GPT. Turbo. Ihr das Feld zu beherrschen schien, sehen wir jetzt die Antworten insbesondere von Google. Aber die nächsten Antworten werfen auch ihre Schatten schon voraus. Insbesondere von Amazon können wir auch noch etwas näher skizzieren. Das ist faszinierend. Und drittens sind natürlich auch die Anwender, unsere Kunden nicht stehen geblieben, sondern haben sich ihr Bild von der Situation gemacht und haben versucht zu überlegen Was mache ich denn? Wie kann ich meine Unternehmensdaten auf eine sichere Art und Weise, aber dennoch mit den Fähigkeiten der großen Sprachmodelle der generativen KI miteinander verbinden, um den möglichst größten Nutzen für mich zu erzeugen? Und da ist wahnsinnig viel passiert. Und wenn man das mal in Zahlen illustrieren wollen ein paar sind ganz spannend. Also 92 % der Fortune 500 Companies nutzen oder artverwandte Produkte, das heißt von 500 Firmen.

[00:06:30.010] - Sprecher 2
WELT Spitzenfirmen nutzen 460 Firmen. Und generative KI hat selbst stabil 200 Millionen monatliche Nutzer und ungefähr 1,8 Milliarden monatliche Besuche. Das hat sich wieder stabilisiert. Da gab es ein kleines Sommerloch, da ging das auf 1,5 Milliarden runter. Aber jetzt im November sehen wir wieder stabile Besuche von 1,8 Milliarden, um das mal in Vergleich zu setzen. Google die die Suchseite, die hat weltweit 85 Milliarden Besucher, das ist etwa Faktor 50. Aber wenn ich jetzt mal die Funktionsweise von beiden ins Verhältnis setze, ist das dennoch eine tolle Zahl. Die Open Source Bewegung hatte ich angesprochen, dort vielleicht die die bemerkenswerteste Entwicklung das Bereitstellung von Llama Llama eins, Llama zwei Llama Code oder Code Llama. Das ist eine Entwicklung von Meta früher Facebook sehr zaghaft, zuerst nur das Modell freigestellt, dann die Parameter dazu veröffentlicht, dann die Bedingungen angepasst und inzwischen ist es wirklich eine fast vollwertige Open Source Lösung. Und wozu hat das geführt? Dieses Llama Modell wurde inzwischen über 100 Millionen mal downgeloadet und Entwickler haben 42.000 abgeleitete Modelle von Llama entwickelt. Da merkt man natürlich auch die Kraft des Open Source Gedankens.

[00:08:03.010] - Sprecher 2
Der setzt sofort Kreativität. Das setzt Produktivität, das setzt Innovation frei. Und wenn man auch sonst vielleicht nicht viel von Meta hält und seinen Standardanwendung mit Llama haben Sie wirklich mal was Großes und was Tolles in die Welt gesetzt. Also tolle, tolle Entwicklung, die wir dort sehen.

[00:08:22.480] - Sprecher 3
Also wir sehen eine Menge an Firmen, die jetzt damit arbeiten wollen. Die Anwendungsfälle sind jetzt aber noch nicht so tiefgehend oder haben sich nicht so tiefgehend weiterentwickelt. Wir kennen natürlich das System, dass wir jetzt Fragen stellen können, das wir natürlich gerade für Klassenarbeiten oder an der Universität das ganze Arbeiten von KI quasi gemacht werden. Aber gibt es neue, andere Anwendungsfälle, dass du uns mal ein bisschen mit auf die Reise nehmen kannst, wohin sich das entwickeln wird?

[00:08:54.310] - Sprecher 2

Also nehmen wir mal den Strang zuerst. Was machen die Anwender? Für die steht natürlich nach wie vor die Frage Was kann ich von diesem trainierten Modell verwenden und wie viel muss ich es noch mit eigenen Informationen, mit eigenen Daten, mit eigenen Erkenntnissen anreichern, damit ich selbst was Positives daraus zurückbekomme? Und in der Tat, ich glaube, die Situation ist gut beschrieben, wenn man sagt Ja, das ist ein vorsichtiges aneinander herantasten und auch dort überwiegen am Anfang die spielerischen Versuche, bis man eben zu den Kosten einsparen, den Potenzialen kommt oder zu den Produktivitätssteigerungen Potenzialen kommt. Und das darf man aber nicht unterschätzen. Man braucht nur sich überlegen, welche Unternehmensprozesse alle in Textinterface oder in Text, Output oder auch in text textbasierten Workflow haben. Ja, also der Harari, der Autor von Homo Deus, der hat gesagt, das Betriebssystem des Menschen ist gehackt worden, die Sprache. Und ich, ich glaube, da drin schwingt das große Potenzial mit, das in jedem Einzelfall zu identifizieren und zu heben. Das ist eine Aufgabe, die noch zu gehen ist.

[00:10:14.770] - Sprecher 2
Und dafür sind dann aber auch Firmen wie die Telekom MMMS da, um das gemeinsam mit den Kunden zu erarbeiten, in Use Cases umzuwandeln, zu implementieren, zu schauen, ob sich die Produktivitätsvorteile erheben. Um ein paar Indikationen anzugeben. Im textbasierten Umfeld kann man sich mit Ausschreibungen beschäftigen. Man nimmt die Ausschreibungsbedingungen rein. Gerade im öffentlichen Bereich sehr viele Ausschreibungen und erzeugt automatisch über generative KI eine Ausschreibung. Ganz toll, oder? Man will über seine Produkte, Preise und Dienstleistungen informieren. Wer tut das nicht? Das tun alle Telekommunikationsunternehmen, das tun alle Logistikunternehmen. Das tun alle Unternehmen, die überhaupt einen Geschäftskunden oder Privatkunden Kontakt haben. Hier hat man es auch leicht, weil die Daten, Leistungen, Preise, Services sind sowieso frei im Internet verfügbar. Warum soll ich die nicht auch in einem natürlich sprachlichen Anfragemodus rund um die uhr 24 mal sieben zur Verfügung stellen und mich dort der Leistung eines großen Sprachmodells bedienen? Das sind also Anwendungen, die sehr schnell auf der Hand liegen und die wirklich Spaß machen. Wie schon in Teil eins gesagt habe.

[00:11:34.170] - Sprecher 2
Wir dürfen natürlich uns nicht nur auf Text zu Text sozusagen versteifen. Dann lassen wir 80 % der Möglichkeiten von generativer KI liegen. Gerade das Multimodale kommt ja jetzt erst so richtig in Schwung. Das heißt die Fähigkeit dazu, auch Videos, Audios, Bilder zu verarbeiten, zu verstehen und sowohl in den Eingabeprozess als auch partiell in den Ausgabeprozess schon zu integrieren. Und die fortgeschrittensten Unternehmen, die überlegen Was ist denn überhaupt meine Kreativleistung? Was die seine? Ich denn? Nehmen wir mal die Halbleiterindustrie, die hier auch allgegenwärtig ist. Und die sagt natürlich, wie komme ich schneller zu neuen Chipentwürfen, zu neuen Chipdesigns? Kann ich das nicht auch mit generativer KI zu unterstützen? Und da hilft mir natürlich nicht Text zu Text, sondern ich würde eine Spezifikation per Text geben und ich brauche natürlich dann ein Ablaufplan, einen Schaltplan etc. etc. pp. Aber warum nicht so weit gehen? Das wird vorgedacht. Das dauert aber noch ein bisschen, das wir auch dort hinkommen.

[00:12:48.330] - Sprecher 3
Du hast eben Google auch erwähnt als einen dieser Hauptplayer. Denen ist es aber beispielsweise jetzt noch nicht gelungen, eine Verknüpfung mit ihrer Suchmaschine herzustellen, wenn ich das richtig sehe. Bzw. Sehen Sie darin ein Problem, Wahrscheinlich sogar eine Kannibalisierung Ihres Geschäftsmodells. Kannst du dazu mal was sagen?

[00:13:05.340] - Sprecher 2
Das war, glaube ich, die große Herausforderung überhaupt für Sie. Das alte Geschäft lief ja so gut, die Suche und da ist ja auch schon KI definitiv drin gewesen, wenn vielleicht auch nicht generative KI. Aber warum sollte man ein existierendes, funktionierendes Geschäftsmodell usurpieren, gefährden oder irgendwie in das Balance bringen? Ich glaube, das war das Problem was Google hatte. Man nennt das ja auch das Innovators Dilemma. Alle Basistechnologie sind von Google entweder erfunden oder weiterentwickelt worden. Gerade die Transfer Transformer Modelle als auch die Attention Mechanismen. Und trotzdem ist ihnen Microsoft einfach davongeflogen mit OpenAI und und gibt und hat auch sehr schnell eine geschäftliche Verknüpfung gefunden. Das heißt, wenn ich das heute bewerten sollte, würde ich sagen Integration von generative KI in das existierende Geschäftsmodell

Microsoft. Eine eins mit Sternchen und Google muss noch mal auf die Sitzbank und und nacharbeiten. Aber das hat sie nicht davon abgehalten. Zunächst mit Bart und jetzt mit ihrer neuesten Ankündigung, wo auch die Resultate schon veröffentlicht sind. Mit Gemini sozusagen wieder den den den Thron zu erklimmen.

[00:14:25.620] - Sprecher 2
Was die Leistungsfähigkeit betrifft Im Vergleich zwischen Gemini und GPT. Ihr in in der multimodalen Leistungsfähigkeit, also wenn es darum geht, Text, Bild, Video, Audio als Input zu verarbeiten, schlägt Gemini vier in neun von zehn Vergleichen, wenn auch immer nur so um 334 Prozentpunkte. Aber sie liegen damit vorn. Zumindest nach dem eigenen Research Report, den wir auch verlinken können in den Beimaterialien unseres Podcasts. Wir müssen das jetzt auch mal selber uns anschauen dürfen. Nach Europa kommt wahrscheinlich etwas später. Das hängt auch mit den Regulierungsund gesetzlichen Regelungen zusammen. Seit 13:12 soll, wo für Unternehmen eingeschränkter Zugriff auf Gemini existieren. Also das muss man noch mal ein bisschen nebeneinander wägen. Auf jeden Fall kann man sagen das Imperium in dem Fall Google. Hat zurückgeschlagen, hat mit Yamina eine unheimlich leistungsfähige Maschine auf die Erde gesetzt und unheimlich leistungsfähig ist es auch im Trillionen Bereich. Wobei ich jetzt die amerikanische Trillion verwende, also zehn hoch 15. Um Missverständnissen vorzubeugen im Trillionen Parameter Bereich und auch etwa 30.000.000.000.000.000.000 Token in der Trainingsmenge. Das sind unvorstellbare Größen.

[00:15:54.340] - Sprecher 2
Es fällt uns schwer, das nachzuvollziehen. Auf jeden Fall sind das Datenmengen, die ein Mensch im Laufe seines Lebens definitiv niemals wird aufnehmen können, verarbeiten können. Und da merkt man natürlich schon, dass so das Ungleichgewicht sich auch dort weiter verstärkt. Aber das ist vielleicht die Spitze des Eisbergs. War vorhin unser Thema. Jetzt haben wir da zwei Spitzen unseres Eisbergs mit Gemini und mit vier fünf wird nicht lange auf sich warten lassen. Das wird also spannend, was dort an Neuem dazukommt. Und dann hat sich noch ein dritter Großer auf den Weg gemacht, nämlich Amazon. Einerseits sind die schon über Entropic investiert und haben Cloud und Cloud zwei als Modelle am Start. Aber jetzt wollen sie noch mal ein richtig großes 2.000.000.000.000.000.000 Parameter Modell an den Start bringen. Codename Olympus wird also also 24 wird auf jeden Fall auch ein ganz spannendes Jahr, wenn man dieses Rennen dort an der Spitze praktisch betrachtet. Vielleicht noch ein Wort zu den Ressourcen dazu einzubringen. Das Gemini Paper listet alle Core Contributors und Contributors zu dieser Arbeit auf. Man kommt dort etwa auf 1000 Namen.

[00:17:16.840] - Sprecher 2
Das heißt, wir reden hier wirklich von 1000 Personen, die an dem Projekt gearbeitet haben. Dann merkt man, welche unheimliche Ressourcensituation man anstreben muss, in der Lage sein muss, Ressourcen an den Start zu bringen, um ein derartig leistungsfähiges Modell praktisch zu erzeugen.

[00:17:37.660] - Sprecher 3
Du hast im Vorgespräch auch mal von den glorreichen Sieben erzählt und natürlich diesem Vorsprung, der anscheinend dort in US amerikanischen Unternehmen jetzt auch herausgearbeitet wurde. Kannst du uns noch mal erklären, was du damit meinst und ob es da überhaupt noch eine Chance gibt, die wieder einzuholen?

[00:17:52.870] - Sprecher 2
Also die glorreichen Sieben. Das nimmt natürlich Bezug auf einen wunderbaren Western aus den 60er Jahren mit Jules Brynner und Horst Buchholz, wenn ich mich recht erinnere, der da den jugendlichen deutschen Liebhaber spielen durfte. Oder als deutscher Schauspieler, den den jugendlichen Liebhaber spielen durfte, die dort ein mexikanisches Dorf vor immer wiederkehrenden Plünderern beschützen. Die glorreichen Sieben. Jetzt verwendet man den Begriff, um wirklich die Vorreiter dieser Welt zu benennen, und da hat man es mit guten Bekannten zu tun. Drei fangen allein mit A an, also als Eselsbrücke Amazon Alphabet und Apple als die ersten der drei. Dann haben wir zwei mit einem das wäre Microsoft und Meta known as Facebook. So, und dann haben wir einen mit n Nvidia, der natürlich von unten das Feld aufmischt. Und ohne Nvidia wäre ein Großteil dieser Entwicklung auch nicht denkbar. Und man zählt häufig Tesla noch dazu, weil die natürlich was Anwendungen von KI,

Autonomes Fahren Stichwort sozusagen betrifft, auch als aber auch moderne Energieversorgung Energiesysteme natürlich in diesem Kontext gesehen werden so und glorreiche sieben Deswegen wenn man sich die Entwicklung des Aktienmarkts anschaut, dann stellt man fest, dass er zumindest bis November Dezember Anfang Dezember, wo er fast nur von diesen glorreichen Sieben getragen, also die positive Entwicklung und der Rest des Marktes hing praktisch wie ein Schluck Wasser in der Kurve und entwickelte sich.

[00:19:34.300] - Sprecher 2
Und deswegen eben diese glorreichen. Und wenn man es jetzt aber speziell auf was können die im Umfeld KI wirklich mal runterbricht, dann merkt man, das ist substanzieller Fortschritt, was dort erreicht ist. Auch da ist das Gemini Paper sehr hilfreich zu lesen. Wenn man allein mal die Architektur, die da dahinter liegt, welche Hardwarebeschleunigung verwendet werden muss, welche territoriale Ausdehnung zwischen den Rechenzentren besteht, die alle in das Training integriert werden. Also man kann wirklich sagen, hier wird kontinental übergreifend das Training fortgeführt, um die 30 Millionen Tokens zu verarbeiten und in ein Modell von einer Trillion Parameter 1,5 Millionen Parameter sozusagen zu zu pressen und Google Nutzer seine eigene Spezialhardware, also die. Version vier und CPU Version fünf, das heißt Tensor Flow Processing Unit. Eine ganz spezielle Hardwarebeschleunigung. Und die wird aber auch dann schon wieder in speziellen Cube Topologien zusammengefasst, das heißt jeweils eine TP bildet sozusagen Stichwort Cube eine Seite eines Würfels und darunter hängt die nächste und an der Seite die dritte und dann an der anderen Seite die vierte. Das heißt, wir haben es mit absoluten Hochbeschleunigungstechnologien dort zu tun und in dem Umfeld ihnen am nächsten kommt dann eigentlich Amazon, die auch in der Lage sind, solche hervorragenden Architekturen cloudbasierte Architekturen anzubieten.

[00:21:12.340] - Sprecher 2
Und das fällt natürlich durchaus schwer, das so nachzuholen. Es ist allerdings so, wenn man in unbekanntes Gelände geht, dann kostet das erstmal mehr Aufwand. Den tragen eben die glorreichen sieben und sie können den auch tragen. Aufgrund ihrer hervorragenden finanziellen Performance. Wenn man jetzt als Early Follower hinterher geht, dann ist ein Großteil des Geländes schon kartiert und man kann dann durchaus mit besserer finanzieller, mit geringerem finanziellen Aufwand auch sehr schnell zu guten Resultaten kommen. Den Vorsprung und die Erfahrung wettzumachen, das ist im Moment schlechterdings unmöglich.

[00:21:52.510] - Sprecher 3
Wie sieht denn dein Blick auf Deutschland oder wie betrachtest du Deutschland in dem Zusammenhang? Wir haben ja auch ein paar Unternehmen, die sich mit dem Thema beschäftigen. Es wird jetzt natürlich auch versucht, von der Politik auf der einen Seite zu regulieren, aber auch zu fördern. Hast du ein bisschen Optimismus für uns mitgebracht?

[00:22:06.880] - Sprecher 2
Also ein bisschen Optimismus hilft immer. Damit können wir uns ja zumindest erst mal anfreunden. Und in der Tat stehen uns natürlich zumindest diese Forschungsergebnisse partiell zur Verfügung und man kann darauf auch aufsetzen. Wir haben in Deutschland die Firma Aleph Alpha, die mit dem Luminus Modell auch ein großes Sprachmodell anbietet. Dort hat sich auch eine Finanzierungsgruppe gefunden, die auch 500 Millionen in die Hand genommen haben, um Aleph Alpha zu stützen. Und ich denke, nie war das Wort von digitaler Souveränität besser angebracht als in so einem Umfeld. Und ich? Ich glaube, wir können die Entwicklung zumindest partiell dort nachvollziehen. Daneben gibt es tolle universitäre Forschungsergebnisse vom Professor Schmidbauer oder vom Professor Reid Huber, wo wirklich tolle Basisarbeit schon vor Jahren gelegt worden ist, von der wir immer noch profitieren. Oder auch im Umfeld der latent Diffusionsmodelle, die die LMU in München nach wie vor ein ganz starker, ganz starker Player. Das heißt, wir haben vereinzelt die Fähigkeit, das jetzt auf das Level der glorreichen Sieben zu heben. Ist es nicht möglich, auch aufgrund der finanziellen Abstände muss man sagen.

[00:23:31.210] - Sprecher 3
Wenn wir stärker in Open Source investieren würden.

[00:23:35.320] - Sprecher 2
Das ist auf jeden Fall eine Variante. Also entweder von der globalen Open Source Bewegung zu profitieren. Über Llama hat man schon gesprochen 100 Millionen Downloads. Wir haben hier auch in der Firma Projekte, wo wir auf Llama aufsetzen, können wir vielleicht später noch ein bisschen untersetzen. Aber dort gibt es auch Projekte von Fraunhofer Instituten. Insbesondere das Fraunhofer AIS, also Institut für Automatisierung und Informationssysteme oder Analyse und Informationssysteme hat eine Zweigstelle hier in Dresden. Die arbeiten auch an einem Open Access Modell und das soll ein 70 Milliarden Parameter Modell, also durchaus schon der gehobenen Klasse, könnte man sagen werden. Und wir haben uns mit den Kollegen hier in Dresden auch getroffen. Unsere generative KI Gruppe und wir werden die Zusammenarbeit fortsetzen und sind sehr gespannt, welche Resultate dort kommen. Also mit anderen Worten, wir können nicht die zehn oder 12 Milliarden von Microsoft kopieren. Wir können nicht den den großen Ansatz, den Amazon oder Gemini mit 1000 Personen an einem Projekt. Das ist nicht möglich, aber wir haben durchaus vereinzelt spannende Ansätze und können das auch stärker auf Europa zuschneiden, auf die deutsche Sprache zuschneiden, auf weitere europäische Sprachen zuschneiden und mit den europäischen Regulierungsbedingungen in Einklang bringen.

[00:25:05.830] - Sprecher 2
Das sind die Optionen, die wir haben. Und wie gesagt, wenn man hinter der gepflasterten Weg Straße den die den, die die Großen gebaut haben herlaufen kann, dann spart das natürlich auch schon Aufwände.

[00:25:18.520] - Sprecher 3
Also ich glaube, du meinst Professor Schmidhuber.

[00:25:20.920] - Sprecher 2
Danke, dass du mich korrigierst. Ehre, wem Ehre gebührt. Professor Schmidhuber, einer der ganz frühen KI Pionier. Deutschlands hat viel zur Aufklärung des verschwindenden Gradienten, Problems oder der explodierenden Gradienten beigetragen und ist auch heute noch sehr visionär unterwegs. Auf die Frage, ob KI uns mal gefährlich wird, war seine Vision die als allererstes wird die auf den Merkur auswandern, weil sie dort bessere Energie Verwertungsbedingungen hat. Finde ich zumindest sehr inspirierend. Meine große Hochachtung für Herrn Schmidhuber und seine Leistungen in der Vergangenheit und seine visionäre Kraft.

[00:26:02.190] - Sprecher 3
Ja, ich habe auch mal eine Idee wegen Geschäftsmodellen, was mir noch nicht so ganz klar ist, wenn ich das mal so vergleiche. Also wenn du die Größe dieser Sprachmodelle immer ansprichst, also wie viel Items haben sie, wie viele Parameter benutzen die und das so auch plastisch hier uns vor führst, habe ich immer so ein bisschen den Gedanken an so einen Formel eins Wagen, der unheimlich viel PS hat, aber das muss man ja auf die Straße bringen. Also es zählt ja auch zum Beispiel wie ist das Gewicht verteilt, wie windschnittig ist der usw und natürlich der Fahrer am Ende. Also ich verstehe das Geschäftsmodell in dieser ganzen Größe und diesen ganzen Umfang dieser Modelle noch nicht so ganz. Wenn ich daran denke, brauchen wir überhaupt diese Größe, wenn ich nachher aus oder spezialisierte Anwendungsfälle habe?

[00:26:49.050] - Sprecher 2
Also das ist eine sehr gute Frage, sehr berechtigt. Aber wie so oft hat darüber Google und Gemini auch nachgedacht. Und haben sie überlegt Ja, braucht denn jeder das High End? Braucht jeder die 1.000.000.000.000.000.001,5 Millionen Parameter? Oder ist für gewisse Anwendungen auch ein anderer Ansatz gut? Und deswegen kamen sie von Anfang an mit einer dreistufigen Klassifikation raus und haben gesagt Gemini wird es einmal als Ultra geben. Das ist The biggest thing on earth sozusagen. Und da ist klar, damit das überhaupt selbst nach dem Training, es ist ja dann vor trainiert und weiß alles, was man wissen kann. Aber um dieses Wissen abzurufen, selbst dazu brauche ich diese Spezialarchitekturen, diese Cube Topologie usw und so fort. Ja und wem ist das schon gegeben? Also bis auf in einer Cloud Installation von Google selber werden das wenig sonst replizieren können abrufen können. Aber kommt Gemini mit der nächsten Idee, dann lass mich in eine kleinere Variante Faktor zehn, vielleicht kleiner auch bereitstellen. Das ist Gemini Pro und dann geht

es aber noch einen Schritt weiter und dann Gemini Nano und das wäre die kleinste und die Gemini Nano, die wir auch in lokal ablaufen Geräten sozusagen einsetzbar.

[00:28:12.450] - Sprecher 2
Das heißt, die läuft mal auf einem Raspberry, die läuft auf einem Handy oder die läuft überhaupt auf einem lokalen Gerät. Stichwort Edge Computing, wo man sowas machen kann. Und ich glaube, da sind die verschiedenen Anforderungen wirklich sehr gut klassifiziert worden. Und man kann das eben in drei Ausspielvarianten Deployment und damit den unterschiedlichen Anwendungen Rechnung tragen. Also Formel eins wäre die Ultra und ganz am Ende Nano. Ja.

[00:28:43.530] - Sprecher 3
Vielleicht Deutsche Tourenwagen, Meisterschaft.

[00:28:46.560] - Sprecher 2
DTM oder ähnliches. Und es ist interessant, die Leistungsfähigkeit dann zu vergleichen. Also haben sie natürlich auch gemacht. Und klar ist das Ultra liegt ungefähr 20 bis 25 % vor dem Pro und ungefähr 40 % vor dem Nano. Das heißt, wenn und das ist schon erheblicher Leistungsverlust. Das heißt, wenn ich jetzt von dem Nano alles will, dann wird er einfach 40 % schlechter arbeiten als der Ultra. Das würde ich also nicht machen. Wenn ich aber sage, der Nano, der ist nur für dieses spezielle Thema Text, sumorizing oder translating oder für ein ganz eingeschränktes Gebiet, da kann der hervorragende Arbeit leisten. Ist ja Faktor 1000 preiswerter zu pflegen, weniger Energie intensiv im Betrieb. Und so weiter und so fort. Also das macht Sinn. Und da betritt Gemini auch wirklich Neuland, sozusagen, das so stark zu klassifizieren. Ich meine, die anderen haben das auch gemacht, die haben Llama gibt es als sieben B 13 B und 64 B, wobei die B immer für Billion Milliarden sozusagen steht. Das ist auch schon so eine klasse Klassifizierung gewesen, aber ich glaube Gemini dreht das noch ein kleines Stückchen weiter nach vorne und hat sich genau überlegt, was sind denn klassische Anwendungsfälle und wie?

[00:30:16.020] - Sprecher 2
Wie geht das Deployment Procedure? Dementsprechend also und.

[00:30:20.130] - Sprecher 3
Wir werden uns in Deutschland nachher mit Google oder Microsoft auseinandersetzen müssen über Lizenzmodelle, dass wir diese KI basierten. Ja Modelle dann benutzen dürfen.

[00:30:30.870] - Sprecher 2
Das ist ja jetzt schon.

[00:30:32.220] - Sprecher 3
Dafür bezahlen wir 30 $ als Privatperson. Das kann ja jetzt nicht das riesen Geschäftsmodell sein.

[00:30:38.670] - Sprecher 2
Na ja, aber warum nicht? Also die Integration bei Microsoft zum Beispiel sieht genauso aus. Da werden es auch die 30 $ oder die 30 € pro Account sein. Wenn ich die volle Integration der KI Möglichkeiten in meine Microsoft 365 in Teams in Office usw haben möchte und das ist ein sehr klares Geschäftsmodell und Unternehmen kann sich dann überlegen will ich das für alle Mitarbeiter ausrollen? Wenn ich 1000 Mitarbeiter habe, dann weiß ich, das kostet mich mal 30.000 € im Monat mehr sozusagen. Aber ich habe es dann für alle, also im Jahr drei 160.000 € mehr für eine fiktive 1000 Personen Firma. Das muss man wissen. Gut, wir haben die Alternativen. Wir können in den Open Source Bereich ausweichen, aber dann muss ich natürlich selber Deployment selber betreiben, betreuen usw.

[00:31:33.450] - Sprecher 3
Das interessiert mich Jetzt noch mal ganz kurz, warum Amazon da für sich ein Geschäftsmodell drin sieht. Ich meine, die machen das ja nicht aus philanthropischen Gründen.

[00:31:42.510] - Sprecher 2

Meta meinst du mit Meta? Genau. Na ich denke, für Meta ging es insbesondere darum, Kompetenz erst mal zu demonstrieren. Ja, ihr könnt uns viel erzählen, OpenAI und Microsoft. Aber wir haben auch Kompetenz und das haben wir ganz zweifelsfrei unter Beweis gestellt mit Lana. Und sie haben durchaus auf die Kraft einer solchen Open Source Bewegung gesetzt, weil dadurch natürlich die Innovation der Welt eingefangen werden kann. Und die kann man dann wieder in eigene Services übersetzen. Also ein konkurrierender Ansatz, aber nicht nicht unbedingt weniger erfolgsversprechend.

[00:32:23.640] - Sprecher 3
Du bist ja auch Berater von der Telekom MMMS Weiterhin auch was ja früher CTO auch in der Firma. Und du hast ja eben schon mal so ein paar Kundenanfragen mögliche dann jetzt auch genannt. Aber was mache ich denn jetzt als Digital Experience Provider, wie die MMS einer ist, wenn ich mich dort weiterentwickeln kann. Und ich brauche ja auch Investitionsvolumen, weil ich das erst mal alles ausgestalten will. Es ist ja nicht sofort, sind ja nicht 100 Firmen da, die sagen Ja, bitte, macht das für mich und wir bezahlen das auch.

[00:32:51.270] - Sprecher 2
Ja, also ich sehe zwei Hauptstoßrichtung für die MMS. Das eine ist natürlich, sie muss ihr klassisches Portfolio noch mal mit mit dieser Fragestellung einfach neu durcharbeiten, durchdringen und sagen Wie sieht Customer Experience unter den Bedingungen generativer KI oder KI allgemein aus? Wie sieht digitale Zuverlässigkeit unter den Bedingungen von generativer KI KI allgemein aus? Wie sieht die digitale Organisation die digitale Zusammenarbeit der Zukunft unter Bedingungen der Generative von KI und KI allgemein aus? Also das wäre so Schritt eins und es ist ganz klar eine überlegene und überragende Customer Experience zum Beispiel für die Zukunft, die ohne generative KI im Sprachmodus im Chatbot Modus arbeitet. Die wird. Die wird keiner mehr brauchen, die wird keiner mehr wollen. Das heißt, ich muss zwingend mein klassisches Portfolio um diese Themen anreichern. Ein anderes Beispiel Die MMS hat viel im Barrierefreiheit Bereich gemacht und im Barrierefreiheit Bereich jetzt auf Möglichkeiten generativer Textgenerierung, Audiogenerierung, Videogenerierung zu verzichten wäre absurd. Das geht also gar nicht. Aber da merkt man schon, wie wie das praktisch ineinander geht. Ein großes Versprechen generativer KI.

[00:34:24.120] - Sprecher 2
Wenn ich jetzt in die Kodierung Programmierung schaue, ist natürlich diesen Prozess der Programmierung der Codierung massiv zu verändern, die Produktivität ganz steil ansteigen zu lassen, die Kosten diesbezüglich zu senken. Die letzten Messungen auch im Umfeld von sowohl am Code als auch im Umfeld von Gemini, also Code Productivity, sprechen, das etwa 55 % Code Productivity mehr entsteht. Und das ist natürlich ein Versprechen. Das kann ein Softwarehaus nicht ungetestet lassen und unversucht lassen, es. Es muss dieses Potential für sich erkennen. Es muss dieses Potenzial heben und auf seine Art eigene Weise interpretieren und in Anwendung bringen. Und da bin ich froh, dass die MMS da schon die ersten Schritte gegangen ist, um das praktisch durchzuführen. Auch da bieten sich Open Source Ansätze natürlich zunächst an, aus mehreren Gründen. Erstens Kostengründe, aber zweitens auch die Datensicherheit und Privatheit meiner eigenen Codebasen, die ich dann mit diesem Sprachmodell integriere. Das habe ich immer noch in meiner eigenen Gewalt. Und dann versuche ich natürlich, dort meine Codegenerierung zu beschleunigen und insbesondere für exotische Sprachen, die aber für meine Kunden einfach noch relevant sind.

[00:35:56.780] - Sprecher 2
Da sind die Vortrainierten selbst Code Llama nicht nicht wirklich gut. Code Llama ist gut für Python und angrenzende Sprachen, aber für spezielle Skriptsprachen. Da kann Code Llama mal eben gar nichts. Und da ist natürlich dann wichtig, dieses Wissen über diese Skriptsprachen, die ja dennoch für meine Kunden gebraucht werden, mit Code Llama zu integrieren. Da gibt es verschiedene Ansätze, wie man das machen kann. Man kann Feintuning, wie man ein Nachtraining eines solchen Modells nennt. Man kann aber auch einfach die, die die größeren Kontext Input Bedingung nutzen, die Codenummer bietet, um praktisch das ganze Repository als Kontext sozusagen zu integrieren und dann mit diesem Repository als Kontext neuen Code zu generieren. Aber da gibt es noch tausende offene Fragestellungen und ich bin aber froh, dass wir sowohl die die Hardware Voraussetzung um sowas zu machen. Training ist aufwendig, wissen wir ja, als auch die inhaltlichen Arbeiten dazu angestoßen haben. Denn das ist neben Beratung und Service Geschäft ist natürlich

Softwareentwicklung in Projekten unser Hauptgeschäft oder auch für Produktentwicklung, wenn Software dort eine Rolle spielt.

[00:37:16.970] - Sprecher 2
Deswegen müssen wir diesen Prozess ganz stark stützen. Und die 55 % das Versprechen Das darf man nicht ignorieren, das muss man auf Herz und Nieren prüfen und für sich selbst nutzbar machen.

[00:37:28.310] - Sprecher 3
Er macht aber auch einen großen Teil eures Geschäfts mit Tests und Integration. Also das ist ja ein wichtiger Teil. Kann man sich vorstellen, dass das auch automatisiert jetzt in Zukunft durch KI Anwendungen passieren wird? Weil ich habe ja im Endeffekt ja immer die gleichen Parameter, nach denen das quasi systematisch dann auch getestet wird.

[00:37:45.470] - Sprecher 2
Also da gibt es drei Denkrichtungen, die man dazu haben kann. Das eine ist das klassische Testen natürlich KI gestützt zu machen. Ich nehme an, ich muss eine neue Oberfläche testen, ob der Prozess genau in der Oberfläche so abgebildet ist wie spezifiziert. Das können wir. Da haben wir ein Produkt dafür, das heißt KI for test oder AI for test. Das können wir mit Erkennungstechnologien und dann stechnologien sind wir dazu heute schon in der Lage. Das ist so die die eine, die klassische Denkrichtung. Die zweite ist, da gehe ich auf den Codeansatz zurück. Wenn ich einmal in der Lage bin, neuen Code zu generieren über solche Sprachmodelle, dann fällt es mir natürlich auch leicht, gleich Entwickler Tests mit dazu zu generieren oder überhaupt Tests dazu zu generieren oder Dokumentation dazu zu kreieren. Zweite Denkrichtung und dritte Denkrichtung. Und das ist auch eine sehr vielversprechende. Wir haben Regulierung von KI Produkten, KI Produkte, KI Lösungen, Leistungen müssen gewisse Vorgaben entsprechen, müssen einer gewissen Risikoklasse entsprechen etc. pp. Und die Fähigkeit das zu testen, ob das gegebene KI Anwendung hat, ist auch noch eine Fähigkeit an sich.

[00:39:08.600] - Sprecher 2
Und da sehe ich auch großes Potenzial, sich sowas zu erschließen. Also in der Lage zu sein, eine gewisse Introspektion einer KI Anwendungen durchzuführen. Und allen erster Schritt dazu wäre, die allgegenwärtigen Benchmarks sozusagen wirklich hernehmen zu können. Also sagen wir mal im multimodalen Language Understanding, da ist es der Malu, das sind 57 Fragegebiete und die müsste ich einfach aus einer Test Datenbasis rausziehen. Und dann jage ich die gegen das neueste Sprachmodell und notiere die Testergebnisse oder Code Code Llama. Dann hole ich die. Die Fragestellung aus dem Human heraus ist ein Benchmark für Codierungsfragen. Arbeite automatisch diese Testfragen ab und notiere das Ergebnis. Oder heller Zweck oder Vino Grande oder 1000 andere Benchmarks, die in dem Umfeld existieren. Das heißt, diese Fähigkeit, Benchmarks automatisch gegen eine KI Anwendung laufen zu lassen und das Ergebnis zu notieren, würde ich als dritte Denkrichtung im Test Umfeld sehen.

[00:40:23.810] - Sprecher 3
Für mich jetzt auch noch mal eine gute Überleitung für ein anderes Thema, was ich gerne noch mit dir besprechen. Nämlich das Thema Regulierung, also nicht nur KI, zu testen, ob sie gut funktioniert und auch das Richtige macht, was wir uns vorher vorgenommen haben, sondern insgesamt natürlich aber auch darüber nachzudenken, Welche Risiken bestehen denn auch bei KI Anwendungen? Wir haben jetzt den Eindruck, kurz vor Abschluss wird ja noch heiß gerungen um Einzelheiten, wie stark wir regulieren wollen. Deutschland, Frankreich und Italien wollen sich gerade jetzt ein bisschen dort rausziehen, eine stärkere Regulierung anzunehmen, sondern sie sagen Nein, brauchen wir nicht. Wie stehst du demgegenüber? Können wir das regulieren und wofür? Und was sollten wir regulieren?

[00:41:03.120] - Sprecher 2
Also ich denke, es braucht eine Regulierung. Ich gehöre aber auch zu den Anhängern, wo weniger mehr ist. Und man muss, glaube ich auch akzeptieren, es gibt völlig verschiedene Ansätze zu zu regulieren. Wir brauchen nicht zu sagen, dass der risikobasierte Ansatz vielleicht der einzige ist. Und das ist ja genau der Denkweg, den die EU jetzt in ihrem Trilog eingefangen hat und hat sich darauf geeinigt. Ist auch valide, erstmal zu sagen Ja, wir wollen gewisse Anwendungsklassen, dort wollen wir

nicht, dass KI eingesetzt wird. Flächendeckende Gesichtserkennung zum Beispiel, Social Scoring zum Beispiel soll verboten sein und bleiben. Finde ich auch erst mal völlig richtig. Aber das wäre der risikobasierte Ansatz. Und dann geht die Pyramide nach unten weiter, bis man in so risikolose oder scheinbar risikolose Klassen kommt. Aber man kann natürlich auch anders regulieren. Man kann zuerst mal den Mensch in den Mittelpunkt stellen und sagen was sind die Human rights? Und wenn diese Human rights durch die KI nicht eingeschränkt werden, dann sollte man machen. Also es gibt völlig andere Philosophien zur Regulierung.

[00:42:14.400] - Sprecher 2
Und wenn man dann natürlich weltweit einen Flickenteppich an Regulierung haben, dann wird es natürlich auch schwer für die Produkte und. Und nicht umsonst sagt Gemini Na ja, so schnell können wir jetzt nicht kommen. Wir müssen erst mal alle folgen, so ein bisschen überprüfen. Also mein Fazit Man braucht Regulierung, man braucht aber auch Demut vor der Regulierung. Was kann ich wirklich regulieren, wenn ich es nicht testen kann? Wenn ich es nicht überprüfen kann, dann ist auch eine Regulierung recht fragwürdig, das muss man sagen. Und es gibt auch sehr viele Vereinfachte, die einfach sagen Ja, lasst uns doch vom Menschen her denken und darauf die dann die Regulierung sozusagen zuschneiden. Also es ist ein komplexes Gebiet. Wir dürfen nicht vergessen, wir haben OECD Regulierung, wir haben zwei verschiedene Ansätze in Europa. Wir haben die US Regulierung und die machen es natürlich schon erheblich komplex. Und über China und seine Leistungsfähigkeit in KI haben wir heute noch gar nicht gesprochen. Ich will es auch nicht vertiefen. Fakt ist, dass die auf der Ebene der Großen mitschwimmen können, was die Leistungsfähigkeit betrifft.

[00:43:23.850] - Sprecher 2
Hier zum Beispiel nur zu erwähnen ja, aber klar sind da Anwendungs wie Social Scoring und flächendeckende Gesichtserkennung sind in aller Munde und das auch schon und werden benutzt. Und das entspricht natürlich nicht unserem Wertesystem.

[00:43:39.750] - Sprecher 3
Jetzt haben wir aber auf der anderen Seite eine Entwicklung auch in den USA, mit der das Altman, der Gründer von OpenAI, sogar gekündigt hatte. Man vermutet wegen Cluster und Sicherheitsbedenken und die Chefentwickler mitgegangen sind, jetzt wieder zurückgekommen ist. Das nährt ja auch diese Verschwörungstheorien, was da gerade los ist und dass die Entwickler selbst und Entwicklerinnen selbst Angst haben von dem, was sie dort bauen und kreieren, nicht mal positiveres Storytelling auch hinbekommen, wo es uns helfen kann, in der Wissenschaft, in der Gesundheitstechnologie, Entwicklung in dem Bekämpfen des Klimawandels usw und diese Angst aus den Köpfen der Menschen herausbekommen.

[00:44:23.430] - Sprecher 2
Ja, also vielleicht eine zweifache Antwort Ich bin totaler Anhänger dieser optimistischen Betrachtung, wo, wo kann es uns helfen? Und sowohl in den Anregungen von Bill Gates, Du hast bis am Anfang darauf eingegangen ist diese positive Sichtweise Was können wir damit alles lösen in der Medizin, bei der Behandlung des Klimawandels? Und auch Dr. Alan Thompson, der wirklich hervorragende Statistiken zu dieser ganzen Fragestellung mit führt. Er sieht auch die zwei grundlegend positiven Sachen. Diese Technologie kann uns einen intelligenten Begleiter, also jedem Individuum, einen intelligenten Begleiter, einen Agenten an die Seite stellen, der uns hilft, der uns besser macht, der uns einfach weiterentwickelt, der uns in Situationen weiterhilft. Also wirklich der persönliche Begleiter schlechthin. Aber er sieht auch die Dimension. Ja, wir haben geopolitische Probleme, wir haben wirtschaftliche Probleme, wir haben den Klimawandel, wir brauchen neue. Energielösung. Und warum nicht auch mit dieser gewaltigen Intelligenz, die kein einzelner Mensch mehr lesen kann? Ich bin am Anfang darauf eingegangen. Warum nicht auch diese Intelligenz für die Lösung dieser Probleme?

[00:45:39.480] - Sprecher 2
Man könnte sagen einen gesellschaftlichen Agenten sozusagen einzusetzen, das sehe ich positiv. Und vielleicht noch ein Wort zu deiner Einführung mit Cuesta. Das war natürlich eine. Ja, im Nachgang würde man fast sagen lächerliche Rochade als Mann raus als zu Microsoft, als man wieder rein, weil sonst die gesamte Firma wahrscheinlich verschwunden wäre. Das konnte nicht im Interesse von Microsoft sein. Wenn ich 12 Milliarden investiere, will ich, dass die Firma performt und und nicht, dass

ich alle zu mir holen muss. Jedenfalls nicht im ersten Schritt. Ich habe. In den USA habe ich auch die Frage gestellt, wie ich jetzt dort war. Und es scheint eine Mischung zu sein aus persönlichen Eitelkeiten und persönlichen Befindlichkeiten, die jetzt zu der Lösung, wie wir sie jetzt haben, geführt haben. Ich sage mal Ende gut, alles gut. Ob Cuesta, das war ja eine Spekulation, die von Altmann aber selber angeheizt worden ist, wirklich einen entscheidenden Beitrag bei diesem Hin und Her geleistet hat, bleibt abzuwarten. Wenn dort jetzt nicht noch mal nach gewaschen wird oder scharf nachgelegt wird, kann man es glaube ich als Episode abtun.

[00:46:54.870] - Sprecher 2
Falls doch Und ein paar Indikationen gibt es ja, dass man dort insbesondere Reinforce und Learning wirklich auf ein Neues Level gebracht hat und damit die Schlussfolgerung Fähigkeit, die reasoning capability, dass die ein völlig neues Niveau erreicht haben. Und zwar hat man eben das Feedback für das Bestehende. Lernen gibt man nicht erst am Ende eines Riesenschritts, sondern gibt man für jeden Teil Teilabschnitt des Schlussfolgerns. Und wenn man das macht, dann verbessert sich der gesamte Prozess erheblich. Und dazu gibt es auch ein Paper von OpenAI. Das kann man nachlesen, das kann man sich ziehen und das scheint zumindest ein Bestandteil von Cuesta zu sein. Ob noch mehr dahinter hängt, das wird die Zukunft zeigen. Im Moment sind alle Geister wieder beruhigt und dann schauen wir mal, wie es dort weitergeht.

[00:47:49.170] - Sprecher 3
Ja, apropos Zukunft Ich glaube, wir können auch mal so in die Schlussgerade einbiegen unseres heutigen Podcasts. Ich würde gerne noch mal ein Thema mit dir besprechen, nämlich Du bist ja auch Professor an der Universität. Wie gehst du mit diesem ganzen Thema Fälschungen durch KI um? Nur mal so ganz persönlich gefragt ist das, weil das ist ja jetzt ein konkreter Anwendungsfall, den viele fasziniert und den sie auch benutzen können.

[00:48:10.110] - Sprecher 2
Ja, in der Tat, wenn man die Studenten auch einfach fragt, das ist eines der häufig eingesetzten Tools, sozusagen noch nicht mal im Sinne einer Fälschung, sondern zur Beschleunigung und Verbesserung etc.. Und ich, ich glaube, das ist auch erst mal der richtige Weg. Das ist, ich glaube, der Fluch jeder Technologie, dass es zum Wohl eingesetzt werden kann, dass es aber auch zum wer eingesetzt werden kann. Und ich glaube, da müssen sich die Menschen an die eigene Nase fassen, um da die richtige Schlussfolgerung daraus zu ziehen.

[00:48:49.170] - Sprecher 3
Also regulieren, glaubst du, wird nicht funktionieren?

[00:48:51.540] - Sprecher 2
Nicht? Nicht wirklich. Alles, was ich nicht durchsetzen kann, überprüfen kann, habe ich wenig Chancen.

[00:48:58.440] - Sprecher 3
Oder wir lassen schriftliche Arbeiten in Zukunft und es gibt nur noch mündliche Prüfungen.

[00:49:01.830] - Sprecher 2
Na, ich denke schon. Zu solchen Auswüchsen. Warum Auswuchs? Also zu solchen Regelungen wird es schon kommen. Man muss das anders angehen. Ich erinnere mich an eine Geschichte meiner Schwester, die Lehrerin ist und in der Pandemie Zeit Hat Sie von Ihren Schülern verlangt, dass die praktisch mit verbundenen Augen mit ihr sprachen? Sozusagen, damit keine anderen Informationen im Hintergrund verwendet werden konnten. Also zu so irgendwelchen ähnlichen Sachen wird es kommen. Ja, die Gefahr durch Fakes ist groß, die ist immanent. Aber ich sage immer man man will dann die die Technologie dafür verantwortlich machen und man erkennt nicht, dass es Menschen sind, die die Fakes produzieren und dann einsetzen.

[00:49:46.110] - Sprecher 3
Also dann würde ich dich natürlich noch gerne fragen, wenn wir hoffentlich nächstes Jahr hier wieder sitzen, vielleicht sogar schon früher. Mal gucken, wie schnell sich das alles entwickelt. Über was

reden wir denn dann bei diesem Thema?

[00:49:57.720] - Sprecher 2
Ja, ich denke, wir werden dann ein Bild haben. Wer ist denn an der Spitze? Sozusagen. Also wie verhält sich GPT fünf zu Gemini 2:00 null? Welche? Welchen Status hat Olympus dazu erreicht, dass wir werden so die drei Spitzenkandidaten, die ich sehe. Wir werden weitere Entwicklung und Verstetigung im OpenSource Bereich sehen und ich denke, wir werden, was die Anwendung betrifft dort. Wird es ein paar Killerapplikationen geben, die für jedes Unternehmen unverzichtbar sind. Und das wird sich bis dahin herausgebildet haben. Das kann durchaus sehr, sehr schnell gehen und darüber können wir dann sprechen. Also mein. Ich denke, Gemini hat eine gute Chance, sehr weit nach vorne zu kommen. Mit 1000 man 1000 Personen kommt man natürlich schon ein paar Schritte weit. Im OpenSource Bereich sind die Llama Derivate natürlich sehr stark, aber gibt es viele andere interessante Entwicklung. Ich hatte schon mal auf die arabischen Entwicklungen Falcon, da wird jetzt ein Falcon 180 B Modell kommen, also auch sehr, sehr spannend. Also es bleibt dynamisch und wie gesagt, im Unternehmensbereich diese Auskunftssysteme als natürlich sprachliche Interfaces, das wird kommen.

[00:51:22.520] - Sprecher 2
Und ja, MMS hat auch spannende Angebote heute schon so eine Art GPT in einer geschützten Umgebung, wo wo Kunden wirklich sinnvoll damit arbeiten können. Also wir werden sehen, wie das Raum greift.

[00:51:38.480] - Sprecher 3
Es bleibt auf jeden Fall spannend und wir freuen uns auch jetzt schon drauf. Auf was ich mich auch freue, ist noch mal von dir zu hören, wie du denn solche Sprachmodelle benutzt. Persönlich setzt du das jetzt für Weihnachtsgrüße ein oder hast du uns auch vielleicht ein paar Tipps mitgebracht, was man in welcher Kombination am besten suchen sollte und wie man das ganze Thema angeht, Da gibt es eine Menge an Videos, mittlerweile ja auch im Internet, die so einem Tipps geben.

[00:52:03.260] - Sprecher 2
Genau. Also die neue Disziplin des PromptEngineering wurde ja schon häufig beschworen. Wie baue ich, wie, wie kitzelig? Meisten die die tollen Eigenschaften der Sprachmodelle heraus. Ich habe mich eher auf die andere Seite auch mal begeben. Während des DLCs unsere Innovations Konferenz hatte ich ein kleines Quiz erarbeitet und da habe ich mich der Benchmarks bedient. Also das finde ich faszinierend. Durch die Benchmarks mal durchzugehen und dann Kandidaten zu finden, die aus diesen Benchmarks kommen oder sehr ähnlich sind und dann trotzdem die großen Sprachmodelle trotz ihrer Allmacht noch mal aufs Glatteis zu führen. Es gibt inzwischen auch eine schöne Website, wo eben Aufgaben gestellt werden, die ein Mensch mit überschaubarem Aufwand lösen kann und wo alle GPS dieser Welt bisher kläglich scheitern. Interessant Website kann ich auch hier schreiben. Die Ursachen und das Grundprinzip, das ist das Überraschende dort ist eigentlich immer das gleiche. Man verklausuliert die Frage in ein Rätsel und das Rätsel ist aber relativ einfach für einen Menschen zu lösen und dann kann er auch die richtige Antwort geben und aber diese indirekt so'n Rätsel und daraus dann eine Schlussfolgerung zu ziehen, das scheint den großen Sprachmodellen immer noch ein bisschen schwer zu fallen.

[00:53:28.310] - Sprecher 2
Also solche Sachen machen natürlich.

[00:53:30.320] - Sprecher 3
Genau das werden wir auch in einem Jahr überprüfen, ob es diese Website dann noch gibt oder oder ob die dann quasi nicht mehr up to date ist. Frank herzlichen Dank, dass du heute hier zu Gast warst.

[00:53:40.520] - Sprecher 2
Ja, danke auch in die Runde und gerne wieder.

[00:53:43.040] - Sprecher 3
Ja, vielen Dank, dass Sie sich auch heute wieder Zeit für unseren Podcast genommen haben. Und in

den Shownotes finden Sie noch weitere Links zu den heutigen Themen und wir haben da auch noch weitere Infos für Sie bereitgehalten. Und wenn Sie in Zukunft keine weitere Folge verpassen wollen, dann würden wir uns sehr freuen, wenn Sie uns bei Spotify oder bei Apple Podcast abonnieren. Bis dahin alles Gute.