Kontrolle und Validierung (generativer) künstlicher Intelligenz

Rainer Glaser

Eine wichtige Voraussetzung für die praktische Nutzung und die Ausschöpfung ihres Potenzials.

Insbesondere seit der Veröffentlichung von ChatGPT [https://openai.com/blog/chatgpt] ist die generative künstliche Intelligenz (GenAI) in den Mittelpunkt des Interesses von Unternehmen weltweit gerückt. Die Unternehmen sehen hier ein breites Anwendungsspektrum in sämtlichen Funktionsbereichen (3 lines of defense für Bankunternehmen eingeschlossen).

Die Branche muss sich allerdings noch auf einen Standard für die Validierung und Kontrolle von (Gen)AI-Modellen einigen, und wir erwarten in diesem Zusammenhang eine rasche Entwicklung, da sich die Technologie zur Messung von Genauigkeit und Zuverlässigkeit weiterentwickelt.

Die Fähigkeit zur Kontrolle, zur Messung der Genauigkeit und Zuverlässigkeit und zur Schaffung von Vertrauen in den Output ist eine wichtige Voraussetzung, um den Wert voll auszuschöpfen. Hierdurch entstehen jedoch Herausforderungen, die von Unternehmen häufig noch nicht gelöst wurden, was die Entwicklung verlangsamt – dies gilt sowohl für die generative als auch für die traditionelle künstliche Intelligenz.

Auf der Grundlage unserer Erfahrung bei der Entwicklung leistungsstarker analytischer Modelle, die zudem praxiserprobt sind, haben wir übergreifende Rahmenwerke und Ansätze entwickelt, die eine angemessene Nutzung und Validierung der Modelle gewährleisten. Dies gilt auch für unsere neueren Modelle, die (Gen)AI und LLM-Funktionalitäten (Large Language Models) nutzen. Wir konnten nachweisen, dass die Kontrolle und Validierung von (Gen)AI nicht nur unerlässlich ist, sondern auch effizient und konsequent umgesetzt werden kann.

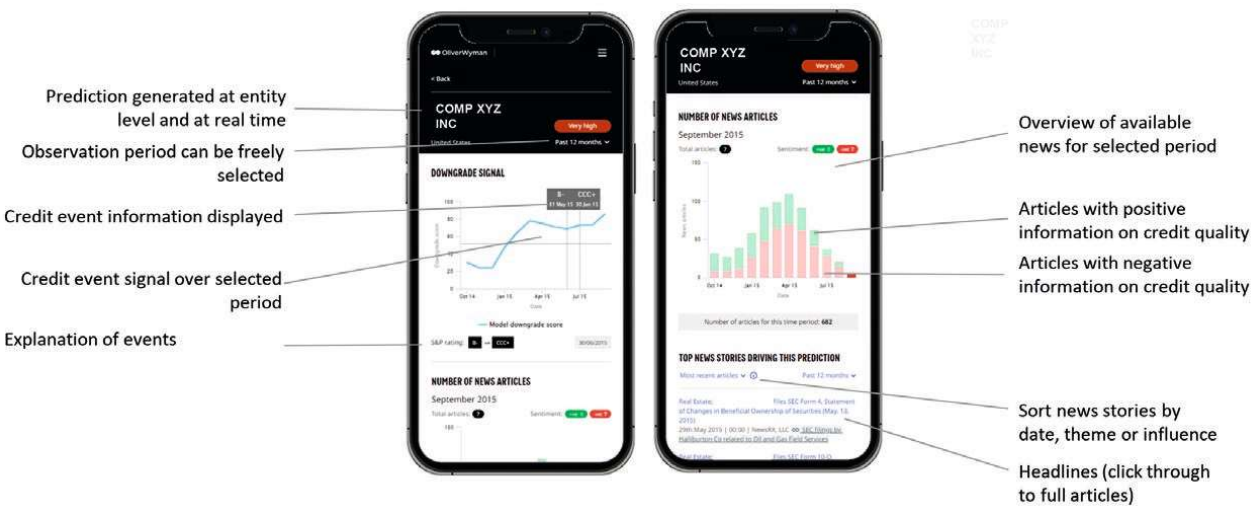
Wir sind uns bewusst, dass das Thema insgesamt sehr umfangreich ist und tiefgreifende Überlegungen und technische Analysen erfordert. Daher soll dieser Artikel lediglich einen Überblick geben, wie der Aufbau, die Kontrolle und die Validierung von (Gen)AI in der Praxis funktionieren können. Dieser Überblick ist nicht allgemein gehalten, sondern zeigt vielmehr einige detaillierte Beispiele, wie unser Validierungsrahmenwerk auf eine aktuelle Lösung (NewsTrack) angewendet wurde, die verschiedentlich eingesetzt wird.

NewsTrack – Eine kurze Einführung

NewsTrack ist eine einzigartige Lösung, die LLMs (Large Language Models) und GenAI nutzt, um anhand einer Vielzahl von Nachrichten-Feeds in Echtzeit Vorhersagen über negative Ereignisse für Unternehmen (z. B. Herabstufung, Zahlungsausfall/Konkurs, Betrug) zu erstellen (▶ Abb. 01).

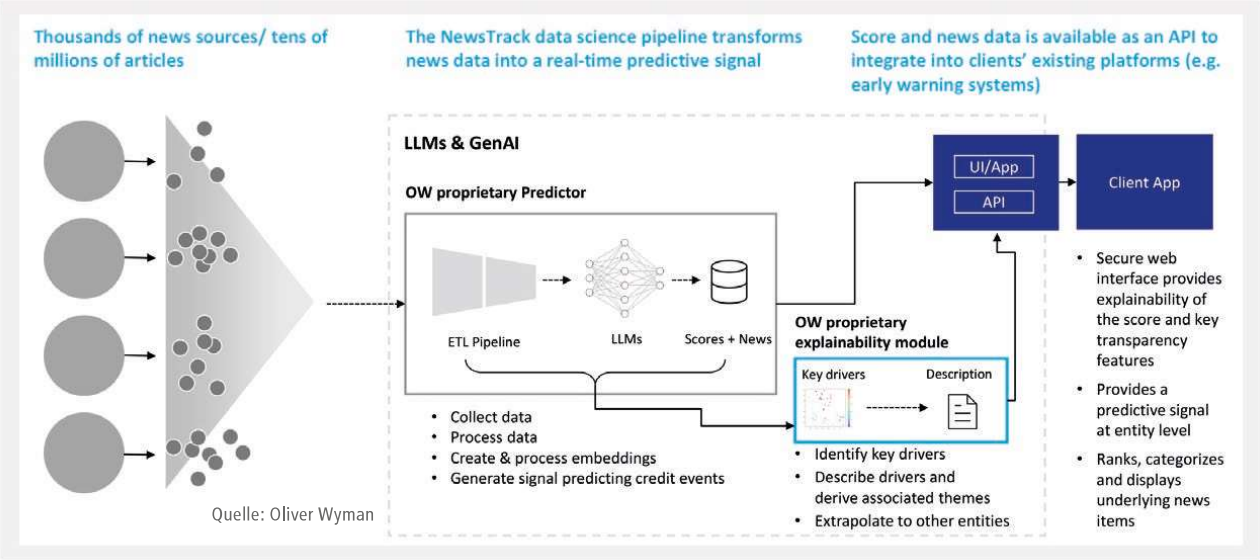
NewsTrack nutzt für die Verarbeitung von Textinputs hochmoderne (vortrainierte) LLMs. Es liefert nicht nur zuverlässige Vorhersagen, sondern auch die zugehörigen Begründungen und bietet damit unübertroffene Transparenz und Erklärbarkeit. Da es für einen spezifischen Zweck trainiert wurde und vollen Zugriff auf die zugrunde liegenden Inputs und deren Interpretation erlaubt, ist es einzigartig

Abb. 01: Output-Beispiel und Benutzeroberfläche von NewsTrack.



Quelle: Oliver Wyman

Abb. 02: Schematische Darstellung unserer proprietären NewsTrack-Pipeline.



Quelle: Oliver Wyman

und unterscheidet sich grundlegend von generischen Sentiment-Scores. Der zugrunde liegende Trainingsansatz ist hocheffizient, erfordert keine menschliche Interaktion und kann daher effizient gepflegt und verfeinert werden.

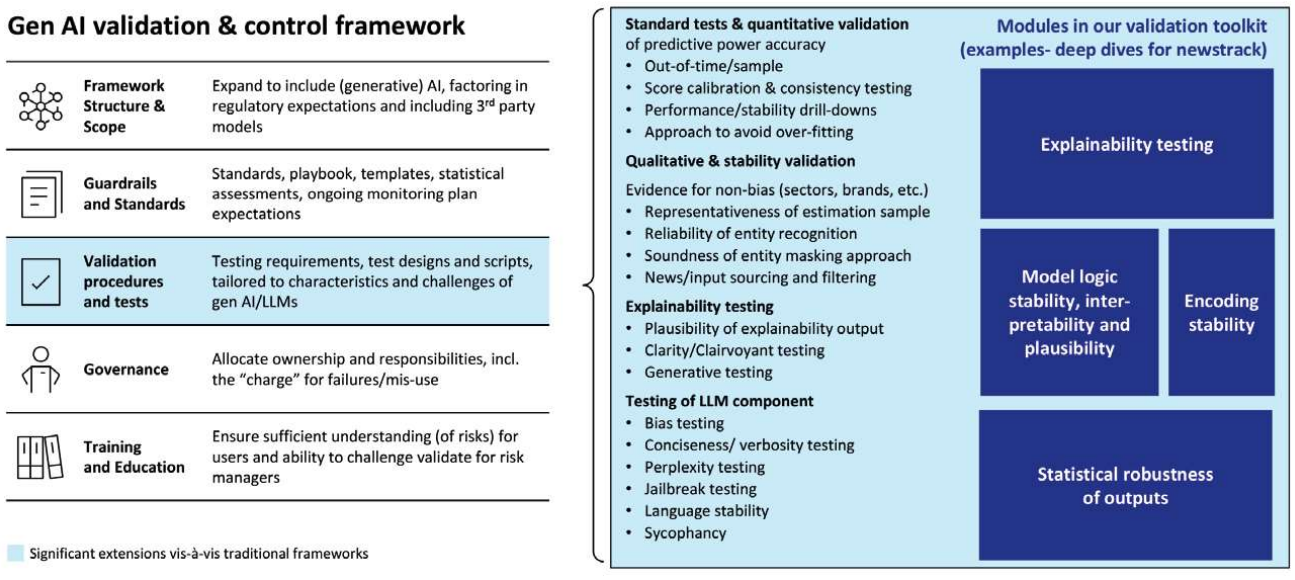
NewsTrack hat sich in einer Vielzahl von Anwendungsfällen bewährt. Hier einige Beispiele:

- Kredit-/Ausfallrisiko: Reduzierung der Kreditverluste/Wertberichtigungen in den Kreditbüchern um ~10 %
- Anlageoptimierung: Leistungssteigerung von ~200 Basispunkten für Anlageportfolios
- Lieferkettenrisiko: Antizipation von Risikoereignissen 3–6 Monate vor ihrem Eintreten

Beim Generieren der zugrunde liegenden Entwicklungspipeline konnten wir diverse Erfolgskriterien erfüllen:

- „Ausbildung" eines LLM zu einem „Fachexperten": Typische LLMs wurden trainiert, um Sprache zu „verstehen" und zu generieren. Unsere Pipeline geht einen Schritt weiter und integriert Fachwissen (z. B. über Kreditrisiken) durch eine geeignete Kombination aus LLM-Feintuning und maßgeschneiderten Prädiktoren.

Abb. 03: Überblick über unser (Gen)AI-Validierungs- und Kontrollrahmenwerk und detaillierte Darstellung der LLM-spezifischen Aspekte.



Quelle: Oliver Wyman

- Selbstständiges Herausfiltern des „richtigen Signals" aus einer beliebigen Kombination strukturierter und unstrukturierter Daten: Unsere Pipeline macht eine manuelle Kennzeichnung oder Filterung von Trainingsdatensätzen überflüssig. Diese wäre nicht nur mit erheblichem manuellem Aufwand verbunden, sondern sie wäre auch anfällig für Bias und menschliches Urteilsverhalten und somit nicht objektiv.
- Überwindung des Blackbox-Problems: Große und tiefgehende Modelle werden häufig als intransparent und undurchschaubar bezeichnet. Dies ist grundsätzlich richtig, ein integraler Bestandteil unserer Pipeline und des Validierungsansatzes ist allerdings das „Erklärbarkeitsmodul", das das Modell zwingt, sich selbst zu erklären.

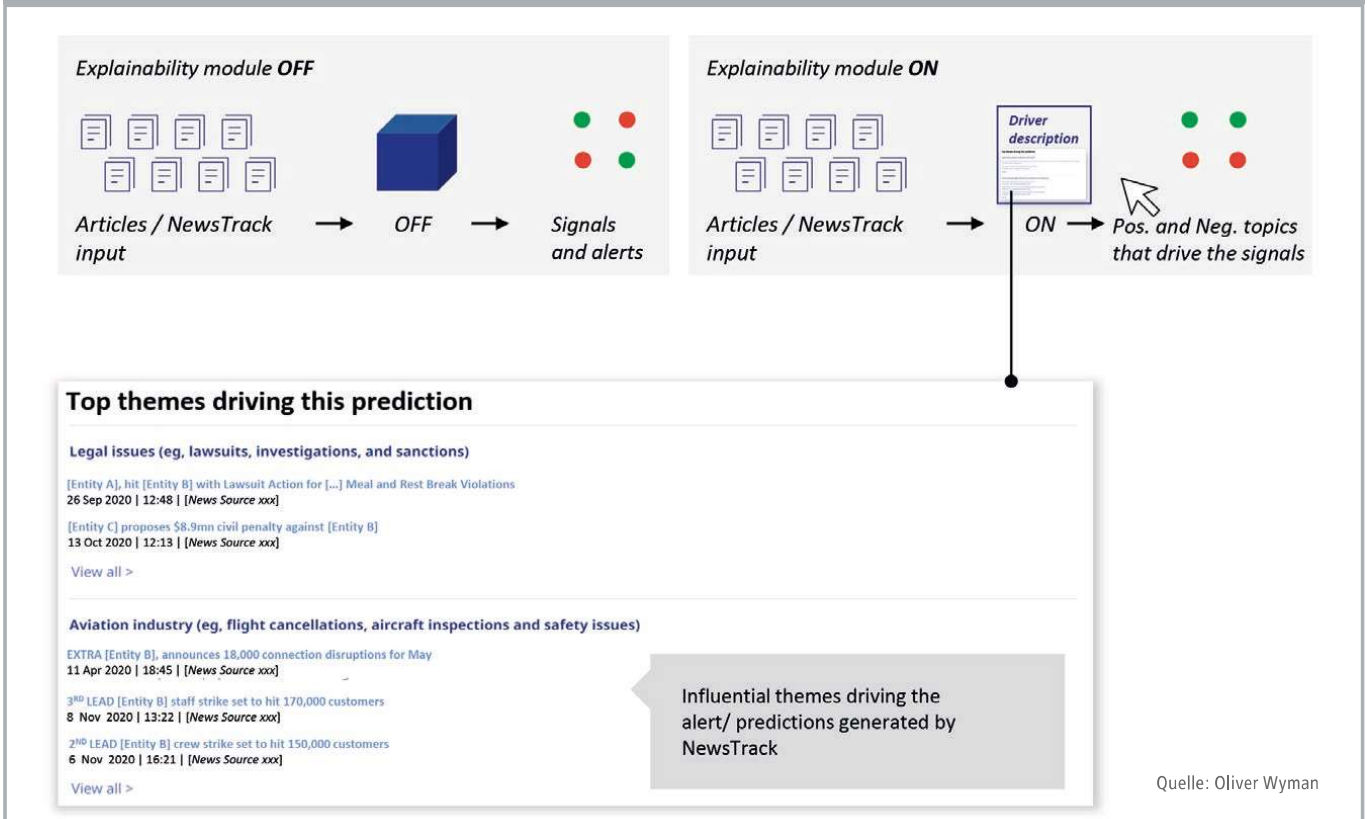
Überblick unseres (Gen)AI-Validierungs- und Kontrollrahmenwerks und Highlights seiner Anwendung für NewsTrack

Eine der wichtigsten Voraussetzungen für die erzielte Wirkung ist die Möglichkeit, das Modell zu kontrollieren und zu validieren. Auch wenn unser Validierungs- und Kontrollrahmenwerk einen viel breiteren Kontext als NewsTrack abdeckt, konzentrieren wir uns hier auf dieses Beispiel, um die Funktionsweise in der Praxis genauer zu veranschaulichen.

Das Rahmenwerk in ▶ Abb. 03 hat hinsichtlich der Struktur Ähnlichkeit mit einem traditionellen Validierungsrahmenwerk, allerdings sind in allen Bereichen Erweiterungen und Anpassungen erforderlich. Da zu erwarten ist, dass GenAI in einem breiten Spektrum von Anwendungsfällen zum Einsatz kommt und von einer großen Zahl von (nicht fachkundigen) Nutzern entwickelt/genutzt wird, muss besonderes Augenmerk auf Governance, Normen, Anwendungsbereich usw. gelegt werden.

(Quantitative) Testverfahren stellen häufig eine besondere Herausforderung dar. Daher möchten wir uns in diesem Artikel auf die quantitativen Aspekte unseres Validierungsrahmenwerks konzentrieren. Da an dieser Stelle nicht alle relevanten Tests vorgestellt

Abb. 04: Das auf NewsTrack angewandte Erklärbarkeitsmodul.



Quelle: Oliver Wyman

werden können, konzentrieren wir uns auf einige aufschlussreiche Beispiele im Zusammenhang mit NewsTrack, die zeigen, dass das Rahmenwerk konsequent und effizient umgesetzt werden kann.

Testen der Erklärbarkeit

Das auf NewsTrack angewandte Erklärbarkeitsmodul schafft ein Verständnis für die Treiber, die den Vorhersagen des Modells zugrunde liegen: Im Falle eines signifikanten Signals wird der Benutzer nicht nur über den Input (Nachrichtenmeldungen) informiert, der das Signal ausgelöst hat, sondern auch über die Themen, die das Modell als signifikante Treiber identifiziert hat, siehe ▶ Abb. 04.

Das Erklärbarkeitsmodul ist eine wesentliche Funktion, da es nicht nur die Plausibilitätsbewertung der Funktionsweise des Modells ermöglicht, sondern auch ein wichtiger Ausgangspunkt für eine Reihe weiterer Validierungsverfahren ist. Diese beruhen auf der Tatsache, dass die relevanten Themen von einem speziellen Algorithmus in automatisierter, konsistenter und generativer (d. h. für den Menschen lesbaren) Form identifiziert werden. Folgende Beispiele veranschaulichen dies:

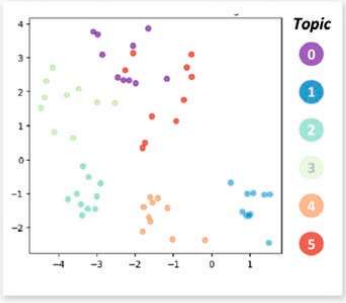
- Stabilität der identifizierten Treiber: Nur wenn die identifizierten Themen unter Störungen (invariant in Bezug auf den Inhalt) der Inputs und der Entwicklungsdatensätze stabil sind, können die Modellvorhersagen als zuverlässig und aussagekräftig angesehen werden (und sind nicht z. B. durch Fehlwahrnehmungen bedingt).
- Die Vollständigkeit der Modelltreiber kann durch Sensitivitätsanalysen in Bezug auf die Granularität der identifizierten Themen bewertet werden: Ist die Identifizierung der Treiber zu granular, sind die Themen nicht mehr aussagekräftig; ist die Identifizierung zu stark aggregiert, werden mit einem bestimmten Treiber zu viele Themen assoziiert.

Diese Überlegungen sind in der Reihe von Tests formalisiert, die im folgenden Abschnitt beschrieben werden.

Abb. 05: Beispiel-Output des Moduls für Stabilität, Interpretierbarkeit und Plausibilität der Modelllogik.

Embedding distribution visualisation across topics

Each colour represents a classified topic



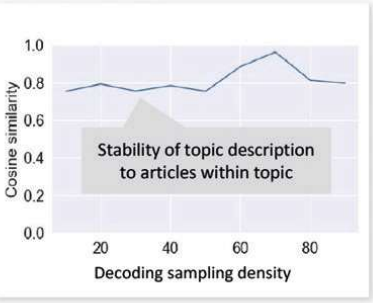
Quantitative evaluation of topic classification

In embedding space

Metric	Value
Homogeneity Score	0.79
Completeness Score	0.87
Adjusted Rand Index	0.66
Fowlkes-Mallows score	0.74

Plot of topic description similarities with decoding sampling density

For one illustrative topic



Verbosity test of topic descriptions

	Verbosity 100%	Verbosity 90%	Cosine Similarity	Verbosity 80%	Cosine Similarity
1	Technology companies competing in consumer markets	Branding and marketing in the tech industry	0.8528	Technology companies expanding into new markets and offerings	0.8415
2	Stock Performance of Various Industries	Mixed performance of top companies in various sectors	0.8016	Performance of top companies in various industries	0.8544
3	Sustainable practices in the fashion industry	Sustainable practices in the fashion and hospitality industries	0.9689	Sustainable practices in global fashion industry	0.9668

Quelle: Oliver Wyman

Stabilität, Interpretierbarkeit und Plausibilität der Modelllogik
Das genannte Modul ermöglicht eine Bewertung der Eignung der vom Erklärbarkeitsmodul identifizierten Themen. Dabei wird auch bewertet, wie gut eine Auswahl von identifizierten Modelltreibern

- ein einzelnes Thema im Vergleich zu einer Gruppe von Themen darstellt (Homogenitätsbewertung),
- alle relevanten Treiber abdeckt (Vollständigkeitsbewertung).

Dies wird durch eine grafische Darstellung für eine visuelle Kontrolle unterstützt, siehe ▶ Abb. 05.

Um die Stabilität der generativen Beschreibung der generierten Treiber zu bewerten, wird eine Sensitivitätsanalyse in Bezug auf die Dichte der die Dekodiersamplingrate (ein Parameter des generativen Ansatzes zur Beschreibung der Treiber) durchgeführt.

Darüber hinaus werden Verbositätstests repräsentativer Treiberbeschreibungen durchgeführt, um zu prüfen, ob die generative Beschreibung der Treiber stabil ist.

Statistische Belastbarkeit des Outputs
Bei den meisten Analysemodellen ist die Fähigkeit, die statistische Leistung des Modells zu testen, zu messen und nachzuweisen, von wesentlicher Bedeutung – dies gilt auch für die Mehrzahl der LLM- und GenAI-Anwendungen. Welche Leistungskennzahlen in Frage kommen, ist abhängig vom jeweiligen Anwendungsfall und gegebenenfalls von der verwendeten Modellstruktur – das Validierungsrahmenwerk deckt eine breite Auswahl relevanter KPIs ab. Siehe ▶ ▶ Abb. 06 für einige Beispiele.

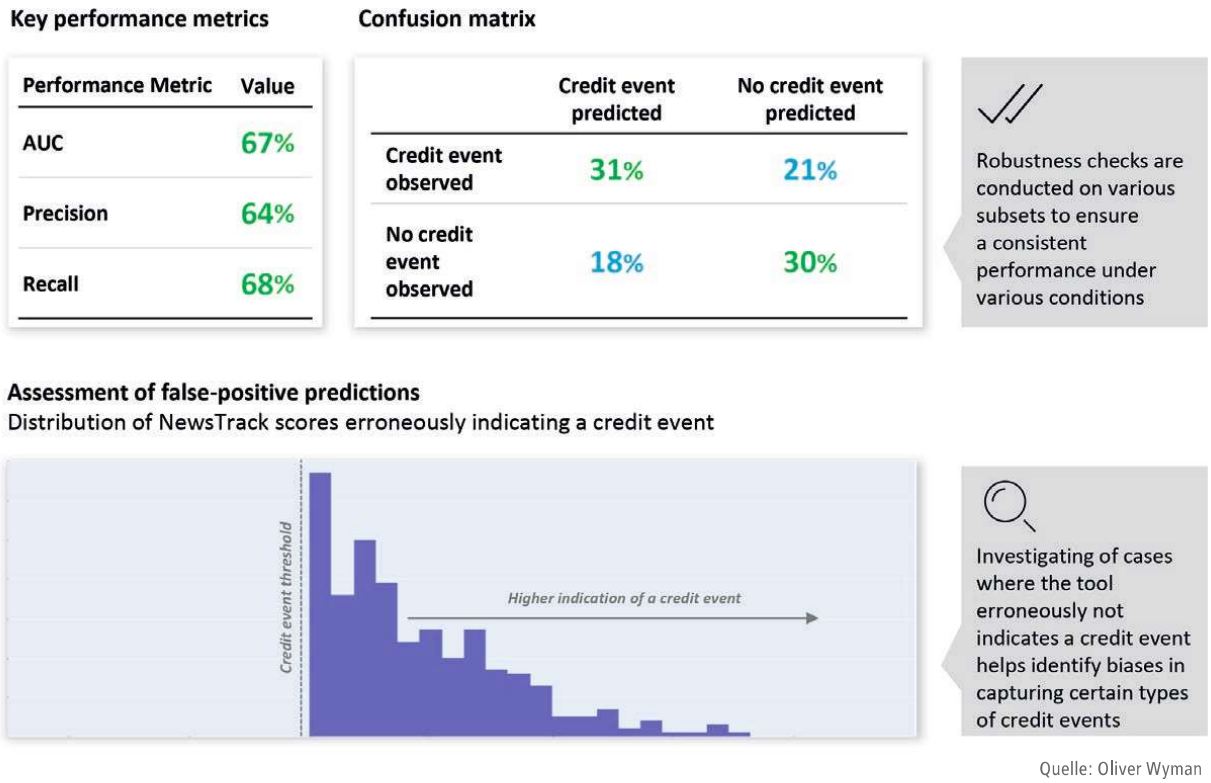
Das Rahmenwerk beinhaltet Tests auf ungesehenen Daten, um festzustellen, ob das Modell überangepasst oder zu vereinfachend ist, um die zugrunde liegenden Muster zu erfassen (z. B. Out-of-Time-Tests/Stichproben, Kreuzvalidierung, Störung von Inputs, Maskierung/Austausch von Entitäten usw.).

Störungsstabilität
Bei Tests zur Störungsstabilität wird geprüft, ob das Modell auf einem tatsächlichen Verständnis des Inhalts beruht oder ob es Artefakte aufgreift. Dies ermöglicht insbesondere eine Überprüfung, ob das Modell für einen bestimmten Anwendungsfall irrelevante Informationen ignoriert und somit Fehlwahrnehmungen entsprechend identifiziert (siehe ▶ Abb. 07).

Das Rahmenwerk unterstützt die Durchführung aller oben genannten Tests unter Störungen (invariant in Bezug auf Bedeutung/Inhalt) der Inputs und/oder der Trainings-/Testsätze. Eine spezielle Bibliothek ermöglicht eine effiziente Durchführung von Textstörungen, z. B. durch Ersetzen mit Synonymen, (Neu-)Übersetzung, Ersetzen von Schlüsselwörtern, Hinzufügen von Zufallsrauschen usw.

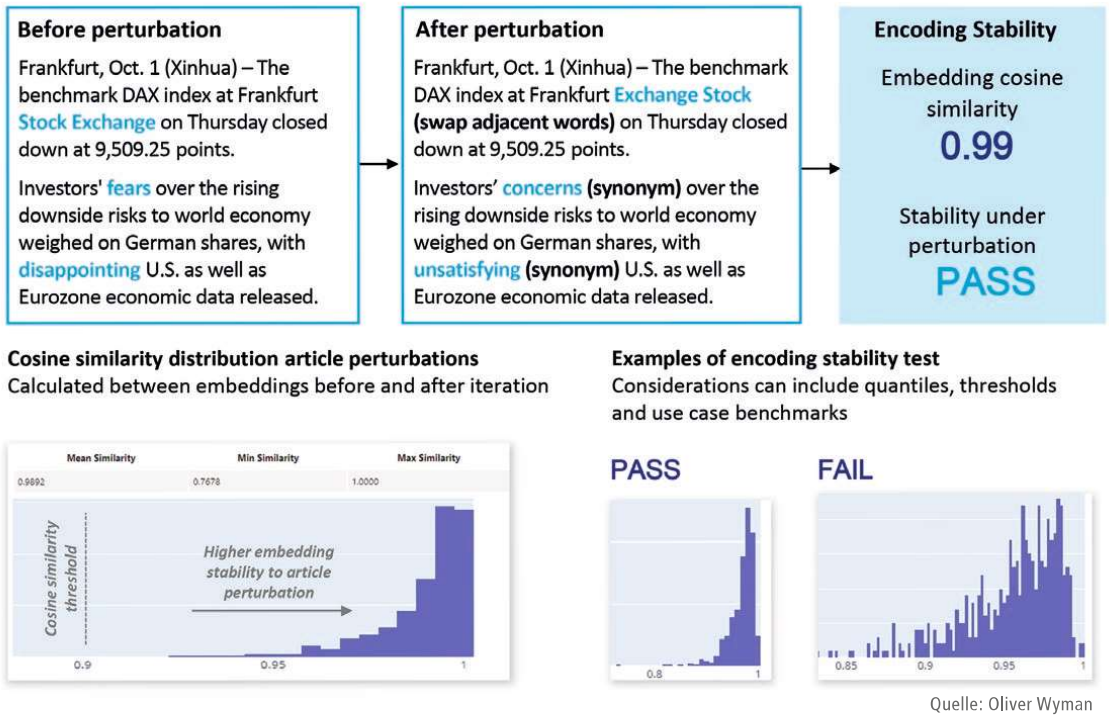
Fazit
Die Fähigkeit, den von einem (Gen)AI-Modell erzeugten Output zu verstehen und zu kontrollieren, ist nicht nur eine Voraussetzung im Hinblick auf Überwachungsaspekte, sondern auch für den prak-

Abb. 06: Beispiel-Output des Moduls für statistische Belastbarkeit.



Quelle: Oliver Wyman

Abb. 07: Anschauungsbeispiele für Prüfungen der Störungsstabilität.



Quelle: Oliver Wyman

tischen Einsatz im Rahmen von Geschäftsentscheidungen: Nur wenn der Benutzer verstehen kann, woher eine bestimmte Vorhersage / ein bestimmter Modell-Output stammt, kann das Modell eine effiziente, informierte Entscheidungsfindung ermöglichen, und der Benutzer wird dem Modell-Output vertrauen und diesen erklären/begründen können.

Die Festlegung des geeigneten Niveaus und der richtigen Ansätze für die Kontrolle und Validierung von (Gen)AI-Modellen ist eine anspruchsvolle Aufgabe, die generalisiert nur schwer gelöst werden kann. Die Konzentration auf eine Reihe relevanter Anwendungsfälle ermöglicht jedoch die Entwicklung eines robusten Validierungs- und Kontrollrahmenwerks, einschließlich einer effizienten Implementierung und Ausführung. Schlüsselkomponenten dieser Rahmenwerke sind die Identifizierung und Beschreibung der Treiber sowie die Bewertung ihrer Plausibilität und Stabilität.

Am Beispiel von NewsTrack konnten wir anhand einiger Elemente aufzeigen, wie ein konsequentes Validierungsrahmenwerk aussieht und in der Praxis effizient umgesetzt werden kann. Der Weg dorthin erfordert sowohl die Entwicklung der richtigen Methoden und Ansätze für die Tests als auch eine effiziente Plattform zur Unterstützung der Durchführung.

Autor
Dr. Rainer Glaser
Partner Finance & Risk / Digital,
Oliver Wyman,
München