

**EQS** | **B C M** | Berufsverband  
der Compliance Manager

Benchmark-Report

# **KI-PERFORMANCE IM BEREICH COMPLIANCE & ETHIK**

Vollständiger Bericht

# Inhalt

<b>Vorworte</b> .....	3
<b>Zusammenfassung für Führungskräfte</b> .....	5
Zentrale Erkenntnisse des EQS-Benchmark-Reports .....	6
Detailierte Gesamtmodellleistung .....	7
Der Leistungsunterschied, den nur neun Monate ausmachen können .....	8
Stärken und Schwächen von KI-Modellen .....	9
Die Modellwahl ist in mehrdeutigen Bereichen wichtig .....	10
Je komplexer die Aufgabe, desto größer die Leistungsunterschiede .....	11
Zuverlässigkeit und Halluzinationsraten .....	12
<b>Praktische Erkenntnisse für Compliance- und Ethik-Profis</b> .....	13
<b>Auswirkungen auf KI-Agenten und die Automatisierung im Bereich Compliance &amp; Ethik</b> .....	16
<b>Ausblick: Was die Zukunft bringt</b> .....	17
<b>Beispiele für Ergebnisse von Benchmark-Aufgaben</b> .....	19
Erkennen potenzieller Vergeltungsmaßnahmen gegen namentlich bekannte Hinweisgeber .....	19
Bewertung der Wirksamkeit einer Sensibilisierungskampagne anhand von Quiz-Ergebnissen (GPT-5 vs. GPT-4o) .....	21
<b>Die Modelle im Detail</b> .....	24
Google Gemini 2.5 Pro: Ein herausragender Allrounder für Compliance & Ethik .....	24
Openai GPT-5: Liebling der Compliance-Beauftragten .....	25
Openai o3: Sprungbrett zu GPT-5 .....	26
Anthropic Claude Opus 4.1: Großartiger Autor, mittelmäßiger Analyst .....	27
Openai GPT-4o: Die blitzschnelle Qualitätsmesslatte von 2024 .....	27
Mistral Large 2: Die „lobende Erwähnung“ aus Europa .....	28
<b>Methodik</b> .....	29
Aufgabenbereich .....	29
Aufgabenarten .....	30
Aufgabenkategorien .....	31
Schwierigkeitsgrad der Aufgaben .....	32
Synthetischer und realer Kontext für Aufgaben .....	32
Alle Aufgaben in der Übersicht .....	33
Die im EQS-Benchmark-Report getesteten KI-Modelle .....	38
Ausführung der Benchmark-Aufgaben .....	39
Bewertung der Modell-Ergebnisse .....	39
Einsatz von KI bei der Erstellung der Benchmark .....	39
<b>Risiken und Schwachstellen</b> .....	40
<b>Schlussbemerkung</b> .....	41
<b>Autoren</b> .....	42
<b>Über EQS Group</b> .....	43
<b>Über den BCM</b> .....	43

# Vorworte

## **KI verändert bereits die Compliance-Arbeit – wenn man weiß, wie man sie nutzt.**

Künstliche Intelligenz ist längst keine ferne Zukunftsvision mehr – sie verändert bereits heute die Art und Weise, wie Compliance-Arbeit erledigt wird. Die Frage ist nicht ob, sondern wie wir sie nutzen.

Bei der EQS Group setzen wir auf verantwortungsvolle Innovationen. Aus diesem Grund haben wir den EQS-Benchmark-Report erstellt: den ersten Report, der sich vollständig auf die realen Herausforderungen im Bereich Compliance & Ethik konzentriert. Wir haben keine theoretischen Fragen gestellt oder Nischenanwendungen getestet, sondern wir haben die derzeit leistungsfähigsten KI-Modell mit genau den Aufgaben getestet, die Ihr Team jeden Tag zu bewältigen hat.

Die Ergebnisse sind beeindruckend. Die Modelle sind leistungsfähiger, als viele erwarten – jedoch nur, wenn sie wohl überlegt auf die richtigen Probleme angewendet werden. Ebenso wichtig ist, dass sie uns zeigen, wo noch die Grenzen liegen – und menschliches Urteilsvermögen nach wie vor unersetztlich ist.

Der Report ist mehr als eine Benchmark. Er ist ein Werkzeug für strategisches Denken. Er hilft Compliance-Verantwortlichen zu verstehen, was KI leisten kann, und zeigt, wie sie bereits in der Praxis eingesetzt wird – so lassen sich die bevorstehenden größeren Veränderungen gezielt vorbereiten.

Ich bin stolz darauf, dass die EQS Group diese Diskussion mit Transparenz und Substanz anführt. Mein Dank gilt den Kundinnen und Kunden, Partnern und Expertinnen und Experten, die ihre Erkenntnisse beigesteuert haben – und ganz besonders BCM für die fortwährende Unterstützung.

Ich hoffe, dieser Bericht stößt in Ihrem Unternehmen produktive Diskussionen an und hilft Ihnen und Ihrem Team, nicht nur compliant zu bleiben – sondern auch in einer Welt relevant zu bleiben, die sich schneller denn je verändert.



**ACHIM WEICK**  
GRÜNDER & CEO, EQS GROUP

## **Als Compliance-Experten müssen wir unseren Tools vertrauen**

Künstliche Intelligenz verändert die Art und Weise, wie Organisationen arbeiten – und damit auch die Erwartungen an Compliance. Das ist in unserem Fachgebiet kein Thema, das wir ignorieren können. Compliance-Verantwortliche können KI nicht einfach unkritisch folgen, aber sie können sich ihr auch nicht verschließen. Beide Ansätze würden den Anforderungen unserer Rolle nicht gerecht.

Stattdessen sind Kompetenz und Vertrauen gefragt: die Fähigkeit zu erkennen, wo KI Effizienz und Erkenntnisse bringen kann und wo ihr Einsatz klar begrenzt sein muss. Das erfordert mehr als technisches Know-how. Es verlangt Reflexion, Dialog und professionelles Urteilsvermögen. Unsere Verantwortung besteht darin, aktiv mitzugestalten, wie KI in der Praxis eingesetzt wird – um ihr Potenzial für Compliance substanzial nutzbar zu machen und zugleich die ethischen und rechtlichen Grenzen zu wahren, die nicht überschritten werden dürfen.

Dies ist kein externes Zusatzthema, sondern ein integraler Bestandteil unserer beruflichen DNA. Die Zukunft der Compliance wird von denen geprägt, die KI nicht einfach akzeptieren oder ablehnen, sondern die Verantwortung übernehmen, ihre Chancen und Risiken mit Klarheit, Integrität und Weitblick zu steuern.

Der EQS-Benchmark-Report steht für die Art von fundierter, praxisnaher Bewertung, die unser Beruf braucht. Statt theoretischer Benchmarks, die von unserem Alltag abgekoppelt sind, testet diese Initiative KI-Systeme an den komplexen, nuancierten Herausforderungen, denen wir beim Schutz der Integrität von Organisationen begegnen. Der Berufsverband der Compliance Manager hatte die Ehre, die professionelle Expertise unserer Mitglieder in diese Bewertung einzubringen.

Die Ergebnisse liefern uns etwas Unschätzbares: ein evidenzbasiertes Verständnis dafür, wo KI tatsächlich zur Effektivität beitragen kann und wo menschliches Urteilsvermögen unersetzt bleibt. Dieses Wissen befähigt uns und gibt uns die Grundlage, die Einführung von KI in unseren Organisationen mit derselben Sorgfalt zu steuern, die wir auch in allen anderen Compliance-Fragen an den Tag legen.

Ich lade Sie ein, diesen Bericht mit Neugier und dem Geist proaktiver Mitgestaltung zu lesen. Studien wie der EQS-Benchmark-Report dienen nicht nur dazu, technologische Veränderungen zu beobachten, sondern sie aktiv zu gestalten. Sie helfen sicherzustellen, dass neue Fähigkeiten auf eine Weise integriert werden, die Vertrauen stärkt, Integrität wahrt und den unersetzlichen Wert menschlicher Expertise in der Compliance bewahrt.



**DR. GISA ORTWEIN**  
PRÄSIDENTIN DES BERUFSVERBANDS  
DER COMPLIANCE MANAGER (BCM),  
CHIEF COMPLIANCE OFFICER  
BEI NORMA GROUP



# Zusammenfassung für Führungskräfte

**KI ist kein fernes Ziel mehr – sie ist da und verändert, wie Compliance-Arbeit erledigt wird.**

Doch bislang wurde die Diskussion über KI in der Compliance mehr von Hype als von belastbaren Fakten bestimmt. Der EQS-Benchmark-Report ändert dies. Zum ersten Mal testen wir die derzeit leistungsfähigsten Modelle anhand der Aufgaben, mit denen Compliance-Profis tatsächlich konfrontiert sind: von der Einstufung der Risiken und der Gewinnung von Informationen bis zum Erstellen von Schulungsinhalten und der Analyse von möglichen Repressalien gegen hinweisgebende Personen. Die Ergebnisse zeigen klare Stärken und hartnäckige Schwächen – und liefern darüber hinaus Hinweise, wo menschliche Expertise weiterhin eingebunden bleiben muss.

Diese Modelle sind von großer Bedeutung. Weltweit verfügt nur eine Handvoll Unternehmen im KI-Sektor über das Kapital und die Talente, um Modelle von Weltklasse zu entwickeln. Und indem sie ihre KI-Modelle den IT-Teams in den Unternehmen und Softwareanbietern zur Verfügung stellen, definieren sie defacto die Fähigkeiten, mit denen die ganze Welt arbeiten kann – einschließlich der Compliance- und Ethikfunktionen. Daher ist es entscheidend, nicht nur die aktuellen Stärken und Schwächen der Modelle zu verstehen, sondern auch deren wahrscheinliche Entwicklung.

Um diese herauszufinden, haben wir sechs führende KI-Modelle anhand von 120 Aufgaben in 10 zentralen Compliance- und Ethik-Bereichen bewertet. Jede Aufgabe wurde so gestaltet, dass sie reale Szenarien widerspiegelt, und jedes Modell wurde hinsichtlich Genauigkeit, Zuverlässigkeit und praktischem Nutzen beurteilt. Wir haben dabei praxisnahe Inhalte und Daten unserer Kunden und Partner einbezogen, um sicherzustellen, dass der Report die tatsächliche Komplexität, Sprache und Entscheidungsfindung widerspiegelt, die Compliance-Experten täglich begegnen. Das Ergebnis ist der erste umfassende Benchmark-Report der Branche, der zeigt, wie gut KI heute bereits die Compliance-Arbeit unterstützt – und was das für die Zukunft des Berufsbildes bedeutet.



## Zentrale Erkenntnisse aus dem EQS-Benchmark-Report

### **Erkenntnis 1: Heutige KI-Modelle sind deutlich leistungsfähiger, als Compliance-Fachleute erwarten**

Bei strukturierten Aufgaben wie Ranking, Extraktion und Klassifikation lieferten die führenden KI-Modelle konstant Ergebnisse, die unseren Erwartungen entsprachen – und diese oft sogar übertrafen. Was viele noch als futuristisch ansehen, ist bereits Realität: KI kann heute einen bedeutenden Teil der wiederkehrenden Compliance-Arbeit übernehmen und so die Experten entlasten, sodass diese sich auf wertvolle Entscheidungen und Strategien konzentrieren können.

### **Erkenntnis 2: Unterschiede zwischen Modellen sind entscheidend – und können sich innerhalb von Monaten drastisch verändern**

Die Leistungsunterschiede zwischen den Modellen sind nicht marginal – in einigen Kategorien sind die besten Modelle zwei- bis dreimal stärker als die schwächsten. Und die Rangliste ist dynamisch: Ein Modell, das in einem Quartal führt, kann im nächsten überholt werden. Compliance-Teams können es sich daher nicht leisten, „KI“ als eine einzelne Fähigkeit zu behandeln – sorgfältige Auswahl, Tests und eine kontinuierliche Evaluierung sind unerlässlich.

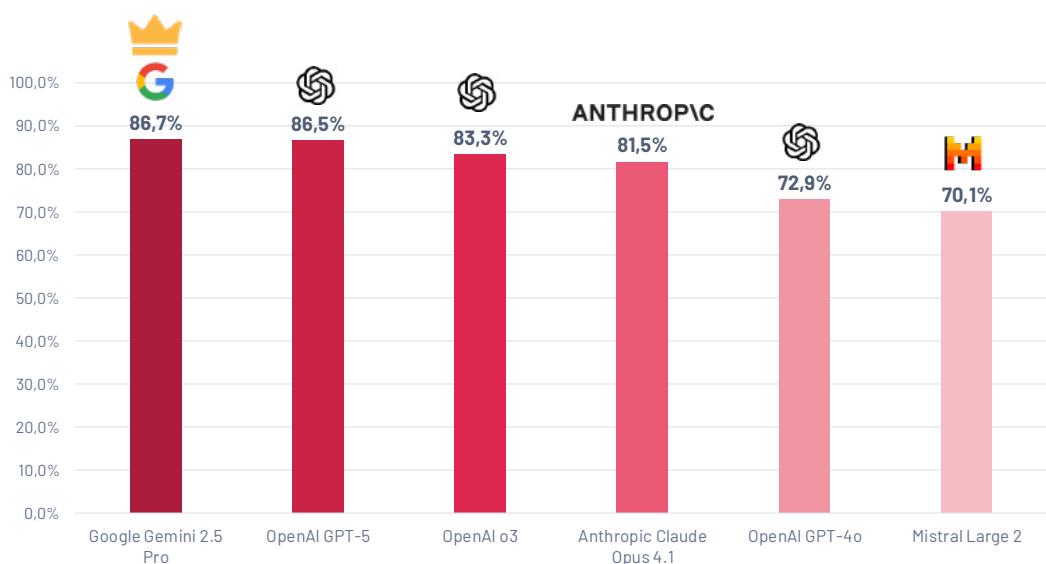
### **Erkenntnis 3: Wirkungsvolle und zuverlässige KI-Agenten sind im Compliance-Bereich nicht mehr weit entfernt**

Der Benchmark-Report zeigt ein klares Muster: Modelle glänzen bei gut strukturierten Teilaufgaben, stoßen jedoch an ihre Grenzen, wenn die Ambiguität steigt. Genau in diesem Umfeld werden sogenannte KI-Agenten – die mehrere Teilaufgaben unter menschlicher Aufsicht verketteten – besonders erfolgreich sein. Angesichts des aktuellen Trends sollten sich Compliance-Teams jetzt auf KI-Systeme vorbereiten, die nicht nur Fragen beantworten, sondern aktiv mitarbeiten. Die richtige Gestaltung der Interaktion zwischen KI-Systemen und Menschen wird entscheidend sein. Der EQS-Benchmark-Report liefert fundierte Einblicke, wo die menschliche Prüfung weiterhin unverzichtbar ist.

## Detaillierte Gesamtmodellleistung

Unter den sechs von uns getesteten KI-Modellen erreichte **Google Gemini 2.5 Pro** mit **86,7 %** das beste Ergebnis. Das Modell zeigte eine starke Leistung über alle Compliance-Bereiche sowie Arten und Kategorien von Aufgaben hinweg. Dicht dahinter folgte mit **86,5 %** **GPT-5 von OpenAI**, das in einigen Kategorien gleichauf oder sogar besser als Gemini abschnitt – ein deutlicher Hinweis darauf, wie schnell sich die Fähigkeiten in der Spur angleichen können. Der Vorgänger, **OpenAI o3**, erreichte **83,3 %** und verdeutlicht sowohl den Fortschritt von GPT-5 als auch den rasanten Innovationszyklus bei den KI-Modellen.

Auf Platz vier erreichte **Claude Opus 4.1 von Anthropic** **81,5 %**. Zwar beträgt der Abstand zu den Top-3 nur wenige Prozentpunkte, doch der EQS-Benchmark-Report zeigte konkrete



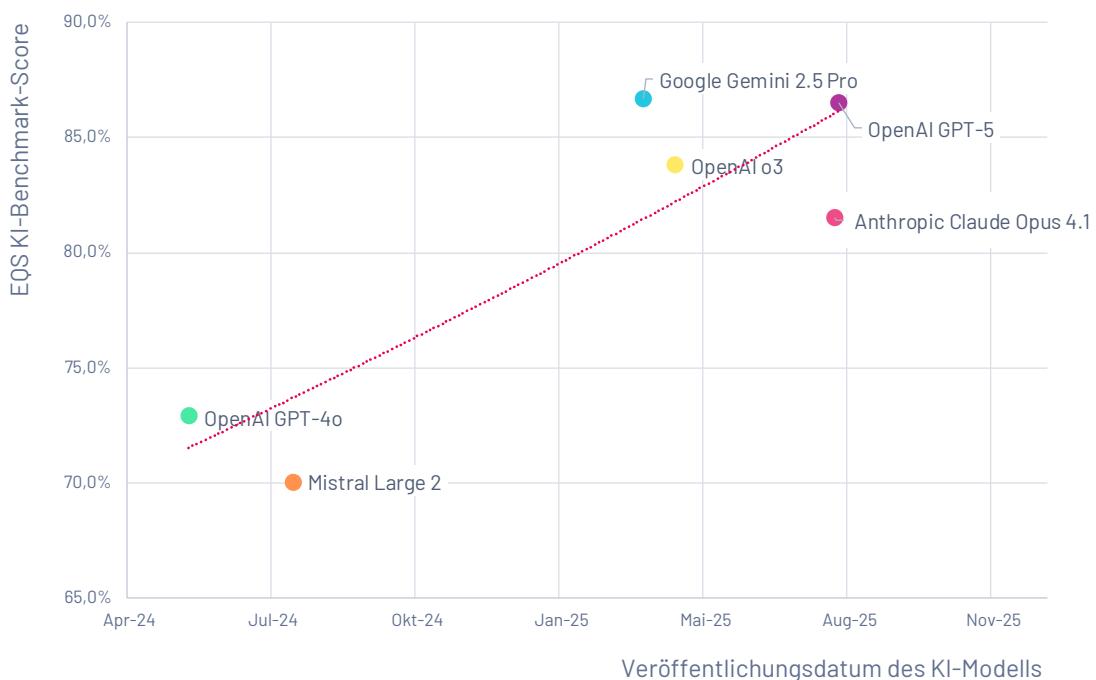
Aufgabenbereiche, bei denen sich dies in spürbar schwächeren Ergebnisse niederschlug. Das verdeutlicht, dass kleine Unterschiede in der Gesamtbewertung in der Praxis große Unterschiede bedeuten können.

Am unteren Ende der Rangliste lagen **GPT-4o (72,9 %)** und **Mistral Large 2 (70,1 %)**. Dieses Ergebnis spiegelt den Generationssprung zwischen den Modellen wider, die 2024 veröffentlicht wurden (wie GPT-4o und Mistral Large 2), und jenen, die 2025 auf den Markt kamen. Die Kluft ist bemerkenswert: Bis August 2025 war GPT-4o das Standardmodell in ChatGPT, das von Hunderten Millionen Fachleuten weltweit genutzt wurde. Die Benchmark-Ergebnisse zeigen jedoch, wie schnell selbst weit verbreitete Modelle veraltet sein können, wenn neuere Generationen erscheinen.

# **Der Leistungsunterschied, den nur neun Monate ausmachen können**

Die rasante Entwicklung im KI-Bereich zeigt sich auch in den häufigen Veröffentlichungen neuer Modelle durch die führenden Labore wie Google, OpenAI oder Anthropic. Kaum ein Quartal vergeht, ohne dass ein neues KI-Modell der Spitzenklasse vorgestellt wird, das erneut Leistungssteigerungen verspricht.

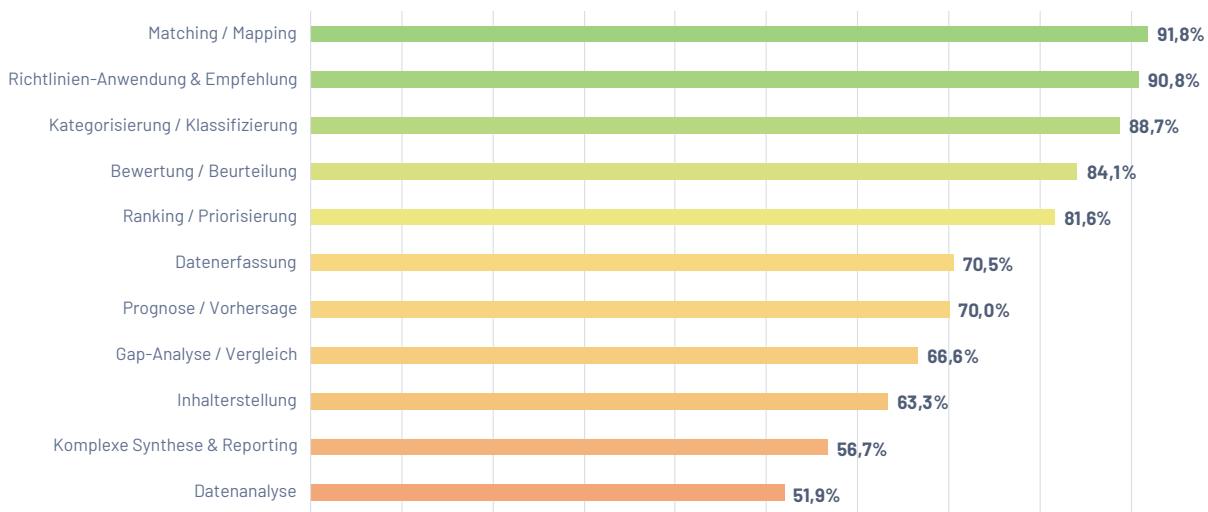
Im Kontext von Compliance & Ethik werden die Fortschritte besonders deutlich: Modelle, die 2025 veröffentlicht wurden, übertreffen tatsächlich die 2024er-Modelle deutlich. So schnitt Mistral Large 2, herausgekommen am 24. Juli 2024, mehr als 17 Prozentpunkte schlechter ab als das Siegermodell Gemini 2.5 Pro, das nur neun Monate später erschien – ein enormer Leistungsunterschied, der in konkreten Aufgaben besonders deutlich wird.



## Stärken und Schwächen von KI-Modellen

Um die Stärken und Schwächen der KI-Modelle zu verstehen, wurden alle Benchmark-Aufgaben kategorisiert. Eine vollständige Aufschlüsselung dieser Kategorisierung ist unter „Methodik“ zu finden. Der Vergleich zeigt klare Stärken der heutigen KI-Modelle: Einfache Aufgaben wie Matching/Mapping von Datensätzen, Entscheidungsfindung (bei Vorlage einer Situation und eines Regelwerks oder einer Richtlinie), Ranking/Priorisierung sowie Kategorisierung/Klassifizierung stellen für die heutige KI-Technologie kaum noch eine Herausforderung dar.

Durchschnittliche Leistung der getesteten Modelle in einzelnen Aufgabenbereichen



Gleichzeitig werden aber auch Schwächen deutlich: Selbst die leistungsfähigsten Modelle tun sich schwer mit komplexer Daten- oder Gap-Analyse und anspruchsvolleren Bewertungs-/Beurteilungsaufgaben. Dies sind auch die Bereiche, in denen die größten Leistungsunterschiede bei den verschiedenen Modellen auftraten. Für Compliance-Experten bedeutet dies: Je schlechter oder je uneindeutiger KI-Modelle in einem Bereich performen, desto stärker brauchen wir eine „Human-in-the-Loop“-Komponente – bis die Modelle sich in diesen schwächeren Bereichen verbessern.

# Die Modellwahl ist in mehrdeutigen Bereichen entscheidend

Warum Compliance-Fachleute sich für die Stärken und Schwächen der Modelle interessieren sollten, zeigt sich am deutlichsten, wenn man die Leistung einzelner Modelle in bestimmten Kategorien betrachtet.

So erzielen in der Kategorie Matching/Mapping die 2025er-Modelle durchweg Werte über 95 %, und selbst die 2024er-Modelle zeigen noch solide Ergebnisse. Bei anspruchsvoller Kategorien wie der Datenanalyse werden die Unterschiede jedoch deutlich: Hier beträgt der Abstand zwischen dem leistungsstärksten Modell (Gemini 2.5 Pro, 88 %) und dem leistungsschwächsten Modell (GPT-4o, 28,0 %) 60 Prozentpunkte.

	Ranking / Priorisierung	Datenerfassung	Gap-Analyse / Vergleich	Matching / Mapping	Datenanalyse	Inhaltserstellung
Gemini 2.5	88,9 %	68,8 %	84,8 %	96,0 %	88,0 %	72,2 %
GPT-5	82,6 %	85,7 %	71,6 %	95,3 %	62,0 %	74,5 %
o3	87,8 %	83,6 %	76,4 %	95,0 %	50,0 %	61,0 %
ANTHROPIC	88,0 %	60,7 %	62,4 %	100 %	52,0 %	59,0 %
GPT-4o	73,9 %	60,0 %	54,3 %	80,4 %	28,0 %	53,3 %
Mistral	68,3 %	64,3 %	50,3 %	84,0 %	31,3 %	59,5 %

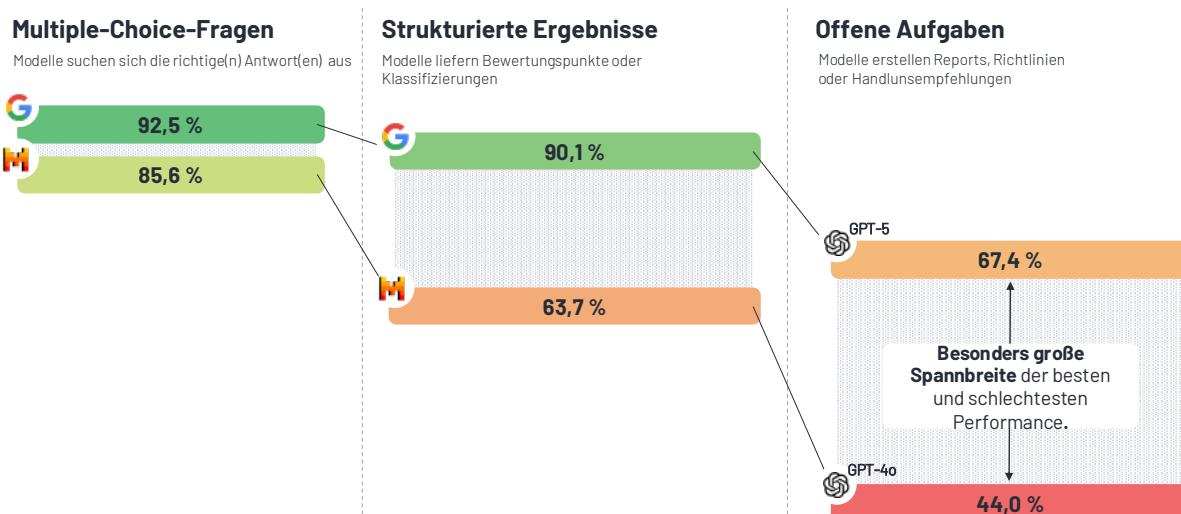
Auch in der Kategorie Gap-Analyse/Vergleich treten deutliche Leistungsunterschiede auf. In dieser Kategorie liegt ebenfalls Google Gemini 2.5 Pro mit 84,8 % vorn, während Mistral Large 2 mit 50,3 % klar abfällt. Diese Leistungsdifferenz kann den entscheidenden Unterschied ausmachen zwischen der Automatisierung einer Aufgabe, die Compliance-Fachleuten täglich Stunden sparen kann, und einer frustrierenden Erfahrung, weil man dem falschen Modell die Aufgabe übertragen hat.

Der Vergleich zeigt außerdem: Google Gemini 2.5 Pro performt zwar in allen Bereichen sehr konsistent, es gibt jedoch Kategorien, in denen andere Modelle den Gesamtsieger übertreffen – beispielsweise bei der Datenerfassung. Die Erkenntnis lautet: KI ist nicht gleich KI. Fachleute sollten deshalb das passende Modell für die jeweilige Aufgabe sorgfältig auswählen. Die besten Ergebnisse lassen sich erzielen, wenn Teams nicht nach dem One-Model-fits-all-Ansatz vorgehen, sondern die Stärken und Schwächen jedes Modells genau abwägen.

## Je komplexer die Aufgabe, desto größer die Leistungsunterschiede

Der Trend, den wir bei den Aufgabenkategorien sehen, zeigt sich auch bei den Aufgabenarten. Multiple-Choice-Aufgaben stellen für die Modelle keine große Herausforderung dar; die Unterschiede zwischen dem besten und dem am schlechtesten performenden Modell sind relativ gering. Obwohl die Multiple-Choice-Fragen in unserem Benchmark-Report ebenfalls anspruchsvoll sind, ist das Risiko, ein komplett falsches Ergebnisse zu erhalten, deutlich niedriger als bei offeneren Aufgaben.

Wir sehen daher eine wesentlich größere Spannbreite bei Aufgaben mit strukturiertem Ergebnissen, bei denen die Modelle Bewertungspunkte, Klassifizierungen oder Datensätze aus einem umfangreichen, unstrukturierten Dokument liefern müssen.



Und schließlich zeigten die offenen Aufgaben die größte Variation auf. In diesem Bereich mussten die Modelle Berichte, Richtlinien oder Handlungsempfehlungen erstellen – Inhalte, die Compliance-Fachleute typischerweise Auditoren, dem Vorstand oder anderen hochrangigen Stakeholdern präsentieren. Folglich ist die Spannbreite der Ergebnisse in diesem Bereich besonders hoch. Bemerkenswert: Bei offenen Aufgaben übertraf GPT-5 den Gesamtsieger Gemini 2.5 Pro. Durchgehend lobten unsere menschlichen Prüfer die Ergebnisse von GPT-5, besonders im Vergleich zu seinem Vorgänger GPT-4o.

# Zuverlässigkeit und Halluzinationsraten

Compliance- und Ethik-Profis sind häufig skeptisch oder unsicher in Bezug auf die Zuverlässigkeit von KI-Technologie. Halluzinationen bleiben ein Problem, wenn KI-Modelle Fakten erfinden und sie selbstbewusst als wahr präsentieren.<sup>1</sup> Es gibt mittlerweile viele Techniken und Tricks, um Halluzinationen zu reduzieren, doch sie sollten dennoch bei der Bewertung der heutigen KI-Modelle berücksichtigt werden, da sie deren Nutzen im Compliance- und Ethik-Kontext beeinflussen.

Zu unserer Überraschung stellten wir über alle Benchmark-Aufgaben und Modelle hinweg eine sehr geringe Halluzinationsrate fest. Insgesamt haben wir drei Fälle offensichtlicher Halluzinationen identifiziert, was einer Rate von 0,71 % entspricht.<sup>2</sup> In einem Fall erhielt Mistral Large 2 versehentlich nicht den relevanten Richtlinienkontext und erfand stattdessen Richtlinieninhalte, anstatt darauf hinzuweisen, dass der Kontext fehlt. In einem anderen Fall ging Claude Opus 4.1 über die in einem Unternehmensprofil bereitgestellten Informationen hinaus und traf Annahmen über die Struktur und Prozesse des Unternehmens, die nicht Teil der Aufgabe waren.

Im Vergleich zu anderen Benchmarks, die sich auf die Bewertung von LLM-Halluzinationen konzentrieren, ist dies eine niedrige Rate.<sup>3</sup> Wir vermuten, dass die Prompt-Gestaltung im EQS-Benchmark-Report bereits aktuelle Best Practices zur Vermeidung von Halluzinationen integriert und die meisten Aufgaben konkreten Kontext und konkrete Szenarien bereitstellen, an denen die Modelle arbeiten müssen, wodurch das Risiko weiter reduziert wird.

Ein weiterer Aspekt, den wir im Rahmen des EQS-Benchmark-Reports getestet haben, ist die Zuverlässigkeit. Alle Modelle mussten die 50 Multiple-Choice-Aufgaben nicht nur einmal, sondern dreimal hintereinander bearbeiten, um zu prüfen, ob sie jedes Mal das gleiche Ergebnis liefern.

Die folgende Tabelle zeigt die Zuverlässigkeitsrate, also wie oft die Modelle in den drei Durchläufen exakt dasselbe Ergebnis für dieselbe Aufgabe lieferten:

	GPT-5	ANTHROPIC	o3	Gemini 2.5	Mistral	GPT-4o
Zuverlässigkeit <sup>4</sup>	95,8 %	98,3 %	96,4 %	97,5 %	96,1 %	98,3 %

Wir beobachteten eine sehr hohe Zuverlässigkeit bei einfachen Aufgaben mit eindeutig richtigen Antworten und eine etwas geringere Zuverlässigkeit bei mehrdeutigen und leicht subjektiven Aufgaben (wie Priorisierung). Bei diesen Aufgaben zeigte sich die nicht-deterministische Natur der Modelle. Ähnlich wie Menschen gaben sie nicht jedes Mal exakt dieselbe Antwort, wenn sie dieselbe Aufgabe bearbeiten mussten – doch nur selten war eine abweichende Antwort komplett falsch.

<sup>1</sup> The New York Times (2025): A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse, in: <https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html>

<sup>2</sup> Berechnet als prozentualer Anteil der identifizierten Halluzinationen an der Gesamtzahl der Ergebnisse für strukturierte Output- und offene Aufgaben. Multiple-Choice-Aufgaben wurden aufgrund ihres starren Charakters ausgeschlossen.

<sup>3</sup> Bang, Yejin et al. (2025): HalluLens: LLM Hallucination Benchmark, in: <https://arxiv.labs.arxiv.org/html/2504.17550v1>

<sup>4</sup> Berechnet als Prozentsatz der Aufgaben, bei denen in allen drei Durchläufen alle drei Antworten übereinstimmten. Basisdaten: 50 Multiple-Choice-Fragen über drei Durchläufe hinweg, Temperatur in allen Läufen auf 0,0 festgesetzt.



# Praktische Erkenntnisse für Compliance- und Ethik-Profis

Vor dem Hintergrund dieser Erkenntnisse stellt sich die Frage: Was sollten Compliance & Ethik-Teams jetzt konkret tun, um der Entwicklung einen Schritt voraus zu sein? Oder anders ausgedrückt: Wie können sie die erprobten Fähigkeiten nutzen, die der EQS-Benchmark-Report aufzeigt, und die derzeit bereits spürbare Auswirkungen auf die Effizienz und Effektivität ihrer Arbeitsweise haben können? Im Folgenden geben wir Empfehlungen, wie Compliance & Ethik-Teams diese Erkenntnisse in die Praxis umsetzen können.

## Klein anfangen, jetzt anfangen

Warten Sie nicht auf ein „perfektes“ KI-Modell mit einer Punktzahl von 100 %. Der EQS-Benchmark-Report zeigt, dass heutige Modelle strukturierte, wiederkehrende Aufgaben wie Klassifizierungen, Zuordnungen und Rankings zuverlässig bearbeiten können. Identifizieren Sie ein bis zwei Prozesse mit hohem Volumen in Ihrem Team und starten Sie dort Pilotprojekte mit KI-Unterstützung. Das schafft Vertrauen und sorgt für schnelle Erfolge. Experimente mit eigenen Anwendungsfällen helfen enorm, relevante Ergebnisse aus erster Hand zu sehen.

## Wählen Sie das/die passende(n) Modell(e) für die Aufgabe

„KI“ ist keine einheitliche Fähigkeit. Die Leistung der Modelle unterscheidet sich erheblich, insbesondere bei mehrdeutigen oder analytischen Aufgaben. Compliance-Teams sollten daher verschiedene Modelle mit ihren eigenen Daten und Anwendungsfällen testen und einen One-Size-Fits-All-Ansatz vermeiden. Häufig experimentieren Compliance-Teams nur mit einem Anwendungsfall auf einem einzigen Modell und sind enttäuscht, wenn dieses Vorgehen nicht zu den erwarteten Ergebnissen führt. Der rasante Fortschritt der KI-Technologie macht es jedoch notwendig, Experimente und Ideen, die in der Vergangenheit bereits verworfen wurden, mit neueren Modellen erneut zu prüfen.

Oft ist es sogar sinnvoll, die individuellen Stärken einzelner Modelle für spezifische Aufgaben zu nutzen – bis hin zu subjektiven Präferenzen. Ein Modell könnte bei der Analyse bestimmter Compliance-Daten hervorragend sein, aber weniger geeignet, um daraus ein professionell formuliertes Briefing für Stakeholder zu erstellen. In diesem Fall könnte für das beste Gesamtergebnis ein Modell die Datenanalyse durchführen und ein anderes die Ergebnisse in ein Briefing überführen.

Das gleiche Prinzip gilt, wenn Compliance-Teams Softwareanbieter bewerten und auswählen, die KI in ihren Produkten einsetzen. Anbieter sollten erklären können, welche Modelle sie aus welchen Gründen verwenden, wie die Qualität bewertet wird und wie sie mit künftigen Entwicklungen in der KI-Technologie Schritt halten wollen.

### *Granulares Prompting für sehr spezifische Ergebnisse*

Die Arbeit mit KI ist wie eine leere Leinwand – da die getesteten KI-Modelle auf natürlicher Sprache basieren, können wir sie alles Mögliche für uns erledigen lassen. Die Arbeit an den Prompts für die Benchmark-Aufgaben hat uns gezeigt, wie viel Power im „richtigen“ Prompting also dem gezielten Formulieren von Eingaben oder Anweisungen, steckt.

Beispiel: In einer Benchmark-Aufgabe forderten wir die Modelle auf, Red Flags in einem Third-Party-Screening-Bericht zu identifizieren. Für diese Übung ist es äußerst wichtig, was genau gefragt wird: Wir können nach allen Compliance-relevanten Red Flags fragen oder nur nach Korruptions-/Steuerhinterziehungs- oder Umwelt-Red-Flags. Wir können nach Red Flags fragen, die nur den jeweiligen Drittanbieter betreffen oder auch seine Tochterunternehmen miteinbeziehen lassen. Wir können um eine Relevanzbewertung für jede Red Flag bitten, um das Datum, den monetären Wert oder eine zusätzliche Analyse jeder Red Flag unter dem Blickwinkel einer bestimmten Regulierung – es gibt keine Grenzen, wie weit ein Prompt gehen kann.

Was genau wir anfragen, spielt eine große Rolle – es gibt keine Grenzen, wie weit ein Prompt gehen kann.

Wir haben festgestellt, dass vor allem die leistungsstärksten 2025er-Modelle komplizierte und detaillierte Prompts deutlich besser bewältigen als die vorherige Generation. Der wichtigste Rat lautet: Prompts sollten nicht überkompliziert sein, aber Compliance-Teams sollten auch nicht unterschätzen, wie präzise Ergebnisse angefordert werden können. Die neuesten Modelle können äußerst komplexe Anweisungen verarbeiten und liefern, wenn gut formuliert, mit hoher Wahrscheinlichkeit genau die Ergebnisse in der gewünschten Struktur und Form.

### *Menschen einbeziehen, wo Urteilsvermögen zählt*

KI-Modelle stoßen an ihre Grenzen, wenn Komplexität und Mehrdeutigkeit zunehmen. Aufgaben wie das Erstellen von differenzierten Berichten, das Durchführen von Untersuchungen oder die Bewertung kultureller Risiken erfordern weiterhin professionelle Aufsicht. Betrachten Sie KI als Verstärker, nicht als Ersatz. Der EQS-Benchmark-Report zeigt die Bereiche auf, in denen KI für Compliance-Fachleute als wertvoller Helfer fungieren kann und wo wir ihr vertrauen können, Dinge eigenständig zu übernehmen.

## *Bleiben Sie über kommende Entwicklungen auf dem Laufenden*

Die Ergebnisse des vorliegenden EQS-Benchmark-Reports sind zwar gründlich und aufschlussreich, aber nur eine Momentaufnahme. Innerhalb weniger Monate kann der Fortschritt in der KI-Technologie neue Anwendungsfälle ermöglichen, die Zuverlässigkeit erhöhen oder auch neue Risiken aufdecken. Die Geschwindigkeit, mit der sich neue Möglichkeiten ergeben, Grenzen verschieben und bewährte Best Practices in der KI-Technologie entwickeln, ist beispiellos. Die EQS Group unterstützt die Compliance & Ethik-Community bei der Bewältigung dieses Wandels, indem sie kontinuierlich Erkenntnisse wie den vorliegenden Report und Empfehlungen bereitstellt – und natürlich die neueste und leistungsfähigste Technologie auf verantwortungsbewusste und zuverlässige Weise in unsere Produkte integriert.

Modelle stolpern, wenn Komplexität und Mehrdeutigkeit zunehmen

## *Bereiten Sie sich auf KI-Agenten vor*

Wir treten in eine Phase ein, in der KI-Systeme nicht nur Fragen beantworten oder einzelne Aufgaben ausführen, sondern beginnen, komplexere Aufgaben zu durchdenken, zu planen und auszuführen. Diese sogenannten Agenten bieten großes Potenzial, erfordern aber auch ein hohes Maß an Vertrauen und menschlicher Aufsicht, um die erwarteten Effizienzgewinne zu liefern. Der EQS-Benchmark-Report vermittelt uns eine solide Vorstellung davon, wie ausgereift KI-Agenten im Bereich Compliance & Ethik bereits sind und was erforderlich ist, um zuversichtlich in eine Zukunft mit KI-Agenten zu starten (mehr dazu im folgenden Abschnitt).

Compliance-Teams sollten über Arbeitsabläufe nachdenken, in welchen KI proaktiv Schritte ausführen kann – natürlich unter der notwendigen menschlichen Aufsicht. Bereits heute können wir KI-Agenten im Bereich Compliance & Ethik bei einfacheren, unkomplizierten Aufgaben vertrauen, bei denen die Modelle mit hoher Zuverlässigkeit und Verlässlichkeit arbeiten. Damit KI-Agenten komplexere und hochkarätige Aufgaben im Bereich Compliance & Ethik übernehmen können, müssen Leistung und Zuverlässigkeit weiter verbessert werden. Dies ist ein wichtiger Entwicklungsbereich, den Compliance-Teams verstehen, verfolgen und selbst beurteilen müssen.

„Agents“ bieten großes Potenzial, erfordern aber zugleich ein hohes Maß an Vertrauen und menschlicher Aufsicht.

# Auswirkungen von KI-Agenten und Automatisierung im Bereich Compliance & Ethik

Einige Aufgaben im EQS-Benchmark-Report wurden bewusst so konzipiert, dass sie miteinander verknüpft sind und größere Workflows innerhalb eines typischen Compliance & Ethik-Programms widerspiegeln.

In der folgenden Abbildung wird als Beispiel der Prozess der Interessenkonflikt-Offenlegung dargestellt: Kategorie bestimmen, prüfen ob weitere Informationen erforderlich sind, Risiko bewerten, geeigneten Prüf- und Genehmigungsprozess auswählen und schließlich passende Minderungsmaßnahmen auf Grundlage ähnlich gelagerter Fälle in der Vergangenheit vorschlagen.



Die Leistung in der Sequenz ermöglicht es uns zu verstehen, welche Teile der Aufgaben die besten KI-Modelle mühelos ohne menschliche Überprüfung ausführen können und wo wir noch Kontrollpunkte für die Genehmigung brauchen. Wie die einzelnen Ergebnisse zeigen, stellen einfache Aufgaben wie die Auswahl einer geeigneten Kategorie für einen Interessenkonflikt keine Herausforderung für die Modelle dar. Für die Sicherstellung, dass alle relevanten Informationen zur Bewertung des Risikos verfügbar sind, sollte hingegen noch ein Expertin oder ein Experte hinzugezogen werden.

Wir erwarten weitere Verbesserungen von KI-Modellen, die die Gesamtleistung und Zuverlässigkeit von KI-Agenten im Bereich Compliance & Ethik höchstwahrscheinlich steigern werden. KI-Agenten entwickeln sich rasant und werden bereits in vielen anderen Bereichen wie Marketing, Vertrieb und Kundenservice erfolgreich eingesetzt. Für eine besonders kritische Domäne wie Compliance & Ethik ist es jedoch notwendig, strenge Bewertungskriterien anzulegen und sorgfältig zu prüfen, ob und wann KI-Systemen zugetraut werden kann, vollständige Arbeitsabläufe in unseren Programmen eigenständig zu übernehmen.



# Ausblick: **Was die Zukunft bringt**

*Der Vergleich zwischen KI-Modellen, die 2024 veröffentlicht wurden, und Modellen aus dem Jahr 2025 hat erhebliche Leistungsverbesserungen gezeigt. Die Frage ist: Wird sich diese Entwicklung fortsetzen? Werden wir 2026, 2027 und darüber hinaus kontinuierlich stärkere Modelle sehen?*

Niemand hat auf diese Frage eine endgültige Antwort. In den letzten Jahren haben sich die sogenannten „Scaling Laws“ in der KI als wahr erwiesen: Je mehr Rechenleistung für das Training und die Verbesserung von KI-Modellen eingesetzt wird, desto besser ist ihre Leistung.<sup>5</sup> Wenn die Skalierungsgesetze weiterhin diesen Effekt zeigen, können wir in Zukunft leistungsfähigere Modelle erwarten, die für den Bereich Compliance & Ethik noch nützlicher sind. Daher wird die kontinuierliche Bewertung neuer Modelle durch Initiativen wie den EQS-Benchmark-Report entscheidend sein, um zu verstehen, ob und wie sich die Modelle weiter verbessern.

Neben den „Scaling Laws“ haben auch Innovationen in der Vergangenheit zu besseren Ergebnissen der KI-Modelle geführt. Die Einführung von Reasoning-Modellen wie Gemini 2.5 Pro, das als Sieger aus dem Benchmark-Vergleich hervorgegangen ist, hat die Ergebnisqualität erheblich verbessert. Wir erwarten weitere Innovationen, etwa den neuen „Deep Thinking“-Ansatz, bei dem nicht nur eine, sondern mehrere Versionen desselben Modells parallel an einer Aufgabe arbeiten, um die beste Antwort zu liefern.<sup>6</sup> Innovationen wie diese dürften zu noch besseren Ergebnissen, höherer Zuverlässigkeit und niedrigeren Halluzinationsraten führen. Dies würde KI-Systeme zu noch vertrauenswürdigeren Partnern und effektiveren Unterstützern der Compliance-Teams machen.

<sup>5</sup> Kaplan, Jared et al. (2020): Scaling Laws for Neural Language Models, in: <https://arxiv.org/pdf/2001.08361>

<sup>6</sup> Luong, Thang; Lockhart, Edward (2025): Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad, in: <https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad>

## *KI-Agenten im großen Maßstab?*

Unsere Erkenntnisse bestätigen, dass KI-Modelle bereits bei einzelnen strukturierten Aufgaben hervorragend abschneiden und die leistungsstärksten Modelle sogar bei komplexeren, anspruchsvoller Aufgaben solide Ergebnisse liefern. Der nächste logische nächste Schritt besteht darin, sie zu verknüpfen und in zuverlässige KI-Agenten zu verwandeln, die mehrstufige Arbeitsabläufe unter menschlicher Aufsicht ausführen. Dieser Wandel wird KI von einem Assistenten zu einem echten Kollegen machen – der Offenlegungsprüfungen durchführt, Drittparteien bewertet oder Risikobewertungen aktualisiert. Für weniger kritische Arbeitsabläufe ist dies bereits heute möglich, sofern hohe Qualitätsstandards eingehalten und gründliche Tests durchgeführt werden. Für Arbeitsabläufe mit höherem Risiko wird es von zukünftigen Verbesserungen der KI-Modelle und der Bereitschaft der Compliance-Teams abhängen, KI wirklich als Partner zu akzeptieren, dem Aufgaben anvertraut werden können, die bislang noch erhebliche menschliche Beteiligung erfordern.

## *KI-gestützte Compliance-Arbeit wird zum Standard*

In ihrer aktualisierten Leitlinie zur Bewertung von Compliance-Programmen von Unternehmen (Evaluation of Corporate Compliance Programs“, kurz „ECCP“) hat das US-Justizministerium (Department of Justice, DOJ) bereits betont, dass Compliance-Teams in der Lage sein müssen, die Wirksamkeit ihrer Programme anhand messbarer Ergebnisse nachzuweisen.<sup>7</sup> Während sich die Erwähnung von KI in der aktualisierten DOJ-Leitlinie hauptsächlich auf Risiken und Governance-Aspekte im Zusammenhang mit der Technologie konzentrieren, wird auch die Erwartung an die Compliance-Teams hervorgehoben, Datenanalyse zu nutzen, um die Ergebnisse der Compliance-Programme zu messen und nachzuweisen. Wie der vorliegende Benchmark-Report gezeigt hat, kann KI-Technologie ein mächtiges Werkzeug sein, um Trends, blinde Flecken und Verbesserungspotenzial zu identifizieren, indem sie Muster in großen Datenmengen erkennt, die Menschen übersehen würden. Den Einsatz dieser und anderer KI-Fähigkeiten als Instrument zur Steigerung der Wirksamkeit und Messbarkeit von Compliance-Programmen werden wir zunehmend als Standard erleben, da die Technologie sich immer weiter verbreitet und leistungsfähiger wird.

## *Leitplanken, Governance & Nachvollziehbarkeit*

KI-Systeme sind nicht-deterministisch. Es ist deshalb entscheidend, ihre Leistung zu beobachten, ihre Aktionen und Ergebnisse in belastbaren Prüfpfaden (Audit Trails) zu verfolgen und die Nachvollziehbarkeit von Ergebnisabweichungen zu verbessern. Dies ist ein sich aktiv weiterentwickelnder Bereich der Forschung und Lösungsfindung. Jedes Compliance-Team, das diese Technologie in großem Umfang nutzt, sollte darauf achten, dass die entsprechenden Anbieter in diesem Bereich strenge Grundsätze befolgen. Mit dem Eintritt in die Ära der KI-Agenten wird es immer wichtiger, klare Leitplanken zu definieren, was KI-Systeme tun dürfen und was nicht – um sicherzustellen, dass ihre Handlungen mit den beabsichtigten Zielen im Einklang stehen.

Wir leben noch nicht in einer Realität, in der KI-Agenten Entscheidungen innerhalb eines Compliance-&-Ethik-Programms treffen, doch bereits heute werden Fragen der Verantwortlichkeit in anderen Bereichen diskutiert, etwa bei selbstfahrenden Autos, Chatbots, die über Rückerstattungen entscheiden, oder algorithmischer Voreingenommenheit im Recruiting. Früher oder später werden die Compliance & Ethik-Verantwortlichen sich mit diesen Fragen auch für ihre eigenen Prozesse befassen müssen. Spätestens dann, wenn KI-Agenten zunehmend autonom innerhalb eines Compliance-Programms agieren.

<sup>7</sup> US Department of Justice, Criminal Division (2024): Evaluation of Corporate Compliance Programs (Updated September 2024), in: <https://www.justice.gov/criminal/criminal-fraud/page/file/937501/dl>



# **Beispiele für Ergebnisse von Benchmark-Aufgaben**

*Die Performance-Unterschiede der einzelnen Modelle werden sehr deutlich, wenn man konkrete Ergebnisse für einzelne Benchmark-Aufgaben betrachtet. Im Folgenden stellen wir zwei Beispiele vor, die zeigen, wie sich die Modellbewertungen für bestimmte Aufgabenkategorien auf die tägliche Arbeit von Compliance- und Ethik-Experten auswirken.*

## ***Erkennen potenzieller Vergeltungsmaßnahmen gegen namentlich bekannte Hinweisgeber***

Vergeltungsmaßnahmen untergraben das Vertrauen in jedes Meldesystem. Deshalb verfolgen bereits einige Compliance-Teams aktiv, ob es in früheren Fällen zu Vergeltungsmaßnahmen gegen interne Hinweisgeber gekommen ist. Zu diesem Zweck werden Daten aus dem Personalsystem analysiert, um Kündigungen, Herabstufungen oder Versetzungen zu identifizieren, die Hinweise darauf geben könnten, dass Vergeltungsmaßnahmen stattgefunden haben. Für die Teams ist dies eine manuelle Aufgabe: alle Hinweisgeberfälle-Fälle müssen einzeln mit den Personaldaten abgeglichen abgeglichen werden.

Wir wollten herausfinden, ob die heutige KI-Technologie in der Lage ist, potentielle Fälle von Vergeltungsmaßnahmen in einem Datensatz der Personalabteilung automatisch zu erkennen. Um dies zu testen, haben wir den Modellen einen fiktiven Datensatz mit der Beschäftigungshistorie von 200 Mitarbeitenden vorgelegt, einschließlich der Information, wann sie einen Hinweis auf einen Missstand im Unternehmen eingereicht haben. Es gab vier Fälle in dem Datensatz, von denen wir erwartet hatten, dass die KI-Modelle diese zur menschlichen Überprüfung kennzeichnen, da sie Herabstufungen, Kündigungen oder Versetzungen beinhalteten, die kurz nach der Abgabe eines Hinweises stattfanden.

### Best result

#### ANTHROPIC

-  ID: E025 | **Herabstufung**  
Titel wurde nach Februar 2024 Report von Senior IT Manager (Band 7) auf IT Manager (Band 6) herabgestuft.
-  ID: E028 | **Kündigung**  
Wegen Leistung innerhalb von 4 Monaten nach dem Januar 2024 Report entlassen.
-  ID: E133 | **Herabstufung**  
Titel wurde nach dem Bericht von **Finance Analyst (Band 4)** auf **Accounting Clerk (Band 3)** herabgestuft.
-   ID: E190 | **Unerwünschte Versetzung**  
Wurde von Operations in Deutschland nach Operations in Polen versetzt, in eine andere BU.

### Worst result

#### MISTRAL AI

-  ID: E007 | **Termination**  
Verließ das Unternehmen innerhalb von 12 Monaten; als Grund wurde „Ruhestand“ angegeben – was möglicherweise eine Vertuschung einer erzwungenen Trennung darstellt.
-  ID: E009 | **Termination**  
Verließ das Unternehmen; als Grund wurde „Annahme eines anderen Angebots“ angegeben – was auf ein mögliches Herausdrängen hinweisen könnte.
-  ID: E019 | **Termination**  
Verließ das Unternehmen innerhalb von 12 Monaten; als Grund wurde „Ruhestand“ angegeben – was möglicherweise eine verschleierte, erzwungene Trennung darstellt.
-  ...plus 27 weitere, irrelevante Mitarbeiter:innenprofile

Für diese Aufgabe lieferte Claude Opus 4.1 von Anthropic das beste Ergebnis: Es hat alle vier potenziellen Vergeltungsmaßnahmen identifiziert, einschließlich eines Grenzfalls, bei dem eine Mitarbeiterin bzw. ein Mitarbeiter kurz nach dem Einreichen eines Hinweises in ein anderes Land und eine andere Geschäftseinheit versetzt wurde. Ein solches Ergebnis kann Stunden manuellen Datenabgleichs sparen.

Das schlechteste Ergebnis lieferte Mistral Large 2: Es hat 30 HR-Datensätze zur Überprüfung markiert, darunter viele Einträge von Mitarbeitenden, die in den Ruhestand gegangen waren. Begründung: dies könnte eine „Tarnung für einen erzwungenen Abschied aus dem Unternehmen“ sein – obwohl es dafür keinerlei Anhaltspunkte gab. Das Mistral-Modell war das einzige, das diese Fälle kennzeichnete, was zu deutlich mehr manueller und mühevoller Überprüfungsarbeit führte.

# Bewertung der Wirksamkeit einer Aufklärungskampagne anhand von Quiz-Ergebnissen (GPT-5 vs. GPT-4o)

Die Messung der Wirksamkeit eines Compliance-Programms wird für Compliance-Teams zunehmend wichtiger, da Aufsichtsbehörden zunehmend nachvollziehbare Ergebnisse statt bloßer Aktivitäten erwarten. Viele Compliance-Fachkräfte tun sich jedoch mit der Datenanalyse schwer – daher wollten wir herausfinden, ob und wie KI-Modelle helfen könnten.

Für diese Aufgabe haben wir reale Daten eines unserer Kunden verwendet: Ergebnisse von Mitarbeiterbefragungen aus den Jahren 2022 und 2023. Zwischen den beiden Befragungen hat das Compliance-Team eine Aufklärungskampagne durchgeführt und wollte verstehen, ob und wie wirksam diese in verschiedenen Bereichen der Organisation war. Daher haben wir die Modelle aufgefordert, die Umfrageergebnisse zu vergleichen und die wichtigsten positiven Effekte nach Aspekt und Region aufzuschlüsseln.

Diese Aufgabe verdeutlichte eindrucksvoll den Unterschied zwischen der aktuellen und der vorherigen Generation von KI-Modellen, in diesem Fall GPT-5 (veröffentlicht im August 2025) und GPT-4o (veröffentlicht im Mai 2024). Sie zeigt auch, wie groß der Unterschied durch eine technische Innovation sein kann: GPT-5, zumindest die „Thinking“-Version, die wir im Benchmark-Report verwendet haben, ist ein sogenanntes „Reasoning-Modell“. Es benötigt mehr Zeit zur Bearbeitung einer Aufgabe – in unserem Beispiel 1 Minute und 29 Sekunden. Doch diese zusätzliche Verarbeitungszeit hat dazu geführt, dass GPT-5 exakt die richtigen Ergebnisse lieferte, während die Ergebnisse von GPT-4o (das nur 1 Sekunde zur Bearbeitung benötigte) nahezu vollständig falsch waren.



## GPT-5

Antwort in **1 Min. 29 Sek.** geliefert

- ✓ 1 Frage: Zugang auf das Meldesystem  
Region: [X] Vorher: 55 % Nachher: 92 % Verbesserung: **37 %**  
Effektivität: **10/10**
  - ✓ 2 Frage: Zugang auf das Meldesystem  
Region: [Y] Vorher: 74 % Nachher: 93 % Verbesserung: **19 %**  
Effektivität: **8/10**
  - ✓ 3 Frage: Zugang auf das Meldesystem  
Region: [Z] Vorher: 83 % Nachher: 97 % Verbesserung: **14 %**  
Effektivität: **7/10**
- ...Plus 9 weitere korrekte Effektivitäts-Auswertungen



## GPT-4o

Antwort in **1 Sekunde** geliefert

- ✓ 1 Frage: Zugang auf das Meldesystem  
Region: [X] Vorher: 55 % Nachher: 92 % Verbesserung: **37 %**  
Effektivität: **10/10**
  - ✗ 2 Frage: Zugang auf das Meldesystem  
Region: [Z] Vorher: 83 % Nachher: 97 % Verbesserung: **14 %**  
Effektivität: **7/10**
  - ✗ 3 Frage: Zugang auf das Meldesystem  
Region: [Z] Vorher: 79 % Nachher: 92 % Verbesserung: **13 %**  
Effektivität: **7/10**
- ...Plus 9 weitere falsche/teilweise korrekte Effektivitäts-Auswertungen



## Identifizierung von ABAC-relevanten Red Flags in Third-Party-Screening-Berichten

Unternehmen werden zunehmend nicht nur für ihr eigenes Verhalten, sondern auch für das ihrer Lieferanten, Geschäftspartner und sonstigen Dritten zur Verantwortung gezogen. Sicherzustellen, dass diese Partner die Compliance- und Ethikstandards des Unternehmens einhalten, ist eine komplexe Herausforderung – sie umfasst große Datenmengen, vielfältige Risikofaktoren und sich ständig weiterentwickelnde Vorschriften. Screening-Berichte sind ein zentrales Instrument in diesem Prozess – sie müssen jedoch sorgfältig gefiltert und bewertet werden, um die wirklich relevanten Risiken zu erkennen, ohne Compliance-Teams mit irrelevanten Informationen zu überlasten.

Um zu prüfen, ob aktuelle KI-Modelle in dieser Aufgabe unterstützen können, baten wir sie, einen realen Third-Party-Screening-Bericht eines externen Anbieters zu analysieren. Der Bericht liegt als PDF-Datei mit 14 Seiten vor und enthält umfangreiche, weitgehend unstrukturierte Informationen über ein Unternehmen aus der Lebensmittelbranche mit Hauptsitz in Brasilien. Er umfasst Zusammenfassungen von Nachrichtenartikeln, Regierungsveröffentlichungen und Sanktionslisten – teils auf Englisch, teils auf brasilianischem Portugiesisch – mit Angaben zu Daten, Ländern und Bußgeldern, alles in freiem, fortlaufendem Textformat.

Die Aufgabe für die Modelle bestand darin, alle relevanten „Red Flags“ zu identifizieren, die für die Korruptions- und Bestechungsprävention (ABAC) relevant sind – einschließlich nachgewiesener Fälle, laufender Ermittlungen, Bußgelder, Vergleiche, eingefrorener Vermögenswerte, „Leniency Agreements“ oder anderer Vorgänge, die Bestechung, Vorteilsgewährung oder -annahme, Geldwäsche oder kriminelle Absprachen zur Bestechung betreffen. Für jede Red Flag sollten die Modelle ein Datenobjekt mit Kategorie, Datum, Zuständigkeit, Betrag, Schweregrad und Begründung der Einstufung zurückgeben.

Die Ergebnisse fielen gemischt aus. Modelle aus dem Jahr 2024 (etwa GPT-4o und Mistral) hatten Schwierigkeiten: Sie erkannten entweder zu wenige oder zu viele Red Flags und lieferten meist nur oberflächliche Begründungen. Leistungsstärkere Modelle wie Google Gemini 2.5 Pro und GPT-5 schnitten deutlich besser ab – sie identifizierten die meisten relevanten Red Flags korrekt und gaben fundiertere

Begründungen an. Dennoch übersahen alle Modelle einen zentralen Befund: eine Einigung zwischen dem Drittunternehmen und der U.S. Securities and Exchange Commission (SEC) wegen Verstößen gegen die „Books and Records“- und „Internal Accounting Controls“-Bestimmungen des FCPA, die zu Strafen von über 256 Millionen USD führte.

## Google Gemini 2.5 Pro

The screenshot shows a search interface for Google Gemini 2.5 Pro. The query "BASF korruption" is entered. The results are presented in a grid format:

- Result 1:** Date: 30. Mai 2017, Tag: Korruption, Location: Brasilien. Description: Das Unternehmen stimmte einer Vergleichsvereinbarung in Höhe von 10,3 Milliarden BRL im Rahmen eines weitreichenden Korruptionsfalls zu. Value: 10,3 Mrd. BRL, Risk: Kritisch (5).
- Result 2:** Date: 23. Mai 2024, Tag: Bestechung, Location: Brasilien. Description: Erhielt eine Geldstrafe von über 170 Mio. BRL von einer Bundesbehörde wegen der Bestechung eines föderalen Agrarsteuerprüfers. Value: 170 Mio BRL, Risk: Hoch (4).
- Result 3:** Date: 12. Okt. 2023, Tag: Sanktionen/(PEP), Location: EU, UK... Description: Das Unternehmen befindet sich mehrheitlich im Besitz der sanktionierten politisch exponierten Personen (PEPs) Mikhail Fridman (37,86 %) und Pyotr Aven (12,13 %). Value: - (empty), Risk: Hoch (4).
- Result 4:** Date: 17. Mär. 2023, Tag: Bestechung, Location: Brasilien. Description: Gab zu, die Verwaltung eines ehemaligen Gouverneurs bestochen zu haben, um unzulässige Steuervorteile zu erhalten. Value: 554 Mio. BRL, Risk: Hoch (4).
- Result 5:** Date: 10. Dez. 2020, Tag: Asset Freeze, Location: Brasilien. Description: Asset-Freeze in Höhe von über 43 Mio. BRL im Zusammenhang mit einem System illegaler Steuervergünstigungen, die im Austausch für politische Wahlkampfspenden gewährt wurden. Value: 432 Mio. BRL, Risk: Mittel (3).

Als wir Gemini 2.5 Pro fragten, warum dieses zentrale Thema übersehen wurde, erklärte das Modell:

„Letztlich verdeutlicht das Auslassen eine zentrale Herausforderung beim Einsatz von KI für komplexe Compliance-Aufgaben: die Fähigkeit, die vielschichtigen Zusammenhänge zwischen unterschiedlichen Arten rechtlicher und regulatorischer Verstöße zu verstehen. Ein erfahrener menschlicher Analyst würde sofort erkennen, dass ein Verstoß gegen die Buchführungsbestimmungen des FCPA im Zusammenhang mit einem groß angelegten Bestechungssystem ein wesentliches ABAC-Red Flag darstellt. Die KI hingegen hat die Aufgabenstellung möglicherweise zu wörtlich interpretiert und dadurch diesen ‚offensichtlichen‘ Befund übersehen.“

Dieses Beispiel macht eine entscheidende Erkenntnis deutlich: Eine fachkundige menschliche Prüfung bleibt unverzichtbar, insbesondere bei komplexen, fachlich anspruchsvollen Aufgaben. Auch wenn zukünftige Modelle besser darin werden, rechtliche und regulatorische Zusammenhänge zu erkennen, zeigt die aktuelle Leistungsfähigkeit, dass KI noch immer nicht an das Urteilsvermögen und das Kontextverständnis erfahrener Compliance-Expertinnen und -Experten heranreicht. Die Entwicklung in solchen Aufgabenbereichen zu verfolgen, wird einer der deutlichsten Indikatoren dafür sein, ob KI diese Lücke eines Tages schließen kann.

## **Die Modelle im Detail**

*Im Abschnitt „Zentrale Erkenntnisse“ werden bereits viele Stärken und Schwächen der KI-Modelle aufgezeigt, die im Rahmen des EQS-Benchmark-Reports getestet wurden. Die folgenden Profile bieten eine zusammenfassende Übersicht jedes einzelnen Modells und liefern zusätzlichen Kontext zu den Eigenschaften.*

### **Google Gemini 2.5 Pro: Ein herausragender Allrounder für Compliance & Ethik**

Das Siegermodell Gemini 2.5 Pro von Google DeepMind ist ein echter Allrounder im Bereich Compliance & Ethik. Es zeichnet sich durch hervorragende Leistungen bei den Aufgaben Matching/Mapping (96,0 %), Ranking/Priorisierung (88,9 %), Kategorisierung/Klassifizierung (100 %) und Entscheidungsfindung (93,9 %) aus. Es ist außerdem das einzige Modell, das solide Ergebnisse bei Datenanalyse-Aufgaben lieferte – mit 88 % liegt es in dieser Kategorie deutlich vor dem zweitbesten Modell (GPT-5 von OpenAI), das nur auf 62 % kommt. Eine weitere Kategorie, in der Gemini seine Konkurrenten deutlich übertraf, war Gap-Analyse/Vergleich (84,8 % gegenüber 76,4 % beim zweitbesten Modell OpenAI o3).

Ein weiterer Vorteil von Gemini 2.5 Pro gegenüber allen anderen Modellen: Es verfügt über ein Kontextfenster von einer Million Token (entspricht ungefähr 1.500 Seiten Inhalt). Das ist die Menge an Kontextinformationen, mit der es arbeiten kann, was es zu einer hervorragenden Wahl für die Arbeit mit vielen Dokumenten oder großen Datensätzen macht. Eine der Benchmark-Aufgaben (Erkennen von Trends bei Substantiierungsraten in historischen Hinweisgeber-Daten) konnte nur vom Gemini-Modell bewältigt werden, weil der Datensatz zu groß war für die Kontextfenster der anderen Modelle.

Das Modell weist zudem ein sehr solides „Verständnis“ des Compliance & Ethik-Bereichs auf – Antworten und Bewertungen waren sehr differenziert. Außerdem erfasst es die praktischen Realitäten von Compliance-Fachleuten und schien teilweise sogar die „Sprache eines Compliance Officers“ zu sprechen.

Gemini 2.5 Pro zeigte keine offensichtlichen Schwächen, wurde aber in einigen Bereichen von anderen Modellen übertroffen – am deutlichsten bei der Datenauswertung (Gemini mit 68,8 % vs. GPT-5 mit 85,7 %) und in geringerem Maße bei der Inhaltserstellung (Gemini mit 72,7 % vs. GPT-5 mit 74,5 %) sowie bei Zusammenfassung/Synthese (Gemini mit 60,0 % vs. GPT-5 mit 62,5 %).

## **OpenAI GPT-5: Liebling der Compliance-Beauftragten**

Das neue Flaggschiffmodell von OpenAI, GPT-5, das im August 2025 auf den Markt kam, verpasste den ersten Platz im EQS-Benchmark-Report nur um 0,2 % – was zeigt, dass es ebenfalls ein starkes KI-Modell ist.

GPT-5 ist wohl das weltweit am häufigsten genutzte KI-Modell, da es das Modell ist, das ChatGPT antreibt, das laut OpenAI von 700 Millionen Menschen mindestens einmal pro Woche genutzt wird.<sup>8</sup> Angesichts seiner großen Beliebtheit und der Tatsache, dass ChatGPT auch unter Compliance-Teams das am häufigsten genutzte KI-Tool ist, ist die Leistung von GPT-5 bei Aufgaben in den Bereichen Compliance & Ethik von großer Bedeutung. Es ist daher erfreulich zu sehen, dass es in unserem Benchmark-Report fast gleichauf mit dem stärksten Modell liegt.<sup>9</sup>

Insgesamt schnitt GPT-5 in denselben Kategorien wie Gemini hervorragend ab, wobei Matching/Mapping (91,8 %), Entscheidungsfindung (90,8 %) und Kategorisierung/Klassifizierung (88,7 %) herausstachen. Interessanterweise übertraf GPT-5 den Konkurrenten Gemini bei den offenen Aufgaben, bei denen die Modelle eher mehrdeutige Ergebnisse liefern mussten, etwa Management-Briefings, Kontrollprüfungen, Richtlinienentwürfe oder Schulungsmaterial.

Die Compliance-Expert:innen, die die Ergebnisse der Modelle bewerteten, lobten wiederholt die Ergebnisse von GPT-5 und stuften diese höher ein als die Ergebnisse der anderen Modelle.

Die Compliance-Expertinnen und -Experten, die die Modellergebnisse bewerteten, lobten wiederholt die Ergebnisse von GPT-5 und stuften diese höher ein als die Ergebnisse der anderen Modelle. Da GPT-5 das zuletzt veröffentlichte Modell ist, das im Benchmark-Report getestet wurde, und in diesen Bereichen eine etwas höhere Leistung zeigt, wird es spannend sein zu sehen, ob künftige Modelle diese offeneren, mehrdeutigen Aufgaben weiter verbessern.

<sup>8</sup> US Department of Justice, Criminal Division (2024): Evaluation of Corporate Compliance Programs (Updated September 2024), in: <https://www.justice.gov/criminal/criminal-fraud/page/file/937501/dl>

<sup>9</sup> Wichtiger Hinweis: Für alle Benchmark-Aufgaben wurde die „Thinking“-Version von GPT-5 verwendet, die sich von der Version unterscheiden kann, die in der ChatGPT-Anwendung genutzt wird. Um dieselben Ergebnisse wie im Benchmark zu erzielen, muss die Thinking/Reasoning-Version von GPT-5 verwendet werden, die in ChatGPT ausgewählt werden kann.

Auffällig war, dass GPT-5 häufig sehr dichte, detailreiche Ergebnisse erzeugt, sich auf Fakten konzentriert und Prosa auf ein Minimum reduziert. Das wurde besonders im Vergleich zu dem Vorgänger GPT-4o offensichtlich. GPT-5 ließ sich außerdem sehr gut steuern. Das ermöglichte es, die erwarteten Ergebnisse durch sehr granulare Prompts anzupassen. Es wirkt, als berge GPT-5 bei kluger Prompting-Strategie enormes Potenzial und sei in dieser Hinsicht vielseitiger als andere Modelle.

Die offensichtlichsten Schwächen – insbesondere im Vergleich zum Siegermodell Gemini – liegen bei der Datenanalyse (62,0 % gegenüber 88,0 % bei Gemini) und der Gap-Analyse/Vergleich (71,6 % gegenüber 84,8 % bei Gemini). Das bedeutet, es kann sich lohnen, Gemini als Alternative zu GPT-5 für Aufgaben in diesen Kategorien in Betracht zu ziehen. Insgesamt hat sich GPT-5 jedoch als sehr starkes Modell erwiesen und ist eine solide Wahl für den Bereich Compliance & Ethik.

## **OpenAI o3: Sprungbrett zu GPT-5**

Auf dem dritten Platz landete o3, das letzte Reasoning-Modell von OpenAI, bevor es durch GPT-5 ersetzt wurde. Wir wollten herausfinden, ob bei GPT-5 im Vergleich zu o3 (das im April 2025 erschien, nicht lange vor GPT-5) spürbare Fortschritte erkennbar sind. Und tatsächlich wurde o3 von GPT-5 um 3,2 Prozentpunkte übertroffen – kein überwältigender Sieg, aber dennoch eine bemerkenswerte Weiterentwicklung. In jedem Fall würden die meisten Unternehmen und Compliance-Teams, die mit OpenAI-Technologie arbeiten, jetzt GPT-5 vorziehen, weshalb wir nicht näher auf die Eigenschaften von o3 eingehen.

GPT-5 lag um 3,2 Prozentpunkte vor o3 – kein überwältigender Sieg, aber dennoch ein spürbarer Fortschritt.

Fun Fact: o3 war das einzige Modell, das sich strikt weigerte, eine der Benchmark-Aufgaben zu bearbeiten – die Aufgabe „Workarounds finden, um bestehende Kontrollen zu umgehen“. Es antwortete knapp: „Sorry, dabei kann ich nicht helfen“, da es annahm, wir wollten die Kontrollen tatsächlich umgehen – dabei war die Übung dazu gedacht, unsere selbst entwickelten Kontrollen kritisch zu hinterfragen, um mögliche Schlupflöcher zu finden. Das ist eine interessante Modelleigenschaft, die auch für andere Anwendungsfälle relevant werden könnte, wenn Compliance-Teams „Was könnte schiefgehen?“-Szenarien erkunden.

## **Anthropic Claude Opus 4.1: Großartiger Autor, mittelmäßiger Analyst**

Claude Opus ist das Flaggschiff-Reasoning-Modell von Anthropic, einem eher weniger bekannten Anbieter von KI-Modellen. Die Modelle von Anthropic werden aber zunehmend beliebter im Bereich der Softwareentwicklung, da sie viele der neuesten Tools unterstützen, die Softwareingenieure verwenden. Auch ihr Chatbot „Claude“ hat eine treue Fangemeinde, obwohl er weit weniger beliebt ist als das in dieser Kategorie dominierende ChatGPT.

Claude Opus 4.1 erschien fast zur gleichen Zeit wie GPT-5, schneidet jedoch im Benchmark-Report deutlich schlechter ab. Es zeigte solide Leistungen in den Kernbereichen, in denen auch die Gewinner-Modelle stark waren, fiel jedoch besonders bei analytischeren Aufgaben ab (52,0 % bei der Datenanalyse, 62,4 % bei der Gap-Analyse/Vergleich, 55 % bei der Zusammenfassung/Synthese). Das Modell ist sehr gut im Schreiben, hat aber Schwierigkeiten, Informationen so zu verarbeiten, dass hochwertige Berichte entstehen. Es war auch das einzige Modell, das die Aufgabe „Auswahl des richtigen Prüfwegs für eine Offenlegung von Interessenkonflikten“ falsch bearbeitete, was zeigt, dass Claude Opus 4.1 zwar sehr kreativ ist, aber Schwierigkeiten mit strukturierten und faktenbasierten Aufgaben hat.

Das Modell ist ausgesprochen gut im Schreiben, hat jedoch Schwierigkeiten, Informationen so zu verarbeiten, dass daraus qualitativ hochwertige Berichte entstehen.

Kurz gesagt: Claude Opus ist insgesamt eine solide Wahl und glänzt besonders beim Formulieren von Berichten oder Briefings – doch es gibt einen spürbaren Abstand zu den stärksten Modellen in unserem Benchmark-Report.

## **OpenAI GPT-4o: Die blitzschnelle Qualitätsmesslatte von 2024**

Als 2023 und 2024 GPT-4 und anschließend GPT-4o herauskamen, waren viele Nutzer verblüfft, wie viel besser diese Modelle im Vergleich zu ihrem Vorgänger GPT-3.5 waren, dem ersten Modell in ChatGPT. Nun wiederholt sich die Geschichte: Gegen GPT-5 und Gemini 2.5 Pro hat GPT-4o keine Chance und schneidet über 13 Prozentpunkte schlechter ab.

Ein wesentlicher Grund für diesen Leistungsunterschied ist die Geschwindigkeit. GPT-4o liefert eine Antwort innerhalb von Sekunden (oft sogar schneller), während GPT-5 und Gemini 2.5 Pro manchmal mehr als zwei Minuten benötigen, um ihre endgültige Antwort zu geben (ein Prozess, der „Reasoning“ genannt wird). GPT-4o war das einzige Nicht-Reasoning-Modell, das wir in die Benchmark aufgenommen haben, um zu testen, welchen Unterschied diese Technologie im Bereich Compliance & Ethik macht.

GPT-4o war das einzige Nicht-Reasoning-Modell, das wir in den Benchmark aufgenommen hatten.

Die Unterschiede sind dramatisch, besonders in analyseintensiven Bereichen. Bei Aufgaben zur Datenanalyse schneidt GPT-4o um 60 Prozentpunkte schlechter ab als Google Gemini 2.5 Pro und bei Aufgaben zur Gap-Analyse/Vergleich um 30 Prozentpunkte. Bei diesen Aufgaben zahlt sich die zusätzliche Zeit, die Reasoning-Modelle für die Verfeinerung ihres Ansatzes und ihrer Antwort benötigen, wirklich aus. Bei vielen dieser Aufgaben war der Mangel an Detailgenauigkeit von GPT-4o und die fehlende Tiefe der Analysen in den Antworten direkt zu erkennen.

Manchmal sind Geschwindigkeit und schnelle Ergebnisse wichtig, aber bei den meisten Compliance-Aufgaben sind Genauigkeit und differenzierte Bewertungen weitaus wichtiger. Daher werden Nicht-Reasoning-Modelle wie GPT-4o für diese Aufgaben immer nur unterdurchschnittliche Ergebnisse liefern – vor allem im Vergleich zu den neuesten Spitzenmodellen. .

## **Mistral Large 2: Die „lobende Erwähnung“ aus Europa**

Mistral Large 2 ist ein Modell, das vom französischen Startup Mistral entwickelt und im Juli 2024 veröffentlicht wurde. Es ist das einzige europäische Modell, das wir im Rahmen des EQS-Benchmark-Reports getestet haben; alle anderen Modelle wurden von US-Unternehmen entwickelt (OpenAI, Google DeepMind und Anthropic). Obwohl es insgesamt als das stärkste KI-Modell aus Europa gilt, hat der vorliegende Benchmark-Report – ebenso wie andere Benchmarks – leider gezeigt, dass selbst das beste europäische KI-Modell nicht annähernd mit den führenden US-Modellen mithalten kann.

Es ist das einzige europäische Modell, das wir im Rahmen des EQS AI Benchmark getestet haben – alle anderen Modelle wurden von US-amerikanischen Unternehmen entwickelt.

Obwohl es sich um ein Reasoning-Modell handelt, hat Mistral in wichtigen, analyseintensiven Bereichen wie Datenanalyse (31,3 %), Gap-Analyse/Vergleich (50,3 %) oder Zusammenfassung/Synthese (52,5 %) deutlich unterdurchschnittlich abgeschnitten. Seine stärksten Kategorien sind Matching/Mapping (91,8 %), Entscheidungsfindung (90,8 %) und Kategorisierung/Klassifizierung (88,7 %), aber selbst in diesen Bereichen lag die Leistung unter der der Spitzenreiter.

Es war zudem das einzige Modell, das Kontext halluzinierte, wenn versehentlich der tatsächliche Kontext für eine Aufgabe fehlte. So erfand es eine Richtlinie, die nicht bereitgestellt worden war. Im Allgemeinen zeigte das Modell eine ähnliche mangelnde Tiefe und Detailgenauigkeit wie GPT-4o und unterschied sich deutlich von Spitzmodellen wie GPT-5 oder Gemini.

Erwähnenswert ist jedoch: Bei einigen der offen formulierten Aufgaben bewerteten unsere menschlichen Compliance-Expertinnen und -Experten die Ergebnisse recht hoch – die Art und Weise, wie Mistral den Umfang definiert und seine Antworten formuliert, scheint bei Compliance-Teams häufiger Anklang zu finden als bei Anwendungen in anderen Bereichen.

# Methodik

Der EQS-Benchmark-Report ist praxisorientiert. Der Fokus liegt auf den realen Herausforderungen, die den Alltag der Compliance-Teams widerspiegeln, und weniger auf theoretischen Überlegungen. Wir wollten verstehen, wie gut die neuesten und leistungsstärksten KI-Modelle diese Anforderungen meistern. Das Aufgabenset für den Benchmark-Report wurde speziell für dieses Ziel konzipiert.

Um dies zu erreichen, haben wir uns mit mehreren unserer Kundinnen und Kunden zusammengeschlossen. Sie gaben uns Feedback zu unserem Aufgabenset und haben konkrete Beispiele aus ihrem Arbeitsalltag geteilt, die wir in das Aufgabenset für den EQS-Benchmark-Report aufgenommen haben.

Insgesamt schafften es in das finale Benchmark-Set mehr als 20 Dokumente und Beispiele unserer Kundinnen und Kunden. Diese waren aufgrund der unterschiedlichen Unternehmensgrößen und geografischen Herkunft sehr vielfältig. So war eine breite Darstellung typischer Herausforderungen möglich, sodass die getesteten KI-Modelle an realen Compliance-Aufgaben gemessen werden konnten..

## Aufgabenbereich

Um im EQS-Benchmark-Report ein breites Spektrum an Compliance-Arbeiten abzudecken, haben wir zehn Themen berücksichtigt, die die wichtigsten Komponenten eines wirksamen Compliance-Programms abbilden. Die Kategorisierung der Aufgaben ermöglicht es uns außerdem, Einblicke in die Leistungsfähigkeit der KI-Technologie in den einzelnen Bereichen zu gewinnen.

**Die zehn Themenbereiche, die in unserem Benchmark-Aufgabenset abgedeckt werden, sind:**

- Regulatorische Analyse & Change-Management
- Risikobewertung & Analyse
- Implementieren und Überwachen von Kontrollen
- Richtlinienerstellung & -aktualisierung
- Third Party Management
- Hinweisgebermeldungen und interne Ermittlungen
- Schulung & Sensibilisierung
- Compliance-Berichterstattung & Dokumentation
- Umgang mit Interessenkonflikten / Geschenken & Einladungen
- Bewertung der Effektivität des Compliance-Programms

Jedes der Themengebiete besteht aus zwölf Benchmark-Aufgaben, die nochmal in Aufgabenarten unterteilt sind (Erläuterung im nächsten Abschnitt).

# Aufgabenarten

Der Benchmark-Report umfasst sowohl verschiedene Arten als auch unterschiedliche Kategorien von Aufgaben, um sicherzustellen, dass die KI-Modelle auf eine Vielfalt an Fähigkeiten getestet werden. Es gibt drei verschiedene Aufgabenarten, mit denen im Benchmark-Report gearbeitet wird:

Art der Aufgabe	# der Aufgaben im Benchmark-Datensatz
Multiple-Choice-Aufgaben	50
Strukturierte Aufgaben	50
Offene Aufgaben	20
<b>Gesamtzahl der Aufgaben</b>	<b>120</b>

## Multiple-Choice-Aufgaben

Dies sind einfache Fragen zu Compliance und Ethik mit vier bis zehn Antwortoptionen, von denen die Modelle eine bis vier auswählen müssen. Multiple-Choice-Fragen haben den Vorteil der Determiniertheit, was bedeutet, dass es objektiv richtige Antworten gibt und die Auswertung leicht ist (richtig / nicht richtig / teilweise richtig).

Obwohl strukturell einfach, prüfen die Multiple-Choice-Aufgaben in diesem Benchmark nicht nur das Wissen, sondern konfrontieren die Modelle auch mit realistischen neuen Szenarien, die sie korrekt einschätzen müssen.

## Strukturierte Aufgaben

Dies ist eine komplexere Aufgabenklasse, bei der die KI-Modelle ihre Antworten in einer vordefinierten Struktur liefern sollen. Dies kann eine Bewertung sein (z. B. Einschätzung eines potenziellen Risikos auf einer Skala von 1 bis 5) oder das Zurückgeben spezifischer Objekte (z. B. das Extrahieren von ABAC-bezogenen Risiko-Flags aus einem Screening-Bericht mit vordefinierten Metainformationen).

Antworten in einer vordefinierten Struktur zu liefern, ist eine entscheidende Fähigkeit, die bei KI-Modellen getestet werden sollte, denn zuverlässige, strukturierte Ergebnisse lassen sich viel leichter in Workflows mit Agenten oder auch in regulären Compliance- & Ethik-Berichten und Auswertungen verwenden.

## Offene Aufgaben

In dieser Aufgabenklasse sollen die KI-Modelle Ergebnisse liefern, die keiner festen Struktur folgen. Zu den Aufgaben gehören das Verfassen von Textentwürfen, Berichten, Antworten auf Fragen und Briefings. Bei einigen Aufgaben sollen die Modelle ihren Output an eine vordefinierte Vorlage oder einen bestimmten Stil anlehnen, sie haben bei ihren Antworten jedoch deutlich mehr Flexibilität als bei Multiple-Choice- und strukturierten Aufgaben.

# Aufgabenkategorien

Viele Nutzerinnen und Nutzer von KI-Technologie verwenden die Modelle heute für einfache Aufgaben, etwa das Erstellen von Textentwürfen, die Grammatikprüfung, Übersetzungen oder Zusammenfassungen.

Mit diesem Benchmark-Report möchten wir die KI-Modelle bei komplexeren Aufgaben herausfordern, die für den Arbeitsalltag der Compliance & Ethik-Fachleute relevanter sind. Daher haben wir Aufgaben integriert, die sich auf Bewertung/Evaluierung, Entscheidungsfindung, Gap-Analyse, Datenanalyse und viele weitere Bereiche konzentrieren – die nachstehende Tabelle gibt einen Überblick.

Aufgabentyp	Beschreibung	# Aufgaben
Bewertung / Beurteilung	Beurteilung der Qualität, Wirksamkeit, Vollständigkeit oder Schwere eines Sachverhalts	52
Entscheidungsfindung	Auswahl der am besten geeigneten Maßnahme, Reaktion oder Vorgehensweise aus den Optionen	16
Rangliste / Priorisierung	Elemente nach Risiko, Relevanz, Dringlichkeit usw. ordnen	9
Inhaltserstellung	Erstellung neuer schriftlicher oder strukturierter Inhalte wie Richtlinien, Berichte, Quizfragen	8
Datenextraktion	Spezifische Informationen aus Texten oder strukturierten Eingaben extrahieren	7
Gap-Analyse / Vergleich	Unterschiede, Überschneidungen oder Widersprüche zwischen Eingaben erkennen	7
Kategorisierung / Klassifizierung	Zuweisen vordefinierter Kategorien, Tags oder Labels zu Eingaben	6
Abgleich / Zuordnung	Entsprechungen oder Beziehungen zwischen Elementen erkennen	5
Datenanalyse	Ableitung von Erkenntnissen, Trends oder statistischen Beobachtungen aus Datensätzen oder strukturierten Eingaben	5
Prognose / Vorhersage	Schätzung künftiger Trends, Risiken, Verhaltensweisen oder Ergebnisse	3
Zusammenfassung / Synthese	Zusammenfassen oder Verdichten von Informationen über mehrere Eingaben hinweg	2

Die Kategorisierung der Aufgaben ermöglicht auch einen Vergleich der Leistung der KI-Modelle bei bestimmten Aufgabenarten. Dies ist wichtig, um zu verstehen, für welche Aufgaben bestimmte KI-Modelle mit hoher Zuverlässigkeit eingesetzt werden können und bei welchen Aufgaben die Ergebnisse eher gemischt ausfallen.

Die Anzahl der Aufgaben ist nicht gleichmäßig auf die verschiedenen Kategorien verteilt, da die Gesamtheit der Aufgaben ein repräsentatives Bild des typischen Aufgabenprofils der Compliance & Ethik-Abteilungen vermitteln soll. Aus diesem Grund und weil wir an das Prinzip „Human in the Loop“ glauben, ist die Kategorie „Bewertung/Evaluierung“ überproportional vertreten, da sie einen großen Teil der täglichen Arbeit von Compliance- und Ethik-Expertinnen und- Experten ausmacht.

## **Schwierigkeitsgrad der Aufgaben**

Die Meßlatte liegt hoch im EQS-Benchmark-Report, der die KI-Modelle mit komplexen und vielfältigen Herausforderungen im Bereich Compliance & Ethik konfrontiert. Wir haben die Aufgaben bewusst so schwer gestaltet, sodass wir nicht erwarteten, dass ein KI-Modell 100 % erreicht; vielmehr soll der Benchmark-Report einen hohen Maßstab darstellen, an dem sich auch künftige Modelle messen lassen können – und so die Grenzen dessen verschieben, was KI bei der Compliance & Ethik-Arbeit realistisch an Unterstützung leisten kann.

Interessanterweise ist es nicht einfach, den Schwierigkeitsgrad der Aufgaben zu beurteilen, bevor der Benchmark-Report durchgeführt worden ist und man sieht, wie die KI-Modelle damit zureckkommen. Bei einigen Aufgaben waren wir uns ziemlich sicher, dass die meisten Modelle gute Leistungen erbringen würden, bei anderen waren wir uns nicht sicher, ob die Aufgabe eine (zu) große Herausforderung darstellt oder ob die Modelle sich mit guten Ergebnissen auszeichnen können. Oft waren wir überrascht, wie gut die Modelle selbst bei scheinbar komplexen und schwierigen Aufgaben abschneiden.

## **Synthetischer und realer Kontext für Aufgaben**

Wie oben erwähnt, haben wir mit mehreren Unternehmen, die zu unseren Kunden zählen, zusammengearbeitet, um Feedback sowie konkrete Beispiele und Daten aus deren täglicher Arbeit zu erhalten. Diese sind in viele Benchmark-Aufgaben eingeflossen. Da wir für andere Aufgaben entweder große Datenmengen oder sensible bzw. personenbezogene Daten benötigten, umfasst der Benchmark-Report auch synthetische Daten, die von KI generiert wurden<sup>10</sup>. Sämtliche von KI generierten Daten wurden von Fachleuten mit spezieller Expertise im Bereich Compliance & Ethik geprüft, um sicherzustellen, dass sie realistisch und relevant sind.

Durch die Kombination von synthetischen und tatsächlichen Daten unserer Kundinnen und Kunden können wir Aufgaben mit einem breiten Spektrum an Kontext einbeziehen. Insgesamt umfassen die Benchmark-Aufgaben Hunderte von Seiten mit Kontextinformationen, darunter Richtlinien, Offenlegungen, behördliche Veröffentlichungen, Screening-Berichte, Fragen von Stakeholdern, Interviewtranskripte und vieles mehr.

Die Einbeziehung großer Mengen an Kontext ist wichtig, um realistische Herausforderungen für die KI-Modelle zu integrieren, da viele der täglichen Herausforderungen im Bereich Compliance & Ethik ebenfalls die Arbeit mit und die Auswertung großer Mengen an Informationen umfassen.

<sup>10</sup> Synthetische Inhalte wurden von OpenAI o3 und Google Gemini 2.5 Pro generiert.

# Alle Aufgaben in der Übersicht

Die Beschreibung aller 120 Aufgaben des EQS-Benchmark-Reports ist nachfolgend aufgelistet. Die Übersicht soll nicht nur einen transparenten Einblick in den Umfang unserer Untersuchung liefern, sondern auch als Anregung für konkrete Anwendungsfälle dienen, bei denen die KI-Technologie heute bereits die Compliance- und Ethik-Community bei ihrer Arbeit unterstützen kann.

## 1. Regulatorische Analyse & Change Management

- Priorisierung von Anti-Korruptionsvorschriften
- Priorisierung von Anti-Korruptionsvorschriften (basierend auf einem Unternehmensprofil)
- Identifizierung vertrauenswürdiger Quellen für regulatorische Updates
- Abgleich regulatorischer Anforderungen mit dem Unternehmensprofil
- Analyse sich überschneidender Vorschriften
- Recherche der drei neuesten offiziellen FCPA-Updates
- Bewertung der Auswirkungen einer neuen Vorschrift auf einer Skala von 1-5
- Bewertung der Auswirkungen einer neuen Vorschrift auf einer Skala von 1-5
- Identifizierung der am stärksten betroffenen Abteilungen bei neuen Vorschriften
- Bewertung der Auswirkungen neuer Vorschriften auf die bestehende Risikolandschaft
- Erstellung eines Executive Briefings zu den Auswirkungen neuer Vorschriften
- Recherche der drei für ein Unternehmen (basierend auf seinem Profil) bedeutendsten regulatorischen Entwicklungen

## 2. Risikobewertung & Ad-hoc-Analyse

- Vorschlag geeigneter Maßnahmen zur Risikominderung im Compliance-Bereich
- Identifizierung von Erleichterungszahlungen, die gegen ABAC-Vorschriften verstößen
- Grenzüberschreitende Risikokartierung (FCPA + UKBA)
- Sapin-II-konforme Reaktion auf Umgehung von Kontrollen
- Analyse von Vergeltungsmustern gegenüber Hinweisgebenden
- Auswahl des zutreffenden Risikoniveaus auf Basis einer Risikobeschreibung
- Bewertung von Eintrittswahrscheinlichkeit und Schweregrad in Prozentwerten auf Basis einer Risikobeschreibung
- Extraktion zentraler Q&A-Informationen aus einer E-Mail-Korrespondenz
- Extraktion wesentlicher Risiko-Metadaten aus einer unstrukturierten Beschreibung
- Bewertung des verbleibenden Korruptionsrisikos eines Lieferanten nach Mitigationsmaßnahmen
- Vorschlag eines Risikoanalyse- bzw. -kartierungsprozesses im Einklang mit Sapin II
- Formulierung einer Antwort auf eine Compliance-bezogene Anfrage aus einer Fachabteilung

### **3. Kontroll-Design & Monitoring**

- Bewertung potenzieller Schwächen bestehender Kontrollen
- Bewertung von Bedrohungen für das bestehende Kontrollsyste
- Auswahl der wirksamsten Maßnahme zur Minderung von Schwächen im Offboarding-Prozess
- Identifizierung der effektivsten Verbesserung bestehender Kontrollen
- Klassifizierung bestehender Kontrollen zur Erkennung doppelter Rechnungen
- Einordnung bestehender Kontrollen in präventive, detektive und korrektive Kontrollen
- Bewertung der Wirksamkeit bestehender Third-Party-Due-Diligence-Kontrollen
- Identifizierung von Überschneidungen und Widersprüchen zwischen Kontrollen
- Zuordnung bestehender Kontrollen zu den passenden COSO-Komponenten
- Klassifizierung bestehender Kontrollen als Key oder Secondary im Rahmen von SOX
- Entwicklung präventiver, detektiver und korrekter Kontrollen zur Verhinderung betrugs- und bestechungsbezogener Zahlungen an Lieferanten
- Aufzeigen möglicher Umgehungswege bestehender Kontrollen

### **4. Richtlinienentwicklung & -überprüfung**

- Rollout von Richtlinien nach Risikoprofil der Abteilungen
- Bewertung der Auswirkungen neuer Vorschriften auf bestehende Richtlinien
- Identifizierung von ISO-37002-relevanten Richtlinienanpassungen
- Auswahl zutreffender Aussagen zum Verhaltens- und Ethikkodex von Disney für Direktoren
- Auswahl zutreffender Aussagen zu den öffentlichen Kartellrechtsregeln von adidas
- Extraktion von Metadaten aus bestehenden Richtlinien
- Identifizierung zentraler Unterschiede zwischen eigenem und konkurrierendem Verhaltenskodex
- Vorschlag erforderlicher Änderungen einer bestehenden Richtlinie aufgrund neuer oder geänderter Vorschriften
- Durchführung einer Gap-Analyse zwischen bestehendem Glossar und undefinierten Begriffen in einer Richtlinie
- Identifizierung wesentlicher Änderungen zwischen verschiedenen Versionen derselben Richtlinie
- Identifizierung von Überschneidungen und/oder Widersprüchen zwischen Unternehmensrichtlinien
- Erstellung einer neuen Richtlinie auf Basis von Beispielen früherer Richtlinien

## **5. Drittparteien-Management**

- Auswahl der Drittparteien mit dem höchsten Risiko
- Priorisierung von Ländern für Anti-Korruptionskontrollen
- Auswahl von Lieferanten mit dem niedrigsten Korruptionsrisikoprofil
- Zuordnung der korrekten NACE-2-Kategorie basierend auf einer Beschreibung
- Zuweisung der passenden Risikokategorie zu einem Lieferanten basierend auf ERP-Daten
- Bewertung von Drittparteien-Fragebögen auf ABAC-Risiken
- Identifizierung unzureichend beantworteter Fragen in Drittparteien-Fragebögen
- Bewertung der Qualität interner Antworten auf Drittparteien-Fragebögen
- Identifizierung von Red Flags zu einer bestimmten Drittpartei auf Basis ihres Screening-Berichts
- Identifizierung und Risikobewertung von ABAC-bezogenen Red Flags in zwei Screening-Berichten von Drittparteien
- Erstellung von Kommentaren oder Nachfragen zu unzureichend beantworteten Drittparteien-Fragebögen
- Bewertung des Bedarfs an zusätzlichen Informationen basierend auf ausgefüllten Drittparteien-Fragebögen

## **6. Speak-up & Untersuchungsunterstützung**

- Zuordnung geeigneter Ermittler auf Basis von Kompetenzprofilen
- Kapazitätsbewertung für neue Falluntersuchungen
- Identifizierung der durch eine Speak-up-Meldung verletzten Richtlinien
- Vorschlag geeigneter Korrekturmaßnahmen im Kontext einer Speak-up-Meldung
- Auswahl von Erstzeugen für Untersuchungsinterviews
- Begründung der durch die Speak-up-Meldung identifizierten Richtlinienverstöße
- Begründung vorgeschlagener Korrekturmaßnahmen als Folge einer Speak-up-Untersuchung
- Korrekte Identifizierung beteiligter Mitarbeitender trotz unvollständiger Namensangaben
- Identifizierung einzelner Vorwürfe in einem komplexen Speak-up-Bericht
- Identifizierung potenzieller Vergeltungsmaßnahmen gegen namentlich genannte Hinweisgebende
- Erstellung eines Fallabschlussberichts anhand der Unternehmensvorlage
- Vorschlag der nächsten sinnvollen Schritte in einer internen Untersuchung

## **7. Schulung & Sensibilisierung**

- Priorisierung von Abteilungen für eine Sensibilisierungskampagne zu Geschenken & Bewirtung
- Vorschlag der geeignetsten Zertifizierungsmethode für Richtlinien
- Auswahl zutreffender Q&A-Paare basierend auf der Anti-Korruptionsrichtlinie von Bayer
- Auswahl der wirksamsten Schulungen im Anschluss an eine Speak-up-Meldung
- Auswahl der geeignetsten Schulungen für ein spezifisches Mitarbeiterprofil
- Priorisierung relevanter Awareness-Themen auf Basis von Q&A- und Speak-up-Daten
- Priorisierung von Abteilungen für Präsenzschulungen basierend auf Richtlinieninhalten
- Qualitätsprüfung von Übersetzungen für Kampagneneinladungen
- Prüfung, ob die Formulierung einer Richtlinie für bestimmte Zielgruppen geeignet ist
- Festlegung der Dringlichkeit für Anti-Korruptions-Auffrischungstrainings in verschiedenen Szenarien
- Erstellung von Multiple-Choice-Fragen auf Basis einer bestehenden Richtlinie
- Erstellung praxisnaher Beispilszenarien für spezifische Abteilungen auf Basis einer bestehenden Richtlinie

## **8. Compliance-Berichterstattung & Dokumentation**

- Umgang mit widersprüchlichen Anforderungen zur Datenaufbewahrung
- Auswahl relevanter KPIs für ESRS-Berichtspflichten
- Priorisierung von ESRS-Standards für die doppelte Wesentlichkeitsanalyse
- Bewertung der Qualität eines Falluntersuchungsberichts
- Prüfung der Vollständigkeit der Dokumentation zum UK Modern Slavery Act
- Identifizierung personenbezogener Daten nach DSGVO
- Erstellung einer jährlichen Budgetschätzung auf Basis historischer Daten
- Prognose neuer Risikobereiche für die nächsten 12-24 Monate
- Prognose geografischer Regionen und Geschäftsbereiche mit der höchsten Wahrscheinlichkeit zukünftiger Vorfälle
- Ausfüllen von Kundenfragebögen auf Basis bestehender Richtlinien
- Vorschlag relevanter Reporting-KPIs zur Darstellung des Mehrwerts der Compliance-Funktion auf Basis von Unternehmens- und Compliance-Profil
- Erstellung eines Vorstandsberichts zur aktuellen Interessenkonflikt-Landschaft basierend auf COI-Metadaten

## **9. Interessenkonflikte / Geschenke & Bewirtung**

- Auswahl der zutreffenden Kategorie für eine Interessenkonflikt-Offenlegung
- Auswahl korrekter Aussagen zu einer komplexen Interessenkonfliktsituation
- Entscheidung über eine Geschenk-Offenlegung basierend auf der G&H-Richtlinie eines Unternehmens
- Entscheidung über eine Spendenanfrage basierend auf einer spezifischen Richtlinie
- Entscheidung über einen Interessenkonflikt gemäß DOJ-Vorgaben
- Bewertung der Schwere von Interessenkonflikten basierend auf bestehender Richtlinie
- Auswahl des passenden Prüfwegs für eine Interessenkonflikt-Offenlegung
- Entscheidung über Geschenke- & Bewirtungsanfragen basierend auf einer E-Mail-Schilderung und geltenden Regeln
- Zuordnung der korrekten Drittparteien zu Offenlegungen trotz Namensähnlichkeiten
- Aufsplittung einer großen Firmenveranstaltung in einzelne und aggregierte G&H-Anfragen
- Identifizierung zusätzlicher Informationen zur Bewertung des ABAC-Risikos einer COI-Offenlegung
- Vorschlag geeigneter Maßnahmen zur Minderung eines Interessenkonflikts basierend auf historischen Mitigationsmaßnahmen

## **10. Programmevaluierung**

- Bewertung der Wirksamkeit des Drittparteien-Programms
- Bewertung der Wirksamkeit des Schulungsprogramms
- Best Practices zur Bewertung der Effektivität von Hinweisgeber-Hotlines
- Definition von Folgeaktionen bei unzureichender Kommunikation zu Policy-Updates
- Adressierung wahrgenommener Verantwortlichkeitslücken bei der Nachverfolgung gemeldeter Fälle
- Identifizierung der relevantesten ABAC-Schwachstellen basierend auf einem Unternehmensprofil (z. B. JP Morgan Chase)
- Vergleich von Speak-up-Metadaten mit einem Branchen-Benchmark
- Bewertung der Wirksamkeit einer Sensibilisierungskampagne anhand von Quiz-Ergebnissen
- Analyse der Bearbeitungszeiten von G&H- und COI-Offenlegungsdaten
- Identifizierung der größten Fehlverhaltens-Hotspots innerhalb einer Organisation
- Erkennen von Substantiierungstrends in historischen Speak-up-Metadaten
- Synthese offener Rückmeldungen von Mitarbeitenden zu möglichen Verbesserungen im Speak-up-Programm

# Die im EQS-Benchmark-Report getesteten KI-Modelle

Der EQS-Benchmark-Report hat das Ziel, neben den neuesten Möglichkeiten auch die Grenzen der KI-Technologie im Kontext von Compliance & Ethik zu verstehen – sowie realistische Erwartungen für die Zukunft zu definieren. Daher war es für uns entscheidend, den Report mit den neuesten und (laut anderen Benchmarks) besten KI-Modellen der Branche durchzuführen.

Da die Entwicklung und Bereitstellung von KI-Modellen mit hohen Investitionen in Rechenleistung, technische Expertise und Infrastruktur verbunden ist, haben sich wenige globale Akteure als führende KI-Technologieanbieter herausgebildet. Ein großer Teil der heute verfügbaren KI-Software und -Funktionen basiert auf den Modellen dieser Anbieter. Daher verschafft uns die Kenntniss der Fähigkeiten dieser KI-Basismodelle auch ein eingehendes Verständnis der aktuellen technologischen Grenzen.

Die drei weltweit führenden Anbieter mit den leistungsstärksten KI-Modellen sind OpenAI, Google und Anthropic. Ihre Chatbot-Produkte ChatGPT, Gemini und Claude werden täglich von Millionen von Nutzern verwendet und ihre Modelle erzielen in anderen etablierten Benchmarks durchweg die besten Ergebnisse. Wir haben daher die neuesten Flaggschiff-Modelle dieser Anbieter in den EQS-Benchmark-Report aufgenommen. Um die Bandbreite der getesteten Modelle zu erweitern, haben wir auch ein Modell des europäischen Anbieters Mistral in unsere Betrachtungen miteinbezogen.

Modellname	Modellverleger	Veröffentlichungsdatum	Kontextfenster	MMLU-Pro Score <sup>11</sup>
GPT-5	OpenAI	7. August 2025	400k Token	87,6 %
Claude Opus 4.1	Anthropic	5. August 2025	200k Token	82,7 %
o3	OpenAI	16. April 2025	200k Token	85,6 %
Gemini 2.5 Pro	Google DeepMind	25. März 2025	1 Million Token	84,1 %
Mistral Large 2	Mistral AI	24. Juli 2024	128k Token	69,7 %
GPT-4o	OpenAI	13. Mai 2024	128k Token	74,1 %

Dass OpenAI mit mehreren Modellen in der Endauswahl vertreten ist, hat zwei Gründe:

Erstens ist GPT-4o als Standardmodell in ChatGPT das weltweit am häufigsten verwendete KI-Modell und damit ein guter Vergleichspunkt für Nicht-Fachleute. Zweitens ist GPT-4o im Gegensatz zu o3 von OpenAI kein Reasoning-Modell – und ermöglichte damit den Vergleich zwischen Reasoning und Nicht-Reasoning-Modellen. Die Leistungsvergleich dieser beiden Modellarten war ebenfalls ein Ziel des Benchmark-Reports.

Kontextfenster beziehen sich darauf, wie viel Kontext das Modell in einer Eingabeaufforderung (Prompt) verarbeiten kann. Ein Token ist ein Inhaltselement, in das KI-Modelle Text aufteilen. Ein Token entspricht ungefähr drei Vierteln eines englischen Wortes, eine Million Token sind damit etwa 1.500 DINA4-Seiten in 12-Punkt-Schrift.

Das bedeutet zum Beispiel, dass Google Gemini 2.5 Pro in einem einzigen Prompt 1.500 Seiten an Inhalten verarbeiten kann, OpenAI GPT-4o dagegen „nur“ etwa 190 Seiten. Für die meisten alltäglichen Anwendungsfälle ist beides mehr als ausreichend, aber bei Aufgaben, wie dem Vergleichen von Richtliniendokumenten oder der Analyse großer Datenmengen, können die Grenzen des Kontextfensters relevant werden.

<sup>11</sup> Alle Benchmark-Ergebnisse stammen von vals.ai, in: [https://www.vals.ai/benchmarks/mmlu\\_pro-04-18-2025](https://www.vals.ai/benchmarks/mmlu_pro-04-18-2025)

## Ausführung der Benchmark-Aufgaben

Es wurde die Bewertungsplattform „LangSmith“ verwendet, um alle 120 Benchmark-Aufgaben auf allen sechs KI-Modellen auf standardisierte und skalierbare Weise auszuführen. Die Plattform ermöglicht die automatisierte Ausführung von Prompts über große Datensätze hinweg unter Verwendung der APIs der Modellanbieter. Für den EQS-Benchmark-Report wurden ausschließlich die offiziellen APIs von OpenAI, Google, Anthropic und Mistral verwendet.

## Bewertung der Modell-Ergebnisse

Wie unter „Aufgabenarten“ beschrieben, gibt es für einige Aufgaben im EQS-Benchmark-Report eine objektiv richtige Antwort, was die Bewertung der Ergebnisse vereinfacht, während vor allem die offenen Aufgaben eine subjektivere, stärker am Menschen orientierten Bewertungsansatz erfordern. Daher verfolgen wir je nach Art der Aufgabe zwei verschiedene Methoden zur Bewertung der Modellergebnisse:

### Automatisierte Bewertung

Für Multiple-Choice-Aufgaben und die meisten strukturierten Aufgaben wird eine automatisierte Bewertung verwendet, da es eine objektiv „richtige“ Antwort gibt. Die Bewertungen können dennoch variieren, da ein Modell möglicherweise teilweise richtig, aber nicht vollständig antwortet, sodass eine Teilbewertung vergeben wird. Wenn ein Modell beispielsweise vier richtige Antworten geben muss, aber nur drei richtige Optionen auswählt, erhält es für diese Aufgabe eine Bewertung von 0,75.

### Menschliche Bewertung

Bei allen offenen und einigen ausgewählten strukturierten Aufgaben erfolgte die Beurteilung durch eine Jury; jede Aufgabe wurde von mindestens zwei Personen bewertet.

Die Mitglieder der Jury sind entweder Mitglied des Berufsverbands der Compliance-Manager (BCM), der größten deutschen Branchenvereinigung für Compliance-Fachkräfte, Compliance-Expertinnen und -Experten aus dem Kundenkreis der EQS Group oder Mitarbeitende der EQS Group mit nachgewiesener Expertise und/oder Erfahrung im Bereich Compliance & Ethik.

## Einsatz von KI bei der Erstellung des EQS-Benchmark-Reports

Neben der Nutzung von KI zur Generierung synthetischer Daten für kontextintensive Aufgaben (wie in „Synthetischer und realer Kontext für Aufgaben“ erklärt) wurde KI auch an anderer Stelle eingesetzt: Konkret wurden GPT-4o, o3, Gemini 2.5 Pro und Gemini 1.5 Flash unter anderem verwendet, um bei der Ideenfindung relevanter Themen zu helfen, konkrete Aufgabenaufforderungen und Ergebnisstrukturen auszuarbeiten sowie Benchmark-Aufgaben vorzutesten.

# Risiken und Schwachstellen

Wie bei jedem Projekt dieses Umfangs mussten wir einige Kompromisse eingehen und Risiken akzeptieren, die in diesem Abschnitt beschrieben werden. Wir sind überzeugt, dass keiner dieser Punkte die Relevanz der Gesamtergebnisse oder die Gültigkeit der gewonnenen Erkenntnisse wesentlich beeinträchtigt. Dennoch möchten wir sie transparent darlegen.

## Modellauswahl

Wir mussten Einschränkungen bei der Anzahl der getesteten KI-Modelle vornehmen, um die Durchführbarkeit des Projekts zu gewährleisten. Daher haben wir uns auf die aktuellsten und leistungsfähigsten Modelle konzentriert, die laut anderen Benchmarks von den weltweit führenden Anbietern stammen. Dennoch mussten wir relevante Modelle wie Grok (von xAI), die Llama-Familie von Meta oder DeepSeek ausschließen. Die nächsten Ausgaben des EQS-Benchmark-Report könnten es aber ermöglichen, Modelle dieser Anbieter oder auch neue Marktteilnehmer einzubeziehen.

Ebenso haben wir die „Deep Thinking“-Versionen von Modellen wie Gemini oder Grok (xAI) nicht berücksichtigt, da diese Varianten noch experimentell, kostenintensiv und nicht über APIs verfügbar sind – was ihre Integration in einen Benchmark-Report wie diesen erschwert.

## Anwendbarkeit auf Herausforderungen der realen Welt

Die Benchmark-Aufgaben wurden, wie bereits beschrieben, bewusst so realitätsnah wie möglich gestaltet – bis hin zur Verwendung von echten Daten und Inhalten aus der Zusammenarbeit mit Kunden und Partnern. Für einige Aufgaben mussten wir jedoch auf synthetischen Kontext zurückgreifen, da echte Daten aufgrund ihrer Sensibilität nicht verfügbar waren. Die Nutzung synthetischer Daten ist in Benchmarks allerdings gängige Praxis. Wir haben auch sichergestellt, dass Struktur und Inhalt dieser Daten realistische Szenarien aus dem Arbeitsalltag von Compliance-Teams abbilden.

## Einschränkungen durch einfache Prompts

Um die Durchführung des Benchmark-Reports effizient zu gestalten, wurde nur begrenzt Zeit in die Optimierung einzelner Prompts investiert. Mit verfeinerten Eingabeaufforderungen wäre es wahrscheinlich möglich, noch bessere Ergebnisse zu erzielen. Compliance-Teams, die KI in ihrem Arbeitsalltag einsetzen möchten, sollten daher gezielt Zeit in das Feintuning ihrer Prompts investieren. Fortgeschrittene Techniken – etwa die Kombination mehrerer Modelle, der Einsatz von RAG (Retrieval-Augmented Generation) oder andere komplexe Ansätze – könnten in bestimmten Bereichen ebenfalls zu besseren Ergebnissen führen, wurden jedoch bewusst nicht in den Umfang dieser Studie aufgenommen. Das bedeutet: Auch wenn ein einzelnes Modell mit einem einfachen Prompt nicht immer das perfekte Ergebnis liefert, kann die KI-Technologie insgesamt mit ausgereifteren Methoden dennoch dorthin gelangen – nur eben außerhalb des Anwendungsbereichs des vorliegenden Benchmark-Reports.

## Multimodalität: Keine Tests von Bild-, Video- und Audioinhalten

Einige der getesteten Modelle können auch Bilder, Videos oder Audioinhalte verarbeiten und erzeugen. Die Einbeziehung dieser multimodalen Fähigkeiten hätte jedoch dem Benchmark-Report eine völlig neue Dimension verliehen und ihn deutlich umfassender gemacht. Zwar sehen wir potenziell interessante Anwendungen für multimodale Funktionen in der KI, etwa die automatisierte Analyse von Bildmaterial als Beweismittel oder die Auswertung von Sprachaufnahmen. Allerdings glauben wir, dass der wichtigste Hebel für Compliance-Teams derzeit in der Verarbeitung von Textinformationen liegt.



# Schlussbemerkungen

## **Compliance & Ethik müssen mit am Tisch sitzen**

Künstliche Intelligenz ist gekommen, um zu bleiben! Die Technologie hat bereits erhebliche Auswirkungen auf unseren Alltag. Schon vor der Einführung von ChatGPT und den leistungsfähigeren Large-Language-Modellen war KI eine mächtige Technologie hinter den Kulissen zahlreicher Empfehlungs- oder Mustererkennungsmaschinen.

Wir gehen davon aus, dass die Reichweite und der Einfluss dieser Technologie in Zukunft weiter zunehmen werden. Selbst wenn der Fortschritt der Technologie ins Stocken geraten sollte – wofür es derzeit keine Anzeichen gibt – befinden wir uns noch in einer frühen Phase der Einführung. So nutzen beispielsweise die meisten Compliance & Ethik-Teams noch nicht die neuesten KI-Funktionen.

Aus Sicht von Compliance & Ethik stehen immer zwei Dinge zur Diskussion:

- Welchen neuen Risiken und Governance-Anforderungen müssen C&E-Teams aufgrund von KI bewältigen?
- Wie können die Compliance & Ethik-Verantwortlichen KI nutzen, um effizienter und effektiver zu werden?

Diese beiden Bereiche werden oft getrennt voneinander betrachtet, doch sie müssen miteinander verknüpft werden: Nur durch die Anwendung und ein solides Verständnis für Künstliche Intelligenz können Compliance-Teams die richtigen Fragen stellen und die richtige Governance entwickeln, um die sehr realen Sicherheitsrisiken und ethischen Herausforderungen anzugehen, die die KI-Technologie mit sich bringt.

In den kommenden Jahren werden noch viele weitere ethische Herausforderungen im Zusammenhang mit KI auf uns zukommen. Deshalb brauchen wir starke und kompetente Compliance & Ethik-Expertinnen und -Experten, die mit am Tisch sitzen, wenn es darum geht, diese Herausforderungen zu meistern.

Wir sind mit unseren Partnern bestrebt, die Compliance & Ethik-Community auf diesem Weg zu unterstützen – durch wertvolle Erkenntnisse, wirkungsvolle Produkte und eine Plattform, auf der wir gemeinsam all diese Herausforderungen diskutieren können.

**Lassen Sie uns die Zukunft von Compliance & Ethik gemeinsam gestalten!**

# AUTOREN



**MORITZ HOMANN**

DIRECTOR PRODUCT  
INNOVATION & AI,  
EQS GROUP



**STEPH HOLMES**

DIRECTOR OF ETHICS &  
COMPLIANCE STRATEGY,  
EQS GROUP



**SABELA PÉREZ GARCÍA**

DIRECTOR  
COMPLIANCE & ETHICS,  
EQS GROUP

# Über EQS Group

Die EQS Group ist ein führender internationaler Cloud-Softwareanbieter für Compliance & Ethics, Datenschutz, Nachhaltigkeitsmanagement und Investor Relations.

Mehr als 14.000 Unternehmen in über 80 Ländern nutzen die Produkte der EQS Group, um Vertrauen aufzubauen – indem sie komplexe regulatorische Anforderungen zuverlässig und sicher erfüllen, Risiken minimieren und transparent über ihre Unternehmensleistung sowie deren Auswirkungen auf Gesellschaft und Umwelt berichten.

Die EQS Group wurde im Jahr 2000 in München gegründet und beschäftigt heute weltweit mehr als 600 Mitarbeitende.

# Über den BCM

Als führender Berufsverband ausschließlich für Compliance-Beauftragte in Unternehmen, Verbänden und anderen Organisationen vertritt der BCM die Interessen seiner Mitglieder gegenüber Politik, Wirtschaft und Gesellschaft.

Im Mittelpunkt seiner Arbeit stehen Information, Vernetzung und die Stärkung des Berufsbildes Compliance.

Der Verband bietet eine Vielzahl kostenloser Services an, die darauf abzielen, seine Mitglieder über aktuelle Compliance-Themen auf dem Laufenden zu halten sowie den Wissenstransfer innerhalb des Netzwerks zu fördern und kontinuierlich weiterzuentwickeln.





CREATING TRUSTED  
COMPANIES

EQS Group AG | Karlstraße 47 | 80333 München  
Tel.: +49 (0) 89 444430-340

[www.eqs.com](http://www.eqs.com)  
[www.integrityline.com](http://www.integrityline.com)