

The background features a vibrant, abstract geometric pattern of overlapping triangles and polygons in shades of purple, teal, orange, and yellow. A stylized rocket ship is depicted in the upper right, launching upwards with a large, multi-colored flame trail. The overall aesthetic is modern and tech-oriented.

DER STATE OF AI CODING REPORT 2026

BRANCHENEINBLICKE

Eine von Hanover Research durchgeführte New Relic Studie befragte Führungskräfte im Technologiebereich (N = 200) ab Manager:innen-Ebene in Unternehmen, die generative oder Agentic AI für die Software- und Anwendungsentwicklung nutzen.

Aktuelle Branchen-Benchmarks geben eine klare Richtung für den Einsatz von KI in der Software- und Anwendungsentwicklung vor. Microsoft berichtet, dass KI mittlerweile 30 % seiner Codebasis schreibt ([CNBC](#)), während Salesforce-CEO Marc Benioff schätzt, dass AI Agenten derzeit zwischen 30 % und 50 % des Workloads übernehmen ([Bloomberg](#)). Meta hat für 2026 ein Ziel von 50 % signalisiert ([Mashable](#)). Die Telemetrie von GitHub Copilot lässt auf einen durchschnittlichen Code-Anteil von 46 % in der gesamten Benutzer:innenbasis schließen ([Larridin](#)). Und vielleicht am bemerkenswertesten ist, dass Google kürzlich bekannt gab, dass 75 % seines Codes KI-generiert sind ([Business Insider](#)).

In einer neuen wegweisenden Studie hat New Relic herausgefunden, dass zwei Drittel der befragten IT-Entscheider:innen (67 %) angeben, dass KI mittlerweile zwischen 51 % und 75 % der wöchentlichen Code-Produktion ihres Unternehmens generiert oder einem signifikanten Refactoring unterzieht. Das durchschnittliche Unternehmen arbeitet heute mit einer Codebasis, die größtenteils nicht von Menschen geschrieben wird.

Dieser Wandel hat Konsequenzen, die dieser Bericht im Detail behandelt. Einige sind positiv. Die Feature-Bereitstellung erfolgt schneller, die Deployment-Häufigkeit ist gestiegen und eine deutliche Mehrheit der Führungskräfte hält KI-generierten Code für qualitativ hochwertiger als von Menschen geschriebenen Code.

Bei manchen ist das nicht der Fall. Incidents in der Produktion, technische Schulden, Microservices-Wildwuchs und die Zeit für Nacharbeiten von Senior Engineers haben alle gleichzeitig zugenommen.

Und die vier häufigsten Arten von Produktionsfehlern durch KI-generierten Code (Integrationsfehler, Compliance-Issues, Datenintegritätsprobleme und Sicherheitsschwachstellen) haben in den letzten sechs Monaten jeweils rund drei von zehn Unternehmen betroffen.

DIE INTERESSANTESTE ERKENNTNIS

Die Lücke zwischen diesen beiden Bildern

KI-generierter Code wird im Moment der Überprüfung als qualitativ hochwertiger eingestuft und liefert messbar schlechtere Ergebnisse, sobald er in die Produktion gelangt. Genau um diese Lücke geht es in diesem Bericht, und sie ist die strategische Chance für die Kategorie Observability im Jahr 2026.

WICHTIGSTE ERKENNTNISSE

ES BESTEHT EINE DISKREPANZ ZWISCHEN DER BEWERTUNG VON KI-CODE IM REVIEW UND SEINEM VERHALTEN IN DER PRODUKTION.

Dies ist der zentrale Widerspruch der KI-gestützten Ära. 94 % bewerten KI-generierten Code beim Review als qualitativ hochwertiger als von Menschen geschriebenen Code; doch sobald er ausgeliefert wird, berichten 78 % von mehr Incidents, 86 % von mehr Notfalleinsätzen durch Senior Engineers, 74 % geben an, dass mindestens 25 % des KI-generierten Codes erheblich überarbeitet werden muss, und 82 % haben in den letzten sechs Monaten bereits mindestens einen damit verbundenen Produktionsausfall erlebt.

↑ 94 % Halten KI-Code für hochwertiger als menschlichen Code

↓ 78 % Melden Sie einen Anstieg von Incident in der Produktion

VIBE CODING IST MAINSTREAM, PRODUKTIONSREIF UND NÄHERT SICH GOOGLES VOLUMENKURVE AN

Zwei Drittel der Unternehmen geben nun an, dass 51 % bis 75 % des Codes KI-generiert sind, und 88 % haben Vibe Coding in formelle Produktionsrichtlinien aufgenommen. Nur 5 % beschränken es auf Nicht-Produktionsumgebungen und niemand verbietet es.



Zwei Drittel geben an, dass der Großteil ihres Codes KI-generiert ist

UNGEPRÜFTES VERTRAUEN IN DIE KI-GENERIERUNG KÖNNTE ZU EINER NACHGELAGERTEN PRODUKTIONSKRISE FÜHREN

Der Kernwiderspruch der Vibe-Coding-Ära könnte auf ein tiefgreifend fehlgeleitetes Vertrauen früh im Lebenszyklus zurückzuführen sein. Fast zwei Drittel der Technologie-Führungskräfte (62 %) geben an, dass ihre Teams KI-generiertem Code oft so sehr vertrauen, dass sie ihn ohne manuelle, zeilenweise Überprüfung ausliefern. Diese weit verbreitete, ungeprüfte Akzeptanz erklärt möglicherweise, warum die wahrgenommene Codequalität bei ersten Reviews künstlich hoch bleibt, im produktiven Einsatz jedoch einbricht.

62 % Gaben an, dass ihre Teams KI-Code vor der Überprüfung vertrauen

OBSERVABILITY IST MITTLERWEILE UNVERZICHTBAR FÜR KI-CODE UND TELEMETRIE WIRD ALS PROMPT IN DIE AUSGABE EINGEFÜGT.

96 % der Führungskräfte bewerten Observability für KI-generierten Code als sehr oder extrem wichtig, und 0 % bewerten sie als wenig oder nicht wichtig. 78 % der Teams fordern KI mittlerweile per Prompt auf, Telemetrie (Logs, Traces, Kund:innen-Metriken) als Teil des generierten Codes selbst einzubinden.

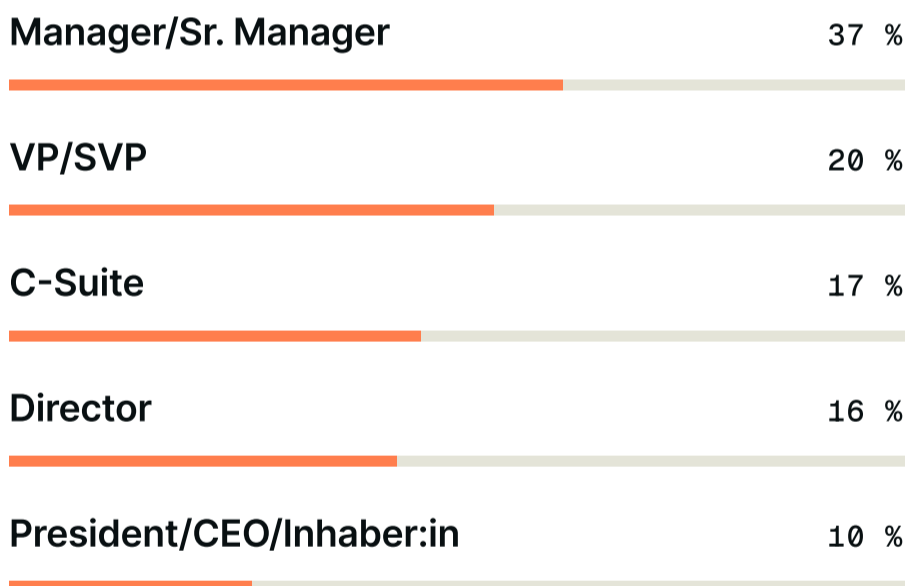
NULL Bewerten Observability als unwichtig für KI-Code

METHODIK

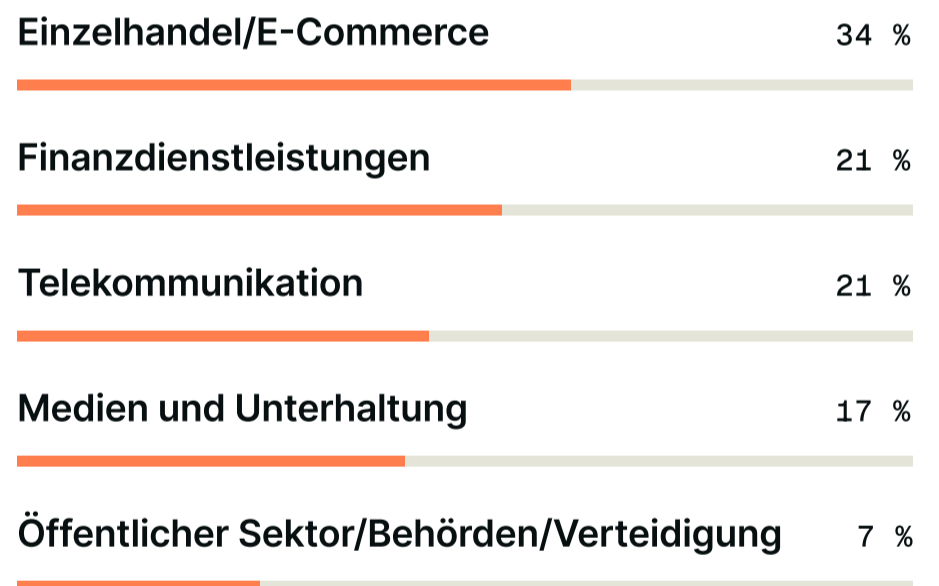
New Relic gab eine Umfrage in Auftrag, die 2026 von Hanover Research online durchgeführt wurde. Es wurden 200 in den USA ansässige Technologie-Entscheider:innen aus den Bereichen IT und Engineering gesampelt, die alle in Vollzeit bei Unternehmen beschäftigt sind, die generative KI für die Software-Entwicklung nutzen. Alle Befragten sind auf Managementebene oder höher tätig und alle verfügen über maßgebliche Entscheidungsbefugnis beim Softwarekauf (65 % primäre Entscheidungsträger:innen, 34 % Mitentscheider:innen).

Die Zahlen in diesem Bericht werden als Prozentwerte auf Basis von n = 200 dargestellt. Zentrale Statistiken im Fließtext verweisen auf das aussagekräftigste Segment einer Distribution; Diagramme zeigen die vollständige Distribution hinter jedem Datenpunkt.

KARRIERESTUFE DER BEFRAGTEN



BRANCHENABDECKUNG



DER UNTERNEHMENSUMSATZ KONZENTRIERT SICH AUF DEN GEHOBENEN MITTELSTAND UND GROSUNTERNEHM

92 %

ORGANISATIONEN HABEN MEHR 500 MITARBEITENDE

18 %

VERZEICHNEN 250 MIO. USD - 499 MIO. USD JAHRESUMSATZ

37 %

MELDEN 500 MIO. USD - 999 MIO. USD JAHRESUMSATZ

33 %

VERZEICHNEN >1 MRD. \$ JAHRESUMSATZ

DIE NEUE NORMALITÄT

KI-generierter Code ist offiziell fest in den Software-Dev-Lifecycle integriert, anstatt ein Sandbox-Hack zu sein.

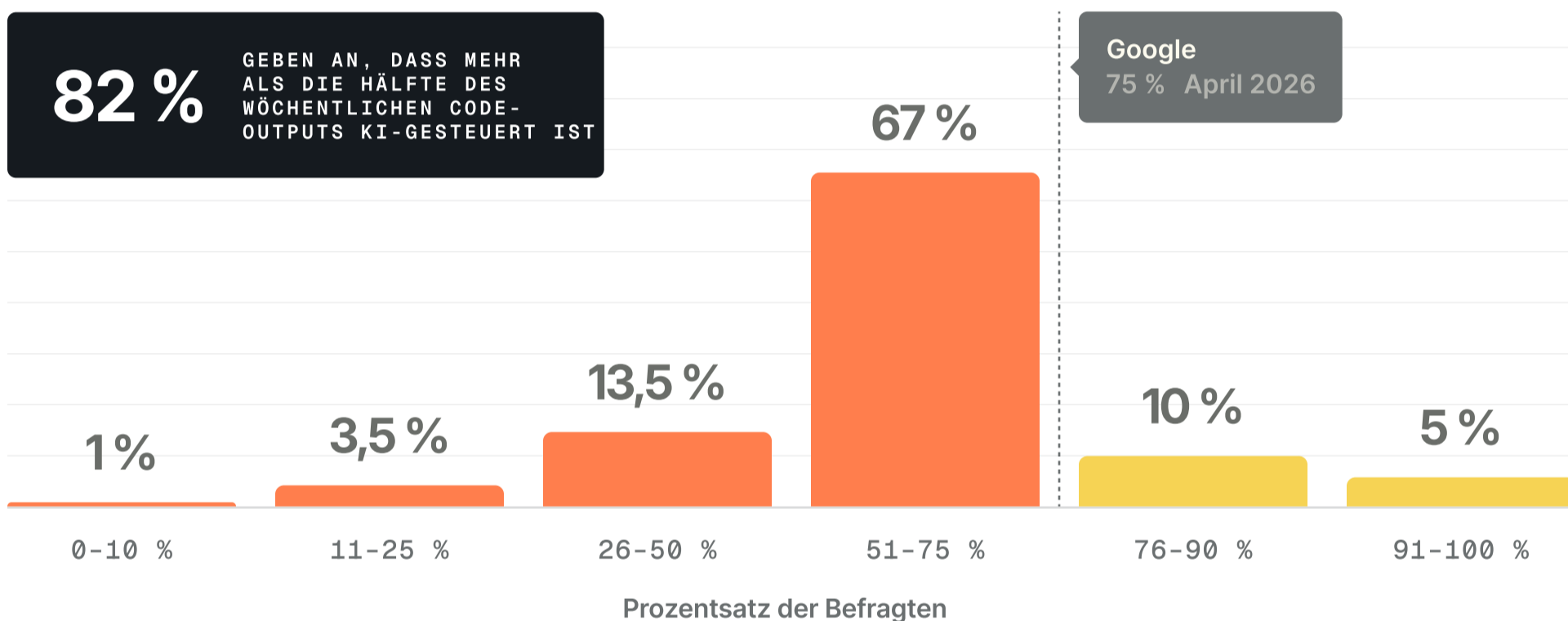


KI-gestützte Programmierung ist mittlerweile zur Norm geworden

67 % der Führungskräfte ordnen den wöchentlichen Output an KI-generiertem Code ihres Unternehmens in den Bereich von 51 % bis 75 % ein.

ANTEIL DES WÖCHENTLICHEN CODES, DER KI-GENERIERT IST ODER DURCH KI EINEM MASSGEBLICHEN REFACTORING UNTERZOGEN WURDE

N = 200



WAS DAS BEDEUTET

Die von Google genannte Zahl von 75 % wurde bei ihrer Veröffentlichung weithin als Ausreißer betrachtet. Die Daten in dieser Studie rücken dies in ein neues Licht. Der Bereich von 51 % bis 75 % ist mittlerweile die häufigste Erfahrung für Technologie-Führungskräfte in den USA. Für die meisten Befragten ist KI der Hauptautor und der Mensch ist der Reviewer.

Bei diesem Aufkommen ist es statistisch unwahrscheinlich, dass der bzw. die auf einen Produktions-Incident reagierende Engineer den Code geschrieben hat, der diesen verursacht hat. Das ändert, was es bedeutet, „das System zu verstehen“.

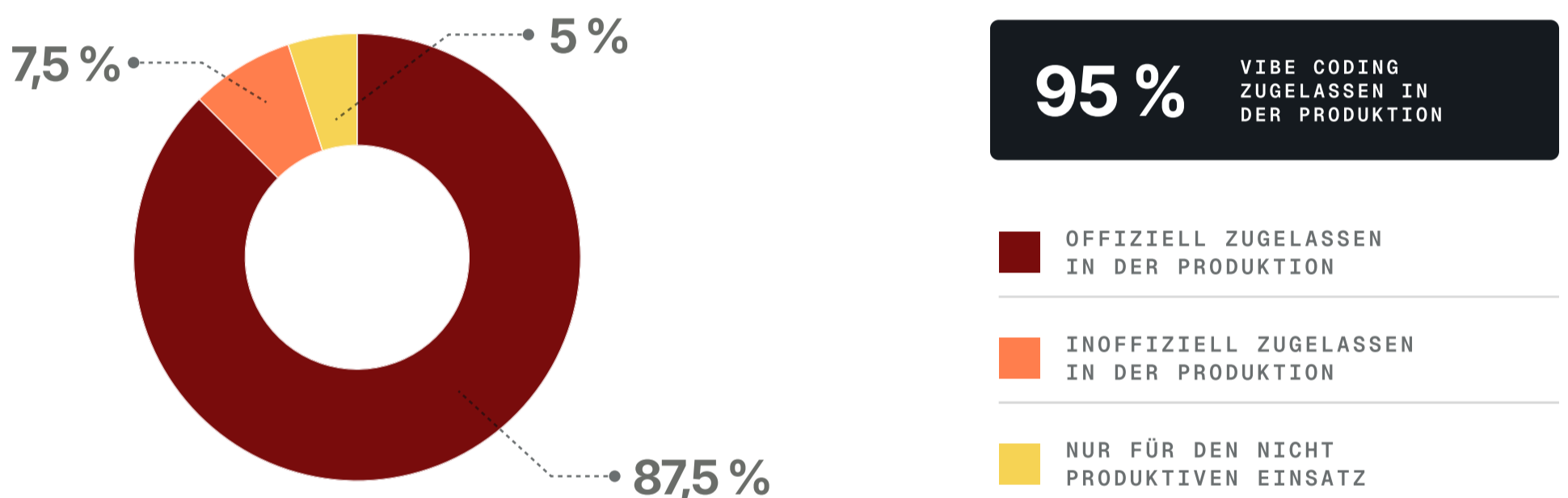
Das Lesen des Quellcodes ist nicht mehr der primäre Weg zum Verständnis. Es gibt zu viel davon, was zu schnell von zu vielen verschiedenen Prompts generiert wird. Laufzeitnachweise (was der Code tatsächlich getan hat, mit welchen Eingaben, in welcher Reihenfolge, mit welcher nachgelagerten Auswirkung) werden zum zuverlässigeren Signal. Engineering-Unternehmen, die ihre Observability-Plattform als das maßgebliche System für „das, was tatsächlich passiert“ betrachten, werden reibungsloser skalieren als jene, die sich weiterhin auf einen Code-Review verlassen.

Vibe Coding hat die Hürde der Produktionsrichtlinien genommen

Nur 5 % der Unternehmen beschränken Vibe Coding auf nicht-produktive Arbeit, und keines verbietet es komplett. Die anderen 95 % erlauben dies für Produktions-Services, entweder formell (87,5 %) oder informell (7,5 %).

UNTERNEHMENSRICHTLINIE ZU VIBE CODING

N = 200



WAS DAS BEDEUTET

Vibe Coding ist keine Nebenbeschäftigung, kein persönlicher Produktivitäts-Hack und keine Sandbox-Technik. Im typischen befragten Unternehmen ist es im Softwareentwicklungs-Lifecycle (SDLC) festgeschrieben. KI-generierter Code trifft auf dieselben Produktionsumgebungen, dieselben umsatzgenerierenden Services und dieselben auf Kund:innen ausgerichteten Endpunkte wie Code, der von Senior Engineers geschrieben wurde, unter denselben SLAs und derselben Ausfalluhr.

Sobald Vibe Coding in der Produktionsrichtlinie verankert ist, muss jedes damit verbundene Governance-System unabhängig von der Urheberschaft gleichermaßen gelten:

Code-Review, Change-Management, Deployment-Schutzmaßnahmen, Incident Response und Audit-Logging.

Die Daten im weiteren Verlauf dieses Berichts zeigen, dass die meisten Unternehmen die Praxis zugelassen haben, ohne die damit verbundenen Kontrollen konsequent zu verbessern.

Führungskräfte, die diese Lücke frühzeitig schließen, indem sie dieselbe Disziplin in den Bereichen Release Health, Change Tracking und SLO-Monitoring, die sie bereits auf von Menschen geschriebenen Code anwenden, auf KI-generierten Code ausweiten, werden den operativen Schock reibungsloser abfedern als diejenigen, die Vibe-codierte Services als separate Kategorie behandeln.

KI-generierter Code erzielt bessere Bewertungen als von Menschen geschriebener Code

61 % der Führungskräfte bewerten KI-generierten Code als von „etwas höherer Qualität“ als von Menschen erstellten Code. Weitere 33 % bewerten es als „viel höher“.

WAHRGENOMMENE QUALITÄT VON KI-GENERIERTEM CODE
IM VERGLEICH ZU VON MENSCHEN GESCHRIEBENEM CODE



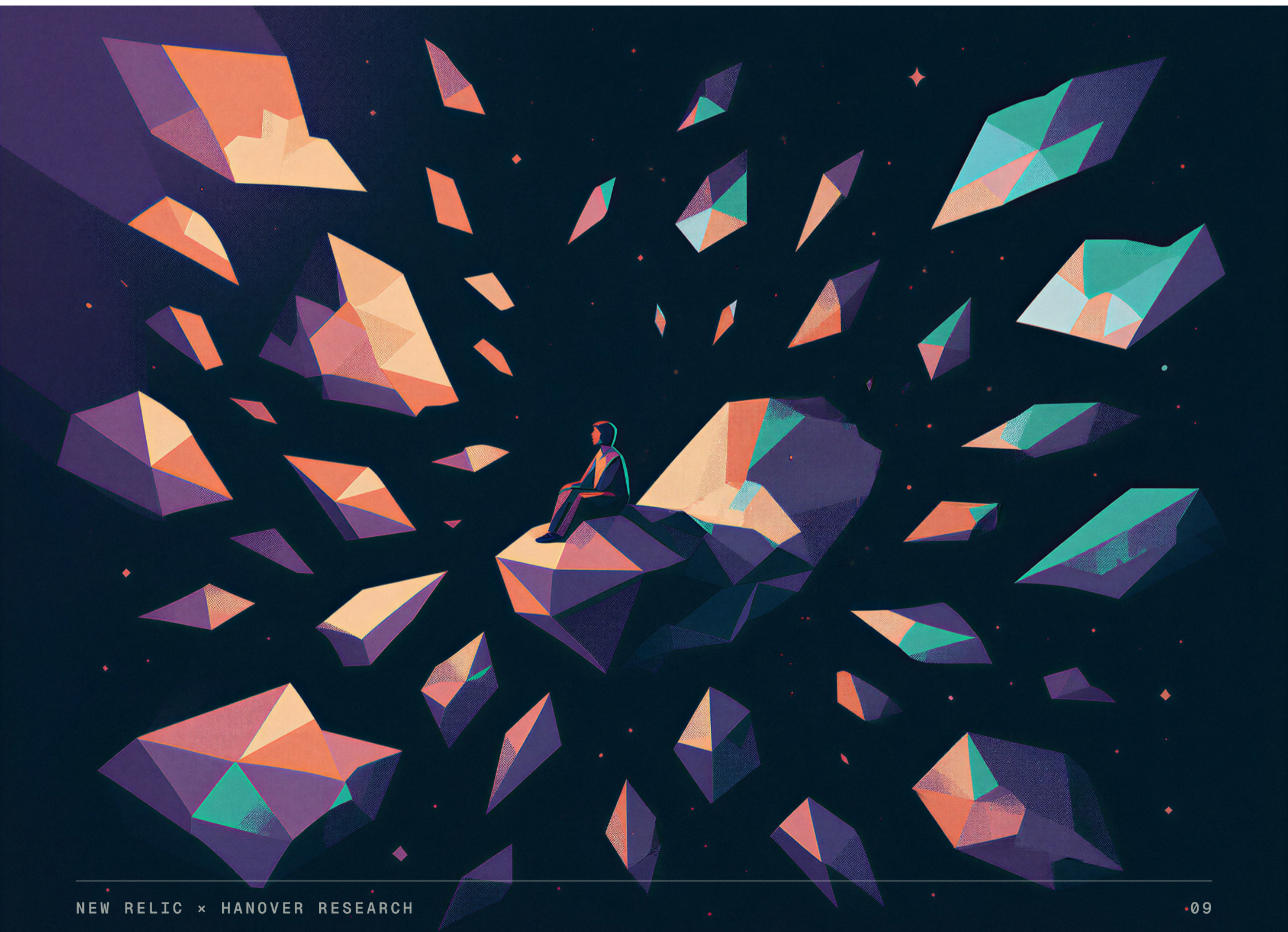
WAS DAS BEDEUTET

Der Hauptvorteil der KI-gestützten Programmierung besteht darin, dass sie ihr Versprechen an dem Punkt einlöst, an dem die meisten Engineering-Unternehmen die Messung der Qualität durchführen: beim Code-Review. Sauberere Muster, einheitlicher Stil, weniger offensichtliche Fehler beim Einreichen. Diese Wahrnehmung ist real, sie ist weit verbreitet und sie ist mit ein Grund dafür, dass Unternehmen bereit sind, die Einführung von Vibe Coding weiter voranzutreiben.

Die wichtige Einschränkung besteht darin, dass die Qualitätsbewertung in diesem Datenpunkt eine Beurteilung ist, die zum Zeitpunkt des Code-Reviews und nicht zum Zeitpunkt der Incident Response vorgenommen wird. Code, der sich gut liest, funktioniert nicht unbedingt gut im Betrieb, und die Metriken, die jeden dieser Aspekte erfassen, unterscheiden sich stark. Die Daten aus dieser Umfrage zeigen, was passiert, sobald derselbe Code unter echtem Traffic, gegen echte Abhängigkeiten und mit echten Kund:innen am anderen Ende ausgeführt wird.

DIE LAUFZEIT- REALITÄT

Die Erstellung von KI-Code hat sich beschleunigt, aber die Aufrechterhaltung der Systemzuverlässigkeit ist teurer und arbeitsintensiver geworden.



Die operative Belastung ist gleichzeitig gestiegen

Die Nacharbeitszeit von Senior Engineers ist in den letzten 12 Monaten in 86 % der Unternehmen gestiegen. Im selben Zeitraum haben der Wildwuchs von Microservices, Produktions-Incidents und technische Schulden bei drei Vierteln der Unternehmen oder mehr allesamt zugenommen.

ANTEIL DES KI-GENERIERTEN CODES, DER ERHEBLICHE ÜBERARBEITUNG ERFORDERT

N = 200

Anzahl der Microservices/Funktionen in der Umgebung



Zeit, die erfahrene und leitende Mitarbeitende für die Fehlerbehebung im Code aufwenden



Anzahl der Produktions-Incident



technische Schulden



■ REDUKTION
 ■ ERHÖHUNG

WAS DAS BEDEUTET

Die vier oben genannten Metriken sind die häufigsten nachgelagerten Kosten der KI-gestützten Entwicklung, und sie haben sich über denselben 12-Monats-Zeitraum, in dem sich die Geschwindigkeit und die wahrgenommene Qualität verbessert haben, im Gleichschritt bewegt. Geschwindigkeit ist günstiger geworden. Zuverlässigkeit nicht. Dieselben Unternehmen, die von den Geschwindigkeitssteigerungen berichten, sind diejenigen, die die hier gezeigte operative Belastung auffangen.

Dies ist der zentrale Kompromiss der KI-gestützten Ära, und er ist nun in den Daten dokumentiert. Die richtige strategische Reaktion ist nicht, das Tempo zu drosseln. Die Geschwindigkeitsgewinne sind real und der Umsatz spiegelt dies wider.

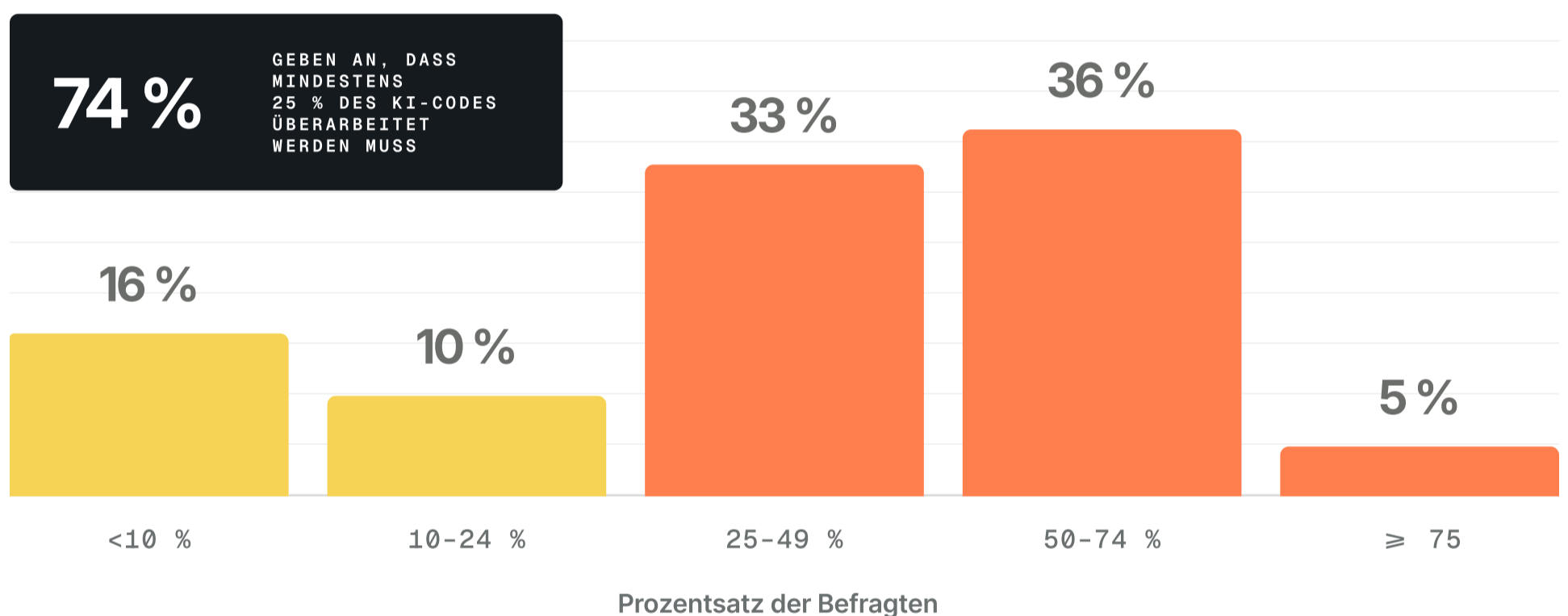
Aufgrund der schnelleren Feature-Bereitstellung berichten 63 % von einer leichten Umsatzsteigerung. Die richtige Reaktion ist, explizit Budget für den operativen Overhead einzuplanen, der mit ihnen einhergeht. Das bedeutet, Incident-Raten, MTTR, Service für technische Schulden und die Nacharbeitszeit von Senior Engineers im selben Kapazitätsplan zu berücksichtigen, der Deployment-Häufigkeit und Feature-Throughput als Ziel hat. Unternehmen, die ihren Führungsteams beide Seiten der Bilanz offenlegen, werden bessere Entscheidungen darüber treffen, wo sie die Geschwindigkeit weiter erhöhen und wo sie pausieren und konsolidieren sollten.

KI-generierter Code führt zu einem Anstieg der Nacharbeit

41 % der Führungskräfte geben an, dass mindestens die Hälfte ihres KI-generierten Codes aufgrund von mangelhaftem Kontext, unvollständigen Daten oder falschen Annahmen eine erhebliche Überarbeitung erfordert. 74 % geben an, dass dies bei mindestens einem Viertel der Fall ist.

ANTEIL DES KI-GENERIERTEN CODES, DER ERHEBLICHE ÜBERARBEITUNG ERFORDERT

N = 200



WAS DAS BEDEUTET

Die „Fix-Forward“-Kosten beim Vibe Coding sind erheblich. Selbst in Unternehmen, in denen KI den Großteil des Codes produziert, geht ein nennenswerter Anteil dieses Outputs zurück zur Korrektur, nicht für den Feinschliff, sondern wegen fehlenden Kontexts, halluzinierter Schnittstellen oder falscher Annahmen über Daten. Diese Nacharbeit wird zunehmend von Senior Engineers durchgeführt, die andernfalls an Problemen mit größerer Hebelwirkung arbeiten würden.

Die Grundursache ist, dass KI-Programmierungstools Code generieren, ohne zu sehen, wie sich das System tatsächlich zur Laufzeit verhält.

Sie können den Quellcode lesen, aber sie können den Trace nicht lesen. Die Unternehmen, die damit begonnen haben, diese Lücke zu schließen (indem sie Produktionstelemetrie, reale Abfragemuster und reale Abhängigkeitsgraphen zurück in den Entwicklungszyklus einspeisen), sind diejenigen, die die niedrigsten Nacharbeitsraten verzeichnen. Den Kreislauf zwischen Laufzeitdaten und Codegenerierung zu schließen, dürfte die bedeutendste Observability-Praxis der nächsten 24 Monate sein, und es ist die Praxis, die den operativen Aufwand am direktesten reduziert.

Unternehmen spüren die Beeinträchtigung durch Fehlermodi

Nur 19 % der Unternehmen geben an, in den letzten sechs Monaten keine Herausforderungen mit KI-generiertem Code gehabt zu haben. Die anderen 82 % geben jeweils im Durchschnitt mehr als zwei Produktions-Probleme an.

DIE GRÖSSTEN HERAUSFORDERUNGEN IM PRODUKTIVBETRIEB DURCH KI-GENERIERTEN CODE IN DEN LETZTEN 6 MONATEN

N = 200

| | |
|----------------------------------|------|
| Integrationsfehler | 30 % |
| Compliance/Governance-Probleme | 30 % |
| Probleme mit der Datenintegrität | 29 % |
| Neu eingeführte Schwachstellen | 28 % |

WAS DAS BEDEUTET

Vier verschiedene Fehlerarten (Integrationsfehler 31 %, Compliance-/Governance-Issues 30 %, Datenintegritätsprobleme 29 % und neu eingeführte Sicherheitsschwachstellen 28 %) haben innerhalb von sechs Monaten jeweils etwa drei von zehn Unternehmen betroffen. Historisch gesehen hat von Menschen geschriebener Code, der von Fachkolleg:innen überprüft wurde, zu stabilen Baseline-Fehlerraten geführt, aber KI-generierter Code verursacht etwa 1,7-mal mehr kritische Laufzeit-Probleme. KI-generierter Code legt die Produktion nicht auf eine einzige, charakteristische Art und Weise lahm. Es verursacht gleichzeitig viele kleine Störungen, und die Ausfälle sind so weit verbreitet, dass die meisten Unternehmen mindestens eine Horrorgeschichte aus den letzten zwei Quartalen haben.

Jeder der vier genannten Fehlermodi weist eine eindeutige beobachtbare Signatur auf.

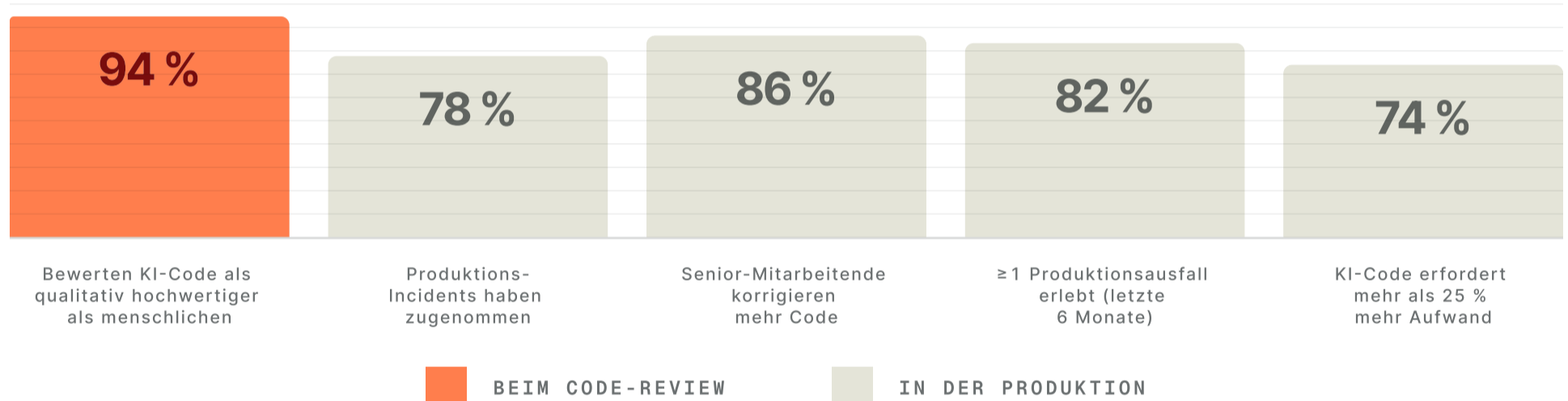
Fehler bei der Integration zeigen sich als Schema-Drift, Vertragsverstöße und steigende Fehlerquoten zwischen Services. Compliance-Issues werden in Audit-Logs und Richtlinienverstößen sichtbar. Datenintegritätsprobleme zeigen sich als Anomalien in der Produktionstelemetrie: Duplizierung, Drift, unerwartete Nullwertraten. Sicherheitsschwachstellen zeigen sich in Trace-Daten, Authentifizierungsmustern und Verhaltensanomalien. Unternehmen, die jeden Fehlermodus einem spezifischen Erkennungsmuster in ihrer Observability-Plattform zugeordnet haben, werden die Lücke zwischen „Ein KI-generierter Fehler hat die Produktion erreicht“ und „Wir haben ihn abgefangen, bevor er echten Schaden anrichten konnte“ viel schneller schließen als jene, die sich noch auf Beschwerden von Benutzer:innen oder geplante Audits verlassen. Dies ist der praktische Anwendungsfall für eine einheitliche Telemetrie-Oberfläche, und es ist einer der Gründe, warum New Relic und Plattformen wie diese steigende KI-Code-Workloads verzeichnen.

Der Widerspruch im Zentrum der KI-gestützten Ära

Dieselben Unternehmen, die KI-generierten Code beim Review als qualitativ hochwertiger bewerten als von Menschen geschriebenen Code (94 %), melden mehr Incidents (78 %), mehr Nacharbeit (74 % geben an, dass mehr als 25 % des Codes Nacharbeit erfordern), mehr Firefighting durch Senior Engineers (86 %) und mehr Produktionsausfälle (82 % in den letzten sechs Monaten) durch denselben Code, sobald er ausgeliefert wird.

KI-CODE SIEHT IM REVIEW BESSER AUS,
SCHNEIDET IN DER PRODUKTION SCHLECHTER AB

N = 200



WAS DAS BEDEUTET

Dies ist das wichtigste Muster in der Studie. Reviewers sehen Code, der sauberer, konsistenter und besser strukturiert aussieht als das, was Menschen normalerweise produzieren. In der Produktion kommt Code zum Einsatz, der häufiger und auf mehr Arten fehlschlägt und dessen Behebung mehr Aufmerksamkeit von Senior Engineers erfordert. Beide Abbildungen sind korrekt. Aber sie messen nicht dasselbe. Die erste führt eine Messung der Lesbarkeit und Oberflächenqualität durch. Die zweite Messung erfasst das Verhalten unter echter Last, echten Abhängigkeiten und echten Edge-Fällen – jenen Bedingungen, in die KI-Programmertools keine Einblicke haben, wenn sie den Code generieren.

Der Widerspruch hat eine einzige Auflösung: Der Grund, warum KI-generierter Code im Review gut aussieht und in der Produktion schlechter abschneidet, ist, dass die KI mit

denn Artefakten arbeitet, die menschliche Prüfer:innen sehen können (Source), und ist blind für die Artefakte, die nur die Produktion erzeugen kann (Trace).

Diese Lücke zu schließen, ist für die nächsten Jahre die strategische Aufgabe der Observability-Schicht. Die Unternehmen, die die KI-gestützte Ära für sich entscheiden, werden diejenigen sein, die ihre Produktionstelemetrie als erstklassigen Input für den Entwicklungszyklus betrachten, nicht nur als nachträgliches Debugging-Tool; als ein Feedback-Signal, das den nächsten Prompt, das nächste Code-Review und die nächste Deployment-Entscheidung verbessert. Genau dafür gibt es den Bereich „Observability“, und die Daten dieser Studie verdeutlichen am besten, warum er so wichtig ist.

Blindes Vertrauen in KI-Generierung führt zu einer Produktionskrise

62 % der Führungskräfte aus der Technologiebranche geben an, dass ihre Engineering-Teams dem von KI generierten Code „oft“ so sehr vertrauen, dass sie ihn ohne manuelle Überprüfung Zeile-für-Zeile in die Produktion übernehmen.

BEFRAGTE, DIE KI-CODE OHNE ZEILENWEISE ÜBERPRÜFUNG VERTRAUEN

N = 200

62 %

HÄUFIG ODER IMMER

WAS DAS BEDEUTET

Der Kernwiderspruch der Vibe-Coding-Ära könnte auf ein tiefgreifend falsch platziertes Vertrauen früh im Dev-Lifecycle zurückzuführen sein. Diese weit verbreitete Akzeptanz könnte erklären, warum die wahrgenommene Codequalität bei ersten Reviews künstlich hoch bleibt, in der Praxis jedoch sofort einbricht.

LLMs sind hervorragend darin, Code zu generieren, der unter isolierten, optimalen Bedingungen perfekt funktioniert. Allerdings lassen sie regelmäßig Grenzfälle, blinde Flecken bei der Parallelität, veraltete API-Abkündigungen und komplexe Zustandsänderungen außer Acht.

Wenn sie ungeprüft ausgeliefert werden, bleiben diese strukturellen Lücken völlig unbemerkt, bis sie auf Live-Traffic von Benutzer:innen treffen.

Die unüberprüfte Übernahme von KI-generiertem Code in die Produktion kann auch einen massiven Anstieg an lokalen Produktions-Incidents auslösen. SRE- und DevOps-Teams sehen sich zunehmend gezwungen, als „KI-Code-Hausmeister:innen“ zu fungieren, wobei sie bis zu einem Drittel ihrer wöchentlichen Arbeitszeit von der Roadmap-Beschleunigung abziehen, nur um ungeprüften maschinellen Output zu sichten und einem Refactoring zu unterziehen.

DAS ZUKÜNFTIGE FRAMEWORK

Technologie-Entscheider:innen stimmen einhellig darin überein, dass Observability eine absolute Notwendigkeit ist, wenn es um die Verwaltung von KI-generiertem Code geht.

Observability ist mittlerweile eine Grundvoraussetzung

39 % der Technologie-Führungskräfte bewerten Observability bei der Arbeit mit KI-generiertem Code als „äußerst wichtig“, die am häufigsten gewählte Antwort. Weitere 57 % bewerten es als „sehr wichtig“. Niemand bewertete es in dieser Umfrage als wenig oder gar nicht wichtig.

DIE BEDEUTUNG VON OBSERVABILITY BEI DER ARBEIT MIT KI-GENERIERTEM CODE

N = 200



WAS DAS BEDEUTET

Die Debatte darüber, ob Observability für KI-generierten Code unerlässlich ist, ist faktisch beendet. Da null Befragte die beiden untersten Skalenpunkte gewählt haben, stellt sich nicht mehr die Frage, ob in Observability investiert werden soll, sondern vielmehr, auf welche Plattform standardisiert werden soll.

Wenn sich die Wichtigkeitswerte an der Spitze so dicht zu einem Cluster verdichten, verlagert sich das strategische Handeln vom Kaufen auf das Konsolidieren. Die meisten der befragten Unternehmen betreiben bereits einen Observability Stack, und viele nutzen mehr als ein Tool.

Die Führungskräfte, die den größten Nutzen aus einem nahezu einstimmigen Konsens ziehen, sind diejenigen, die diesen nutzen, um auf eine einzige Plattform zu standardisieren, die sowohl Observability als auch AIOps abdeckt, anstatt weiterhin für eine fragmentierte Abdeckung zu bezahlen, der keiner der Engineers im Bereitschaftsdienst um 2 Uhr morgens voll und ganz vertraut.

Plattformen wie New Relic, die alle vier Signale unter einem Datenmodell und einer Abfrageoberfläche vereinen, sind für diesen Konsolidierungsschritt bestens gerüstet.

Telemetrie wird in den Prompt der KI geschrieben

56 % der Führungskräfte geben an, dass ihre Teams die KI „oft“ per Prompt auffordern, bei der Code-Generierung spezifische Telemetrie (Logs, Traces, benutzerdefinierte Metriken) einzubeziehen. Weitere 22 % tun dies „immer“. Niemand tut dies selten oder nie.

WIE OFT TEAMS DIE KI PER PROMPT ANWEISEN, TELEMETRIE IN GENERIERTEN CODE ZU INTEGRIEREN

N = 200



78 %

GEBEN AN, DASS IHRE TEAMS OFT ODER IMMER PROMPTS FÜR LOGS, SPANS UND BENUTZERDEFINIESTE METRIKEN VERWENDEN

WAS DAS BEDEUTET

Engineers warten nicht erst, bis der Code geschrieben wurde, um sich über Observability Gedanken zu machen. Sie ziehen sie nach vorne in den Prompt selbst. Telemetrie wird Teil der KI-Definition von „fertig“.

Die Entscheidung darüber, welche Logs und Traces erfasst und worüber Alerts ausgelöst werden sollen, verlagert sich weiter nach vorne, aus dem SRE-Backlog in den Prompt der Entwickler:innen. Daraus ergeben sich zwei Konsequenzen. Erstens hängt die Qualität der Telemetrie eines Unternehmens nun stark von den Prompts und Vorlagen ab, die IT-Engineers verwenden.

Zweitens werden Unternehmen, die auf meinungsstarke, schemabewusste Konventionen (semantische OpenTelemetry-Konventionen, strukturiertes Logging, konsistente Span-Attribute) standardisieren, ein viel schärferes Verhältnis von Signal zu Signalrauschen erzielen als jene, die Entwickler:innen ihr eigenes Format per Prompt festlegen lassen. Wer den Standard in Cursor, Copilot, Claude Code und ChatGPT setzt, legt faktisch den Observability-Standard für die nächste Phase dieser Kategorie fest.

Fazit

Drei Erkenntnisse prägen das restliche Jahr 2026 für Engineering-Unternehmen.

GESCHWINDIGKEIT IST REAL, IHR PREIS WIRD ZUR LAUFZEIT FÄLLIG.

Schnellere Releases, häufigere Deployments und eine höhere Qualität in der Review-Phase halten alle, was sie versprechen. Die Kosten zeigen sich in Incidents, Nacharbeit, Wildwuchs und der Zeit von Senior Engineers. Führungskräfte, die beide Seiten dieser Bilanz in ihre Planung einbeziehen, werden bessere Entscheidungen darüber treffen, wo sie vorantreiben und wo sie verstärken sollten.

DIE LÜCKE ZWISCHEN BEDENKEN UND KONTROLLE IST DIE AM BESTEN UMSETZBARE ERKENNTNIS DER STUDIE.

Führungskräfte sind zu Recht besorgt über die Risiken von KI-generiertem Code. Die Kontrollen, die sie tatsächlich bereitgestellt haben, werden dem Anliegen noch nicht gerecht. Die Unternehmen, die diese Lücke am schnellsten schließen – mit mehrschichtigen Schutzmaßnahmen in den Bereichen Sicherheitsscans, Deployment-Leitplanken, Governance-Richtlinien und Laufzeit-Observability –, werden diejenigen sein, die KI-lastige Codebasen unter Einhaltung von Enterprise-SLA ohne inakzeptables Risiko betreiben.

DER WIDERSPRUCH ZWISCHEN DER QUALITÄT BEIM REVIEW UND DER PRODUKTIONSREALITÄT IST GENAU DER PUNKT, AN DEM OBSERVABILITY IHREN WERT BEWEIST.

KI-Coding-Tools sehen den Quellcode. Sie sehen den Trace nicht. Bis sich diese Schleife schließt, wird KI-generierter Code im Review weiterhin gut abschneiden und Teams in der Produktion überraschen. Die Observability-Plattform, die zum System of Record dafür wird, was KI-Code tatsächlich tut, und dieses Signal in den Entwicklungszyklus zurückspeist, steht im Zentrum dessen, wie Engineering-Unternehmen im KI-gestützten Zeitalter arbeiten. New Relic wurde genau für diesen Kreislauf entwickelt, weshalb dies die Ebene ist, in die es sich zuerst zu investieren lohnt.

New Relic hat diese Studie veröffentlicht, weil die darin beantworteten Fragen genau die Fragen werden, mit denen sich derzeit jedes Engineering-Unternehmen auseinandersetzt. Die Daten zeigen, dass die Ära KI-generierter Software angebrochen ist, dass ihre operativen Konsequenzen real sind und dass Observability zu der Ebene geworden ist, auf der diese Konsequenzen verwaltet werden.

ÜBER HANOVER

Hanover Research ist ein globales Unternehmen für Marktforschung und Analysen, das mit einem Festpreis-Abonnementmodell arbeitet, um maßgeschneiderte quantitative und qualitative Einblicke zu liefern. Das 2003 gegründete Unternehmen beschäftigt über 200 Analyst:innen, Statistiker:innen und Umfragemethodiker:innen. Hanover Research führt Primärmarktforschung, prädiktive Datenmodellierung und Segmentanalysen durch, um strategische, operative und finanzielle Entscheidungen in den Bereichen Wirtschaft, Gesundheitswesen und Bildung zu unterstützen.

ÜBER NEW RELIC

Die intelligente Observability-Plattform von New Relic erfasst, normalisiert und analysiert Telemetriedaten über den gesamten Software-Stack hinweg. Durch die Vereinheitlichung von Metriken, Events, Logs und Traces in einer einzigen Telemetry Data Platform verlagert New Relic das System-Engineering vom reaktiven Debugging zum prädiktiven Performance-Management. Unternehmen, darunter Adidas Runtastic, das Amerikanische Rote Kreuz, Domino's, GoTo Group, Ryanair, Topgolf und William Hill, nutzen daher New Relic zur Überwachung der Infrastruktur-Health und zur Optimierung der Anwendungs-Performance.

ÜBER DIESEN REPORT

Die Daten in diesem Bericht stammen aus einer Online-Umfrage aus dem Jahr 2026, die von Hanover Research durchgeführt und von New Relic in Auftrag gegeben wurde. Es wurden 200 in Vollzeit beschäftigte, in den USA ansässige Technologie-Entscheider:innen in IT- und Engineering-Rollen bei Unternehmen gesampelt, die generative KI für die Entwicklung von Software nutzen.



[Mehr zur New Relic Plattform](#)