

Fallstudie

Kann eine KI ihren eigenen 360-Grad-Feedback-Fragebogen beurteilen

Ein dokumentierter Praxistest mit 24 Aussagen und 14 Gütekriterien

0	5	8	nicht nachgewiesen
Kriterien vollständig erfüllt	Kriterien teilweise erfüllt	Kriterien noch nicht geprüft	Kriteriumsvalidität

Kernaussage

Die KI konnte innerhalb weniger Minuten einen professionell wirkenden Fragebogen erstellen. Sie konnte außerdem erkennen, welche Qualitätsnachweise fehlen. Sie konnte diese Nachweise jedoch nicht selbst erzeugen.

Prof. Dr. Waldemar Pelz

Institut für Management-Innovation | September 2026

Worum es in dieser Fallstudie geht

Ein allgemeines generatives KI-System erhielt den Auftrag, aus seinem gelernten Sprach- und Fachwissen einen neuen Fragebogen für ein 360-Grad-Feedback zu formulieren. Anschließend musste es den eigenen Entwurf anhand einer Checkliste mit 14 Gütekriterien beurteilen.

Die Prüfung folgte einer strengen Regel: Ein Kriterium durfte nur dann als erfüllt gelten, wenn ein konkreter Nachweis vorlag. Professionelle Formulierungen und theoretische Plausibilität galten nicht als Beleg.

Das Ergebnis ist für die betriebliche Praxis entscheidend. Die KI erzeugte einen überzeugenden Entwurf, konnte aber kein einziges Gütekriterium vollständig nachweisen. Besonders wichtig: Es fehlte jeder Nachweis, dass hohe Fragebogenwerte tatsächlich mit erfolgreicher Führung zusammenhängen.

Was die Fallstudie zeigt

- KI kann Formulierungen, Gliederungen und erste Entwürfe erheblich beschleunigen.
- KI kann formale Schwächen benennen, wenn sie ausdrücklich dazu aufgefordert wird.
- KI kann ohne reale Befragungsdaten weder Messzuverlässigkeit noch Trennschärfe oder Faktorstruktur nachweisen.
- KI kann ohne passende Erfolgsdaten nicht beurteilen, ob der Fragebogen wirksame Führung erkennt.
- Die Selbstprüfung durch dasselbe KI-System ist keine unabhängige Validierung.

Einordnung

Die Fallstudie ist kein Vergleich verschiedener KI-Systeme und keine wissenschaftliche Validierungsstudie. Sie dokumentiert einen nachvollziehbaren Praxistest. Der verwendete Fragebogen ist ein KI-Entwurf und darf nicht als validiertes Instrument für Personalentscheidungen verstanden werden.

Die verwendeten Aussagen wurden für diesen Versuch neu formuliert. Es wurden keine Original-Items des Berufsbezogenen Persönlichkeitstests, des Gießener Inventars der Umsetzungskompetenz oder des Gießener Inventars der Transformationalen Führung eingesetzt.

So wurde der Praxistest durchgeführt

Der Versuch bestand aus zwei aufeinanderfolgenden Aufgaben. Zunächst sollte die KI einen vollständigen, geschlechtsneutralen Fragebogen mit konkreten Verhaltensaussagen erstellen. Danach erhielt sie die Checkliste und musste den eigenen Entwurf bewerten.

Auftrag zur Erstellung

Erstelle aus deinem allgemeinen Fachwissen einen Fragebogen für ein entwicklungsorientiertes 360-Grad-Feedback. Verwende 20 bis 30 geschlechtsneutrale Aussagen zu beobachtbarem Führungsverhalten. Formuliere die Aussagen mit dem gemeinsamen Satzanfang ‚Diese Person ...‘. Nutze eine fünfstufige Häufigkeitsskala sowie die Antwortmöglichkeit ‚kann ich nicht beurteilen‘. Verwende keine Items aus geschützten oder validierten Originalinstrumenten.

Auftrag zur Prüfung

Prüfe den von dir erstellten Fragebogen anhand der vorgelegten Checkliste. Bewerte ein Kriterium nicht allein deshalb als erfüllt, weil die Aussagen verständlich oder plausibel klingen. Nenne für jedes Urteil den konkreten Nachweis. Wenn Daten aus realen Befragungen oder statistische Untersuchungen fehlen, markiere das Kriterium als ‚noch nicht geprüft‘ beziehungsweise ‚nicht nachgewiesen‘. Unterscheide ausdrücklich zwischen einer formalen Prüfung der Formulierungen und einem Nachweis der tatsächlichen Qualität des Fragebogens.

Dokumentation

- Generatives KI-System: OpenAI ChatGPT im Codex-Arbeitsmodus
- Durchführung: 22. September 2026
- Fachliche Einordnung: Prof. Dr. Waldemar Pelz
- Prüfgrundlage: Checkliste zur Qualität eines Fragebogens für ein 360-Grad-Feedback
- Untersuchungsgegenstand: neu erzeugter Fragebogen mit 24 Aussagen in sechs Bereichen
- Nicht vorhanden: reale Befragungsdaten, Vergleichswerte, statistische Kennzahlen und externe Erfolgsdaten

Warum die Beweisregel notwendig ist

Ein Sprachmodell kann seine eigenen Formulierungen für verständlich und fachlich plausibel halten. Daraus folgt jedoch nicht, dass verschiedene Personen die Aussagen gleich verstehen, dass die Items statistisch zusammengehören oder dass der Fragebogen erfolgreiche Führung erkennt. Die Beweisregel verhindert, dass fehlende Daten durch sprachliche Sicherheit ersetzt werden.

Fragebogen zum beobachtbaren Führungsverhalten

Bitte denken Sie bei jeder Aussage an dieselbe Führungskraft und an deren Verhalten während der vergangenen sechs Monate. Vervollständigen Sie gedanklich den Satz: Diese Person ...

1	2	3	4	5	0
nie	selten	manchmal	häufig	fast immer	kann ich nicht beurteilen

1 Ziele und Entscheidungen

- 1 erklärt, welche Ergebnisse erreicht werden sollen.
- 2 macht deutlich, welche Aufgaben Vorrang haben.
- 3 nennt nachvollziehbare Gründe für wichtige Entscheidungen.
- 4 passt Prioritäten an, wenn sich die Rahmenbedingungen ändern.

2 Kommunikation und Beteiligung

- 5 gibt relevante Informationen rechtzeitig weiter.
- 6 hört Einwände an, bevor sie Stellung nimmt.
- 7 fragt vor wichtigen Entscheidungen nach abweichenden Einschätzungen.
- 8 klärt nach Besprechungen, wer welche Aufgabe übernimmt.

3 Delegation und Verantwortung

- 9 überträgt Aufgaben mit einem klar beschriebenen Ergebnis.
- 10 macht deutlich, welche Entscheidungen Mitarbeitende selbst treffen dürfen.
- 11 hält vereinbarte Entscheidungsspielräume ein.
- 12 unterstützt bei Hindernissen, ohne die Aufgabe sofort wieder an sich zu ziehen.

4 Feedback und Entwicklung

- 13 spricht beobachtetes Verhalten zeitnah an.
- 14 beschreibt konkret, welche Auswirkungen dieses Verhalten hatte.
- 15 vereinbart einen konkreten nächsten Entwicklungsschritt.
- 16 greift getroffene Entwicklungsvereinbarungen später wieder auf.

5 Zusammenarbeit und Konfliktklärung

- 17 spricht Spannungen an, bevor sie sich verschärfen.
- 18 gibt den Beteiligten Gelegenheit, ihre Sichtweise darzustellen.
- 19 konzentriert sich bei Meinungsverschiedenheiten auf die konkrete Sache.
- 20 hält nach einer Konfliktklärung fest, was vereinbart wurde.

6 Veränderung und Lernen

21 erklärt, warum eine Veränderung notwendig ist.

22 erprobt neue Vorgehensweisen zunächst in überschaubaren Schritten.

23 fragt nach einem Versuch, was funktioniert hat und was nicht.

24 korrigiert Entscheidungen, wenn neue Informationen gegen die bisherigen Annahmen sprechen.

Offene Fragen

1. Welches Verhalten dieser Person trägt besonders zu guter Zusammenarbeit oder guten Ergebnissen bei?

2. Welche konkrete Verhaltensänderung hätte aus Ihrer Sicht den größten Nutzen?

AUSGEFÜLLTE CHECKLISTE

Bewertung der 14 Gütekriterien

Die Bewertungen beziehen sich ausschließlich auf den vorliegenden KI-Entwurf und die tatsächlich verfügbaren Nachweise. Fehlende Daten werden nicht als Qualitätsmangel behauptet, sondern als offene Prüfung ausgewiesen.

Bewertungsschlüssel

Erfüllt	Ein konkreter Nachweis oder eine klare, überprüfbare Begründung liegt vor.
Teilweise erfüllt	Die Anforderung ist nur teilweise belegt oder erkennbar eingeschränkt.
Nicht erfüllt	Die Anforderung wird nachweislich nicht erreicht.
Noch nicht geprüft	Die notwendigen Daten oder Nachweise fehlen. Daraus folgt noch kein negatives Qualitätsurteil.

Prüfgrenze

Die Bewertung ‚noch nicht geprüft‘ ist bewusst von ‚nicht erfüllt‘ zu unterscheiden. Ein Entwurf kann sich später als hochwertig erweisen. Solange die erforderlichen Untersuchungen fehlen, darf seine Qualität jedoch nicht als nachgewiesen dargestellt werden.

Kriterien 1 bis 8

Nr.	Kriterium	Bewertung	Begründung
1	Inhaltsvalidität	Noch nicht geprüft	Es fehlen eine systematische Anforderungsanalyse, ein dokumentiertes Kompetenzmodell und Beurteilungen unabhängiger Fachleute. Die sechs Bereiche wirken plausibel, ihre Vollständigkeit ist aber nicht belegt.
2	Verständlichkeit und Beobachtbarkeit	Teilweise erfüllt	Die meisten Aussagen beschreiben konkretes Verhalten und enthalten einen Bezugszeitraum. Ob alle Beurteilergruppen die Aussagen gleich verstehen und beobachten können, wurde nicht mit realen Personen getestet.
3	Konstruktvalidität	Noch nicht geprüft	Die Zuordnung der Items zu den sechs Kompetenzen beruht nur auf sprachlicher Plausibilität. Es gibt keine Nachweise dafür, dass jedes Item tatsächlich die bezeichnete Kompetenz misst.
4	Faktorielle Validität	Noch nicht geprüft	Es liegen weder explorative noch bestätigende Faktorenanalysen vor. Unbekannt ist, ob sechs unterscheidbare Bereiche entstehen oder Items auf mehreren Bereichen gleichzeitig laden.
5	Reliabilität	Noch nicht geprüft	Ohne Antworten realer Beurteilender lassen sich weder Cronbachs Alpha noch McDonalds Omega oder andere Kennzahlen der Messzuverlässigkeit berechnen.
6	Trennschärfe der Items	Noch nicht geprüft	Es gibt keine Item-Skala-Korrelationen. Daher ist unbekannt, welche Aussagen zwischen hohen und niedrigen Ausprägungen einer Kompetenz unterscheiden.
7	Kriteriumsvalidität	Nicht nachgewiesen	Es liegen keine Daten zu Zielerreichung, Teamleistung, Mitarbeiterbindung, Kundenzufriedenheit oder anderen Erfolgsmerkmalen vor. Der Fragebogen kann deshalb nicht belegen, dass er wirksame Führung erkennt.
8	Objektivität	Teilweise erfüllt	Instruktion, Zeitraum und Antwortskala sind vorgegeben. Regeln zur Auswahl und Gewichtung der Beurteiler, zum Umgang mit fehlenden Werten, zur Auswertung und zur Interpretation fehlen.

Kriterien 9 bis 14

Nr.	Kriterium	Bewertung	Begründung
9	Differenzierungsfähigkeit und Decken- oder Bodeneffekte	Noch nicht geprüft	Ohne Antwortverteilungen ist unbekannt, ob die Items relevante Unterschiede sichtbar machen. Allgemein erwünschtes Verhalten könnte sehr hohe Zustimmungswerte erzeugen.
10	Qualität der Beurteilergruppen	Noch nicht geprüft	Zahl, Auswahl, Beobachtungsdauer und Mindestgröße anonymer Gruppen sind nicht festgelegt. Dieses Kriterium hängt wesentlich von der späteren Durchführung ab.
11	Vergleichbarkeit der Beurteilergruppen	Noch nicht geprüft	Es wurde nicht untersucht, ob Mitarbeitende, Kollegen und Vorgesetzte dieselben Aussagen gleich verstehen und vergleichbare Bewertungsmaßstäbe verwenden.
12	Schutz vor Urteilsverzerrungen	Teilweise erfüllt	Verhaltensnahe Aussagen und die Option ‚kann ich nicht beurteilen‘ helfen. Anonymität, Mindestgruppengrößen und Kontrollen von Milde, Strenge, Halo-Effekt oder Gefälligkeitsurteilen sind nicht geregelt.
13	Fairness und Akzeptanz	Teilweise erfüllt	Die beschriebenen Verhaltensweisen erscheinen grundsätzlich beeinflussbar. Relevanz, kulturelle Verständlichkeit und Akzeptanz wurden jedoch weder mit Betroffenen noch für unterschiedliche Führungsrollen geprüft.
14	Praktischer Entwicklungsnutzen	Teilweise erfüllt	Die Aussagen sind überwiegend handlungsnah und die offenen Fragen fördern konkrete Hinweise. Ob die Ergebnisse akzeptiert werden und zu Verhaltensänderungen führen, ist nicht untersucht.

Ergebnis

Kein Kriterium konnte vollständig als erfüllt nachgewiesen werden. Fünf Kriterien wurden als teilweise erfüllt bewertet, acht blieben ungeprüft und die besonders wichtige Kriteriumsvalidität war nicht nachgewiesen.

Was aus der Prüfung folgt

Besondere Stärken des Entwurfs

- überwiegend konkrete und verhaltensnahe Formulierungen
- einheitlicher Satzanfang und ein klarer Beurteilungszeitraum
- fünfstufige Häufigkeitsskala mit der Option ‚kann ich nicht beurteilen‘
- grundsätzlich brauchbare Ansatzpunkte für ein Entwicklungsgespräch
- zwei offene Fragen, die Stärken und den wichtigsten Veränderungsbedarf konkretisieren

Wesentliche Schwächen und offene Prüfungen

- keine Erprobung mit realen Beurteilenden
- keine nachgewiesene Struktur der sechs Kompetenzbereiche
- keine Kennzahlen zur Zuverlässigkeit und Trennschärfe
- keine Prüfung von Decken- und Bodeneffekten
- keine Prüfung der Vergleichbarkeit der Beurteilergruppen
- kein Nachweis eines Zusammenhangs mit tatsächlichem Führungserfolg

Vorrangiger Verbesserungsbedarf

Zuerst muss geklärt werden, welche Führungsanforderungen im betreffenden Unternehmen und für die jeweilige Rolle erfolgsentscheidend sind. Danach sollte der Entwurf mit Fachleuten und Betroffenen geprüft, in einem Pilotprojekt eingesetzt und anhand realer Antworten überarbeitet werden. Erst mit Daten zu relevanten Erfolgsmerkmalen lässt sich beurteilen, ob hohe Werte tatsächlich wirksame Führung anzeigen.

Einsatzentscheidung

Verwendungszweck	Beurteilung
Ideensammlung und erster Entwurf	Geeignet, sofern alle Aussagen anschließend fachlich geprüft werden.
Pilotprojekt zur Führungskräfteentwicklung	Mit deutlichen Einschränkungen einsetzbar, wenn Zweck, Datenschutz, Anonymität und Auswertung geregelt sind.
Reguläres Entwicklungsinstrument	Erst nach Erprobung, statistischer Prüfung, Überarbeitung und erneuter Prüfung empfehlenswert.
Beförderung, Vergütung oder andere wichtige Personalentscheidung	Nicht ausreichend validiert und in der vorliegenden Form ungeeignet.

Schlussfolgerung

Der Versuch bestätigt die zentrale Grenze KI-generierter Fragebögen: KI kann überzeugend formulieren und einen Entwicklungsprozess vorbereiten. Sie kann ohne geeignete Daten aber nicht beurteilen, ob die erfassten Merkmale zuverlässig gemessen werden und ob sie für erfolgreiche Führung relevant sind.

Ein KI-generierter Fragebogen kann später durchaus alle Gütekriterien erfüllen. Dazu muss er mit realen Beurteilenden erprobt, statistisch untersucht, überarbeitet und möglichst an einer geeigneten Stichprobe erneut geprüft werden. Dann handelt es sich nicht mehr nur um einen KI-generierten Fragebogen, sondern um ein entwickeltes und validiertes Messinstrument, bei dessen Entwicklung KI eingesetzt wurde.

Weiterführende Informationen

Checkliste und Grundlagen zur Fragebogenqualität:

www.360-grad-feedback.net/fragebogen/

Grenzen von KI beim Zugang zu neuem Managementwissen:

www.managerimpulse.de/grenzen-von-ki-und-google/

Autor und fachliche Einordnung

Prof. Dr. Waldemar Pelz ist emeritierter Professor der Technischen Hochschule Mittelhessen und Geschäftsführer des Instituts für Management-Innovation, einer Ausgründung der THM. Zu seinen Arbeitsschwerpunkten gehören Führungskompetenzen, Persönlichkeit, Umsetzungskompetenz, Potenzialdiagnostik und 360-Grad-Feedback. Er hat verschiedene Fragebögen und Testverfahren entwickelt, mit umfangreichen Stichproben geprüft und in Projekten mit Unternehmen eingesetzt.

© 2026 Institut für Management-Innovation