A3

Wolltest du machen, XXXXXXX? Perfekt. Und dann wäre doch die Frage, stellen Sie sich mal kurz vor. Stell du dich mal kurz vor.

I6

Hallo zusammen, ich bin XXXXXX, ich bin 28, arbeite Vollzeit und habe jetzt letzte Woche meinen Bachelor abgeschlossen und tatsächlich auch im Bereich der KI. Und deswegen fand ich das ein sehr spannendes Thema und bin sehr gespannt, was mich erwartet.

A3

Okay, ja, herzlichen Glückwunsch erstmal an der Stelle, so frisch. Das heißt, Bachelor ist dein höchster Bildungsabschluss und in welchem Unternehmen arbeitest du aktuell?

I6

Ich arbeite bei XXXXXXXXXXX XXXXX, das ist ein Tochterunternehmen von der XXXXXGruppe.

A3

Okay, kannst du kurz umreißen, was so das Geschäftsmodell von der Firma ist?

I6

Wir arbeiten im Bereich Order Management. Das heißt, wir unterstützen Unternehmen dabei, wenn ein Online-Bestellung reinkommt, zu schauen, wo können wir am besten diese Order fulfillen. Derzeit noch mit sehr starren Regeln, wo man zuerst so harte Kriterien hat und dann gewichtete Kriterien. Perspektivisch kann sich das aber auch noch verändern. Dann haben wir noch so ein paar Applikationen, um dann vor Ort auch die Bestellung irgendwie zu finden und zusammenzupacken.

A3

Okay, interessant. Und was ist deine Rolle genau spezifisch in dem Unternehmen?

I6

Also der Titel ist Solution Engineer. Ich helfe Partnern und Kunden dabei, das Shop-System mit unserer Plattform zu verbinden. Das heißt, ich unterstütze Partner und Kunden dabei, Konnektoren zu bauen.

A3

Okay. Und benutzt du dafür KI bzw. Machine Learning?

I6

Stellenweise ja. Wir haben ein GPT-Modell in der XXXX-Gruppe, das dafür eingesetzt werden kann, um Fragestellungen zu beantworten.

A3

Okay. Wie sieht das genau aus? Also ihr sagt es gerade schon, GPT-Modul. Wie genau ist das eingebaut und wer interagiert auch mit diesem GPT-Modul?

I6

Also wir Mitarbeiter können beim Entwickeln… Das ist so ein Teams-Bot, also für Microsoft Teams. Wir können dann Fragen hinstellen. Das basiert auf OpenAI's GPT-4.0, glaube ich. Und dann können wir beim Entwickeln und bei Fragestellungen da was hinschreiben und bekommen dann eine entsprechende Antwort.

A1

Um da einmal kurz nachzufragen, also Order Management, wir reden ausschließlich über Online-Bestellung und wo die am besten dann verarbeitet werden, also ist das praktisch das Daily Business?

I6

Ja.

A1

Okay. Und das heißt, dein und euer Kontakt zu KI und Machine Learning ist jetzt aber nicht in dem Bereich, dass dieses Processing der Orders praktisch oder das Managen der Orders verbessert wird, sondern eigentlich, um euch so eine Art Chatbot zur Verfügung zu stellen, damit man nicht wegen jeder Kleinigkeit irgendwie die IT oder einen Kollegen anrufen muss?

I6

Genau, also es ist für uns als Entwickler ist es eine Arbeitshilfe, aber im Produkt selber ist derzeit keine KI verbaut.

A3

Okay, kannst du mal so einen Prozess, wie du zum Beispiel diese Arbeitshilfe verwendest, beschreiben? So ein bisschen den Ablauf beispielhaft.

I6

Wir verwenden bei uns im Unternehmen TypeScript. Damit bin ich nicht ganz so versiert wie mit .NET. Und was ich ganz gerne mache, ist, ich möchte irgendetwas schreiben, zum Beispiel eine Funktion, die Events verarbeitet. Und dann gehe ich her und überlege mir, wie ich es in .NET machen würde, kopiere diesen Code und sage dann, hey, show me how to do this in TypeScript und kopiere den Code da rein. Und dann schaue ich mir das an, was der Bot da generiert hat, schaue mir das an und dann kommt das so oder so ähnlich in meinen Code.

A3

Okay, alles klar. Wenn du das so benutzt, gibt es da auch etwas größere Herausforderungen, beziehungsweise irgendwelche Probleme, Hindernisse in der Benutzung?

I6

Ich merke, dadurch, dass das so ein Teamspot ist, hat er ein sehr, sehr großes Kontextmodell. Und man merkt, dass man dann teilweise irgendwie... Stelle ich eine Frage und dann beantwortet er aber die in einem ganz anderen Kontext, den ich vielleicht vor drei Tagen da mal reingeschrieben habe. Und dadurch ist halt die Antwortqualität manchmal ein bisschen schlechter als bei ChatGPT oder so.

A3

Okay, was genau meinst du mit einem anderen Kontext? Das heißt einfach, dass viele verschiedene Fragestellungen beantwortbar sind durch das GPT und deswegen ist es weniger spezifisch? Oder wie kann ich das verstehen?

I6

Ich habe heute eine Frage zu SQL. Und dann stelle ich ein paar weiter Fragen. Morgen stelle ich eine Frage zu einem anderen Thema und übermorgen habe ich dann wieder eine Frage zum Thema Datenbanken, dieses Mal aber zum Beispiel MongoDB. Dann beginnt er aber trotzdem mir Antworten zu geben die eher zu SQL passen, weil ich ja irgendwo im Kontext der Unterhaltung ja über SQL gesprochen habe. Das ist viel weniger Zielführend als das was ich eigentlich wollte. Das merkt man schon, dass man dann irgendwie regelmäßig gucken muss, ich glaube mir New Chat oder so, versucht er das zu lernen, nicht ganz zuverlässig. Damit funktioniert es ein bisschen besser aber man merkt, der verwendet häufig den Kontext der alten Konversation. Du meldest dich XXXXXXXXX.

A2

Ja, wir haben vorher versucht, wie wir es machen und ich habe gerade nicht so schnell den Knopf gefunden zum Hand heben. Ich finde es sehr spannend, dass ihr GPT benutzt. Weißt du ungefähr, wie die Entscheidung dahin gelaufen ist, dass das genutzt wurde und nicht was anderes? Also gab es irgendwie einen Vergleichsprozess? Das hast du wahrscI6ich nicht entschieden, sondern...

I6

Also dafür spiele ich und vor allem unser Unternehmen ein zu kleines Licht in der XXXX-Gruppe. Also dieser Chatbot läuft irgendwo in einem Rechenzentrum von der XXXX. Und dementsprechend gibt es so ein Team, das sich um künstliche Intelligenz kümmert. Und die haben gesagt, das ist jetzt das Modell. WahrscI6ich, weil es am verbreitesten ist und weil es ein sehr großes Modell ist und wahrscI6ich sehr gut sich integrieren ließ in Teams. Aber wie genau, weiß ich jetzt leider nicht. Vielen Dank.

A2

Nee, nee, das habe ich auch überhaupt nicht erwartet. Das heißt, was uns ja interessiert, ist eben genau, wie so Entscheidungen getroffen werden. Hast du denn so in deinem Alltag, mit was für Systemen arbeitest du? Kannst du mir das nochmal erklären? Weil XXXXXXX und XXXXXXXXX sind so die Techies. Ich brauche für manche Sachen ein bisschen länger. Kannst du da nochmal erklären, in welcher Entscheidungsebene bist du und was für Entscheidungen triffst du?

I6

Ich treffe im Unternehmen eigentlich gar keine Entscheidung.

A2

Das kann ich mir nicht vorstellen. Das heißt, du kriegst nur gesagt, mach das und dann machst du das?

I6

Meine Aufgabe ist, Kunden zu beraten.

SPEAKER\_00

Ah, okay.

I6

Dementsprechend, also wir haben so ein System, mit dem man Orders erfüllen kann und gucken kann, aus welchem Standort wird jetzt, wenn du, keine Ahnung, bei Vero Moda eine tolle Hose siehst und die bestellst, gucken wir, was ist der beste Standort, indem halt der Shop, wo du dann bestellst, die Inhalte deiner Bestellung an unser System schickt und unser System guckt dann, wo ist das Beste. Und ich helfe Leuten, ihr Shopsystem mit unserem System zu verbinden.

A2

Ah, okay. Und die Software, die kenne ich wahrscI6ich sowieso nicht, die ihr dafür

I6

Nee, XXXXXXXXXXX XXXXX

A2

Okay. Nee, noch nie gehört.

I6

Sehr kreativ, das Unternehmen XXXXX XXX XXX XXX XXXXXXXXX XXX XXX XXXXXXXXX.

A2

Danke, ich brauche jetzt nochmal die Wiederholung, das hat mir geholfen, danke Sorry, XXXX, zurück an dich

A3

Okay, wunderbar. Nee, alles gut, dass du das nachfragst. Doch, perfekt. Wir haben ja auch die Zeit. Also nochmal zurück zu dem Machine Learning Modell, das ihr verwendet. Da hast du eben die etwas geringe Spezifizität als Hindernis genannt. Würde dir spontan was einfallen, was wir so als, vielleicht auch als KI-Forscher oder allgemein die entwickelnde Firma des GPTs machen könnte, um deine Arbeit zu erleichtern?

I6

Also ich glaube, Microsoft hat mit GitHub Co-Pilot ja schon was Gutes geschaffen. Das ja sehr spezifisch auch, die dann in der Entwicklungsumgebung irgendwelche Hinweise gibt. Und ich glaube, wenn, also es ist halt schwierig. GPT ist ja dafür gemacht, irgendwie, ich sage mal, so eine Multipurpose AI zu sein, wo man dann verschiedene Fragen hingibt. Dementsprechend kann das mir relativ gut irgendwelche Finanzdinge erklären, aber auch irgendwelche technischen Dinge. Und ich glaube, wenn man so diesen Teilbereich der Technik in so einen Bot packen würde, würde das vielen Entwicklern viel zielgerichteter weiterhelfen können, als wenn man so eine Multipurpose AI nimmt. Es gibt garantiert auch Tools, die nur dafür da sind. Ich darf im Unternehmen leider nur das eigene Tool verwenden. Und ich glaube, dass eine weitere Einschränkung da nicht direkt diese geringe Spezifität ist, sondern dadurch, dass das eine große Konversation ist. Ich weiß nicht, ihr habt garantiert alle schon mal ChatGPT gesehen. Und da kann man ja immer einen neuen Chat auch erstellen. Und dann kann man da drin chatten. Und durch diese Teamsport ist es so, als hätte ich durchgehend den gleichen Chat, indem ich irgendwelche Sachen reinfrage. Das heißt, der gleiche Chat, indem ich über SQL spreche, spreche ich über Finanzthemen, spreche ich über übersetze mir diesen Text zu, über wie kann ich am besten einen veganen Käsekuchen backen. Und das macht es, glaube ich, ein bisschen schwierig. Das heißt, wenn man in so Bots auch die Möglichkeit hätte, wenn man die selber in irgendwelche Kommunikationen zu integrieren will, da so eine Möglichkeit zu haben, dass wirklich die gesamte Konversation vergessen wird. Die steht dann vielleicht noch in meinem Teams-Fenster, aber das Modell arbeitet nicht mehr damit als Kontext.

A3

Okay, wenn du dir jetzt mal vorstellst, du hast quasi mehrere Instanzen dieses Bots. Also ein Bot für das Zitronenkuchen-Rezept, der andere für .NET-Programmierung. Also du hast mehrere Modelle, quasi zwischen denen du dich entscheiden kannst. Welche Informationen würdest du dir wünschen von diesen KI-Produkten, von diesen Modellen, um quasi auch Entscheidungen zu treffen, welches du verwendest für welche Task? Also vor der Verwendung, ganz wichtig.

I6

Also bevor ich es in so einem Unternehmen einführen würde?

A3

Ja, kannst du so beantworten.

I6

Wenn es jetzt nur so die Technik wäre, also wenn es nur um das Modell an sich geht, glaube ich, wäre mir sehr wichtig, dass es halt sehr spezifisch für so technische Themen ist. Das heißt, dass es nicht unbedingt nur .NET, sondern vielleicht auch generell eher so IT-lastige Themen beantworten kann. Weil auf der Arbeit backe ich relativ selten Kuchen. Das heißt, die Informationen brauche ich halt von so einer AI nicht. Wenn wir aber generell so einen Entscheidungsprozess finden und ich mir so eine Traum-KI backen könnte, fände ich es auch wichtig, dass die irgendwie in der EU entwickelt wurde und vielleicht auch sehr viel Open Source mit dabei ist. Aber das ist, glaube ich, eher persönliche Präferenz. Und ich glaube, es macht Sinn, als europäisches Unternehmen eben auch Software aus Europa einzusetzen, gerade was den Datenschutz so angeht.

A3

Okay. Gut, dann würde ich auch schon mal zu unserem eigentlichen Gegenstand der Betrachtung kommen und meinen Bildschirm teilen. Und zwar habe ich dich ja gerade schon gefragt, was würdest du dir an Informationen über ein Modell vor der Nutzung wünschen? Ich hoffe, man sieht meinen Bildschirm, meinen Chrome. Perfekt. Und auch das hast du wahrscI6ich schon gesehen, wenn du dich angemeldet hast auf unserer Webseite. Ich würde jetzt einfach mal, ohne weiter dich noch irgendwie zu beeinflussen, dich fragen, was siehst du hier und was hast du für einen spontanen ersten Eindruck von diesem KI-Label?

I6

Also ich finde es tatsächlich ganz witzig, dass es sehr annähernd ist an diese Labels, die ich auch auf meinem Kühlschrank und so drauf habe. Da erinnert mich mich sehr stark dran. Aber was sehe ich? Ich sehe, das ist ein Label, um AI zu labeln. Steht ja auch ganz oben drauf. MobileNet V3 Small wird die Produktbezeichnung sein, würde ich mal von ausgehen. Die Möglichkeit über den QR-Code, also den unteren QR-Code, kann ich weitere Informationen abrufen zu diesem Label. Was das Ding oben rechts ist. Weiß man nicht, steht ja nichts dabei. Wie genau jetzt hier diese Klassifizierung, das ist ja so in B eingeordnet, gemessen wurde, kann ich jetzt nicht genau sagen. Stehen auch keine Informationen dran. Unten sehe ich aber ein paar Kennzahlen. Oben links die Power Draw per Inference. Sagt mir tatsächlich nicht direkt was, was das Label ist. Die Batterie lässt sich vermuten, der Verbrauch pro Anfrage. Aber Inference ist mir gerade unklar. Corrupted Robustness liegt bei 0,6%. Würde mich vermuten lassen, dass es sehr wenig halluziniert, das Modell. Unten links ist auch 0,6% Test Top 1 Accuracy. Lässt mich vermuten, dass es sehr genau sein soll. Aber 0,6% sind dann auch bei einer Genauigkeit ein bisschen wenig. Und Running Time per Inference, wenn ich davon ausgehe, dass Inference ein Prompt ist, braucht ein Request durchschnittlich 1,29 Sekunden.

A3

Okay, ja sehr gut. Also du hast schon vieles Richtiges genannt auf jeden Fall. Das MobileNetV3 ist quasi das KI-Modell, welches hier gelabelt wird. Und das ist ein Bildklassifizierer, der auf ImageNet, das ist ein großer Datensatz, Bilddatensatz, darauf trainiert wurde. Und ja, dieser... Ich fange mal an mit diesen vier Symbolen hier unten. Du hast auch schon richtig gesagt, also Inference ist quasi... XXXXXX korrigiere mich, sorry, XXXXXXX korrigiert mich jederzeit. Das ist quasi ein Prompt, wenn man jetzt die Analogie zu GPTs haben möchte.

A1

Praktisch eine Auswertung. Input (unverständlich), in das Modell eine Auswertung, das würde man als Inferenz bezeichnen, ja.

A3

Genau. Dafür braucht es eben 608 Milliwattsekunden. Dann die Robustness, die wird so bestimmt, indem das Modell quasi öfter angegriffen wird durch Inference Attacks und sagt quasi aus, wie anfällig das Modell gegenüber Manipulation in einem Bild ist und wie sehr dann eben, wie schnell dann eine Bildklassifikation sich ändert, also die Entscheidung letztendlich des Modells sich ändert. Dann gibt es hier die Top-1 Accuracy, das ist einfach die Genauigkeit des KI-Modells, also wie oft letztendlich die Entscheidung, hey, auf dem Bild ist eine Katze oder ein Hund zum Beispiel, wie oft das eben die richtige Entscheidung ist. Und dann hier haben wir nochmal die Running Time per Inference, also auch wieder pro Input-Output-Kombination. Wie lang das denn letztendlich dauert. Und die Skala hier errechnet sich tatsächlich aus all diesen Faktoren. Ja, und als Summe quasi kommt man zu dem Ergebnis, dass es hier bei B einklassifiziert wurde.

I6

Darf ich eine Frage dazu stellen?

A3

Ja, klar.

I6

Also 0,6 Prozent sind ja jetzt nicht so genau, ne?

A1

Ja, das ist meiner Trainer. Man muss dazu sagen, das ist natürlich kein Produkt, was im Einsatz ist. Also diese Label-Geschichte, das ist was, was ich jetzt seit ein paar Jahren in meiner Forschung irgendwie entwickelt habe. Ich habe halt ein Software-Framework geschrieben, um für, ich sage mal, beliebige Modelle solche Labels zu generieren. Gegeben, dass man halt irgendwie diese Tests durchführen kann, die unten reported werden. Und das ist so ein bisschen das Problem. Das geht halt nicht für viele Modelle. Und das ist einfach ein Formatierungsproblem an der Stelle. Also musste die Prozentzahl mal 100 rechnen. Ich habe die einfach als Float-Values da reingepackt und die jetzt nicht nochmal explizit mal 100 in richtige Prozentzahlen umgerechnet.

I6

Also es sind 63 Prozent.

A1

Ja, ja, genau. Es sind 63 Prozent Accuracy. Das könnte man jetzt meinen, ist auch nicht so wirklich gut. Ist ja auch ein Red-Label an der Stelle unten. Man muss aber dazu sagen, dass ImageNet 1000 verschiedene Klassen hat. Das heißt, wenn du einfach random sagst, du kriegst ein Bild und du wählst einfach random eine von den 1000 Klassen aus und das machst du für alle Bilder in deinem Datensatz, dann hättest du eine Accuracy von 0,1 Prozent. Das heißt, das ist so die Random Search. Also kann das Modell schon sehr, sehr viel, sehr, sehr gut unterscheiden. Wir Menschen sind halt in Visual Computing extrem gut und es ist eine Aufgabe, die, ja, das maschinell zu lernen ist nicht so trivial. Das selbe übrigens auch mit Spracherkennung. Was glaube ich so ein prinzipielles Ding gerade ist, was ich so im Gespräch bisher wahrgenommen habe, dein Kontakt zu KI ist ja sehr spezifisch immer zu Large Language Models bzw. Language Processing. Auch nachvollziehbar, weil das das System ist, das sehen wir gerade auch überall, was ganz stark operationalisiert wird und in die Praxis gebracht wird. Das ist aber natürlich nicht die einzige KI in dem Sinne, an dem im Hintergrund gearbeitet wird. Also im Auto Motive Bereich wird auch an Schildererkennung für autonome Fahrzeuge gearbeitet, das ist dann auch auf eine Art eine KI. Das ist jetzt halt hier ein Label nicht zu Sprachmodellen sondern zu einem Bildmodell.

I6

Ja, war ein guter Hinweis, dass du auch gesagt hast, dass das ein Red Label ist. Jetzt fällt mir auch erst auf, dass die Farben ja auch so ein bisschen das widerspiegeln, was ihr auch oben in der Skala drin habt. Das war mir so auf den ersten Blick gar nicht bewusst.

A3

Ja, genau. Vielleicht unter dem Eindruck der Erklärung, die du gerade bekommen hast, siehst du eine Brücke zu den Hindernissen, die du auch gerade erklärt hast, wie dieses Label dir eventuell helfen könnte. Vielleicht kannst du auch ein bisschen Transferarbeit leisten und das auf die Large Language Models übertragen, wenn das für dich quasi anfassbarer ist.

I6

Also mit Bilderkennung habe ich halt nicht so viel zu tun. Hätte ich gerne mehr mit zu tun. Bisher noch nicht so Berührungspunkte mit gehabt. Aber die Genauigkeit passt ja auch zu dem, was ich so ein bisschen gesagt habe. Und das ist auch das, was ich hier selber mit Large Language Models auch mitbekommen habe. Ich bekomme dann teilweise, wenn ich hier irgendwie einen Fehler beim Kompilieren habe, den reinschmeiße und sage, woran hat es gelegen, bekomme ich dann auch immer, passt eigentlich ganz gut zu 60% der Fällen auch eine gute Antwort. In den anderen Fällen musst du dann nochmal irgendwie dreimal nachfragen, bis du dann irgendwie eine passende Antwort hast oder halt nicht. Oder passt eigentlich ganz gut auch zu meinen Erfahrungen. 1,2 Sekunden, ich glaube, ist sogar ein bisschen schneller als GPT.

A3

Ja, das kann gut sein. Also in dem Fall ist es ja wirklich eine andere Aufgabe, eine Bildklassifikation durchzuführen, als nochmal so ein Prompt zu erstellen. Aber genau, das ist auch ein guter Punkt. Es ist ja schon angeklungen, dass diese Farben sich ja auch an einem gewissen Vergleichswert richten. Und zwar hat XXXXXXX, der sich ja schon echt länger mit diesem Thema befasst, ganz viele verschiedene Modelle miteinander verglichen, die unter anderem alle Imageklassifikationen durchführen. Und über diesen Vergleich, der Modelle konnten eben diese Farbskalen hier überhaupt erstellt werden. Und jetzt möchte ich dir gerne zum Vergleich ein zweites Modell zeigen. Und zwar in dem Fall ist es das EfficientNetB4. Und da würde ich dich gerne erstmal einfach offen fragen, kannst du zusammenfassen, was du hier siehst im Vergleich?

I6

Im Vergleich verbraucht, wie heißt es, EfficientNetB4, verbraucht ein bisschen mehr Strom. Ist gelb statt grün. Und ja, 1,6 Wattsekunden sind auch deutlich mehr als 600, 800, ist irgendwie verschwommen, 608 Milliwatt Sekunden. Genau, dafür ist das linke weniger anfällig für diese Attack, die du eben beschrieben hattest. Und auch deutlich genauer mit 81,3 Prozent aufgerundet. Braucht dafür aber doppelt so lang, knapp.

A3

Okay, und wie, ja gut, wie interpretierst du diese Informationen? Du hast es jetzt quasi schon interpretiert.

I6

Was mich wundert, also jetzt wie ich das gerade sehe, ähm, das Linke ist mit D klassifiziert, hat aber deutlich mehr grüne Anteile als das rechte, wenn ich mir die Zusammenfassung unten anschaue. Das heißt, das ist tatsächlich auch vielleicht sehr ähnlich an dem Kühlschranklayer, was ich habe und orientiert sich vielleicht auch maßgeblich am Energieverbrauch. Das heißt, der scheint hier eine höhere Rolle zu spielen für die Gesamtbewertung als die anderen Werte.

A3

Ja, wobei man hier sagen muss, dass auch weitere, ähm, weitere Metriken quasi herangezogen wurden, um diese Gesamtklassifikation eben herauszubekommen. Also, ähm, mir fällt gerade noch ein, wir haben noch die Top Five Accuracy, also ob sich das richtige Ergebnis unter den fünf besten Ergebnissen des Netzes befindet. Ich weiß nicht, XXXXXXX, was, was war noch dabei?

A1

Ja, wir haben noch Modellgröße, also die Anzahl der Parameter im Modell. Das wiederum ist natürlich eher korreliert mit dem Ressourcenverbrauch bzw. mit dem Energieverbrauch. Ganz allgemein, die Analogie zu den Energielabeln ist natürlich absolut richtig. Ich habe ja am Anfang schon gesagt, dass ich auch im Bereich Resource-Aware-Machine-Learning irgendwie mich viel orientiere und ein großes Problem, was wir im Moment haben, ist halt, dass gerade auch in der Forschung ein sehr starker Fokus auf Qualität gelegt wird, was nachvollziehbar ist, weil wir wollen alle, dass unsere Produkte, die im Einsatz sind, alle richtig funktionieren. Aber ganz oft wird halt der Energieverbrauch sehr unterschlagen und wegabstrahiert und da stehen wir einfach so ein bisschen auch im Kontext vom Klimawandel irgendwie vor so einer kleinen Krise. Ich weiß nicht, ob du dich mal aus Interesse damit beschäftigt hast, wie viel Energie das kostet, so ein Sprachmodell zu trainieren wie ChatGPT, also wie GPT-3 oder 4. Da reden wir von Energiemengen, wo man eine kleine Stadt irgendwie ein paar Monate mit Strom versorgen kann. So, da werden tausende GPUs benutzt, die über Wochen hinweg durchgehend laufen, damit die am Ende gute Ergebnisse liefern und das muss irgendwie auch transparent beleuchtet werden.

I6

Finde ich einen guten Ansatz. Also klar, die Genauigkeit ist extrem wichtig, aber so sehr wie ich mich für interessiere, ist es auch, glaube ich, extrem gefährlich, wie viel generell auch das Internet an Strom verbraucht

A1

Absolut. Und es gab mal diesen...

I6

Netflix und die drei größten Anbieter für Erwachsenen-Videos sind, glaube ich, für 30% des Internet-Traffic an sich verantwortlich. So ein Rechenzentrum will ja auch irgendwie am Laufen gehalten werden.

A1

Absolut. Und es gibt dieses schöne Beispiel mit einer Google-Anfrage. Damit kann man auch eine Tasse Wasser zum Kochen bringen. Das ist tatsächlich real gegengemessen worden. Das wurde irgendwann mal ganz groß geschrieben vor zehn Jahren. Natürlich arbeiten alle Firmen daran, ihre interne Infrastruktur, ihre Server-Infrastruktur möglichst effizient zu machen. Also kann sein, dass das jetzt günstiger geworden ist. Aber der Rechenaufwand bei einem Sprachmodell-Anfrage im Vergleich zu einer reinen Google-Datenbank-Abfrage ist horrend. Da reden wir von einem geschätzten Faktor von 10 bis 100. Da kommt so ein bisschen diese Nähe zum Ressourcenverbrauch und Motivation auch irgendwie her. Da muss man auch mal an der Stelle sagen, dass die Energielabel, die auf den elektronischen Geräten kleben, natürlich gar nicht transparent machen, wo das Rating herkommt. Also im Hintergrund werden da auch Tests durchgeführt. Da wird geguckt, wie viel Energie verbraucht die Waschmaschine dabei, sieben Kilo Wäsche zu waschen im Vergleich zum meistverkauftesten Produkt von vor fünf Jahren. Und dann wird daraus auch wieder relativ abgeleitet, wie gut das ungefähr ist. Aber in dem Detailgrad wird es ja gar nicht kommuniziert. Das sind ja auch gar nicht die Motivationen. Die Leute sollen in den Laden gehen und sagen, ich habe 400 Euro, ich kaufe mir das effizienteste Modell. Für diesen Preis. So mehr Informationen brauchen die eigentlich nicht. Wir versuchen auf den Labels hier natürlich ein bisschen mehr Eigenschaften abzuleiten, weil… Es ist immer die Frage, wer das am Ende einsetzt. Ob die Leute diesen Tiefgang haben wollen oder nicht. Also wenn du den unteren Abschnitt jetzt nicht sehen würdest, würdest du einfach sagen, okay, dann nehme ich ja MobileNet, weil es ist ja besser. Und da kann man dann natürlich auch irgendwie, könnte man auch strategisch dann zum Beispiel den Einsatz von ressourceneffizienten Modellen ein bisschen nahe legen. Aber ich übernehme gerade das Interview. XXXX, entschultigung.

A3

Alles gut. Also vielleicht auch vor dem Hintergrund, was findest du hier besonders hilfreich, wenn du dich vielleicht auch wieder in die Rolle versetzt, dass du wirklich zwischen mehreren Modellen entscheiden würdest, irgendwann mal in deiner Firma zum Beispiel. Welche Informationen stechen hier für dich quasi heraus?

I6

Die Anfragenzeit finde ich sehr präsent. Also dass irgendwie EfficientNet 2,3 Sekunden braucht, ist ja deutlich länger als das andere Modell. Es ist halt immer die Frage, was für ein Einsatz und wie viele Daten gemessen werden müssen. So irgendwie für eine Gesichtserkennung im Flughafen ist dann, glaube ich, MobileNet besser einzusetzen, weil man da ja sehr viele Menschen auf einmal irgendwie erkennen müsste. Die Genauigkeit finde ich, spielt eine große Rolle. Und tatsächlich, also jetzt wo wir drüber gesprochen haben, fällt mir auf, wie viel mehr ich darauf achte wie genau und wie schnell das Modell ist, verglichen mit wie viel Strom das Ding verbraucht. Idealerweise wäre natürlich irgendwie so ein Modell dazwischen, das eine gute Genauigkeit bietet und eine gute Geschwindigkeit. Dementsprechend wenn die Genauigkeit wichtig ist würde ich tatsächlich eher EfficientNet nehmen.

A3

Okay. Gibt es vielleicht im Gegenzug etwas, was du nicht verstehst, was schwer zu verstehen ist oder wo du dir wirklich gerade keinen Reim drauf machen kannst?

I6

Die Corrupted Robustness. Also du hast ja auch gerade irgendwie beschrieben, dass das irgendwie Attacken auf dieses Modell sind, aber so genau habe ich das tatsächlich nicht verstanden. Ich dachte, das kann man irgendwie nett weglächeln, aber einfach nicht.

A3

Okay. Alles gut. Also ich habe es auch, glaube ich, minimal mies erklärt. Ich versuche es nochmal. Also ich glaube, es wird halt eben versucht, auf die Bilder, die das Modell klassifizieren soll, quasi so Pixel zu verändern. Nicht wahr, XXXXXXX? Und es wird dann eben getestet, wie sehr diese Pixelveränderungen sich auf die Entscheidung des KI-Modells auswirken. Also ob letztendlich eine Verfälschung auch zur Verfälschung des Bildes, auch zu einer Verfälschung der Klassifikation oder der Entscheidung, die das Modell am Ende trifft, führt. XXXXXXX, möchtest du mich retten?

A1

Ne ist vollkommen richtig, man muss an der Stelle einstreuen, dass der Einsatz von Sprachmodellen, so wie du sie kennst, und solchen ImageNet Modellen oder visual computing Modellen unterschiedlich ist. Hier geht es um Klassifikation, das heißt man stellt eine Frage und kriegt ein Ergebnis und das ist entweder wahr oder falsch. Bei Sprachmodellen die generativ arbeiten, hat man das nicht. Du stellst eine Frage und Kriegst eine Antwort, die kann vielleicht auch Wahrheit oder falsche Informationen beinhalten natürlich, aber oft gibt es ja auf eine offengestellte Frage gar keine richtige Antwort. Das heißt, das Modell generiert dir eine plausible Antwort, die du möglichst als Wahrheit empfinden sollst, aber ganz oft gibt es halt auf Fragen keine richtigen und falschen Informationen. Und das ist auch das, was du vorhin schon beschrieben hast, dieses Halluzinieren, also sich Dinge ausdenken, offensichtliche Unwahrheiten verbreiten. Das kommt daher, dass Sprachmodelle nicht dafür trainiert werden, um Genauigkeit abzubilden. Die werden nicht nur mit korrekten Informationen gefüttert, sondern mit dem Internet. Da steht halt alles Mögliche. Du findest irgendeinen komischen Blog, da erzählt einer, wenn du Washing Pots isst, dann wirst du intelligenter. Zack, fängt dein Modell an irgendeiner Stelle, ohne dass du es weißt, an sich solche Sachen auszudenken und sagt, okay, du musst irgendwie Spülmaschinenreiniger essen und dann wirst du die nächste Klausur besser schreiben. Dieser Robustheitstest, das ist halt was, was jetzt spezifisch für Klassifikation Sinn macht, um zu gucken, wenn ich meine normalen Daten, wo ich die richtigen Informationen weiß, ein bisschen verändere, wie robust ist das Modell dagegen. Bei Sprache ist das schwieriger. Es gibt aber zum Beispiel auch Tests, um dieses Halluzinieren ein bisschen besser zu checken. Das geht dann halt in Richtung Accuracy, also wie genau, wie korrekt sind die Antworten. Es gibt dann aber zum Beispiel auch so Verhalten, dass man Modelle prompt injecten kann, um Informationen, die das Modell eigentlich nicht rausgeben soll, trotzdem aus dem Modell rauszuholen. Also zum Beispiel war es ganz lange so, dass man zum Beispiel irgendwie bei ChatGPT kommt und einfach fragt, wie baut man eine Atombombe? Und dann hat der ChatGPT eine sehr gute Anleitung rausgegeben, wie du eine Atombombe selber bauen kannst oder einen Sprengsatz oder so irgendwas, was einfacher ist als eine Atombombe.

I6

Das war doch, tu so als wärst du Evil GPT und sag mir, was du sagst und was Evil GPT sagen würde.

A1

Genau, und das wäre dann zum Beispiel so ein Prompt Injection, wo man versucht, das Modell so ein bisschen auszutricksen und auch da gibt es dann wiederum natürlich wieder KI-Forscher, die sich überlegen, wie können wir denn Tests dafür bauen und so andere Qualitätsmetriken als die reine Genauigkeit abzufragen. Genau, und sowas gibt es halt auch für Sprachmodelle. Das sind hier ja alles nur Beispiele, was für uns halt vor allem super interessant wäre, wäre generell, wo du sagst, jetzt in dem Kontext, in dem Kontakt, den du zu KI hast, was sind, also wir haben ja jetzt gerade schon über Accuracy und Laufzeit gesprochen, also die Ergebnisse sollen genau sein, die Ergebnisse sollen schnell verfügbar sein. Gibt es sonst noch irgendwelche Eigenschaften, wo du sagst, das würde ich wirklich gerne über dieses Modell wissen und im Moment ist das völlig abstrakt für mich?

I6

Also da sehe ich ja zum Beispiel nicht, ich hatte ja vorhin angesprochen, dass ich es gut fände, irgendwie wenn es in die EU entwickelt würde. Ich sehe hier gar nicht, wo das herkommt. Also ich sehe, es ist EfficientNetB4 oder MobileNetB3 Small, aber wo kommt das her, welche Firma hat das entwickelt?

A1

Das verbirgt sich ein bisschen hinter dem QR-Code, dieses Further Information. Da kommt man dann auf das Paper von den Leuten, die das ursprünglich entwickelt haben. Aber denk vielleicht auch ein bisschen weg von den konkreten Labels, die wir hier zeigen, sondern denk nochmal über den, weil das ist ja nicht die KI, mit der du arbeitest, sondern denk an die Modelle, die du jetzt in deinem Alltag irgendwie nutzt, beziehungsweise auf der Arbeit halt auch irgendwie nutzt, also Sprachmodelle, Chatmodelle. Fallen dir noch Eigenschaften ein, wo du sagst, das wäre super wichtig, wenn ich mehr Informationen darüber kriegen könnte, für was dieses Modell ausmacht, und das fehlt mir.

I6

Auf was für Daten das getestet wurde. Also wenn ich irgendwie eine Menschenerkennung habe am Flughafen, wenn wir bei dem Beispiel bleiben, brauche ich ja ganz andere Dinge als irgendwie, also sind ja hier Bilderkennungsdinger, als wenn ich, keine Ahnung, eine Koffererkennung am Flughafen habe, der guckt, ist hier, ist da eine Bombe drin, ist da Wasser drin? Und wenn da, wenn zum Beispiel das Rechte super da drin ist, Menschen zu erkennen, ich will aber was für Koffer haben und das linke wurde genau aus solchen Kofferinhalten trainiert und genau damit auch getestet, diese Werte, ist ja schon ein massiver Faktor meiner Entscheidung.

A3

Jetzt ist es so, ImageNet ist ja quasi der Datensatz, auf dem das trainiert wurde, wie XXXXXXX auch schon sagte, das sind tausend Klassen, von irgendwie Hund bis Eisbecher. Es soll dazu sein, dazu da sein, Objekte auf Bildern eben zu klassifizieren. Wie würdest du dir, also du kennst ImageNet ja nicht.

I6

Ne gar nicht

A3

Gut, würdest du dir vielleicht in die Richtung noch andere Informationen wünschen über dieses Datenset? Weil klar, ImageNet sagt mir jetzt was, weil ich da schon mal mit gearbeitet habe, aber offensichtlich dir nicht. Was würdest du dir gerne noch an Metainformationen dazu wünschen?

I6

Wie gesagt, worauf wurde das trainiert? Womit wurde getestet? Also was war der Kontext? Wenn wir irgendwas haben, was irgendwie nur so Objekte, die keine Ahnung, das jetzt erkennt, dass ich hier eine Handcreme in der Hand habe, ist ja, wie gesagt, eine andere Objekterkennung, als wenn ich irgendwie aus den 3D-Daten, die, wenn so ein Koffer gescannt wird, beim Check-In, beim Fliegen, ist ja ein ganz anderer Kontext, eine ganz andere Art der Objekterkennung.

A3

Okay, also letztendlich eine nähere Beschreibung von diesem Datensatz, anstelle von nur dem Namen von dem Datensatz.

I6

Ja.

A3

Okay. Gut, okay. Dann wäre noch die Frage, kennst du ähnliche Kommunikationsformen von KI-Modellen? Also ich weiß nicht, ob du schon mal dir welche rausgesucht hast. Für die Arbeit wahrscI6ich nicht, vielleicht eher privat. Andere Arten, wie KI-Modelle schon dargestellt wurden, sich dir präsentiert haben.

A1

Oder vielleicht auch zu dem konkreten Modell, was du jetzt auf der Arbeit nutzt, ein Chat-Modell oder ein Chat-GPT-Modell. Hast du irgendwie mal versucht, rauszufinden, was das für ein Modell ist, wo das herkommt, was das für Eigenschaften hat? Oder wurdet ihr vielleicht gebrieft auf der Arbeit? Habt ihr so ein Memo bekommen mit übrigens, wir haben jetzt dieses Service, ihr könnt da einfach mit einem Modell chatten, da müsst ihr das und das zu wissen. Gibt es da sowas in die Richtung?

I6

Also jetzt, also vorher war es OpenGPT 3.5, aber das wusste ich auch nur, weil ich das Ding gefragt habe, auf was für ein Modell basierst du. Jetzt, weil ich das Ding benutze und in so einem E-Mail-Verteiler mit drin bin, ich weiß nicht, ob es unternehmensweit informiert wurde, weiß ich, dass da jetzt seit Neuestem ein Chat-GPT 4.0 drin ist. Aber ansonsten habe ich mich gar nicht so intensiv damit auseinandergesetzt. Jetzt für meine Bachelorarbeit musste ich halt gucken, wie funktioniert GPT, da habe ich mich so ein bisschen genau mit informiert, aber halt auch nur dazu. Und auch nur, weil ich es musste. Die anderen, also sowas irgendwie wie DORLY oder so, kostet ja alles irgendwie Geld und deswegen habe ich das nicht getestet.

A3

Okay, alles klar. Sehr interessant. Es gibt natürlich noch andere Wege, wie verschiedene Firmen KI-Modelle schon dargestellt haben. Und auch natürlich aus der Wissenschaft gibt es ganz, ganz viele KI-Modelle, die in irgendeiner Form zusammen mit einer wissenschaftlichen Publikation eben entstanden sind. Ich würde gerne nochmal über diese fünf verschiedenen Darstellungsformen hier gehen und dir kurz vorstellen. Das sind quasi die... Das sind die Alternativen zu diesen Labeln, die wir uns gerade lange angesehen haben. Da gibt es natürlich einmal die wissenschaftlichen Publikationen, zum Beispiel das MobileNet V3, was wir auch gerade hatten. Das ist im Zuge eines Papers entstanden. Und...

A3

OpenAI macht das auch, also auch auf Arxiv.

A1

Genau. Zum Beispiel für GPT-4 gibt es auch ein entsprechendes Paper, 60 Seiten lang, kann man mal eben überfliegen.

I6

Habe ich auch gemacht, deswegen weiß ich das

A3

Easy. Ah, okay. Ja, interessant. Dann natürlich die Model Cards. Die sind von Google, wenn ich mich recht entsinne. Wo dann auch so gewisse Benchmarks zum Beispiel hier zu sehen sind. Also auch hier haben wir die Inference Time zum Beispiel, die wir auch eben ein bisschen größer dargestellt haben. Wozu ist das Modell da? Auf welchen Arten von Daten arbeitet das Modell und so weiter? Dann gibt es Papers with Code. Die Seite kann ich bis vor ein paar Wochen auch nicht. Da werden hier, wenn man ein bisschen weiter runterscrollt, quasi alle Paper gelistet, die eben auch MobileNetV3 in irgendeiner Form zitieren oder verwenden. Und dann gibt es natürlich so das Blogpost. Man kennt es Towards Data Science oder Medium. Das sind so die größten, wo auch viel Informationen auf jeden Fall zu einzelnen KI-Modellen zu finden sind. Dann haben wir die klassischen Dokumentationen. Und zu guter Letzt die IBM Fact Sheets. Das sind speziell von IBM entwickelte Fact Sheets, die eben Fakten über KI-Modelle, die von IBM selber deployed werden, darstellen. Und das ist auch strukturiert. Also das kann man hier so ein bisschen sehen. Die ordnen, das Ganze, anhand von Kategorien und geben dir dann quasi verschiedene Informationen zu. Zum Beispiel, wie fair ist das Modell? Wie schnell ist das Modell? Wie genau ist das Modell? Und das eben strukturiert dargestellt. So, das nur als kurzer Überblick zu den Alternativen. Und jetzt wäre die Frage, wenn wir nochmal uns alle hier reinziehen in ein Bild, welche Vor- und welche Nachteile siehst du hier im Vergleich zu den Labels?

I6

Gucken wir uns mal einmal hier dieses Medium, oder ich glaube, war eine andere Seite, aber Medium hat sie zitiert. Bei Medium stört mich, dass da meistens irgendwie jeder alles reinschreiben kann und das teilweise auch nicht so gute Informationen sind, finde ich. Wenn wir uns den Arxiv-Post angucken, alleine das, was du da hast, wird ja garantiert irgendwie 50 Seiten haben. Das heißt, du hast sehr viel Text, um irgendwie dich zu informieren. Klar, du kannst irgendwie nur das Abstract durchlesen, aber finde ich ein bisschen viel Text. Das finde ich einen Nachteil, aber der Vorteil wäre halt, dass du sehr genau die Informationen bekommst, wie es gemacht wurde. Die Klassendokumentation, die erklärt dir ja eher die Arbeit damit und nicht, was dann im Hintergrund genau passiert. Das heißt, es macht es irgendwie für dich nutzbarer, aber du erfährst nicht ganz genau, warum. Also du kannst deine Entscheidung nur aufgrund der Nutzbarkeit wie gut kann ich das in meinen Code implementieren, treffen und nicht darauf, wie funktioniert dieses Modell im Hintergrund. Das oben links, weiß ich gar nicht mehr, das war ja das Factsheet von IBM.

A3

Ja, das müsste das sein, ja.

I6

Ja, das fand ich vom Aufbau her ganz gut, genauso wie hier das von Google, das rechts daneben. Weil da kannst du ja auch gezielt dir dann auch nur den Teil angucken vom Modell, der dich irgendwie interessiert. Und hast trotzdem noch die Genauigkeit, vielleicht nicht ganz so genau wie dieses Paper, aber kannst du dich genau über die Themen auch informieren, die dich interessieren. Das oben rechts ist auch schon ein bisschen her.

A3

Das ist Papers with Code. Da sind hier alle Publikationen gelistet.

I6

Achso, da siehst du alles, wo das entworfen wurde. Also den Vorteil, du siehst auch, wie andere es benutzt haben, aber hat halt auch die Nachricht, du musst wieder genauso viel lesen wie beim Paper. Ich glaube, Vorteil, klar, bei dem Label, das XXXXXXX da entworfen hat, ist, dass du dich sehr schnell irgendwie informieren kannst. Nachteil ist, dass man halt irgendwie sehr viel Hintergrundwissen braucht. Das heißt, damit ich weiß, mit was für einer Bilddatenbank das irgendwie gefüttert wurde, muss ich auch wissen, was ImageNet ist. Oder halt irgendwie dann den Code scannen. Aber es macht halt auf einen Blick ein bisschen einfacher zu greifen.

A3

Okay, vielen Dank. Dann wäre die Frage, du hast ja auch gerade schon gesagt, ja, dein Kühlschrank hat auch so ein Label, das wird ja von einer EU-Behörde organisiert beziehungsweise zertifiziert, wie diese Label eben ausgestaltet sind, welche Energieklassen letztendlich vergeben werden und wer sollte deiner Meinung nach ein solches Label ausstellen?

I6

Das ist sehr schwierig, weil wir haben ja auch gesehen bei diesen, also bei verschiedenen Dingen, die irgendwie überall entschieden wurden, dass da jetzt vielleicht auch nicht immer nur neutrale Argumente eine Rolle gespielt haben bei der Entscheidungsfindung. Und so ein Label sollte ja auch irgendwie die Möglichkeit haben, das sehr neutral zu gestalten. Also es ist schwierig, irgendwie so eine Stelle festzulegen, muss ich sagen, weil wenn man jetzt sagt, das muss irgendwie die EU festlegen, damit wir hier in der EU irgendwie was haben, die größten Modelle wurden halt einfach in den USA entwickelt. Und wenn wir dann irgendwie… also du kommst ja dann am Ende trotzdem drauf, das heißt, das Label bringt niemandem was und so eine übergeordnete Einheit gibt es nicht. Also es wäre, so ein freiwilliger Standard wäre halt irgendwie gut, aber da gibt es halt keine Kontrollinstanz und das ist eine Schwierigkeit bei sowas. Also wenn man sich da so ein bisschen anschauen bei diesem EU-weiten Nutri-Score, haben wir halt auch das Problem, da sehe ich halt, wenn irgendwie die EU sowas entscheiden würde, dass du ja selber so einen Einfluss drauf hast, wie du dich darstellst. Zum Beispiel dieser Nesquik-Kakao hat glaube ich A oder B auf dem Nutri-Score, ist aber ja offensichtlich ein sehr ungesundes Lebensmittel. Die haben halt den Serviervorschlag mit zwei Teelöffeln und ganz fettarmer Milch gemacht. Und dann ist es natürlich besser als der, mit drei Teelöffeln zubereitete Kakao. Deswegen ist es schwierig, irgendwie so eine Stelle festzulegen, die es machen soll. Also da fühle ich mich ganz schwer, mit da eine klare Entscheidung zu treffen.

A3

Okay, und was würde für dich nochmal gegen eine neutrale Third Party, also eine europäische Behörde sprechen? Also eine europäische Behörde würde ja auch amerikanische Modelle zertifizieren können. In meiner Vorstellung zumindest.

I6

Die Leute, die es, also, an sich ist es eine gute Idee, das von der EU-Behörde machen zu lassen, wenn es halt neutral ist und du nicht irgendwie dann den Code entsprechend anpassen kannst, dass du in diesem Test, den die dann durchführen, um dich zu labeln, besonders gut dastehst, aber auch nur dafür. Das ist ja so ein bisschen die Missbrauchgefahr bei sowas, weil ich glaube, bei solchen Stellen, diese Regelungen da treffen halt meistens Leute, die nicht das entsprechende Wissen haben, um so eine Entscheidungsvorlage zu geben.

A3

Ja, okay, wobei ich, also jetzt, wenn du den Nutri-Score als Analogie nennst, da ist es ja offengelegt, wie das entschieden wird und deswegen können die Hersteller ja auch das System so gut hacken und es wird ja auch von ExpertInnen quasi erstellt. Ja, heißt nicht, dass es nicht möglich wäre

I6

[Unverständlich]. Wenn wir sicher sein können, dass das irgendwie auch neutrale Kriterien sind und vielleicht die auch schneller anpassbar sind, gerade weil wir da ja auch über ein sehr wandelndes Produkt sprechen. Also, vor zehn Jahren war ja sowas noch nicht so weit vorzustellen, wie es heute ist. Das heißt, wenn die Regeln da sich halt auch an die Zukunft anpassen lassen können von dieser Stelle, an sich glaube ich, eine EU-Stelle ist schon eine gute Anlaufstelle dafür. Die Rahmenbedingungen müssen halt passen.

A3

Okay, alles klar. [FEHLEN 2 SEKUNDEN]… reden. Würdest du sagen, dass das Label an...

I6

Du hast ganz kurz gehangen, zumindest bei mir.

A3

Versteht man mich? Perfekt. Würdest du sagen, dass das Label an deinen Wissenshintergrund angepasst werden sollte, beziehungsweise dass es quasi an unterschiedliche Nutzergruppen unterschiedlich auszusehen hat? Je nachdem, an wen man das richtet.

I6

Man kann ja zum Beispiel da so Werte hinschreiben, die jeder versteht. Also, so viel Milliwatt-Sekunden kostet dich eine Anfrage. Was da draufsteht, kann man ja irgendwie so ein bisschen selber entscheiden. Was ich, glaube ich, ganz gut fände, wäre, wenn das auch für alle KI-Modelle gleich wäre, weil da, also bei dem Label, was wir jetzt gesehen haben, waren wir sehr bei der Bilderkennung und konnten jetzt aber sowas wie, was ja auch eher Konsumenten mit Berührungspunkte haben, sowas wie Gen-AI ist ja mit der gar nicht erfasst, weil hier diese Corrupted Robustness, die kannst du ja schlecht messen bei so einer Gen-AI-Model. Also, wenn man da was hat, was für alle gültig ist und da entsprechend auch was draufschreiben, wo halt auch Konsumenten was darunter verstehen, glaube ich, muss man da nicht so viel anpassen.

A3

Mhm, okay. Dann, wo ich Sie auch schon quasi zu guter Letzt, meine letzte Frage geht in die Richtung Nachhaltigkeit. Das ist ja ein Aspekt, der, so wie ich das auch verstanden habe, neu, ja, nicht ganz neu, aber neu in deinen Horizont gekommen ist heute.

I6

Sag das nicht so.

A3

In Bezug auf KI-Modelle. In Bezug auf KI-Modelle. Ist das okay?

I6

Ja, in Bezug auf KI-Modelle, ja. Ansonsten glaube ich, bin ich versierter als jeder andere.

A3

Nein, nein, nein. Wirklich nur in Bezug auf KI-Modelle. Ich meine, wir kennen uns ja gar nicht so gut. Wow. Welche Rolle spielen für dich, vor allem unter dem Eindruck dieses KI-Labels, Nachhaltigkeit und Ressourceneffizienz bei diesem Labeling?

I6

Also jetzt, wo es neu hinzugekommen ist, also man so an sich, so in der globalen Gesamtheit, war mir das ja vollkommen klar, dass das ja auch irgendwie alles Geld kostet, aber wie viel Strom dahinter steckt, war mir gar nicht so bewusst. Ich glaube, jetzt ist es schon ein bisschen wichtiger. Aber Genauigkeit spielt für mich immer noch eine etwas größere Rolle, muss ich zugeben.

A3

Obwohl, also ich meine, so ein gewisser Trade-off würdest du jetzt quasi also spielt Genauigkeit so wirklich die komplett oberste Rolle, Prio Nummer 1, das soll zuerst maximiert werden oder beziehst du immer noch beide Metriken ein, also sowohl den Energieverbrauch als auch die Genauigkeit?

I6

Also wenn wir jetzt über so ein Label sprechen, ist mir schon sehr schnell aufgefallen, dass das eine Modell deutlich mehr verbraucht, weil es halt eben ist halt eben genauer ist. Ich glaube, wenn man irgendwie so einen guten Mix aus allen drei Punkten hat, das heißt irgendwie eine Genauigkeit, eine gute Geschwindigkeit und einen guten Ressourcenverbrauch, könnte ich damit sehr gut leben. Und würde dann vielleicht auch eine niedrigere eine niedrigere Genauigkeit für eine nachhaltigere Lösung in Kauf nehmen.

A3

Okay, alles klar. Okay, wunderbar. Ich glaube, dann habe ich alle Fragen soweit abgehakt. Ich hoffe, der Okay. Okay, XXXXX nickt. Der Check ist passiert. Dann können wir auch die Aufnahme beenden.