// WHITEPAPER

KI-BASIERTE CONTENT-GENERIERUNG IN DER BILDUNGSBRANCHE

TITANOM TECHNOLOGIES

Grundbegriffe

LLMs, Content-Pipeline & Embeddings

Large Language Models (LLMs) sind große neuronale Netzwerke, die in erster Linie zum Erstellen von Texten genutzt werden. Sie können durch das Training mit großen Datenmengen logische und zusammenhängende Sätze bilden. Allerdings erlauben sie nach der Programmierung durch den Hersteller keine individuellen Anpassungen mehr und verfügen nicht über keinen veränderbaren Wissensspeicher.
In der Anwendung fungieren sie ähnlich einem sachkundigen Verkäufer: Sie haben ein tiefgehendes allgemeines Wissen, können dieses aber nicht immer effektiv in spezifischen Situationen nutzen und es mangelt ihnen oft an speziellem Fachwissen in bestimmten Bereichen. Momentan sind LLMs in ihren Fähigkeiten begrenzt und können keine neuen Informationen aufnehmen oder ihr Verhalten anpassen. Die Informationen jenseits eines Eingabeprompts bleiben statisch. Trotzdem ist zu erwarten, dass wir in der nahen Zukunft erhebliche Entwicklungen sehen werden, die fortschrittlichere Lernmethoden zulassen.

Die **Content-Pipeline** kann eine Abteilung, ein Softwareprogramm oder ein Prozess sein, der für die Erzeugung von Text- und Bildinhalten in einem Unternehmen zuständig ist. Dieses System hat immer klare Eingabe- (Input) und Ausgabeformate (Output):

- **Input:** Hierbei handelt es sich um spezifische Anforderungen an den zu erstellenden Inhalt, wie Themengebiet, die zu entwickelnden Fähigkeiten bei Übungsaufgaben oder Ideen für die Darstellung bei Einführungsinhalten. Darüber hinaus wird auch allgemeines didaktisches Wissen und Know-how für die Erstellung qualitativer Inhalte genutzt.

- **Output:** Dies ist der erstellte Inhalt, der als Grundlage für weitere Bearbeitungsschritte dient, bevor er dem Endnutzer gezeigt wird. Es ist wichtig zu beachten, dass dieser Inhalt noch nicht in seiner endgültigen Form vorliegt, sondern weiterbearbeitet und gestaltet wird.

Zusammenfassend ist die Content-Pipeline ein integriertes System, das eine organisierte und effektive Erstellung von Inhalten ermöglicht, die dann für die nächste Stufe der Bearbeitung und Nutzung bereit sind.

Embeddings bieten eine vektorbasierte Darstellung von Texten im multidimensionalen Raum. Die räumliche Nähe dieser Vektoren zeigt die Ähnlichkeit der Texte an: je näher sie sind, desto ähnlicher sind die Texte, und je weiter sie voneinander entfernt sind, desto unterschiedlicher sind sie. Diese Transformation in Vektoren wird durch neuronale Netze, die als Embedder bezeichnet werden, durchgeführt.

Einstieg

In der jüngsten Zeit haben Large Language Models (LLMs), insbesondere ChatGPT, erhebliche Innovationen in der Content-Erstellungsbranche angestoßen. Obwohl es schon vor ChatGPT verschiedene LLMs gab, ermöglichte die GPT-3 Generation mit ihren verbesserten generativen Fähigkeiten völlig neue Anwendungen, die bisher nicht realisierbar waren.
Stand Mitte 2023 können LLMs, vor allem in kreativen Bereichen wie der Erstellung von Bild- oder Audioinhalten, wo qualitative Kriterien weniger streng sind, bereits ein Niveau erreichen, das mit menschlicher Erstellung mithalten kann. Jedoch können sie menschliche Content-Ersteller in textbasierten Bereichen, besonders bei hohen Qualitätsanforderungen und komplexen Richtlinien, nicht vollständig ersetzen.
Es ist daher wichtig, LLMs nicht als kompletten Ersatz, sondern als nützliche Ergänzung in der Automatisierung von Content-Pipelines zu sehen. Sie fungieren vor allem als Beschleuniger, die die Umsetzung einfacher Content-Erstellungsprojekte ermöglichen, welche bisher aufgrund von Zeit- und Kosteneinschränkungen vernachlässigt wurden. Dies eröffnet neue Marktchancen, die durch die bisherigen Beschränkungen in der menschlichen Content-Produktion nicht realisierbar waren.

Chancen

Prozessoptimierung

Unternehmen können ihre Mitarbeiter effektiv unterstützen, indem sie innovative KI-Tools zur Optimierung von Routineaufgaben bereitstellen. Diese Neuausrichtung der traditionellen Methoden erfordert oft eine Anpassung der internen Abläufe, um eine nahtlose Integration zu ermöglichen. Dadurch können die Kosten pro Produkt reduziert werden, was langfristig zu günstigeren Preisen für contentbasierte Produkte führen wird. Dieser Trend bietet Unternehmen die Chance, durch frühzeitige preiswerte Angebote mehr Marktanteile zu gewinnen und somit Umsatz und Gewinn zu erhöhen.

Vollautomatisierung von Routine-Content

Gleichzeitig zeichnet sich eine Entwicklung in Richtung der Vollautomatisierung der Erstellung von Routine-Content ab, die neue Einsatzmöglichkeiten schafft. Die Live-Erstellung von Aufgaben und Inhalten im Produktionssystem kann eine bisher nicht dagewesene Anpassung an die spezifischen

Bedürfnisse der Nutzer ermöglichen. Eine erweiterte Vielfalt bei den Übungsaufgaben erlaubt es, die individuellen Bedürfnisse der Lernenden genauer zu adressieren. Dies kann beispielsweise durch die Förderung einer Fähigkeit innerhalb eines speziell gewählten Themenbereichs oder durch die individuelle Auswahl der Unterrichtsmethode erreicht werden.

ChatBots

Darüber hinaus entwickeln sich ChatBots, die sehr genau auf Nutzereingaben und -probleme eingehen können, ohne sich auf vordefinierte Template-Antworten zu verlassen. Diese Bots könnten komplexe didaktische Konzepte begreifen und dem Nutzer im Zusammenhang mit seiner Fragestellung kompetent erläutern, gestützt auf festgelegte Quellen. Dieser Fortschritt harmoniert mit dem aktuellen Trend, GPT-basierte First-Level Supports in verschiedenen Wirtschaftsbereichen zu integrieren. Der Bildungssektor unterscheidet sich jedoch dadurch, dass er die Möglichkeit bietet, vorhandene didaktische Ansätze fachkundig zu vermitteln, anstatt einfach statische Informationen in natürliche Sprache umzuwandeln.

Adaptives Feedback

Die Bildungsbranche steht tiefgreifenden Neuerungen im Bereich des adaptiven Feedbacks für Lernende. KI-Modelle können bestimmte Fehlermuster erkennen und den Schülerinnen und Schülern auf fundierte Weise in natürlicher Sprache helfen, die richtige Lösung zu finden. Früher war die Fehlerkorrektur eine Routineaufgabe der Lehrer im Offline-Kontext, aber in Lernsoftware war sie wegen Zeitverzögerungen und der Unmöglichkeit, sofortiges Feedback zu geben, nicht umsetzbar. Jetzt erlauben Large Language Models (LLMs) die Automatisierung dieser Aufgabe, was eine neue Ära der sofortigen Unterstützung für Lernende einläutet. Diese Technologie könnte den bisherigen erheblichen Aufwand beim Erstellen statischer Feedbacks für jede mögliche Antwort ersetzen, wodurch ein noch nie dagewesenes Maß an Personalisierung und Effizienz erreicht werden kann.

Herausforderungen

Die Einführung von KI-Werkzeugen im Bildungsbereich birgt verschiedene Herausforderungen, die gemeistert werden müssen. Erstens besteht das Problem der mangelnden Möglichkeit, Quellen anzugeben. Die Large Language Models (LLMs) können manchmal falsche Informationen erzeugen, die nicht auf bestätigte Quellen zurückgeführt werden können. Das stellt ein Zuverlässigkeitsproblem dar, besonders in einem Bildungskontext, wo die Informationen zuverlässig und zitierbar sein müssen.
Zweitens ist die Qualität der generierten Inhalte oft ein Problem, da sie strukturell falsch sein können. Es ist notwendig, fortgeschrittene Qualitätsbewertungsmethoden zu entwickeln und zuverlässige Testprozesse zu etablieren, um Fehler zu minimieren und zu verhindern, dass sie den Nutzer negativ beeinflussen.
Drittens ist der Übergang von einem Prototyp zu einem marktreifen Produkt eine weitere Herausforderung. Der Entwicklungsprozess kann neue Probleme verursachen oder alte wieder aufleben lassen, was die Handhabung kompliziert macht. Dies unterscheidet sich von der herkömmlichen Softwareentwicklung und kann dazu führen, dass einige gute Ideen nur im Stadium des Prototypen bleiben.
Sowohl Entwickler als auch Pädagogen müssen diese Herausforderungen bewusst und strategisch angehen, um eine erfolgreiche Integration von KI-Tools im Bildungssektor zu ermöglichen.

Lösungen

Didaktischer ChatBot

Eine Möglichkeit, die durch aktuelle GPT-Modelle leicht umgesetzt werden kann, ist der „didaktische ChatBot“. Dieses Werkzeug hilft Lernenden, fundierte und mit Quellen belegte Antworten auf ihre Fragen zu bekommen, was das Lernen deutlich verbessert.

Die Architektur eines didaktischen ChatBots ist einfach gehalten und lässt sich daher leicht in verschiedene Systeme integrieren.
Im Herzen des ChatBots befindet sich der Vektorspeicher, eine Art Datenbank, die die Informationen aller verwendeten Quellen speichert. Bei einer Nutzeranfrage sucht das System nach den passendsten Quellen, indem es die Ähnlichkeit Embeddings der gespeicherten Quellen mit dem Embedding der Frage vergleicht. Um die Genauigkeit dieses Prozesses zu gewährleisten, werden während der Entwicklung spezielle Tests durchgeführt, die prüfen, ob der Bot die richtigen Quellen auf Basis spezieller Fragen identifizieren kann.
Außerdem stellt der ChatBot die Antworten individuell zusammen, basierend auf festgelegten Lernanforderungen. Dabei benutzt das System spezielle Lehrinformationen und Hinweise, die gemeinsam mit der Frage und den Quellen in das LLM eingegeben werden, um sinnvolle und lehrreiche Antworten zu geben. So kann beispielsweise definiert werden, ob der Bot nur Probleme analysieren, Hilfestellungen geben oder die Lösungen erklären soll.
Insgesamt stellt diese Technik eine zuverlässige und auf individuellen Content basierende Lehrhilfe dar, die eine bedeutende Ressource im Lernumfeld bietet.

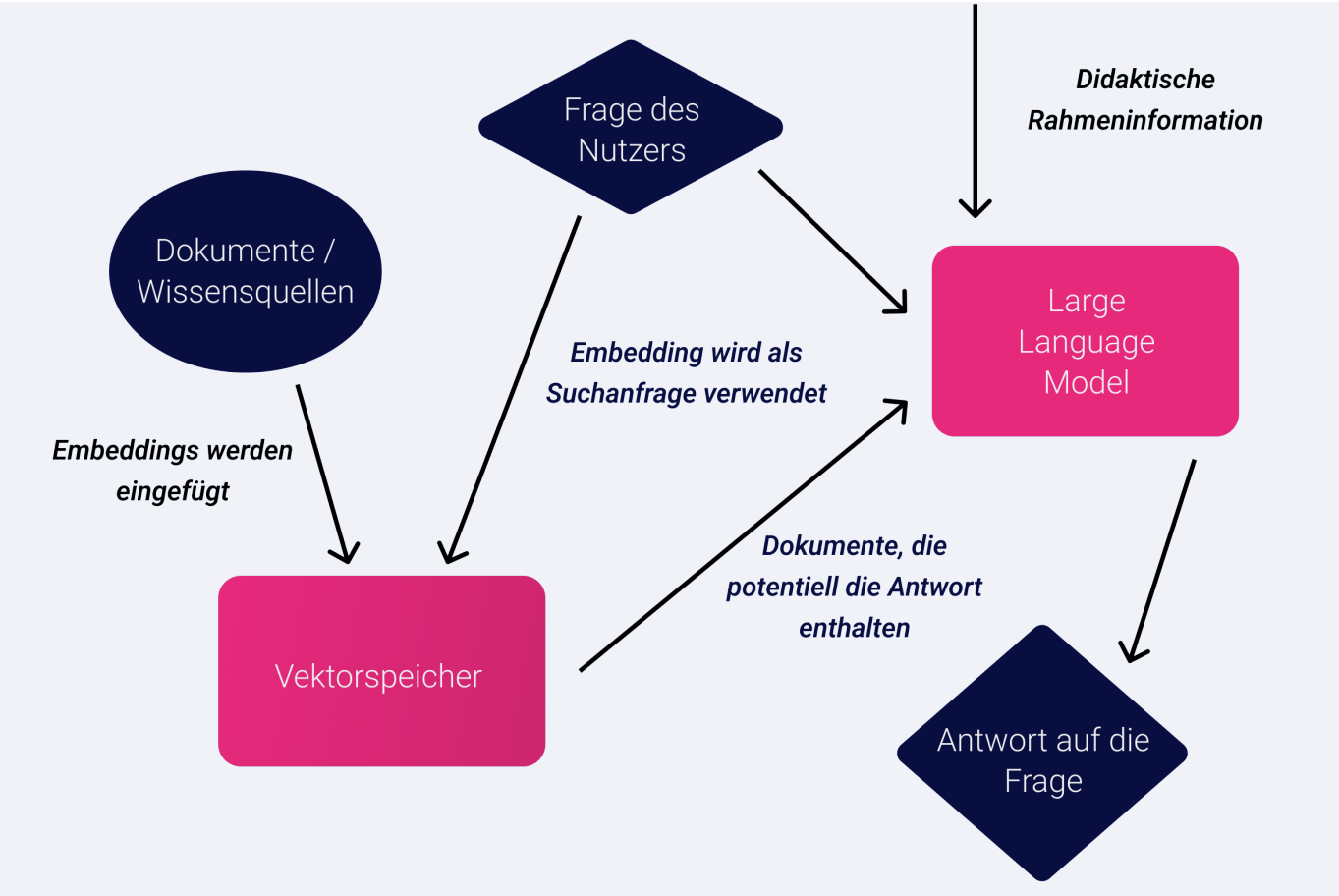

Abb. 1: Architektur eines didaktischen ChatBots

Vollautomatisierte Erstellung von Routine-Content

Der Einsatz von Few-Shot-Sampling und individuellen Filtern ermöglicht die Automatisierung der Erstellung von strukturell ähnlichen Inhalten, was neue Anwendungsgebiete und die Erweiterung von Produkten ermöglicht.
Bei diesem Verfahren bekommt das Large Language Model (LLM) unterschiedliche „Musterlösungen“ vorgelegt, die als Basis für die Generierung neuer, ähnlich strukturierter Feedbacks und Übungen verwendet werden. Jedoch ist die Qualität der so erstellten Inhalte nicht immer ausreichend. Um diese präzise bewerten zu können, muss eine große Menge an Inhalten erzeugt werden (sogenannte Content-Dumps), die dann einzeln auf ihre Richtigkeit überprüft werden müssen, um eine genaue Statistik über die Qualität zu erhalten.
Bei diesem einstufigen Verfahren werden zunächst alle generierten Inhalte als endgültiges Ergebnis betrachtet (als richtig klassifiziert), ohne automatische Filterung durch das System. Basierend auf der Auswertung des Content-Dumps kann nun sehr präzise berechnet werden, wie viele fehlerhafte Inhalte an den Nutzer weitergegeben werden:

	als richtig klassifiziert	*als falsch klassifiziert*	
Generierung ist richtig	*70 %*	*0 %*	*70 %*
Generierung ist falsch	*30 %*	*0 %*	*30 %*
	100 %	*0 %*	*100 %*

Abb. 2: Einstufiges Verfahren

Um die Qualität der erzeugten Inhalte zu verbessern, kann die Eingabeaufforderung (Prompt), die die Generierung steuert, modifiziert werden. Allerdings stößt man bei komplexeren Anforderungen schnell an eine Grenze, die durch einfache Anpassungen des Prompts nicht überwunden werden kann. Um die Qualität dennoch zu erhöhen, müssen nun Filtersysteme eingeführt werden, die die Richtigkeit der Inhalte automatisch beurteilen. Dabei gibt es zwei verschiedene Arten von Filtern:

- **Regelbasierte Filter**: Diese Art von Filter hilft dabei, häufig vorkommende Standardfehler zu erkennen, die durch einfache Bedingungsabfragen identifiziert werden können. Oft können schon wenige solcher Filter die Qualität deutlich verbessern.

- **Probabilistische Filter:** Diese Filter können den erzeugten Inhalt semantisch untersuchen und Fehler erkennen, die mit regelbasierten Filtern nicht gefunden werden können.

Mit dieser zweistufigen Methode kann der Generator (die Kombination aus LLM und Filter) den Output des LLMs effektiv bewerten, wodurch ein Großteil der fehlerhaften Ausgaben gefiltert werden kann. Dies ermöglich schlussendlich eine Steigerung der prozentualen Qualität der an den Nutzer ausgegebenen Ergebnisse:

	als richtig klassifiziert	*als falsch klassifiziert*	
Generierung ist richtig	*50 %*	*20 %*	*70 %*
Generierung ist falsch	*10 %*	*20 %*	*30 %*
	60 %	*40 %*	*100 %*

Abb. 3: Zweistufiges Verfahren

Im Vergleich zur Abbildung 2 ist die Genauigkeit der vom System an den Benutzer gelieferten Inhalte von 70% auf 83% (50% / 60% = 83%) gestiegen. Dies wurde durch die Anwendung der Filtersysteme erreicht. Dieser Ansatz schafft einen klaren Qualitätsmaßstab, der als Leistungsindikator für Verbesserungen am Generator dient. Dadurch wird der Entwicklungsprozess beschleunigt und die Ergebnisse können quantifiziert werden.
Insgesamt handelt es sich bei diesem Verfahren um eine ressourcenschonendere Alternative zur herkömmlichen Erstellung von Routine-Content, die Zeit einspart und neue Märkte eröffnet.

Entwicklungsprozess: Produktisierung von generativer KI

In der Phase der Produktisierung von Generierungslösungen steht die Qualität des erstellten Inhalts im Mittelpunkt. Es ist wichtig zu verstehen, dass Perfektion zwar angestrebt wird, aber praktisch nicht erreichbar ist. Ähnlich wie bei menschlichen Leistungen kann keine fehlerfreie Qualität garantiert werden. Der Unterschied besteht darin, dass es bei Maschinen schwieriger ist, die Verantwortlichkeit für Fehler festzustellen, im Gegensatz zu Menschen, bei denen dies klarer ist. Dennoch bleibt das Endprodukt im Wesentlichen gleich, unabhängig von der Quelle der Fehlerquelle.
Das Hauptziel in diesem Prozess ist die Maximierung der „Korrektheit", die als Prozentsatz der korrekten Inhalte definiert ist, die nach Anwendung von regelbasierten und probabilistischen Filtern an den Benutzer weitergegeben werden. Diese Korrektheit kann nicht exakt berechnet werden, da sie von verschiedenen Eingabeparametern abhängt, die bei jeder Generierung unterschiedlich sind. Daher kann nur eine Wahrscheinlichkeitsverteilung erstellt und Konfidenzintervalle berechnet werden. Das Ziel in der Produktentwicklungsphase für Generierungslösungen ist es, den Erwartungswert der Verteilung näher an 100% Korrektheit zu bringen (siehe Abbildung 4).

In der Entwicklung von Generierungslösungen ist es wichtig, immer die Kosten- und Zeiteffizienz dieser im Auge zu behalten. Hierbei ist es von zentraler Bedeutung, die Rate der Betafehler gering zu halten. Das sind Fälle, in denen korrekt generierte Inhalte fälschlicherweise als fehlerhaft markiert werden, wenn zu strenge Filter angewendet werden. Eine zu hohe Rate solcher Fehler würde zu erhöhten Kosten und längeren Generierungszeiten führen, da unnötig viele Anfragen an den Generator gestellt werden müssten, um ein Ergebnis zu erhalten, dass als richtig klassifiziert wird.
Gleichzeitig ist es in der Praxis nicht möglich eine 100%ige Korrektheit zu erreichen – desto höher die Korrektheit, desto höher ist zwangsweise der Betafehler. Diese Wechselwirkung zwischen der Erhöhung der Korrektheit und der Verringerung der „Landing Rate", also dem Prozentsatz der generierten Ergebnisse, die von den Filtern als „korrekt" bewertet werden, stellt eine zentrale Herausforderung bei der Entwicklung dar. Daher ist es entscheidend, in jedem Anwendungskontext eine angemessene Balance zu finden, die eine adäquate Qualität und Landing Rate gewährleistet.

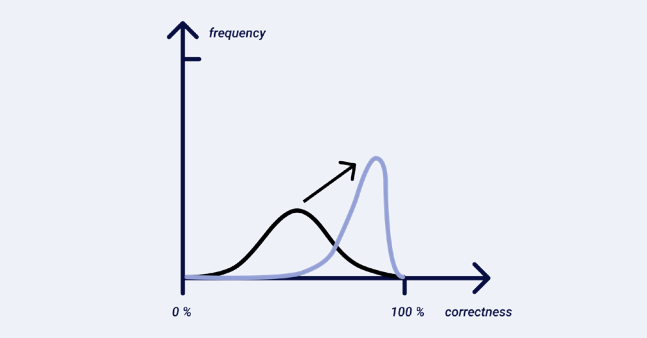

Abb. 4: Steigerung der Qualität

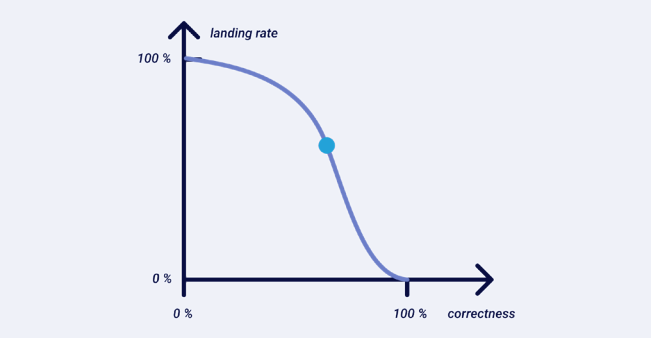

Abb. 5: Wechselwirkung zwischen Qualität und Effizienz

In Abbildung 5 zeigt sich, dass eine beinahe vollständige Korrektheit immer mit einer geringen Anzahl an Inhalten einhergeht, die als korrekt klassifiziert werden. Der Verlauf der Kurve beruht auf Erfahrungswerten – häufig die eine effiziente Optimierung nur bis zu einem gewissen Punkt (in der Graphik eingezeichnet) möglich ab welchem großen Abstriche in der Landing Rate gemacht werden müssen um die Qualität weiter zu erhöhen. Es ist entscheidend, ein Gleichgewicht zu finden, bei dem der Generierungsprozess qualitativ hochwertig, kosteneffizient und zeitsparend ist, um eine nachhaltige und hochwertige Benutzererfahrung zu gewährleisten.

Für die erfolgreiche Optimierung der Correctness ist es essenziell, dass die Entwickler nicht strikt nach den festgelegten Konzeptanforderungen des Produktteams vorgehen. Stattdessen sollten sie die Möglichkeit haben, aktiv mit dem System zu arbeiten, um das didaktische Konzept des Generators vollständig zu verstehen. Dies erfordert eine Fähigkeit, die möglicherweise nicht von jedem herkömmlichen Softwareentwickler erwartet wird. Als Abnahmekriterium in diesem Prozess dient eine quantitative Analyse des Content Dumps, welche auf zuvor festgelegten, den spezifischen Anwendungsfällen entsprechenden Testfällen beruht. Diese Testfälle könnten ihren Ursprung in den Produkt- oder Didaktikteams haben.
Es ist entscheidend, die aufgewendete Zeit für die Optimierung einzelner Generatoren genau zu erfassen. Es muss anerkannt werden, dass es für einige Generatoren möglicherweise nicht realistisch ist, das angestrebte Qualitätsniveau zu erreichen, und daher sollten die Entwicklungs- oder Optimierungsbemühungen frühzeitig eingestellt werden, um Ressourcen zu sparen.
Bei der Behandlung neu identifizierter Fehler ist es ratsam, eine zentrale Dokumentation zu führen, anstatt separate Tickets für jeden einzelnen Fehler zu erstellen, da es oft unmöglich ist, diese Fehler vollständig zu beseitigen. Stattdessen kann versucht werden, ihre Häufigkeit zu reduzieren. Wenn die Korrektheit bei einer nachfolgenden Messung das festgelegte Qualitätsniveau unterschreitet, erfordert dies eine Intervention der Entwickler. Wenn die Korrektheit jedoch das festgelegte Qualitätsniveau nicht unterschreitet, sind keine weiteren Maßnahmen erforderlich. Nach der Prüfung des gesamten Content Dumps kann ein Konfidenzintervall für die Korrektheit ermittelt werden. Es ist wichtig zu beachten, dass ein Content Dump mindestens 30 Inhalte enthalten sollte, um aussagekräftige Konfidenzintervalle für die Korrektheit berechnen zu können.

Über Titanom

Willkommen bei Titanom, Ihrer Wegbereiter für innovative Softwareprodukte in der Bildungsbranche. Unsere Mission ist es, mit Ihnen zusammen als engagierte Technologiepartner sofort einsatzfähige KI-Lösungen zu entwickeln, die Barrieren entfernen und Menschen befähigen. Wenn Sie Interesse an einem unserer Use Cases haben oder weitere Fragen zu unserem Ansatz und unseren Technologien stellen möchten, zögern Sie bitte nicht, sich an uns zu wenden. Wir stehen Ihnen zur Verfügung, um Potenziale zu erkunden, Machbarkeiten zu prüfen und gemeinsam Produkte zu entwickeln, die den Anforderungen einer sich rasch wandelnden Welt gerecht werden.
Wir freuen uns darauf, mit Ihnen zusammenzuarbeiten und gemeinsam die Zukunft der KI-gestützten Technologien zu gestalten.

Andrija Vuksanovic
CEO Titanom Technologies GmbH
andrija.vuksanovic@titanom.com

Leonhard Benkert
Project Manager
leonhard.benkert@titanom.com