Large Language Models auf dem Prüfstand

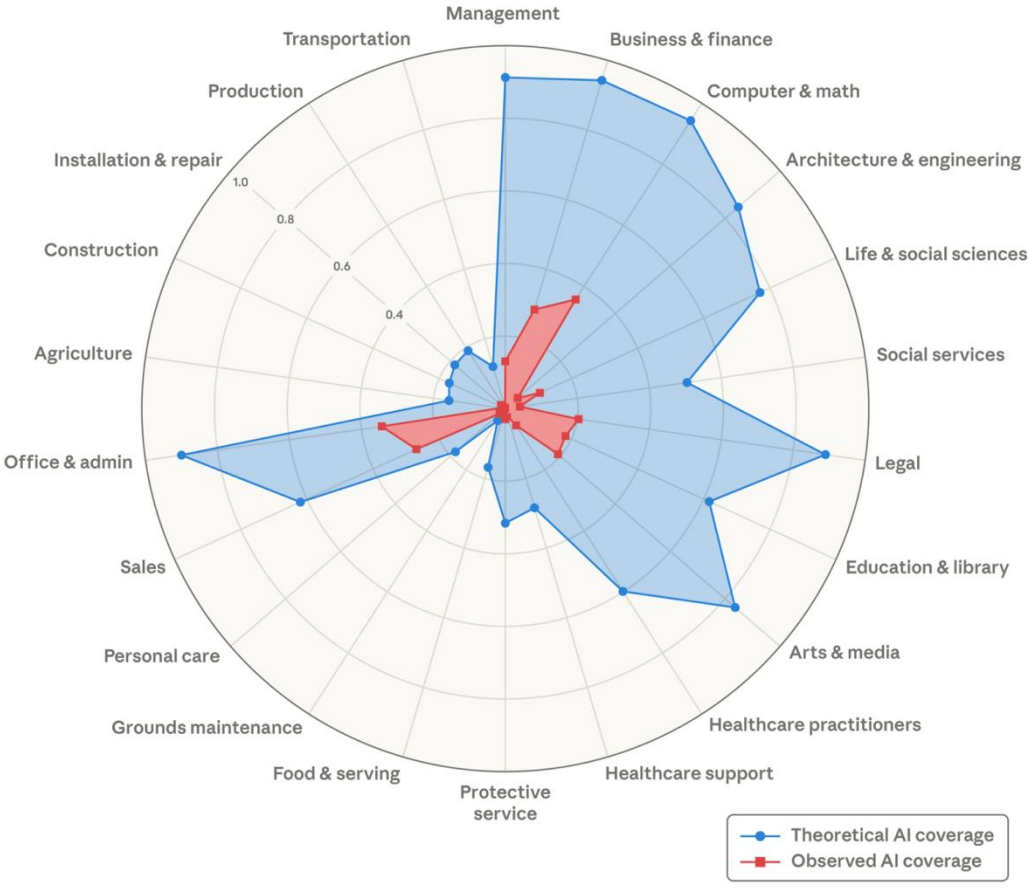
Empirische Befunde zur KI-Leistung in Bilanzierung und Steuerrecht

9. Symposium Steuern & Bilanzen

Manuel Kaburek
WU Wien · Abteilung für Unternehmensrechnung und Revision

AUSGANGSLAGE

Potenzial und tatsächliche Nutzung



Blau: Was ginge

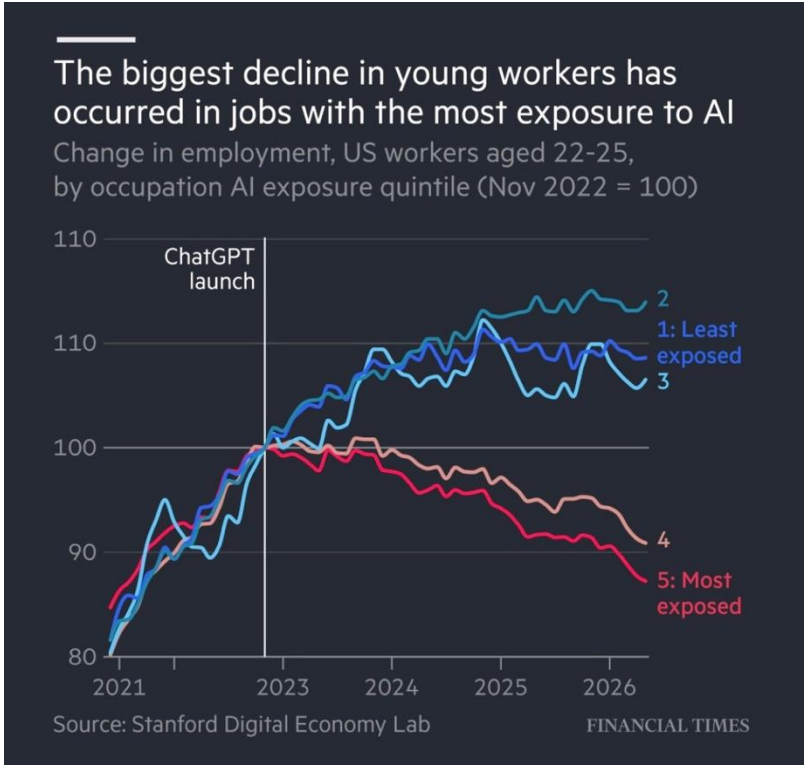
Der Anteil der Tätigkeiten eines Berufs, den ein Sprachmodell grundsätzlich übernehmen könnte. Computer & Math, Business & Finance, Office & Admin und Legal liegen ganz außen.

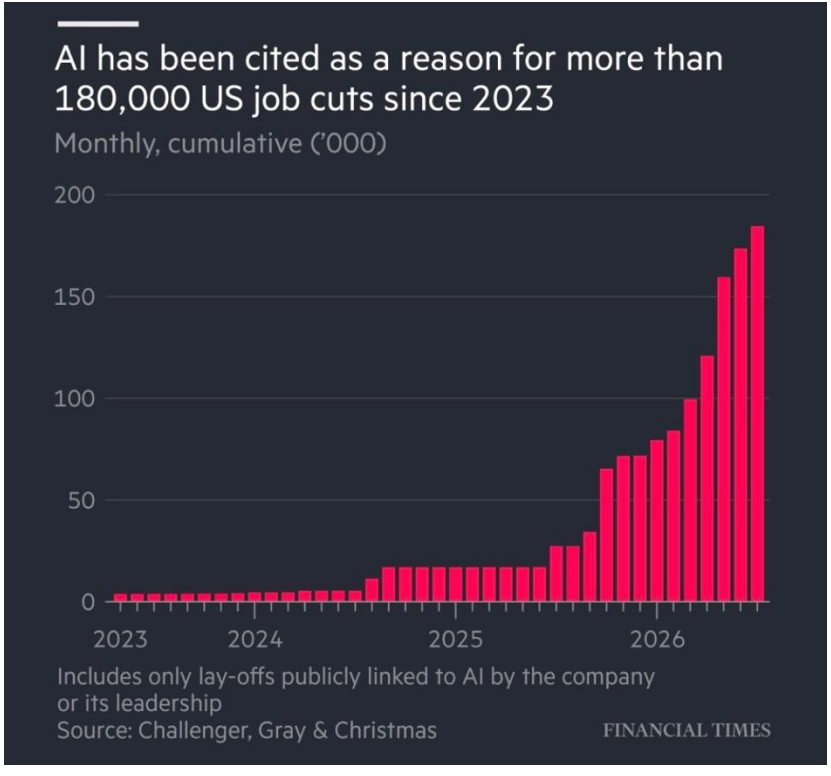
Rot: Was tatsächlich genutzt wird

Die rote Fläche bleibt weit hinter der blauen zurück. Bei Computer & Math sind es 33 % beobachtete Abdeckung.

Abbildung: Massenkoff, M. / McCrory, P. (2026): Labor market impacts of AI. Anthropic, März 2026.

Der Arbeitsmarkt reagiert bereits auf KI





Grafiken: Financial Times. Daten: Stanford Digital Economy Lab (links) und Challenger, Gray & Christmas (rechts).

KI in der Finanzverwaltung

13. AUGUST 2025

OTS

13.08.2025, 09:17:05 / OTS0031

Finanzministerium lukrierte im Vorjahr 354 Mio. Euro Steuermehreinnahmen dank KI-Methoden

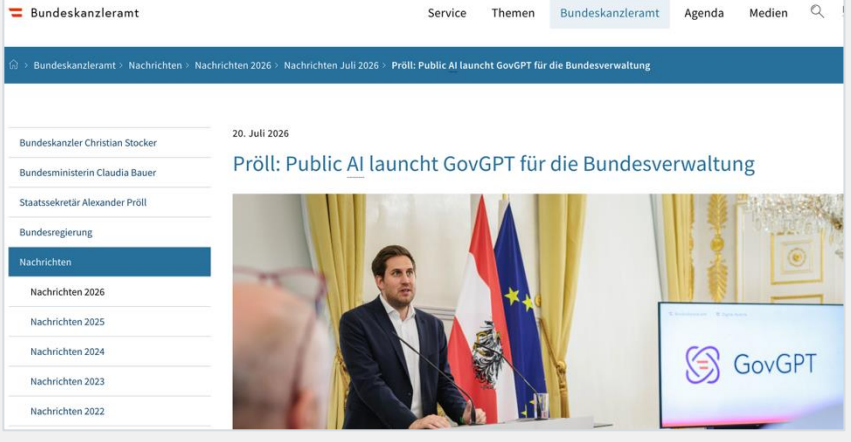
Jahresbilanz 2024 des Predictive Analytics Competence Center im BMF liegt vor

Wien (OTS) - Das Predictive Analytics Competence Center (PACC), eine Spezialeinheit im Finanzministerium, setzt auf künstliche Intelligenz, um Steuerbetrug und Compliance-Verstöße aufzudecken. Im Jahr 2024 wurden durch die Risikomodelle des PACC in Summe rund 6,6 Mio. Fälle aus den unterschiedlichsten Bereichen der Steuerverwaltung überprüft. Dadurch wurden durch die prüfenden Organe der Finanzverwaltung rund 354 Millionen Euro an Steuermehreinnahmen erzielt.

2024 wurden rund 6,6 Mio. Fälle über Risikomodelle ausgewertet — künftig auch mit Sprachmodellen.

Quelle: Bundesministerium für Finanzen, Presseaussendung vom 13. August 2025. https://www.bmf.gv.at/presse/pressemeldungen/2025/august/pacc-ki.html

20. JULI 2026



Bundeskanzleramt

Service Themen Bundeskanzleramt Agenda Medien

Bundeskanzleramt › Nachrichten › Nachrichten 2026 › Nachrichten Juli 2026 › Pröll: Public AI launcht GovGPT für die Bundesverwaltung

Bundeskanzler Christian Stocker
Bundesministerin Claudia Bauer
Staatssekretär Alexander Pröll
Bundesregierung
Nachrichten
Nachrichten 2026
Nachrichten 2025
Nachrichten 2024
Nachrichten 2023
Nachrichten 2022

20. Juli 2026

Pröll: Public AI launcht GovGPT für die Bundesverwaltung

GovGPT

Das BMF startete am 21. Juli als erstes Ressort nach dem Bundeskanzleramt mit dem Rollout.

Quelle: Bundeskanzleramt, Nachricht vom 20. Juli 2026. https://www.bundeskanzleramt.gv.at/bundeskanzleramt/nachrichten-der-bundesregierung/2026/07/proell-public-ai-launcht-govgpt-fuer-die-bundesverwaltung.html

Agenda

01 **KI im Rechnungswesen**
Adoption, konkrete Anwendungsfelder, was sich in der Arbeit verschiebt

02 **Die Forschungsfrage**
Beherrscht ein Sprachmodell österreichisches Bilanz- und Steuerrecht?

03 **AccountingBench**
564 Prüfungsaufgaben, 20 Modelle

04 **Was die Modelle können, und was nicht**
Leistung, Jurisdiktionsgefälle, Selbsteinschätzung, Kosten

05 **Konsequenzen für die Praxis**
Einsatzfelder, Verantwortung

WO DIE PRAXIS STEHT

KI ist im Rechnungswesen bereits im Einsatz

40 %

ORGANISATIONEN NUTZEN GENAI

Ein Jahr davor waren es 22 %. Zugelegt hat vor allem die Nutzung in der laufenden Arbeit.

77 %

ERWARTEN AGENTISCHE KI BIS 2030

Im Einsatz sind solche Systeme bisher bei 15 %. Dort stand generative KI vor zwei Jahren auch.

65 %

DER BEFRAGTEN IN STEUER UND RECHNUNGSWESEN

akzeptieren KI-gestützte fachliche Empfehlungen. In der Rechtsberatung sind es 17 %.

Thomson Reuters Institute: 2026 AI in Professional Services Report, n = 1.514, Feldzeit Oktober–November 2025.
Thomson Reuters Institute, Future of Professionals 2026, n = 1.800, Feldzeit März–April 2026.

Im Steuerbereich steht der Rollout erst bevor

WO KI IM FINANZBEREICH EINGESETZT WIRD (2024)

36 %	**Accounting**
33 %	Financial Planning
25 %	Treasury
18 %	Risiko
18 %	**Steuern**

AKTIVE KI-NUTZUNG IM FINANZBEREICH INSGESAMT

30 % → **75 %**

2024 2026

Die Nachfolgestudie erhebt keine Aufschlüsselung mehr nach Funktionsbereichen.

Links: KPMG Global AI in Finance Report, Februar 2025, n = 2.900, Feldzeit September 2024, Figure 4 (breite oder selektive Nutzung). Rechts: KPMG Global AI in Finance 2026, n = 1.013, Feldzeit März 2026.

Was sich in der Arbeit ändert

	HEUTE	KÜNFTIG
UMFANG	Geprüft wird eine Stichprobe.	Ausgewertet wird alles, die Rechenzeit kostet fast nichts.
ZEITPUNKT	Nachschau nach Ende der Periode.	Laufend, während die Buchungen entstehen.
AUFMERKSAMKEIT	Der Berater bearbeitet sämtliche Fälle.	Das System übernimmt Normalfälle, der Berater fokussiert sich auf Ausnahmen.
WISSEN	Nachschlagen in Normen und Kommentaren dauert.	Wissen ist in Sekunden abrufbar.

Einordnung der KI-Verwendung

Diese Aufgaben können effizienter erledigt werden:

- Vertragsprüfung
- Recherche in Standards und Literatur
- Abschlussunterlagen und Arbeitspapiere
- Entwürfe für Anhangangaben

Das bleibt aber zu prüfen:

- Auf welche Daten sich die Antwort stützt.
- Ob der Standard richtig angewendet wurde.
- Ein Fehler fällt nur auf, wenn ihn jemand sucht.
- Die Haftung und Verantwortung übernimmt am Ende ein Mensch.

DIE FORSCHUNGSFRAGE

Beherrscht ein Sprachmodell österreichisches Bilanz- und Steuerrecht?

Die verfügbaren Belege zur Leistungsfähigkeit von Sprachmodellen stammen fast durchwegs aus dem angloamerikanischen Raum oder beziehen sich auf international einheitliche Standards.

Für nationales Recht — UGB, BAO, EStG, UStG — gab es bisher keine belastbare Messung.

UNSER BENCHMARK

AccountingBench

564

GEPRÜFTE AUFGABEN

Aus Universitäts- und Berufsprüfungen, gegen Expertenlösungen bewertet. Überwiegend Freitextaufgaben: Das Modell muss die Lösung herleiten, nicht nur ankreuzen.

20

SPRACHMODELLE

Die führenden öffentlich verfügbaren Modelle von OpenAI, Anthropic, Mistral, xAI und weiteren Anbietern wurden auf identische Aufgaben getestet.

3

KATEGORIEN

Steuerrecht, externes Rechnungswesen und Kostenrechnung mit dem Schwerpunkt auf Steuerrecht.

METHODIK

Wie eine Bewertung entsteht

01 **Aufgaben aus echten Prüfungen**

Aus Berufsprüfungen und universitärer Lehre und erfasst nach Teilgebiet, Rechtsrahmen und Niveau.

02 **Drei Durchläufe je Aufgabe**

Jedes Modell beantwortet jede Aufgabe dreimal und gibt dabei die eigene Selbsteinschätzung (Confidence) an.

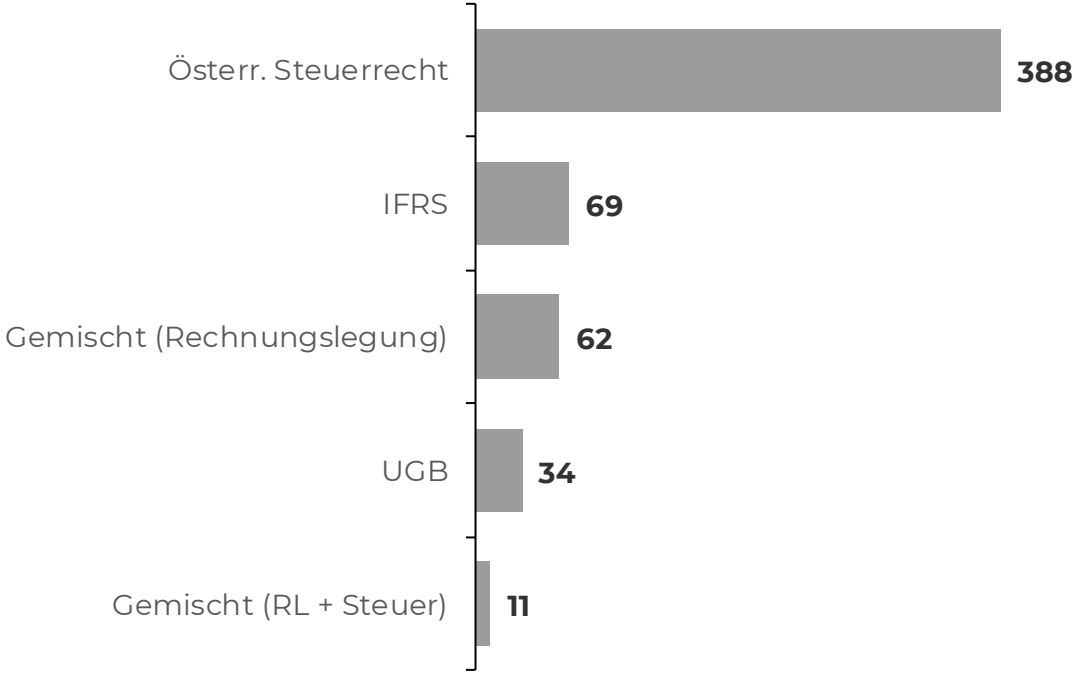
03 **Hybride Bewertung**

Strukturierte Aufgaben per Formel; Freitext von 0 bis 100 gegen hinterlegte Bewertungskriterien; numerische Antworten mit Toleranzgrenze.

04 **Aggregation und Nachvollziehbarkeit**

Auswertung je Teilgebiet und Rechtsrahmen. Jeder Lauf ist protokolliert und reproduzierbar.

Zusammensetzung der 564 Aufgaben



- **Teilgebiet**

Steuerrecht 349 · Externes Rechnungswesen 166 · Internes Rechnungswesen 49

- **Aufgabentyp**

Rechtsauslegung 458 · Berechnung 68 · Buchungssatz 38

DAS GESAMTBILD

Das beste Modell erreicht 83 %

83,2 %

BESTES MODELL

claude-opus-5. Der Durchschnitt über alle Aufgaben, mit Teilpunkten für teilweise richtige Lösungen.

53,0 %

VOLLSTÄNDIG KORREKT

Gut die Hälfte der Aufgaben löst das beste Modell ohne jeden Abzug. Der Rest ist teilweise oder ganz falsch.

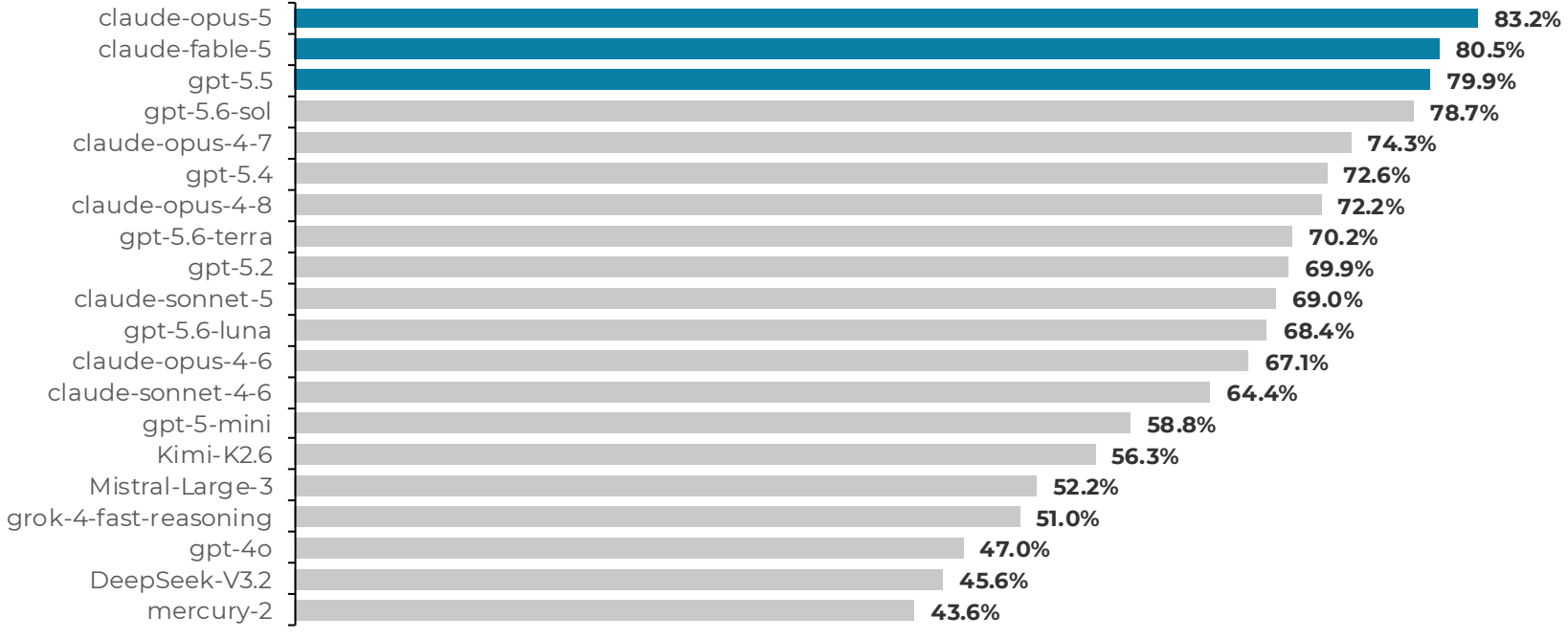
43,6 %

SCHWÄCHSTES MODELL

mercury-2. Zwischen bestem und schwächstem Modell liegen 39,6 Prozentpunkte.

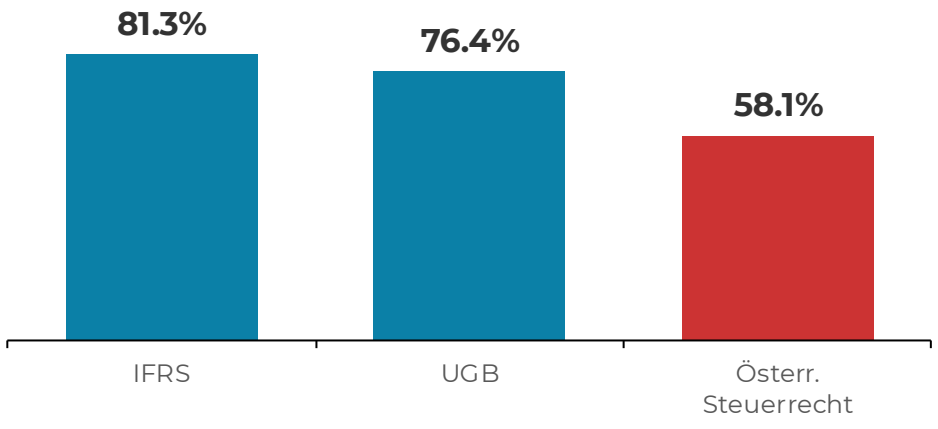
DIE RANGLISTE

Rund 40 Prozentpunkte zwischen den Modellen



claude-sonnet-5: 563 von 564 Aufgaben · Mistral-Large-3: 560 von 564 Aufgaben · grok-4-fast-reasoning: 552 von 564 Aufgaben

Trefferquote je Rechtsrahmen

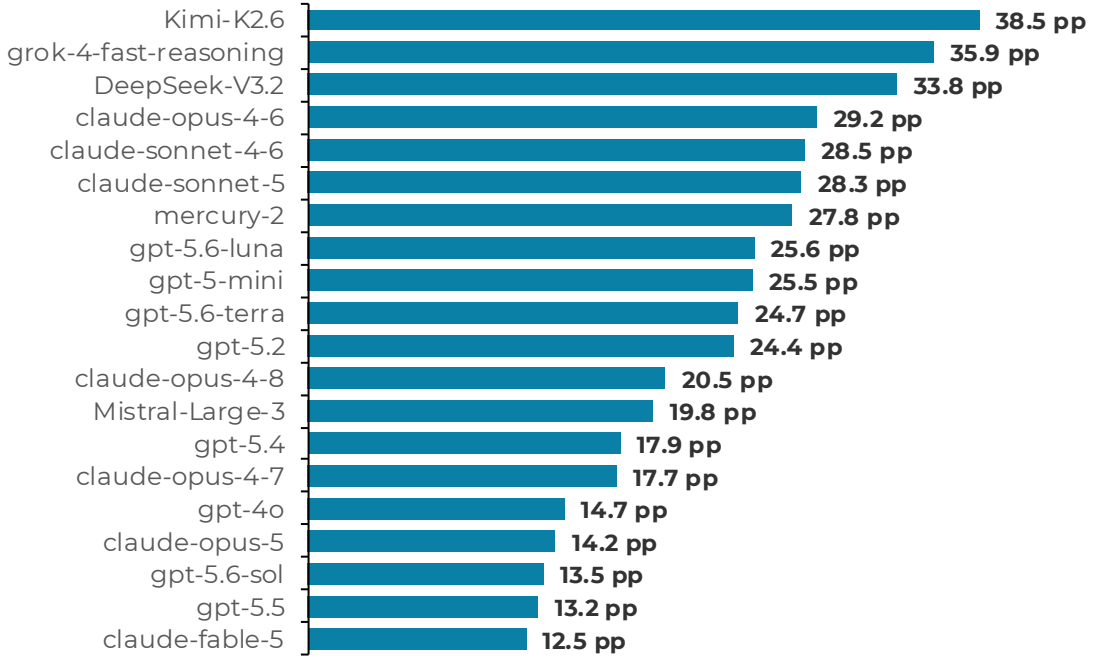


Strukturelles Gefälle

- **IFRS — 81,3 %**
  Weltweit einheitlich, umfangreich dokumentiert.
- **UGB — 76,4 %**
  Nationale Rechnungslegung — nur 4,9 Prozentpunkte unter IFRS.
- **Österr. Steuerrecht — 58,1 %**
  Auslegungslastig und laufend geändert.

Von IFRS zum UGB sind es 4,9 Prozentpunkte, vom UGB zum österreichischen Steuerrecht 18,3. Je lokaler und auslegungsbedürftiger der Stoff, desto schwächer schneidet das Modell ab.

Abstand IFRS zu Steuerrecht



Differenz IFRS zu Steuerrecht

- **claude-fable-5 — 12,5 pp**

Der kleinste Abstand zwischen IFRS und Steuerrecht.

- **Kimi-K2.6 — 38,5 pp**

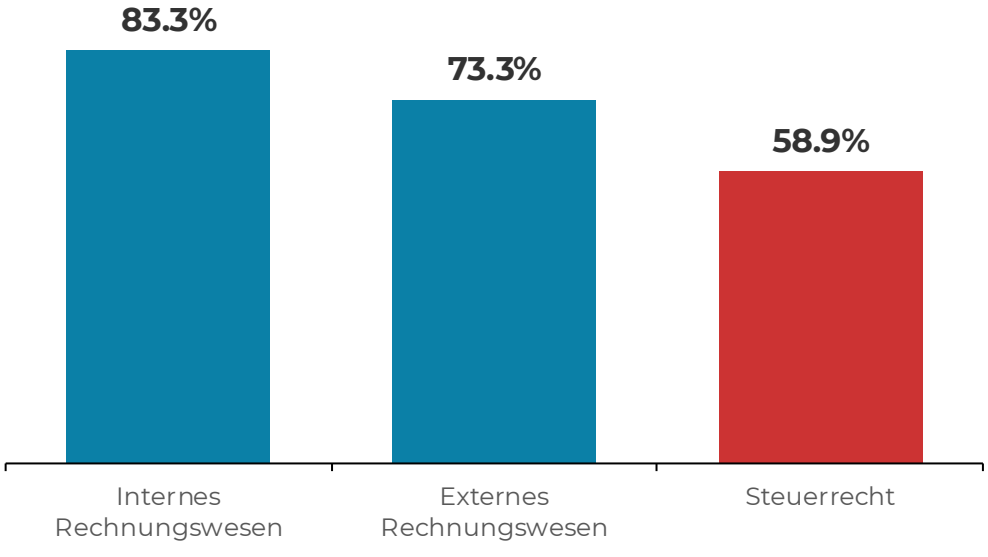
Der größte Abstand, typischerweise bei den schwächeren Modellen.

- **Kein Modell ohne Gefälle**

Auch das beste Modell bleibt im Steuerrecht deutlich hinter seiner eigenen IFRS-Leistung zurück.

NACH TEILGEBIET

Trefferquote je Teilgebiet



- **Internes Rechnungswesen — 83,3 %**

  49 Aufgaben, Kostenrechnung. Klar definierte Verfahren, wenig Auslegung.

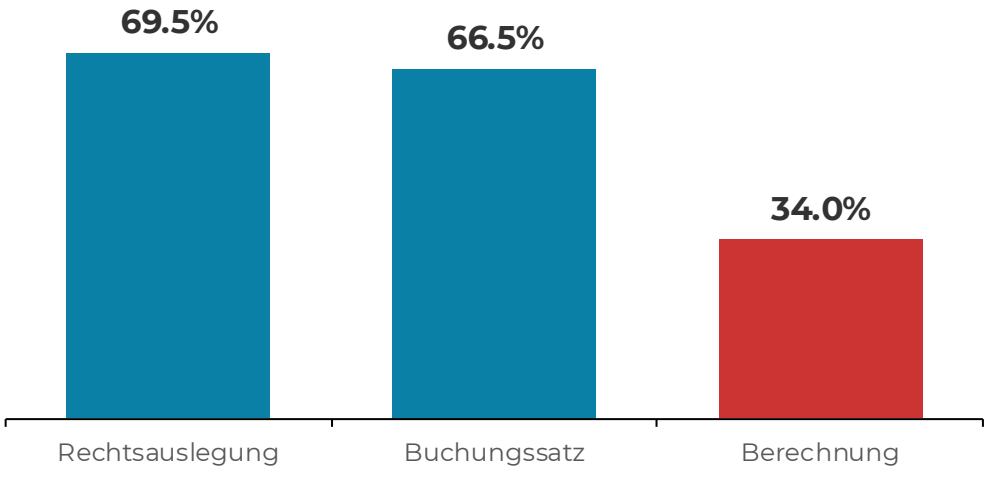
- **Externes Rechnungswesen — 73,3 %**

  166 Aufgaben, IFRS und UGB. Zehn Prozentpunkte darunter.

- **Steuerrecht — 58,9 %**

  349 Aufgaben — der Schwerpunkt des Benchmarks und zugleich das schwächste Teilgebiet.

Berechnungsaufgaben sind am schwächsten



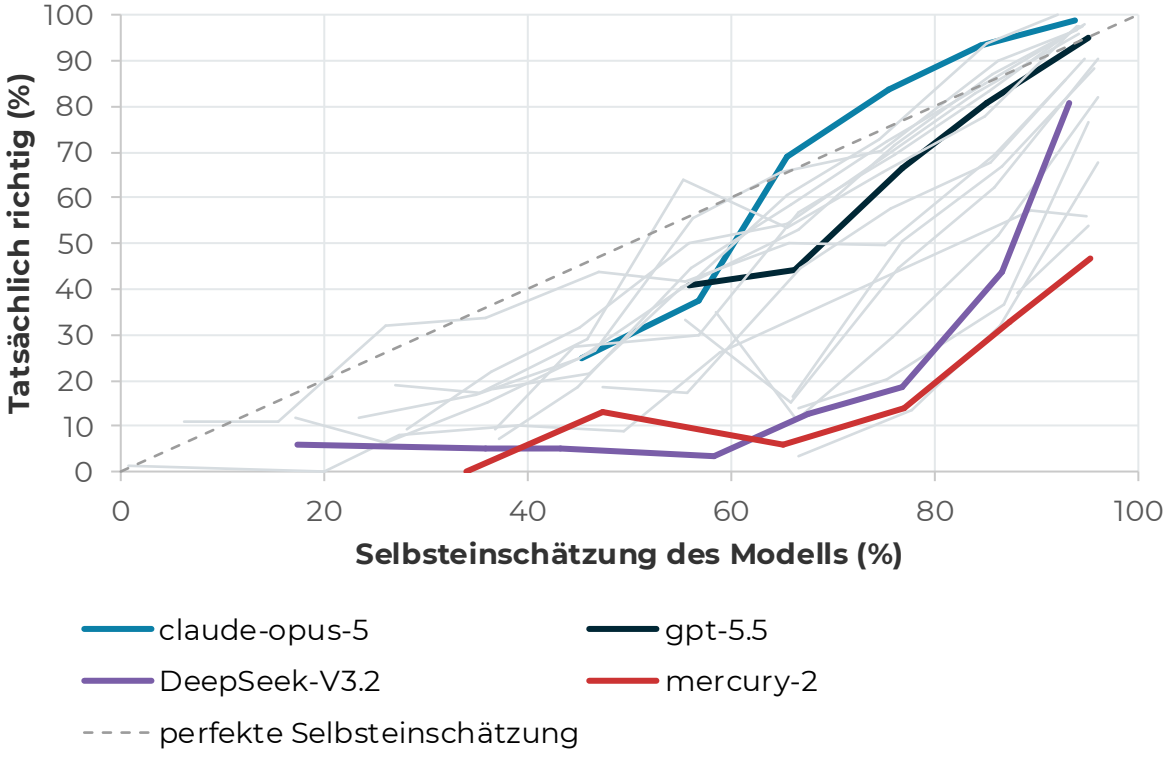
Gegen die Intuition

Man erwartet, dass eine Maschine gut rechnet und beim Auslegen scheitert. Gemessen ist es umgekehrt: Reine Berechnungsaufgaben sind die schwächste Kategorie im Benchmark.

GRÖSSTE SELBSTÜBERSCHÄTZUNG

Bei Berechnungen liegt die Selbsteinschätzung 32,4 Prozentpunkte über der tatsächlichen Trefferquote.

19 von 20 Modellen sind zu selbstsicher

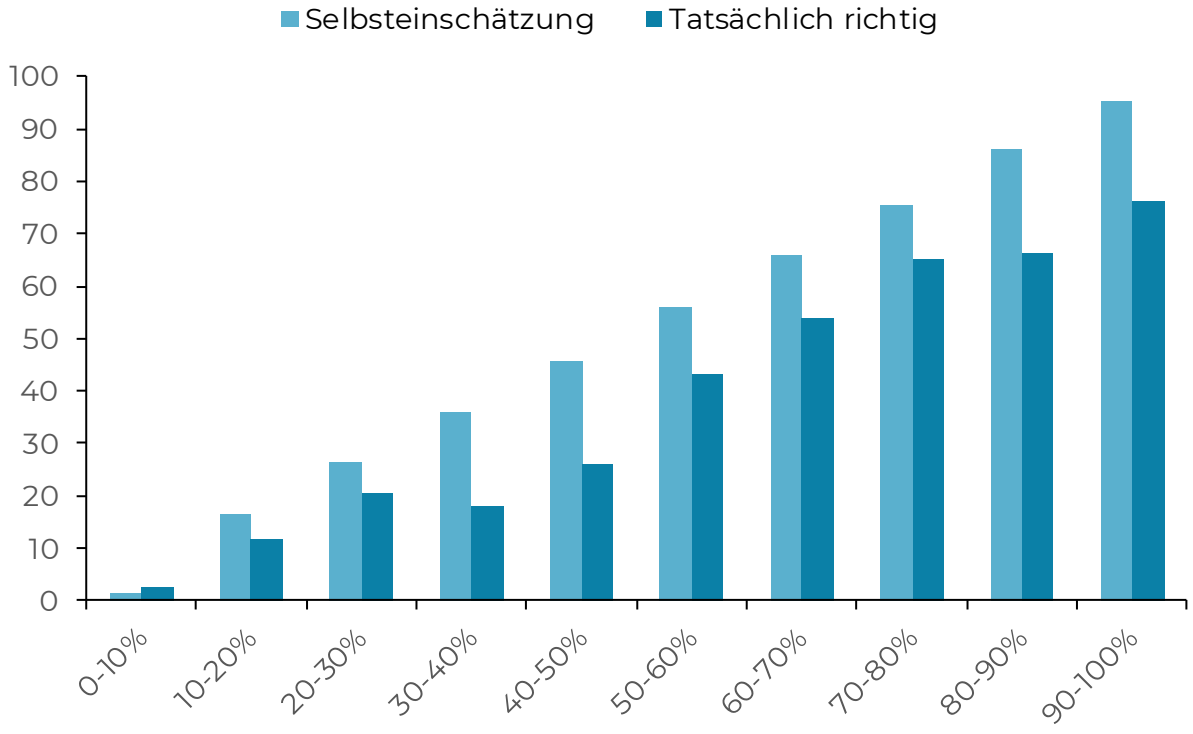


- **Die gestrichelte Diagonale**
  wäre perfekte Selbsteinschätzung
  80 % Sicherheit = 80 % richtig

- **Unterhalb: Selbstüberschätzung**
  im obersten Sicherheitsband gibt mercury-2 95 % an — richtig davon sind nur 47 %.

- **claude-opus-5 liegt darüber**
  die einzige Ausnahme: das beste Modell unterschätzt sich eher

Hohe Sicherheit bedeutet nicht Richtigkeit



- **Band 90–100 %:**

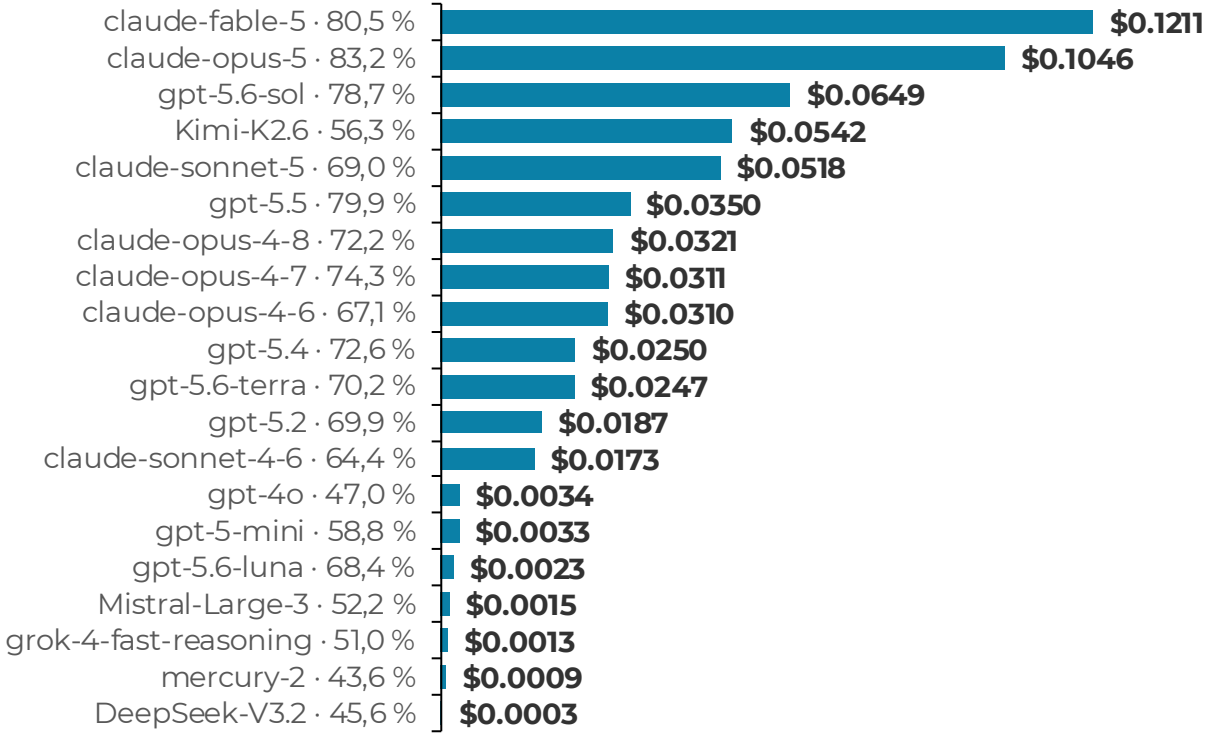
36,6 % aller Antworten fallen in diesen Confidence-Bereich. Tatsächlich richtig sind davon nur 76,3 %.

Linke Säule: die Sicherheit, die das Modell selbst angibt.

Rechte Säule: wie oft die Antwort tatsächlich stimmt.

Gleich hohe Säulen hieße: richtig eingeschätzt - je größer der Abstand, desto stärker die Überschätzung.

Kosten und Nutzen

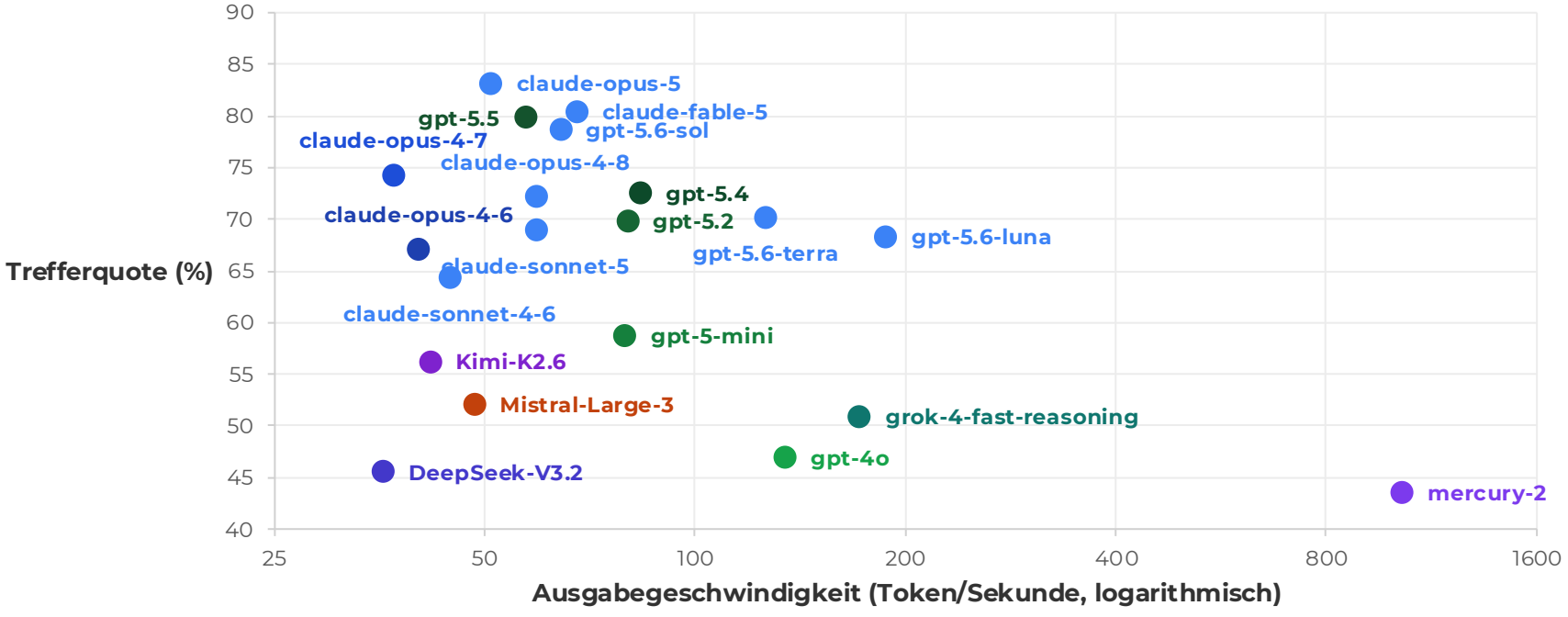


gpt-5.6-luna — 68,4 %
$0.0023
das Preis-Leistungs-Optimum im Test

claude-opus-5 — 83,2 %
$0.1046
45× teurer als gpt-5.6-luna für 14,8 Prozentpunkte mehr

claude-fable-5 — 80,5 %
$0.1211
das teuerste Modell ist nicht das beste

Geschwindigkeit der Modelle vs Trefferquote



Trefferquote: AccountingBench, n = 564. Ausgabegeschwindigkeit: Artificial Analysis, Stand August 2026.

Conclusio

01 Welches Modell verwendet wird, macht rund 40 Prozentpunkte in der Trefferquote aus.

02 Das österreichische Steuerrecht ist die Schwachstelle: 58,1 % gegenüber 81,3 % bei IFRS.

03 Auslegen können die Modelle besser als rechnen. Reine Berechnungsaufgaben liegen bei 34,0 %.

04 Der Preis des Modells ist ein Qualitätsmerkmal.

05 Kontrolliert wird nach Aufgabentyp und Rechtsrahmen. Die Selbsteinschätzung des Modells ist dafür kein Maßstab.

06 Fachwissen bleibt die Voraussetzung. Prüfen kann nur, wer die Sache selbst beherrscht.

Kontakt

E-MAIL

AccountingBench@wu.ac.at

WEBSITE

boardservice.center/research/accountingbench.html

WAS WIR SUCHEN

Praxisfälle: Echte Aufgaben aus dem Kanzleialltag (komplexe Sachverhalte, die mit umfassendem Normverständnis erst beantwortbar werden), die wir anonymisiert in den Benchmark aufnehmen.

Kooperationen: Sowohl Kanzleien als auch Unternehmen, die interessiert daran sind, wie KI schon heute eingesetzt werden kann.

ZUR BENCHMARK-WEBSITE

Rangliste, Auswertungen und Methodik im Detail

VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Manuel Kaburek
Institute for Accounting and Auditing
Financial Accounting and Auditing Group

WU
Vienna University of Economics and Business
Welthandelsplatz 1, 1020 Wien

Quellenverzeichnis

API-Preislisten der Anbieter (2026). OpenAI, Anthropic, Mistral AI, xAI, DeepSeek und Inception Labs. Stand 18. August 2026. Grundlage der Kostenauswertung.

Artificial Analysis (2026). Independent analysis of AI models: Output speed in tokens per second. Stand August 2026. https://artificialanalysis.ai

Brynjolfsson, E., Chandar, B., & Chen, R. (2025, überarbeitet 2026). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab. https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/

Bundeskanzleramt (2026, 20. Juli). Pröll: Public AI launcht GovGPT für die Bundesverwaltung [Nachricht der Bundesregierung]. https://www.bundeskanzleramt.gv.at/bundeskanzleramt/nachrichten-der-bundesregierung/2026/07/proell-public-ai-launcht-govgpt-fuer-die-bundesverwaltung.html

Bundesministerium für Finanzen (2025, 13. August). Finanzministerium lukrierte im Vorjahr 354 Mio. Euro Steuermehreinnahmen dank KI-Methoden [Presseaussendung]. https://www.bmf.gv.at/presse/pressemeldungen/2025/august/pacc-ki.html

Challenger, Gray & Christmas, Inc. (2026). The Challenger Report. https://www.challengergray.com/blog/category/job-cuts-report/

KPMG International (2025). Global AI in finance report. https://kpmg.com/xx/en/our-insights/ai-and-technology/kpmg-global-ai-in-finance-report.html

KPMG International (2026). Global AI in finance 2026 — The decision advantage: How AI is producing value across the finance function.

Lovett, N. (2026). The tragedy of the cognitive commons: How AI could disrupt the regeneration of professional expertise. Human Resource Development Review. Vorab-Onlineveröffentlichung. https://doi.org/10.1177/15344843261470602

Massenkoff, M., & McCrory, P. (2026, 5. März). Labor market impacts of AI: A new measure and early evidence. Anthropic. https://www.anthropic.com/research/labor-market-impacts

Murray, C. (2026, 19. August). Is AI really responsible for recent job cuts? Financial Times. https://www.ft.com/content/0dc14b44-96f6-4b1f-921a-8cba8030eafc

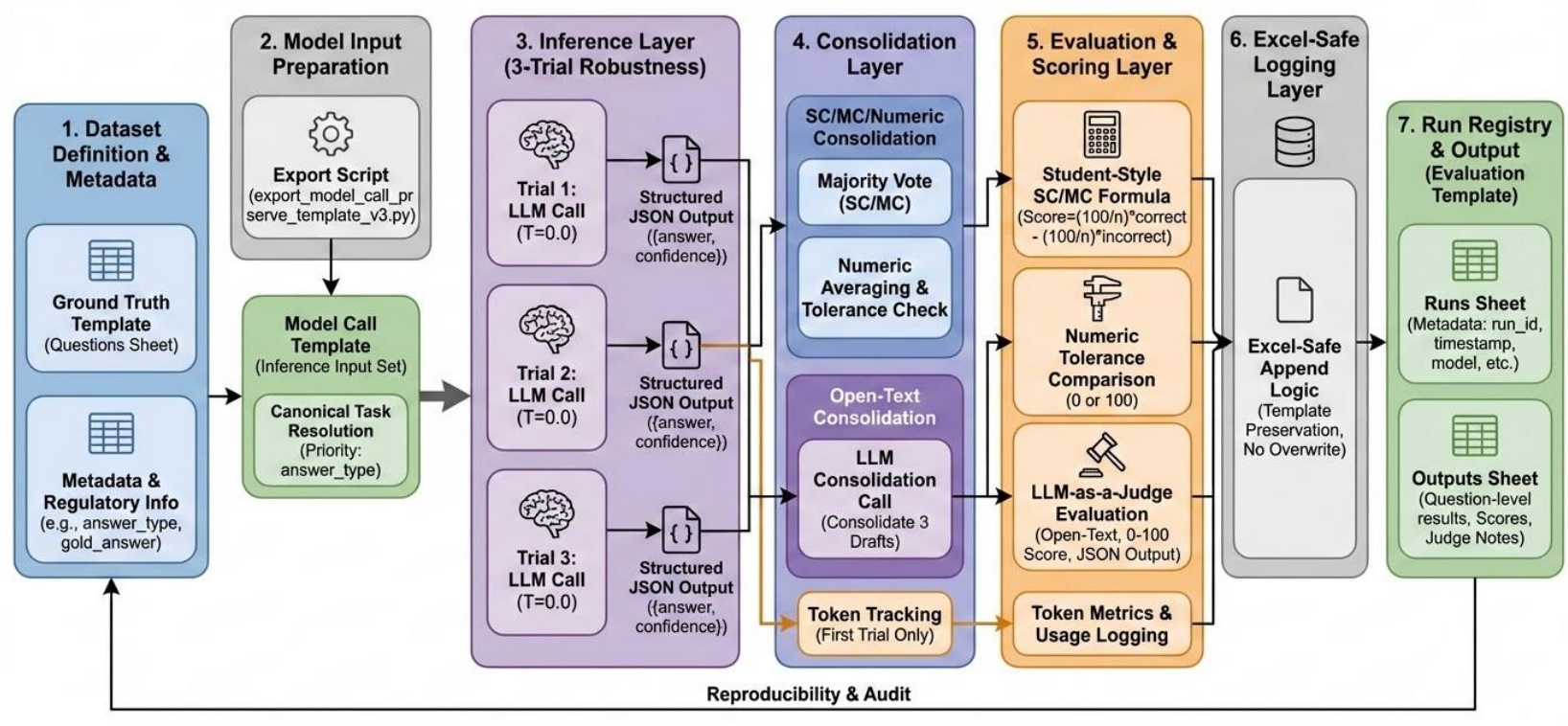
Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Hrsg.), Computers, communications, and the public interest (S. 37–72). Johns Hopkins Press.

Thomson Reuters Institute (2026). 2026 AI in professional services report: GenAI is here, agentic AI is coming — and business model shifts are next. https://www.thomsonreuters.com/en/reports/2026-ai-in-professional-services-report

Thomson Reuters Institute (2026, 22. Juni). Future of professionals 2026: As AI adoption grows, so do the challenges. https://www.thomsonreuters.com/en-us/posts/technology/future-of-professionals-2026/

ANHANG

AccountingBench



Bewertungsverfahren im Detail

01 **Strukturierte Aufgaben**

Single und Multiple Choice werden per Formel bewertet.

$$\text{SC/MC-Score} = \frac{100}{c} \cdot n_c - \frac{100}{c} \cdot n_i \qquad (1)$$

where c denotes the number of correct alternatives, n_c the number of correct marks by the model, and n_i the number of incorrect marks by the model.

02 **Freitext**

Bewertung von 0 bis 100 durch ein Sprachmodell als Judge, gegen die je Aufgabe hinterlegten Bewertungskriterien und die Musterlösung.

03 **Numerische Antworten**

Bewertung mit aufgabenspezifischer Toleranz.

04 **Grenzen des Verfahrens**

Der Judge ist selbst ein Sprachmodell. Bewertungen von Freitextantworten sind daher mit Unsicherheit behaftet — die Rangfolge der Modelle ist robuster als der einzelne Punktwert.

Heat Map

	GESAMT	KATEGORIE			AUFGABENTYP			ANTWORTTYP				RECHTSRAHMEN				NIVEAU		
Modell	**Gesamt**	**Steuer-recht**	**Externes RW**	**Internes RW**	**Rechts-auslegung**	**Berech-nung**	**Buchungs-satz**	**Multiple Choice**	**Freitext**	**Single Choice**	**Buchungs-satz**	**Österr. StR**	**Gemischt**	**UGB**	**IFRS**	**Berufs-prüfung**	**Master**	**BHS**
Aufgaben (n)	564	349	166	49	458	68	32	164	333	44	15	388	62	34	69	392	116	56
claude-opus-5	83,2	81,7	84,8	89,1	88,8	44,3	83,8	94,0	76,8	93,2	87,5	79,6	91,3	87,8	93,8	82,7	94,2	63,9
claude-fable-5	80,5	81,1	78,1	85,1	86,6	41,5	73,6	91,4	75,7	84,1	71,6	78,1	88,2	79,3	90,6	81,1	90,5	55,8
gpt-5.5	79,9	77,7	83,1	84,5	83,8	52,3	79,5	91,7	74,0	84,1	79,6	76,6	87,8	82,0	89,8	77,8	92,2	69,0
gpt-5.6-sol	78,7	76,6	80,8	86,0	84,2	40,2	78,0	90,1	72,2	90,9	73,1	75,7	88,9	75,1	89,2	77,6	89,6	63,7
claude-opus-4-7	74,3	70,4	78,5	88,4	79,2	40,1	73,4	90,5	65,9	81,8	74,0	69,1	89,2	81,9	86,8	71,3	91,6	60,2
gpt-5.4	72,6	68,6	78,0	82,4	78,0	33,5	73,3	90,3	62,9	86,4	71,3	67,2	83,7	82,4	85,1	69,3	90,6	58,3
claude-opus-4-8	72,2	67,6	76,6	89,7	76,4	39,4	76,3	88,5	63,2	77,3	79,9	65,9	90,2	80,4	86,4	69,4	90,3	54,2
gpt-5.6-terra	70,2	63,6	80,2	84,1	74,6	34,4	79,3	89,0	58,8	86,4	81,7	63,1	87,4	81,6	87,8	65,5	92,2	58,3
gpt-5.2	69,9	63,9	78,3	84,3	73,9	37,3	76,7	90,1	58,6	81,8	74,7	62,9	84,4	83,1	87,3	65,8	90,0	57,2
claude-sonnet-5	69,0	63,4	76,5	82,9	73,6	30,5	79,8	89,5	58,1	72,7	87,8	61,4	83,3	82,3	89,7	65,4	90,1	50,0
gpt-5.6-luna	68,4	61,1	78,8	85,3	72,1	36,1	80,4	89,9	55,2	86,4	81,1	60,8	86,7	84,5	86,4	63,3	90,0	59,7
claude-opus-4-6	67,1	59,6	76,6	88,0	70,1	41,3	76,7	88,5	55,8	70,5	76,9	58,8	88,9	76,0	88,0	61,9	89,6	57,0
claude-sonnet-4-6	64,4	55,4	77,0	85,5	67,3	38,3	74,4	87,7	51,0	75,0	79,6	55,5	87,0	80,5	84,0	57,4	90,7	59,2
gpt-5-mini	58,8	51,0	68,9	80,3	62,7	28,0	64,4	83,7	45,7	65,9	68,0	50,7	78,8	76,2	76,2	52,0	85,7	50,9
Kimi-K2.6	56,3	44,8	72,8	82,1	58,6	34,3	63,1	86,1	38,7	75,0	66,3	45,1	82,7	72,9	83,6	47,9	87,4	50,6
Mistral-Large-3	52,2	42,8	62,4	83,2	57,3	23,0	39,4	86,2	33,7	70,5	36,1	42,8	84,5	70,4	62,6	42,4	89,2	42,9
grok-4-fast-reasoning	51,0	40,1	64,3	80,8	52,5	34,2	57,1	83,0	32,0	68,2	62,2	40,4	76,8	62,9	76,3	42,7	80,0	47,5
gpt-4o	47,0	38,4	56,2	77,0	53,3	11,7	34,4	82,2	28,6	63,6	32,8	38,1	78,6	71,9	52,8	36,1	89,3	36,1
DeepSeek-V3.2	45,6	36,2	57,8	71,4	50,1	20,2	32,3	79,1	28,1	65,9	25,2	35,8	74,4	53,5	69,6	37,2	81,2	31,1
mercury-2	43,6	32,9	56,8	75,4	47,8	19,0	34,3	78,2	23,5	70,5	37,5	33,1	76,5	62,8	60,9	32,9	82,4	37,9

Trefferquote ≥ 90 % | 75–90 % | 60–75 % | 50–60 % | < 50 %

Angaben in Prozent, Zeilen nach Gesamttrefferquote sortiert. Aufgabenzahl je Spalte für die 17 Modelle mit vollständigem Lauf; claude-sonnet-5 563 von 564, Mistral-Large-3 560 von 564, grok-4-fast-reasoning 552 von 564. Nicht ausgewiesen: gemischt Rechnungslegung + Steuer (n = 11), Auslegung + Buchungssatz (n = 6), offene Zahlenantworten (n = 8)

Bewertungsbasis je Modell

Modell	tasks_mit_lauf	mit_final_answer	bewertbar
DeepSeek-V3.2	564	564	564
Kimi-K2.6	564	564	564
claude-fable-5	564	564	564
claude-opus-4-6	564	564	564
gpt-5.6-sol	564	564	564
claude-opus-4-7	564	564	564
claude-opus-4-8	564	564	564
claude-opus-5	564	564	564
claude-sonnet-4-6	564	564	564
gpt-4o	564	564	564
gpt-5-mini	564	564	564
gpt-5.2	564	564	564
mercury-2	564	564	564
gpt-5.4	564	564	564
gpt-5.5	564	564	564
gpt-5.6-luna	564	564	564
gpt-5.6-terra	564	564	564
claude-sonnet-5	564	563	563
Mistral-Large-3	564	560	560
grok-4-fast-reasoning	552	552	552

Trefferquote je Modell und Rechtsrahmen

Modell	IFRS	UGB	Gemischt	Ö. Steuer	
claude-opus-5	93,8	87,8	91,3	79,6	
claude-fable-5	90,6	79,3	88,2	78,1	
gpt-5.5	89,8	82,0	87,8	76,6	
gpt-5.6-sol	89,2	75,1	88,9	75,7	
claude-opus-4-7	86,8	81,9	89,2	69,1	
gpt-5.4	85,1	82,4	83,7	67,2	
claude-opus-4-8	86,4	80,4	90,2	65,9	
gpt-5.6-terra	87,8	81,6	87,4	63,1	
gpt-5.2	87,3	83,1	84,4	62,9	
claude-sonnet-5	89,7	82,3	83,3	61,4	
gpt-5.6-luna	86,4	84,5	86,7	60,8	
claude-opus-4-6	88,0	76,0	88,9	58,8	
claude-sonnet-4-6	84,0	80,5	87,0	55,5	
gpt-5-mini	76,2	76,2	78,8	50,7	
Kimi-K2.6	83,6	72,9	82,7	45,1	
Mistral-Large-3	62,6	70,4	84,5	42,8	
grok-4-fast-reasoning	76,3	62,9	76,8	40,4	
gpt-4o	52,8	71,9	78,6	38,1	
DeepSeek-V3.2	69,6	53,5	74,4	35,8	
mercury-2	60,9	62,8	76,5	33,1	

Kalibrierung im Detail

Modell	n	Trefferquote	Selbsteinschätzung	Differenz		
mercury-2	564	43,6	92,3	48,7		
Mistral-Large-3	550	52,2	94,4	42,2		
grok-4-fast-reasoning	552	51,0	91,1	40,1		
DeepSeek-V3.2	564	45,6	84,8	39,2		
Kimi-K2.6	564	56,3	90,1	33,8		
gpt-4o	562	47,0	76,9	29,9		
gpt-5.6-luna	564	68,4	91,0	22,6		
gpt-5.6-terra	564	70,2	87,9	17,7		
gpt-5.4	564	72,6	85,3	12,7		
gpt-5.6-sol	564	78,7	91,3	12,6		
gpt-5-mini	559	58,8	71,4	12,6		
claude-opus-4-6	465	67,1	77,8	10,7		
claude-sonnet-4-6	505	64,4	72,9	8,5		
claude-opus-4-7	564	74,3	79,5	5,2		
gpt-5.5	564	79,9	85,0	5,1		
claude-opus-4-8	564	72,2	76,2	4,0		
claude-fable-5	564	80,5	82,2	1,7		
gpt-5.2	564	69,9	71,6	1,7		
claude-sonnet-5	563	69,0	69,9	0,9		
claude-opus-5	564	83,2	78,0	-5,2		

Trefferquote mit 95-%-Konfidenzintervallen

Modell	n	Trefferquote	CI unten	CI oben	voll korrekt
claude-opus-5	564	83,2	80,9	85,5	53,0
claude-fable-5	564	80,5	78,0	83,1	52,3
gpt-5.5	564	79,9	77,4	82,3	46,6
gpt-5.6-sol	564	78,7	76,2	81,1	45,2
claude-opus-4-7	564	74,3	71,6	77,0	42,6
gpt-5.4	564	72,6	69,8	75,4	42,2
claude-opus-4-8	564	72,2	69,5	74,9	40,6
gpt-5.6-terra	564	70,2	67,4	73,1	39,9
gpt-5.2	564	69,9	67,1	72,8	38,5
claude-sonnet-5	563	69,0	66,0	71,9	41,0
gpt-5.6-luna	564	68,4	65,6	71,3	39,0
claude-opus-4-6	564	67,1	64,2	69,9	37,6
claude-sonnet-4-6	564	64,4	61,4	67,4	35,1
gpt-5-mini	564	58,8	55,8	61,8	29,1
Kimi-K2.6	564	56,3	53,2	59,4	30,9
Mistral-Large-3	560	52,2	49,0	55,3	28,7
grok-4-fast-reasoning	552	51,0	47,8	54,2	28,8
gpt-4o	564	47,0	43,8	50,3	27,1
DeepSeek-V3.2	564	45,6	42,5	48,7	23,6
mercury-2	564	43,6	40,5	46,7	24,3

Kosten je Aufgabe

Modell	Trefferquote	USD/Aufgabe	USD/1000	Token ein	Token aus	Token Reasoning
DeepSeek-V3.2	45,6	$0.0003	0,33	416	263	0
mercury-2	43,6	$0.0009	0,94	358	164	460
grok-4-fast-reasoning	51,0	$0.0013	1,26	398	376	902
Mistral-Large-3	52,2	$0.0015	1,48	404	746	0
gpt-5.6-luna	68,4	$0.0023	2,32	372	1761	1025
gpt-5-mini	58,8	$0.0033	3,28	368	1267	512
gpt-4o	47,0	$0.0034	3,35	368	98	0
claude-sonnet-4-6	64,4	$0.0173	17,26	520	1016	0
gpt-5.2	69,9	$0.0187	18,69	368	1076	449
gpt-5.6-terra	70,2	$0.0247	24,68	372	1706	1024
gpt-5.4	72,6	$0.0250	24,98	368	1305	557
claude-opus-4-6	67,1	$0.0310	30,98	520	1227	0
claude-opus-4-7	74,3	$0.0311	31,07	701	1050	0
claude-opus-4-8	72,2	$0.0321	32,05	711	1124	0
gpt-5.5	79,9	$0.0350	34,98	364	980	516
claude-sonnet-5	69,0	$0.0518	51,82	710	3037	0
Kimi-K2.6	56,3	$0.0542	54,20	458	11820	8256
gpt-5.6-sol	78,7	$0.0649	64,87	372	1736	1284
claude-opus-5	83,2	$0.1046	104,63	711	2854	0
claude-fable-5	80,5	$0.1211	121,12	711	1858	0

Limitationen

Kompetenz

Echtes fachliches Schließen auf einen neuen Sachverhalt wird nicht gemessen.

Training auf Prüfungsformate

Die Anbieter trainieren gezielt auf prüfungsähnliche Aufgaben. Ein Teil der Leistung kann Formatvertrautheit sein.

Bekanntes Material

Ein Teil der Rechtsquellen und Musterlösungen ist öffentlich zugänglich. Erinnern kann fachliches Urteil ersetzen, ohne dass man es an der Antwort erkennt (Datenkontamination).