ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Klimm, Johanna; Müller, Kathrin; Rößle, Felix

Working Paper

Einsatzpotenziale von KI zur Lösung von Controlling-Fallstudien

Rosenheim Papers in Applied Economics and Business Sciences, No. 13/2026

Provided in Cooperation with:
Rosenheim Technical University of Applied Sciences

Suggested Citation: Klimm, Johanna; Müller, Kathrin; Rößle, Felix (2026) : Einsatzpotenziale von KI zur Lösung von Controlling-Fallstudien, Rosenheim Papers in Applied Economics and Business Sciences, No. 13/2026, Technische Hochschule Rosenheim, Rosenheim, https://nbn-resolving.de/urn:nbn:de:bvb:861-opus4-31963

This Version is available at:
https://hdl.handle.net/10419/337476

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Technische Hochschule Rosenheim

Einsatzpotenziale von KI zur Lösung von Controlling-Fallstudien

Johanna Klimm
Absolventin
Technische Hochschule Rosenheim

Kathrin Müller
Wissenschaftliche Mitarbeiterin der Fakultät für Betriebswirtschaft
Technische Hochschule Rosenheim

Prof. Dr. Felix Rößle
Professor für allgemeine Betriebswirtschaftslehre, insb. Controlling
Technische Hochschule Rosenheim
Korrespondierender Autor: felix.roessle@th-rosenheim.de

Rosenheim Papers in Applied Economics and Business Sciences

No. 13/2026

Abstract

Diese Arbeit analysiert die Einsatzfähigkeit von Large Language Models (LLMs) wie Google Gemini und Microsoft Copilot bei der Bearbeitung komplexer Controlling-Fallstudien. Praxisnahe Themen wie Kapitalflussrechnung, Selbstkostenberechnung und Eigen- vs. Fremdfertigung werden in Deutsch und Englisch anhand von fünf Kriterien (Korrektheit, Vollständigkeit, Plausibilität, Anwendbarkeit, Verständlichkeit) geprüft und von und ChatGPT ein zweites Mal geprüft. Die Ergebnisse zeigen, dass LLMs im ersten Schritt strukturierte und plausible Lösungen liefern, jedoch ihre Leistung in der Selbsteinschätzung stark überbewerten. Signifikanztests bestätigen, dass die Selbsteinschätzungen deutlich positiver ausfallen als die Bewertungen durch eine neutrale Künstliche Intelligenz (KI) und den Vergleich mit Musterlösungen. Entgegen der Erwartung können die LLMs englischsprachige Aufgaben nicht besser lösen als deutschsprachige. Auch dieser Zusammenhang erweist sich als statistisch signifikant. Die Ergebnisse verdeutlichen, dass menschliche Expertise und Kontrolle für strategische Entscheidungen weiterhin unerlässlich sind, während LLMs derzeit vor allem als Unterstützungstools geeignet sind.

Inhaltsverzeichnis

Inhaltsverzeichnis I
Tabellenverzeichnis I
Abbildungsverzeichnis I
1. Einleitung ..... - 1 -
2. Literaturüberblick ..... - 1 -
3 Methodik ..... - 3 -
3.1 Auswahl der LLMs ..... - 3 -
3.2 Auswahl der Fallstudien und Bewertungskriterien ..... - 4 -
3.3 Vorgehen der Auswertung ..... - 5 -
3.4 Signifikanztests ..... - 6 -
4 Ergebnisse ..... - 8 -
4.1 Selbsteinschätzung durch KI ..... - 8 -
4.2 Vergleich der Ergebnisse zwischen deutschen und englischen Cases ..... - 10 -
4.3 Vergleich von Gemini und CoPilot ..... - 11 -
5. Schluss ..... - 12 -
Literaturverzeichnis IV

Tabellenverzeichnis

Tabelle 1: Auswertung der Tests nach Schritten ..... - 9 -
Tabelle 2: Auswertung Sprachvergleich ..... - 10 -
Tabelle 3: Auswertung LLM-Vergleich ..... - 12 -

Abbildungsverzeichnis

Abbildung 1: Selbsteinschätzung durch KI ..... - 8 -

1. Einleitung

Die fortschreitende Digitalisierung verändert nahezu alle Bereiche der Unternehmensführung. Besonders im Controlling eröffnen digitale Technologien neue Möglichkeiten, Prozesse effizienter zu gestalten und die Qualität von Entscheidungen zu verbessern. Künstliche Intelligenz nimmt dabei eine zentrale Rolle ein, da sie in der Lage ist, schnell große Datenmengen zu analysieren, Muster zu erkennen und Prognosen zu erstellen.

Rammer (2024) zeigt, dass KI derzeit vor allem in den Bereichen Produktion und Logistik sowie Vertrieb und Marketing eingesetzt wird. Auch in der Medizin wird KI bereits eingesetzt, um beispielsweise Bilddaten auszuwerten und so Diagnosen sowie die Behandlungsplanung zu unterstützen (Kepp et al., 2025). Im Gegensatz dazu befindet sich der Einsatz künstlicher Intelligenz im Controlling noch in einem frühen Stadium. Unternehmen stehen vor der Herausforderung, die Potenziale von KI sinnvoll in bestehende Strukturen zu integrieren und gleichzeitig Unsicherheiten hinsichtlich Zuverlässigkeit und organisatorischer Auswirkungen zu berücksichtigen.

Die zentrale Forschungsfrage dieser Arbeit lautet: Inwieweit können aktuelle KI-Modelle zur Lösung von Controlling-Aufgaben eingesetzt werden? Ziel ist es, empirisch zu untersuchen, ob und in welchem Umfang Large Language Models (LLMs) praxisnahe Controlling-Fallstudien in deutscher und englischer Sprache bearbeiten können und ob dabei Unterschiede in der Lösungsqualität erkennbar sind. Neben der Analyse der Korrektheit der Ergebnisse steht insbesondere der Vergleich der Lösungsqualität zwischen den Sprachen Englisch und Deutsch im Vordergrund.

Die weitere Arbeit ist folgendermaßen aufgebaut. Kapitel 2 gibt einen Überblick über die aktuelle Literatur, Kapitel 3 erklärt die Methodik der empirischen Auswertung, Kapitel 4 zeigt und diskutiert die Ergebnisse und Kapitel 5 fasst zusammen und gibt einen Ausblick.

2. Literaturüberblick

Künstliche Intelligenz zählt zu den bedeutendsten Technologien der digitalen Transformation und beeinflusst zunehmend auch das Controlling. Sie ermöglicht in diesem Bereich beispielsweise die automatisierte Verarbeitung großer Datenmengen, präzisere Prognosen und schnellere Entscheidungsprozesse (Friedl, 2019). Eine Zentrale Frage der Forschung ist derzeit, ob KI eigenständig Aufgaben übernehmen kann oder vor allem als unterstützendes Werkzeug dient (Abbas, 2025).

Der Einsatz von künstlicher Intelligenz im Controlling befindet sich derzeit noch in einer frühen Entwicklungsphase. In anderen Bereichen ist das Potenzial von KI bereits deutlich sichtbarer. Das gilt sowohl für die Forschung als auch für den Praxiseinsatz. So wird sie etwa in der Finanzanalyse eingesetzt, um Muster in großen Datenmengen zu erkennen und Cashflows vorherzusagen (Kureljusic & Karger, 2024). Theoretische Ansätze, wie das Fuzzy Logic Modell (Logik, die zwischen wahr und falsch differenzieren kann), verdeutlichen zudem, wie KI dabei helfen kann, interne Kontrollsysteme zu bewerten und Verbesserungspotenziale aufzuzeigen (Ling, 2015).

Die Forschung im Bereich Controlling beschränkt sich aktuell im Wesentlichen auf Literaturarbeiten. Abbas (2025) gibt mit insgesamt 91 Studien einen umfassenden Überblick über den aktuellen Forschungsstand und zeigt, dass KI-Technologien Prozesse und Rollen im Management Accounting grundlegend verändern, jedoch organisatorische Anpassungen und neue Rollenbilder erforderlich machen. Darüber hinaus zeigt der Autor zentrale Themenfelder auf, die als Ausgangspunkt für zukünftige Forschungsarbeiten dienen können. In diesem Zusammenhang zeigen beispielsweise Emetaram und Uchime (2021), dass KI und RPA (Robotic Process Automation) vor allem regelbasierte, wiederkehrende Aufgaben effizienter gestalten, während Controller zunehmend analytische und überwachende Aufgaben übernehmen. Liang (2025) betont die Chancen von Big Data, KI und Echtzeitanalysen für Transparenz und Effizienzsteigerung in der Kostenrechnung. Gleichzeitig weisen Arkhipova et al. (2024) auf neue Spannungsfelder hin, etwa zwischen menschlich getroffenen und algorithmisch erzeugten Entscheidungen. Insgesamt zeigen die bisherigen Forschungsergebnisse, dass KI derzeit vor allem zur Automatisierung standardisierter Prozesse wie Datenaufbereitung, Reporting oder Anomalieerkennung eingesetzt wird (Aleksandrova et al., 2023).

Empirische Arbeiten sind dagegen im Bereich Controlling bislang nahezu nicht vorhanden. In angrenzenden Bereichen, wie etwa dem Accounting, liegt jedoch bereits eine Vielzahl von empirischen Arbeiten vor. So zeigen etwa Street und Wilck (2023), dass ChatGPT einfache Buchhaltungsaufgaben zuverlässig löst, bei komplexen Aufgaben jedoch fehleranfällig ist. Fotoh und Mugwira (2025) kommen zu einem ähnlichen Ergebnis: LLMs können Routineaufgaben im Audit wie das Erstellen von Arbeitspapieren unterstützen, komplexe Tätigkeiten benötigen aber weiterhin menschliche Expertise. Diese Einschätzung wird auch von Hacker (2023) geteilt, der aufzeigt, dass ChatGPT in der Finanzberichterstattung eine Vielzahl von Routineprozessen automatisieren kann, etwa durch Textgenerierung und Datenanalyse. Komplexe Bewertungen und Prüfungen werden jedoch weiterhin menschliche Kontrolle und Interpretation erfordern. Auch Kenney (2023) betont, dass ChatGPT zwar bereits im Finanz- und Rechnungswesen eingesetzt wird, gleichzeitig aber Risiken wie fehlerhafte Ausgaben und ungenaue Ergebnisse beachtet werden müssen. Ergänzend weisen Kim, Muhn und Nikolaev (2024) nach, dass

LLMs plausible Analysen auf Analystenniveau erstellen können, jedoch bei spezifischer Fachregeln – etwa der Anwendung von Rechnungslegungsvorschriften (GAAP/IFRS) und bilanziellen Kennzahlenlogiken – noch deutliche Schwächen aufweisen. Ein Praxisbeispiel aus der Liquiditätsplanung der Meridian Adhesives Group, verdeutlicht das Potenzial von KI indem sie historische Daten automatisch klassifizierte und wöchentliche Cashflow-Prognosen erstellte (Donato & Mannering, 2024). Gleichzeitig wird in der Literatur hervorgehoben, dass bestehende Studien oft technikzentriert sind und organisatorische Aspekte wie Integration, Erklärbarkeit (Explainable AI) und Akzeptanz zu wenig berücksichtigen (Kureljusic & Karger, 2024).

Eine der ersten Studien, die praxisnahe Controllingaufgaben mithilfe von KI behandelt, stammt von Markl et al. (2026). Die Autoren zeigen darin, dass die von der KI erzeugten Ergebnisse eine tendenziell übersteigerte Selbstwahrnehmung aufweisen. Insgesamt ist das Themenfeld „Controlling und KI“ in der wissenschaftlichen Forschung jedoch bislang nur unzureichend repräsentiert. Hier soll diese Arbeit ansetzen.

3 Methodik

Ziel der Untersuchung ist es, den Einsatz von Künstlicher Intelligenz bei der Bearbeitung komplexer Controlling-Fallstudien zu analysieren und zu bewerten. Dabei wird insbesondere untersucht, ob KI-Systeme in der Lage sind, eigenständig valide Lösungen zu erarbeiten und ihre Ergebnisse realistisch einzuschätzen. Für die Analyse werden drei aktuelle und verbreitete Large Language Models und drei anspruchsvolle Controlling-Fallstudien aus einer umfangreichen Fallstudie verwendet, die jeweils in Deutsch und Englisch getestet werden, um mögliche sprachbedingte Leistungsunterschiede zu identifizieren.

3.1 Auswahl der LLMs

Zur Untersuchung der Einsatzfähigkeit künstlicher Intelligenz in der Bearbeitung komplexer Controlling-Fallstudien wurden drei etablierte Large Language Models ausgewählt: Google Gemini, Microsoft Copilot und ChatGPT. Diese Selektion basiert auf Kriterien wie Marktdurchdringung, praktischer Relevanz und evidenzbasierten Erkenntnissen aus der Fachliteratur zur Leistungsfähigkeit in diversen Kontexten (siehe etwa de Carvalho Souza & Weigang, 2025; Semeraro et al., 2024; Salman et al., 2025). Die Auswahl ermöglicht eine nuancierte Evaluierung der Modelle hinsichtlich Problemlösungskompetenz und Hypothesenüberprüfung in Ergänzung zu den Modellen von Markl et al. (2026).

3.2 Auswahl der Fallstudien und Bewertungskriterien

Für die Analyse wird eine umfangreiche Controlling-Fallstudie von Weber et al. (2016) ausgewählt, die realistische und praxisrelevante Aufgabenstellungen enthält. Aus der Fallstudie zur Firma x-presso AG werden drei Aufgaben verwendet, die möglichst typische Controlling-Szenarien abbilden. Die Cases umfassen dabei die Berechnung einer Kapitalflussrechnung, die Ermittlung von Selbstkosten, sowie eine Analyse der Sinnhaftigkeit von Eigen- oder Fremdfertigung.

Die ausgewählten Aufgaben sind so gewählt, dass die verwendeten KIs die relevanten Werte und Beträge nicht vorgegeben bekommen, sondern sie eigenständig aus einem komplexen Dokument herauslesen müssen. Die gesamte Controlling-Fallstudie hat dabei eine Länge von 39 Seiten. Dabei wird die Firma und der Konzern zuerst allgemein Vorgestellt. Dies erhöhte den Schwierigkeitsgrad erheblich und stellt sicher, dass nicht nur Rechenfähigkeiten, sondern auch Kontextbezug und Verständnis von Tabellen (z.B. Bilanzen und Kostenaufstellungen) und Grafiken (z.B. Darstellungen von Produktionsprozessen oder Produktportfolios) der KIs getestet werden. Um die Qualität der Lösungen in verschiedenen Sprachen vergleichen zu können, werden gezielt Aufgaben ausgewählt, die in beiden Sprachen inklusive Musterlösungen (notwendig für Backtesting) vorliegen.

Die Auswertung der Lösungen erfolgt anschließend anhand von fünf Kriterien, die sich an Markl et al. (2026) orientieren:

- Korrektheit: fachlich richtige Lösung der Aufgabe
- Vollständigkeit: Abdeckung aller wesentlichen Inhalte und Teilaufgaben
- Plausibilität: logische Nachvollziehbarkeit der Argumentation und Ergebnisse
- Anwendbarkeit: Eignung der Lösung für den praktischen Einsatz im Unternehmen
- Verständlichkeit: klare und strukturierte Darstellung der Lösung

Alle Kriterien werden auf einer Skala von 1 bis 10 bewertet (für die Bewertungsmethodik siehe 3.3.), wobei 1 für eine unzureichende Erfüllung der Aufgabe und 10 für eine perfekte Erfüllung steht. Die Auswertung kombiniert quantitative Analysen (Mittelwerte, Standardabweichungen) mit qualitativen Elementen (Analyse von Begründungen und Fehlermustern in Stichproben).

3.3 Vorgehen der Auswertung

Um die empirische Fragestellung in Bezug auf den eigenständigen und validen Einsatz von KI bei der Bearbeitung komplexer Controlling-Aufgaben systematisch zu adressieren, werden drei zentrale Hypothesen formuliert, die als Leitfäden für die empirische Überprüfung dienen:

- H1: Künstliche Intelligenzen sind in einer Selbsteinschätzung in der Lage, Controlling-Fallstudien erfolgreich zu lösen, das heißt sie erzielen durchschnittlich signifikant hohe Bewertungen (> 9 von 10 Punkten) in den relevanten Untersuchungskriterien, was auf eine gute fachliche Kompetenz hindeutet.

- H2: Die durchschnittlichen Gesamtbewertungen unterscheiden sich signifikant zwischen den Schritten (siehe nachfolgend in 3.3.), wobei die Bewertungen in Schritt 1 signifikant höher sind als im Vergleich zur Bewertung der neutralen KI und im Backtesting mit einer Musterlösung – was auf eine mangelnde Bestätigung (d. h. Überbewertung in der Selbsteinschätzung) hinweist.

- H3: Die durchschnittliche Gesamtbewertung bei der Lösung in der Sprache Englisch ist höher als für deutsche Aufgaben (über alle Schritte und KIs hinweg).

Diese Hypothesen leiten sich aus der Literatur ab: H1 und H2 basieren auf Studien zur Prognosefähigkeit von LLMs (vgl. z.B. Losbichler und Lehner (2021) und Kokina et al., (2025)) und berücksichtigen die bekannten Grenzen von KI-Systemen in komplexen Anwendungsbereichen. Insbesondere Losbichler und Lehner (2021) betonen, dass KI-Modelle in unsicheren und vielschichtigen Umfeldern an ihre Grenzen stoßen. H3 basiert auf der Annahme, dass moderne LLMs überwiegend auf einer englischsprachigen Trainingsbasis beruhen, die zu einer besseren Leistung in der Lingua franca der KI-Entwicklung führen könnte (Brown et al., 2020).

Das Vorgehen gliedert sich in drei aufeinander aufbauende Schritte, die eine schrittweise Validierung der Lösungen und eine Bewertung der Selbstreflexionsfähigkeit der LLMs ermöglichen. Diese Schritte orientieren sich an Markl et al. (2026). Jeder Schritt wird pro Kombination (Aufgabe, LLM, Sprache) zehnmal wiederholt, um eine verlässliche Datengrundlage zu schaffen, die Variabilität zu minimieren und robuste Ergebnisse zu erzielen.

Schritt 1: Erstellung der Lösung durch KI und Selbstbewertung. In diesem initialen Schritt erhalten Gemini und Copilot die vollständige 39-seitige Fallstudie (inkl. Tabellen und Kontextdaten) als Input, ohne vorgegebene Zahlen oder Hilfestellungen. Die LLMs werden angewiesen, die Aufgabe eigenständig zu lösen, eine strukturierte Lösung zu formulieren und diese anschließend selbst anhand der fünf Kriterien zu bewerten. Dieser Ansatz testet primär die Problemlösungsfähigkeit (H1) und die Fähigkeit zur metakognitiven Einschätzung. Die Selbstbewertung dient als Basis, um potenzielle Überschätzungen zu identifizieren, wie sie in der Literatur beschrieben werden (Li et al., 2024).

Schritt 2: Bewertung der Lösung durch eine neutrale KI. ChatGPT wird als unabhängige Instanz eingesetzt. Es erhält die Originalaufgabe sowie die Lösungen aus Schritt 1 (separat für Gemini und Copilot) und bewertet diese objektiv nach denselben Kriterien. Dieser Schritt kontrastiert die Selbstbewertung mit einer externen Perspektive und prüft, ob Abweichungen systematisch auftreten. Gleichzeitig ermöglicht er einen ersten Test von H3, indem Sprachunterschiede in den Bewertungswerten verglichen werden.

Schritt 3: Bewertung mit Musterlösung als Referenz. ChatGPT bewertet die Lösungen aus Schritt 1 erneut, nun ergänzt um die fachlich validierte Musterlösung von Weber et al. (2016). Dieser Backtesting-Ansatz simuliert einen realen Controlling-Kontext, in dem Referenzdaten verfügbar sind, und erlaubt eine präzise Kalibrierung der Bewertungen. Er testet H1 und H2 durch einen Vergleich mit der Musterlösung. Die Integration der Musterlösung adressiert potenzielle Bias in der neutralen Bewertung und quantifiziert die Abhängigkeit von externer Validierung.

Die Gesamtauswertung erfolgt in zwei Ebenen: Quantitativ wird ein Signifikanztest durchgeführt, um die Hypothesen zu prüfen. Qualitativ werden ausgewählte Lösungen (z.B. bei Extremwerten) thematisch analysiert, um Einflussfaktoren wie Kontextverständnis oder sprachliche Nuancen zu beleuchten. Diese Methodik gewährleistet eine robuste Überprüfung der Hypothese und das Verfahren liefert damit wichtige Informationen für den praktischen Einsatz von KI im Controlling, insbesondere in Hinblick auf eine reine Unterstützungsfunktionen oder tatsächlicher Entscheidungsqualität.

3.4 Signifikanztests

Zur Überprüfung der formulierten Hypothesen (H1, H2 und H3) werden statistische Signifikanztests durchgeführt. Diese sollen prüfen, ob die beobachteten Mittelwertsunterschiede zufälligen Schwankungen oder systematischen Effekten zuzuschreiben sind.

Es wird der Student’s t-Test durchgeführt (Student, 1908).

Einstichproben-t-Test für H1

Hypothese H1 postuliert, dass die durchschnittliche Gesamtbewertung in der Selbsteinschätzung (Schritt 1: „Selbsttest“) über 9 Punkten liegt. Hierzu wird ein einseitiger Einstichproben-t-Test durchgeführt, der den Stichprobenmittelwert gegen einen festen Hypothesenwert ($\mu_0 = 9$) testet.

Die Formel für die T-Statistik lautet:

$$t = \frac{x - \mu_0}{\frac{s}{\sqrt{n}}}, \tag{1}$$

wobei x der Stichprobenmittelwert, μ_0 der Hypothesenwert (hier 9), s die Standardabweichung der Stichprobe und n die Stichprobengröße ist. Die Freiheitsgrade (df) ergeben sich aus

$$df = n - 1. \tag{2}$$

Paarweise Zweistichproben-t-Tests für H2 und H3

Für H2, die eine mangelnde Bestätigung der Lösungen im ersten Schritt durch eine dritte KI und im Vergleich zu einer Musterlösung annimmt und für H3, die Unterschiede in der Sprache erwartet, werden einseitige Zweistichproben-t-Tests für unabhängige Stichproben durchgeführt (Field, 2024). Diese prüfen, ob die Mittelwerte in Schritt 1 signifikant höher sind als in Schritt 2 („Test durch Dritte KI“) bzw. Schritt 3 („Backtesting mit Musterlösung“).

Die Formel für die t-Statistik lautet:

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{3}$$

wobei x_1 und x_2 die Mittelwerte der untersuchten Gruppen, $s\frac{2}{1}$ und $s\frac{2}{2}$ die Varianzen und n_1 sowie n_2 die Stichprobengrößen sind.

Der p-Wert ergibt sich für alle Tests (H1, H2 und H3) aus der t-Verteilung. Alle Tests werden in Excel durchgeführt und die Analysen basieren auf einem 5% Signifikanzniveau.

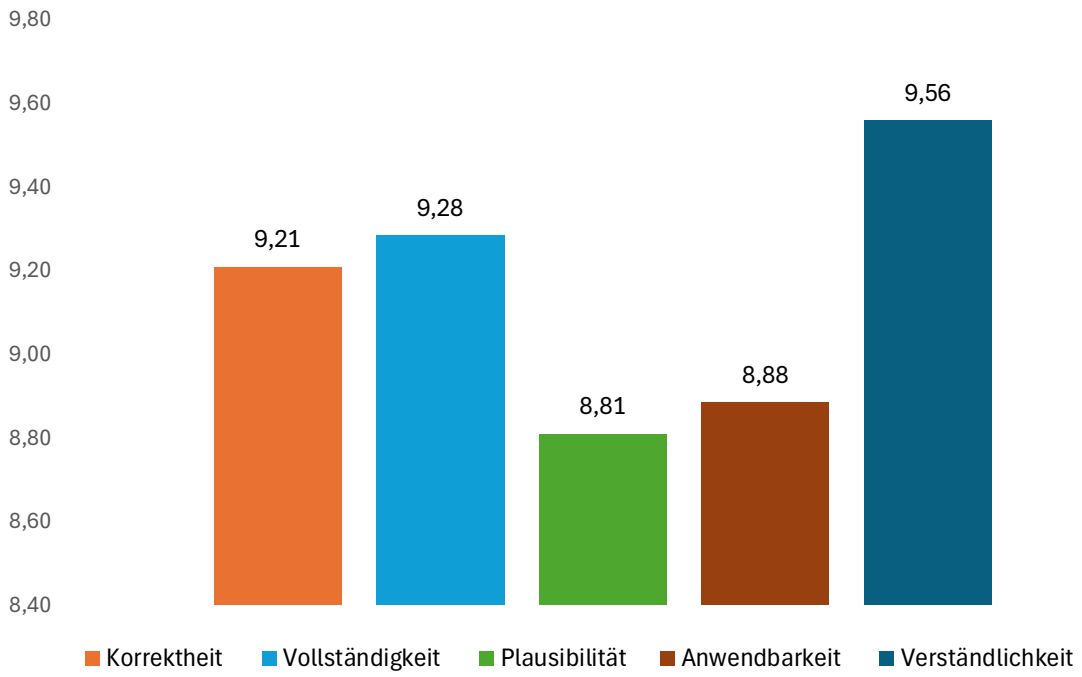
4 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse der Fallstudien untersucht und die Hypothesen getestet.

4.1 Selbsteinschätzung durch KI

Wie bereits im Literaturüberblick erwähnt, zeigt die aktuelle Forschung, dass der Einsatz von KI im Controlling trotz großer Fortschritte noch wesentlichen Einschränkungen unterliegt. Losbichler und Lehner (2021) betonen beispielsweise, dass KI-Systeme besonders in komplexen und unsicheren Umfeldern an ihre Grenzen stoßen. Des Weiteren lassen sich Entscheidungsprozesse nicht vollständig abbilden, da nicht alle relevanten Einflussfaktoren sichtbar oder kontrollierbar sind.

Abbildung 1: *Selbsteinschätzung durch KI*



Diese Abbildung zeigt Schritt 1 mit den 5 Bewertungskriterien Korrektheit, Vollständigkeit, Plausibilität, Anwendbarkeit und Verständlichkeit. Dabei steht 0 für eine sehr schlechte Lösung der Aufgabe durch die KI und 10 für eine perfekte Lösung. Die Werte geben den Mittelwert der jeweils 10 Wiederholungen an. Der Wert 9,21 bei Korrektheit bedeutet beispielsweise, dass die KI ihre eigene Lösung auf einer Skala von 0 bis 10 sehr gut einschätzt.
Quelle: eigene Darstellung

Abbildung 1 zeigt, dass im Gegensatz zu den bisherigen Ergebnissen der Forschung, die Tests zunächst ein sehr positives Bild mit Mittelwerten um die 9 ergeben. In Bezug auf Hypothese 1 zeigt sich, dass die Werte oberhalb von 9 liegen und die Selbsteinschätzung

der KI sehr gut ist. Für H1 ergibt der t-Test einen Mittelwert von 9,15 in Schritt 1 ($t = 2{,}32$; $p = 0{,}99$), was die Nullhypothese nicht ablehnt – die Selbsteinschätzung ist nicht signifikant unter 9, sondern tendenziell hoch. Die KI kann die gestellten Controlling-Aufgaben in Selbsteinschätzung korrekt, nachvollziehbar und strukturiert lösen. Die Ergebnisse aus Schritt 1 bestätigen zudem die Ergebnisse von Markl et al. (2026).

Allerdings beruhen diese Ergebnisse ausschließlich auf der Selbsteinschätzung der KI. Da sie ihre eigene Lösung bewertet, ohne zu wissen, ob diese tatsächlich korrekt ist, sind die guten Bewertungen eher als subjektive Einschätzung der KI zu sehen und müssen entsprechend kritisch hinterfragt werden.

Im darauffolgenden Schritt werden die gleichen Aufgaben und Lösungen von einer neutralen KI bewertet. Hier zeigen sich bereits deutlich niedrigere Bewertungen zum Beispiel bei der Plausibilität der Lösung, die von einem Durchschnitt von 8,81 auf einen Wert von 6,38 sinkt, was darauf hindeutet, dass die ursprüngliche Selbsteinschätzung der ersten KI zu optimistisch ausfällt. In Schritt 3, nachdem ChatGPT zusätzlich die Musterlösung zur Verfügung gestellt wird, verschlechtern sich die Bewertungen teilweise erheblich (z.B. Korrektheit von 9,21 auf 4,11). Das bestätigt, dass die ursprünglichen Lösungen der ersten KI in mehreren Bereichen offensichtlich nicht korrekt sind, obwohl die Selbsteinschätzung zunächst sehr gut ausfällt. Die Ergebnisse von den Schritten 2 und 3 werden in Tabelle 1 dargestellt

Tabelle 1: *Auswertung der Tests nach Schritten*

Schritt	Korrektheit	Vollständigkeit	Plausibilität	Anwendbarkeit	Verständlichkeit
Selbsttest	9,21	9,28	8,81	8,88	9,56
Test durch neutrale KI	6,10	6,31	6,38	6,66	8,19
Backtesting mit Musterlösung	4,11	5,12	5,11	4,96	6,48

Diese Tabelle zeigt die Entwicklung der Bewertungsergebnisse der 3 Schritte mit den 5 Bewertungskriterien Korrektheit, Vollständigkeit, Plausibilität, Anwendbarkeit und Verständlichkeit. Dabei steht 0 für eine sehr schlechte Lösung der Aufgabe durch die KI und 10 für eine perfekte Lösung. Die Werte geben den Mittelwert der jeweils 10 Wiederholungen an.
Quelle: eigene Darstellung

Im Folgenden wird Hypothese 2 untersucht, die annimmt, dass die Lösungen der KI durch eine dritte KI (Schritt 2) und im Backtesting mit Musterlösung (Schritt 3) nicht bestätigt werden, was zu signifikant niedrigeren Bewertungen führt. Die paarweisen einseitigen Zweistichproben-t-Tests bestätigen dies: Für den Vergleich von Schritt 1 vs. Schritt 2 ergibt sich ein p-value von $3{,}16 * 10^{-29}$, was die Nullhypothese ($\mu_Schritt1 \leq \mu_Schritt2$) ablehnt. Ein ähnliches Ergebnis zeigt sich für das Backtesting im Schritt 1 vs. Schritt 3 mit einem p-value von $4{,}66 * 10^{-58}$. Diese Ergebnisse deuten auf eine signifikante Abnahme der Bewertungen hin, was eine mangelnde Bestätigung der initialen Lösungen impliziert. Dieses Ergebnis bestätigt die Argumentation von Losbichler und Lehner (2021). Die Ergebnisse legen nahe, dass die KI in den getesteten Szenarien nicht in der Lage ist, die komplexen Controlling-Zusammenhänge vollständig und fehlerfrei zu erfassen.

4.2 Vergleich der Ergebnisse zwischen deutschen und englischen Cases

Tabelle 2: *Auswertung Sprachvergleich*

Sprache	Schritt	Korrektheit	Vollständigkeit	Plausibilität	Anwendbarkeit	Verständlichkeit
DE	Selbsttest	9,12	9,27	8,85	8,92	9,55
	Test durch neutrale KI	6,70	6,85	6,65	7,13	8,28
	Backtesting mit Musterlösung	4,98	5,75	5,72	5,73	6,73
ENG	Selbsttest	9,30	9,30	8,77	8,85	9,57
	Test durch neutrale KI	5,50	5,77	6,12	6,18	8,10
	Backtesting mit Musterlösung	3,23	4,48	4,50	4,18	6,22

Diese Tabelle zeigt die Entwicklung der Bewertungsergebnisse der 3 Schritte mit den 5 Bewertungskriterien Korrektheit, Vollständigkeit, Plausibilität, Anwendbarkeit und Verständlichkeit aufgeteilt nach Sprache des Cases. Dabei steht 0 für eine sehr schlechte Lösung der Aufgabe durch die KI und 10 für eine perfekte Lösung. Die Werte geben den Mittelwert der jeweils 10 Wiederholungen an.
Quelle: eigene Darstellung

Im Folgenden wird Hypothese 3 untersucht, die annimmt, dass KIs Controlling-Aufgaben in englischer Sprache besser lösen als in deutscher. Die Auswertung (siehe Tabelle 2)

zeigt jedoch auf den ersten Blick kein eindeutiges Ergebnis. In einigen Fällen schneiden die Lösungen der deutschen Aufgaben besser ab, in anderen wiederum die englischen. Im Gesamtbild schneiden die deutschen Lösungen etwas besser ab.

Der einseitige Zweistichproben-t-Test zur Prüfung von H3 ergibt, dass die Hypothese nicht bestätigt werden kann. Da der Mittelwert für Englisch sogar niedriger ist, wird die Nullhypothese ($\mu_{Englisch} \leq \mu_{Deutsch}$) nicht abgelehnt. Es gibt einen signifikanten Unterschied, aber in der entgegengesetzten Richtung. Eine mögliche Erklärung für diese uneinheitlichen Ergebnisse ist, dass die Fallstudien aus einem deutschsprachigen Lehrbuch stammen und von den Autoren ins Englische übersetzt wurden. Dabei gehen möglicherweise inhaltliche Details verloren, was die Verständlichkeit und Genauigkeit der Aufgaben in der englischen Version beeinflusst. Darüber hinaus hängt die Fähigkeit von KI-Modellen, Aufgaben in verschiedenen Sprachen korrekt zu lösen, stark von ihrem Training ab. Wie Huang et al. (2025) beschreiben, werden Large Language Models anhand riesiger Textmengen trainiert, um Sprachmuster zu erkennen und auf dieser Basis Vorhersagen für den nächsten passenden Ausdruck oder Satz zu treffen. Beim Training lernen die Modelle statistische Zusammenhänge und Bedeutungen zwischen Wörtern und Sätzen, ohne jedoch Sprache im menschlichen Sinne zu verstehen. Sie können dadurch zwar Texte in verschiedenen Sprachen verarbeiten und generieren, sind jedoch auf die Qualität und Vielfalt der Trainingsdaten angewiesen. Wenn ein Modell beispielsweise überwiegend mit englischen Texten trainiert wurde, ist seine Leistung in Englisch in der Regel höher als in anderen Sprachen.

Somit bleibt offen, ob ein klarer sprachlicher Vorteil besteht, oder ob die Unterschiede eher auf die Aufgabenformulierung und Übersetzungsqualität zurückzuführen sind.

4.3 Vergleich von Gemini und CoPilot

Zuletzt werden die beiden KI-Modelle Google Gemini und Microsoft Copilot miteinander verglichen. Die Auswertung zeigt, dass Copilot insgesamt etwas bessere Ergebnisse erzielt, insbesondere in den beiden zentralen Kategorien Korrektheit und Anwendbarkeit, die für den praktischen Einsatz besonders hohe Bedeutung haben.

Wie auch in der Literatur beschrieben wird, ist Copilot stärker auf strukturierte, praxisnahe Anwendungen ausgerichtet, während Gemini durch seine breitere, generalistische Ausrichtung zwar vielseitiger einsetzbar ist, bei komplexen analytischen Problemstellungen jedoch schneller an Grenzen stößt (de Carvalho Souza & Weigang, 2025; Salman et al., 2025). Die Ergebnisse dieser Untersuchung bestätigen diese Einschätzung. Wie Tabelle 3 zeigt, erzielt Copilot in den Aufgaben insgesamt höhere Werte, insbesondere in

den Kategorien Korrektheit und Anwendbarkeit. Die Hypothese 2 wird zudem bei beiden LLMs erneut bestätigt. Besonders in der Kategorie Anwendbarkeit zeigt sich die praktische Relevanz dieser Unterschiede: Während Copilot hier einen Wert von 7,28 erreicht, liegt Gemini mit 6,03 deutlich darunter – ein Unterschied, der für den produktiven Einsatz im Controlling von erheblicher Bedeutung ist. Dies deutet darauf hin, dass Copilot nicht nur theoretisch korrekte Lösungen liefert, sondern diese auch in einer Form präsentiert, die unmittelbar in den Arbeitsalltag von Controllern integrierbar ist.

Tabelle 3: *Auswertung LLM-Vergleich*

LLM Model	Schritt	Korrektheit	Vollständigkeit	Plausibilität	Anwendbarkeit	Verständlichkeit
CoPilot	Selbsttest	9,17	9,13	8,73	8,95	9,23
	Test durch neutrale KI	6,77	6,83	6,95	7,28	8,38
	Backtesting mit Musterlösung	4,40	5,33	5,52	5,23	6,48
Gemini	Selbsttest	9,25	9,43	8,88	8,82	9,88
	Test durch neutrale KI	5,43	5,78	5,82	6,03	8,00
	Backtesting mit Musterlösung	3,82	4,90	4,70	4,68	6,47

Tabelle 3 zeigt den direkten Vergleich der beiden verwendeten KI-Modelle, Microsoft Copilot und Google Gemini. Sie stellt deren Leistungsfähigkeit in den drei Untersuchungsschritten gegenüber und verdeutlicht, in welchen Bereichen die Modelle besonders stark oder schwach abgeschnitten haben.
Quelle: eigene Darstellung

5. Schluss

In dieser Arbeit wird untersucht, inwieweit Large Language Models (LLMs) komplexe Controlling-Aufgaben eigenständig lösen können, mit einem zusätzlichen Fokus auf die sprachliche Leistungsfähigkeit in Deutsch und Englisch. Dazu wurden drei umfangreiche praxisnahe Aufgaben einer komplexen Fallstudie bearbeitet, die typische Unternehmensszenarien abbilden. Im Rahmen der Untersuchung wurden eine Kapitalflussrechnung, eine Selbstkostenberechnung sowie eine Analyse zur Eigen- oder Fremdfertigung verwendet. Die Aufgaben wurden von den LLMs in beiden Sprachversionen gelöst, um die sprachabhängige Qualität und Konsistenz der Ergebnisse zu vergleichen.

Die Ergebnisse zeigen, dass LLMs bereits logisch strukturierte Lösungen erzeugen können. Allerdings überschätzen sich die Modelle in der Selbsteinschätzung deutlich. Diese Abweichungen werden statistisch bestätigt durch t-Tests auf einem Signifikanzniveau von 5 %. Die Selbsteinschätzungen liegen signifikant über den Bewertungen der neutralen KI ChatGPT sowie dem Backtesting mit Musterlösungen. Gleichzeitig kann die Hypothese, dass LLMs Controlling-Fälle in englischer Sprache besser lösen als in deutscher, nicht bestätigt werden. Tendenziell zeigen sich sogar bessere Ergebnisse für die deutschsprachigen Aufgaben, was überraschend ist, da die Erwartungshaltung ein besseres Training der KI-Modelle in Englisch erwartet hat. Im direkten Vergleich erweist sich Copilot als das zuverlässigere Modell mit höheren Werten bei Korrektheit und Anwendbarkeit, während Gemini häufiger inhaltliche Lücken aufweist.

Zukünftige Studien könnten vor allem als weiteren Schritt eine menschliche Prüfung der Ergebnisse ergänzen, um Einschränkungen von künstlichen Intelligenzen wie Halluzinationen (KI-Systeme, insbesondere LLMs, generieren selbstbewusst faktisch falsche oder erfundene Informationen, die plausibel klingen, aber keine Grundlage in der Realität haben), begrenzte Reasoning-Fähigkeiten oder Überschätzung der eigenen Fähigkeiten besser auszuschließen.

Für die Praxis bedeutet dies, dass KI-Systeme aktuell vor allem als Unterstützungstools sinnvoll eingesetzt werden können, beispielsweise zur Automatisierung von Standardberechnungen oder zur Aufbereitung von Daten. Bei komplexen, strategischen Entscheidungen ist jedoch weiterhin die fachliche Kontrolle durch menschliche Expertise nötig. Ein hybrider Ansatz, bei dem KI-Vorschläge liefert und der Mensch die finale Entscheidung trifft, erscheint aktuell am sinnvollsten (Kureljusic & Karger, 2024).

Abschließend lässt sich festhalten, dass KI im Controlling ein hohes Zukunftspotenzial besitzt. Derzeit kann sie die Arbeit von Controllern unterstützen und beschleunigen, sie jedoch noch nicht ersetzen. Unternehmen sollten daher schrittweise vorgehen, klare Einsatzbereiche definieren und gleichzeitig die notwendigen Kontrollmechanismen implementieren, um die Potenziale der Technologie optimal zu nutzen und mögliche Risiken zu begrenzen. Die empirischen Ergebnisse dieser Arbeit verdeutlichen zudem, dass die KI-generierten Ergebnisse stets kritisch geprüft werden müssen, bevor sie in Entscheidungsprozesse einfließen.

Literaturverzeichnis

Abbas, K. (2025). Management accounting and artificial intelligence: A comprehensive literature review and recommendations for future research. *The British Accounting Review*. https://doi.org/10.1016/j.bar.2025.101551

Aleksandrova, A., Ninova, V., & Zhelev, Z. (2023). A survey on AI implementation in finance, (cyber) insurance and financial controlling. *Risks, 11*(5), 91. https://doi.org/10.3390/risks11050091

Arkhipova, D., Montemari, M., Mio, C., & Marasca, S. (2024). Digital technologies and the evolution of the management accounting profession: A grounded theory literature review. *Meditari Accountancy Research, 32*(7). https://doi.org/10.1108/MEDAR-07-2023-2097

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*. https://doi.org/10.48550/arXiv.2005.14165

de Carvalho Souza, M., & Weigang, L. (2025). Grok, Gemini, ChatGPT and DeepSeek: Comparison and applications in conversational artificial intelligence. *Inteligencia Artificial, 2*(1). https://doi.org/10.5281/zenodo.14885243

Donato, N., & Mannering, A. (2024). AI-enabled forecasting unifies cash flow visibility for multinational manufacturer. *Armanino*. https://www.armanino.com/-/media/pdf/case-studies/ai-forecasting-unifies-cash-flow-visibility-for-manufacturer.pdf

Emetaram, E., & Uchime, H. (2021). Impact of artificial intelligence (AI) on accountancy profession. *Journal of Accounting and Financial Management, 7*(2), 15–25.

Field, A. (2024). *Discovering statistics using IBM SPSS statistics* (6th ed.). SAGE.

Fotoh, L., & Mugwira, T. (2025). Exploring large language models in external audits: Implications and ethical considerations. *International Journal of Accounting Information Systems, 56*, 100748. https://doi.org/10.1016/j.accinf.2025.100748

Friedl, G. (2019). Künstliche Intelligenz im Controlling. *Controlling, 31*(5). https://doi.org/10.15358/0935-0381-2019-5-35

Hacker, B. (2023). Wird ChatGPT das Rechnungswesen revolutionieren? Der Nutzen von Künstlicher Intelligenz (KI) im Rechnungswesen. *IRZ – Zeitschrift für Internationale Rechnungslegung, 18*(3), 117–121.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X. & Qin, B. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43(2): https://dl.acm.org/doi/10.1145/3703155

Kenney, A. (2023). The promise and peril of ChatGPT. *Journal of Accountancy*. https://www.journalofaccountancy.com/issues/2023/may/the-promise-and-peril-of-chatgpt.html

Kepp, T., Uzunova, H., Ehrhardt, J., & Handels, H. (2025). KI-basierte Anwendungen in der medizinischen Bildverarbeitung. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz*. https://doi.org/10.1007/s00103-025-04093-7

Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial statement analysis with large language models. *arXiv*. https://doi.org/10.48550/arXiv.2407.17866

Kokina, J., Blanchette, S., Davenport, T., & Pachamanova, D. (2025). Challenges and opportunities for artificial intelligence in auditing: Evidence from the field. *International Journal of Accounting Information Systems, 56*, 100734. https://doi.org/10.1016/j.accinf.2025.100734

Kureljusic, M., & Karger, E. (2024). Forecasting in financial accounting with artificial intelligence: A systematic literature review and future research agenda. *Journal of Applied Accounting Research, 25*(1), 81–104. https://doi.org/10.1108/JAAR-06-2022-0146

Li, J., Yang, Y., Zhang, R., & Lee, Y. (2024). Overconfident and unconfident AI hinder human–AI collaboration. *arXiv*. https://doi.org/10.48550/arXiv.2402.07632

Liang, H. (2025). Modern technology's role in accounting cost calculation of industrial enterprises: Informatization as a key strategy to improve management efficiency. *Heliyon, 11*(1), e41523. https://doi.org/10.1016/j.heliyon.2024.e41523

Ling, L. (2015). Research on enterprise internal control financial assessment system based on artificial intelligence. *Revista Ibérica de Sistemas e Tecnologias de Informação, 16*(B).

Losbichler, H., & Lehner, O. M. (2021). Limits of artificial intelligence in controlling and the ways forward: A call for future accounting research. *Journal of Applied Accounting Research, 22*(2). https://doi.org/10.1108/JAAR-10-2020-0207

Markl, K., Polz, M., Diener, M., & Rößle, F. (2026). Empowering financial controlling with artificial intelligence: An empirical evaluation. *Proceedings of the 17th International Conference on Agents and Artificial Intelligence* [Fortcoming].

Rammer, C., Doherr, T., Kinne, J., & Lenz, D. (2024). *KI-Einsatz in Unternehmen in Deutschland: Strategische Ausrichtung und internationale Position*. ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung. https://www.zew.de/publikationen/ki-einsatz-in-unternehmen-in-deutschland

Salman, I., Ameer, O., Khanfar, M., & Hsieh, Y.-H. (2025). Artificial intelligence in healthcare education: Evaluating the accuracy of ChatGPT, Copilot and Google Gemini in cardiovascular pharmacology. *Frontiers in Medicine, 12*, 1495378. https://doi.org/10.3389/fmed.2025.1495378

Semeraro, F., Gamberini, L., Carmona, F., & Monsieurs, K. (2024). Clinical questions on advanced life support answered by artificial intelligence: A comparison between ChatGPT, Google Bard and Microsoft Copilot. *Resuscitation, 195*, 110114. https://doi.org/10.1016/j.resuscitation.2024.110114

Street, D. L., & Wilck, J. (2023). 'Let's have a chat': Principles for the effective application of ChatGPT and large language models in the practice of forensic accounting. *Journal of Forensic and Investigative Accounting, 15*(3), July–December. https://doi.org/10.2139/ssrn.4351817

Student. (1908). The probable error of a mean. *Biometrika, 6*(1), 1–25. https://doi.org/10.1093/biomet/6.1.1

Weber, J., Schäffer, U., & Binder, C. (2016). *Einführung in das Controlling: Übungen und Fallstudien mit Lösungen* (3. Aufl.). Schäffer-Poeschel.