KURZFASSUNG EINER BACHELORARBEIT

LLM-Based Reasoning Assistants in Financial Planning

A Decision Quality Comparison with Professional Planners. Ein KI-Assistent und eine erfahrene Fachperson lösen unabhängig voneinander dieselben zwei Prüfungsfälle der eidgenössischen Berufsprüfung Finanzplaner/in.

RAHMEN

Arbeit	**Bachelorarbeit BSc Betriebsökonomie**
Vertiefung	**Banking & Finance**
Hochschule	**ZHAW School of Management and Law**
Abgabe	**Mai 2026**
Autor	**Lukas Bachmann**
Sprache	**Englisch**

Kurzfassung von innoVorsorge auf Grundlage der englischen Originalarbeit; massgebend ist das Original. Übersetzungen und Paraphrasen sind sinngemäss.

FRAGESTELLUNG

"How does the decision quality of an LLM-based reasoning assistant compare to that of a professional financial planner, and what does this imply for the integration of such tools in financial advisory practice?"

Sinngemäss: Wie schneidet die Entscheidungsqualität eines KI-Assistenten im Vergleich zu jener einer Finanzplanerin oder eines Finanzplaners ab, und was folgt daraus für den Einsatz solcher Werkzeuge in der Beratung?

AUSGANGSLAGE

Allgemeine Sprachmodelle erreichen in mehreren Studien bei Aufgaben mit finanziellem Schlussfolgern eine Genauigkeit von 65 bis 80% (Fieberg et al., 2025; Hean et al., 2025; Hofmann et al., 2026; Zhang et al., 2025). In der Schweizer Studie von Hofmann et al. (2026) mit 14 Modellen und 58 Fragen aus dem MAS Financial Consulting der ZHAW kam das beste Modell auf eine Abdeckungsquote von 0.810, im Bereich Beratungsprozess und Ethik erreichten die Modelle aber nur 0.448. Offen war, wie ein eigens gebautes, auf Schweizer Recht gestütztes Werkzeug gegen eine qualifizierte Fachperson abschneidet, wenn beide denselben vollständigen Fall mit bekannter Musterlösung bearbeiten.

1 Methode

VERSUCHSAUFBAU

- Parallelvergleich: KI-Assistent und Fachperson lösen dieselben Fälle unabhängig, mit denselben Unterlagen, ohne Einsicht in die Lösung der anderen Seite und ohne Musterlösung.
- Zwei offizielle Prüfungsfälle der eidg. Berufsprüfung Finanzplaner/in mit amtlicher Musterlösung: FA15 Fall 4 (Prüfung Juni 2021) und Nullserie 2024 Fall 1.
- Keine Rückfragen erlaubt: Wie an der Prüfung arbeiten beide nur mit dem Falldossier. Unklarheiten löst die KI über ausgewiesene Annahmen, die Fachperson über ihr Urteil.
- Beide liefern ein schlicht gestaltetes PDF, Frage für Frage gegliedert, damit die Bewertung den Inhalt und nicht die Aufmachung misst.

DIE ZWEI TEILNEHMENDEN

KI-Assistent: Claude Opus 4.6 mit einer kuratierten Wissensbasis aus 62 Dokumenten (13'766 Zeilen) zu 1. Säule, 2. Säule, 3. Säule, Steuern (Bund und alle 26 Kantone) und Zusatzthemen wie FIDLEG, Erb- und Güterrecht. Parameter mit Stand 1. Januar 2026, vorwiegend aus amtlichen Quellen. Ein deutschsprachiger Systemprompt richtet sich gegen die in der Literatur beschriebenen Fehlerquellen (nur Schweizer Recht, Wissensbasis vor Trainingswissen, Rechnungen Schritt für Schritt, Annahmen kennzeichnen). Je Fall ein einziger Aufruf; die passenden Dateien wählte der Autor nach festen Regeln aus (13 für den Fall 2021, 15 für den Fall 2024).

Fachperson: selbstständige Finanzplanerin oder selbstständiger Finanzplaner mit rund fünf bis zehn Jahren Berufserfahrung, derzeit im Abschluss der eidg. Berufsprüfung. Ohne KI-Werkzeuge, mit Rechner, Planungssoftware und Nachschlagewerken, ohne Zeitlimit.

BEWERTUNG

Bewertet hat nicht ein menschliches Expertengremium, sondern ein Sprachmodell als Richter ("LLM-as-a-judge" nach Zheng et al., 2023): GPT-4o erhielt die amtliche Musterlösung als Referenz und beurteilte jede Lösung einzeln, ohne die Gegenseite zu sehen (reference-guided single-answer grading). Vergeben wurden ganze Punkte von 1 bis 10 in fünf Dimensionen, je mit kurzer Begründung: *Sachliche Richtigkeit*, *Rechnerische Richtigkeit*, *Vollständigkeit*, *Qualität der Empfehlung* und *Klarheit der Struktur*. Die Gesamtnote ist der ungewichtete Durchschnitt. Gegen bekannte Verzerrungen: Richter aus einer anderen Modellfamilie als der KI-Assistent, Anweisung, den Inhalt und nicht die Länge zu bewerten, und jede Bewertung zweimal (acht Durchläufe für vier Lösungen), wobei Abweichungen über zwei Punkte geprüft worden wären. Zusätzlich wurden die Fehler durch Vergleich mit der Musterlösung in vier Kategorien eingeteilt: Rechnung, Recht/Regulierung, Wissensstand und Individualisierung.

2 Ergebnisse

PUNKTE JE DIMENSION (SKALA 1–10)

	FALL 2021			FALL 2024			DURCHSCHNITT		
Dimension	**KI**	**Planer/in**	**Diff.**	**KI**	**Planer/in**	**Diff.**	**KI**	**Planer/in**	**Diff.**
Sachliche Richtigkeit	8	8	0	9	8	+1	8.5	8.0	+0.5
Rechnerische Richtigkeit	7	7	0	8	7	+1	7.5	7.0	+0.5
Vollständigkeit	7	7	0	9	8	+1	8.0	7.5	+0.5
Qualität der Empfehlung	8	8	0	9	8	+1	8.5	8.0	+0.5
Klarheit der Struktur	9	8	+1	9	8	+1	9.0	8.0	+1.0
Gesamt	**7.8**	**7.6**	**+0.2**	**8.8**	**7.8**	**+1.0**	**8.3**	**7.7**	**+0.6**

Nach Table 5 der Arbeit. Beide Bewertungsdurchläufe ergaben für alle 20 Dimensionspaare dieselben Punkte.

WAS DIE PUNKTE ZEIGEN

Im Durchschnitt liegt der KI-Assistent in jeder Dimension mindestens gleichauf, mit 8.3 gegenüber 7.7 Punkten. Der grösste Abstand betrifft die Klarheit der Struktur (+1.0), was die Arbeit plausibel mit dem vorgegebenen Antwortschema des Systemprompts erklärt. Im Fall 2021 beträgt der Unterschied nur 0.2 Punkte, im breiter angelegten Fall 2024 einen Punkt. Ob das an der Komplexität liegt oder am einzelnen Fall, lässt sich mit zwei Beobachtungen nicht entscheiden.

Die Rechnung ist für beide die schwächste Dimension (KI 7.5, Fachperson 7.0), aber aus verschiedenen Gründen: Die KI rechnete methodisch richtig mit teils falschen Eingabewerten, bei der Fachperson fehlten oft die Zwischenschritte, mit denen sich das Ergebnis hätte prüfen lassen.

BEARBEITUNGSZEIT (NEBENBEFUND)

KI-Assistent, Fall 2021	**314 Sekunden**
KI-Assistent, Fall 2024	**401 Sekunden**
Fachperson, Fall 2021	**79 Minuten**
Fachperson, Fall 2024	**92 Minuten**
Verhältnis KI zu Fachperson	**rund 1:15 und 1:14**

Die Fachperson arbeitete ohne Zeitlimit. Die Arbeit führt die Zeit nur beschreibend auf; über die Eignung entscheidet die Qualität.

PRÜFPUNKTE ZUM AKTUELLEN RECHT

13. AHV-Monatsrente (ab Dezember 2026)	**KI bestanden, Fachperson bestanden**
Abgelehnte BVG-Reform (September 2024)	**KI bestanden, Fachperson bestanden**
Rückwirkender Einkauf in die Säule 3a (ab 2026)	**KI bestanden, Fachperson bestanden**

Nach Table 7 der Arbeit. Die drei Änderungen liegen nach dem Trainingsstand vieler Sprachmodelle und standen deshalb in der Wissensbasis. Kein einziges Mal verwies die KI auf deutsches Recht oder EU-Recht.

FEHLER NACH KATEGORIE

Kategorie	KI-Assistent		Fachperson	
Rechnung	**3**	Fall 2021: AHV-Beitragslücke mit rund vier statt drei Jahren, daher Skala 40 bis 42 statt 41 und Rente CHF 2'173 bis 2'281 statt CHF 2'227 im Monat; anlegbares Vermögen CHF 500'000 bis 600'000 statt CHF 750'000 bis 900'000. Fall 2024: Überentschädigungskürzung der BVG-Renten nicht angewendet (Musterlösung: CHF 78'400 auf CHF 73'080).	**2**	In beiden Fällen Rechenschritte zu knapp dokumentiert, Zwischenwerte teils fehlend; nicht vollständig nachprüfbar.
Recht/Regulierung	**0**	Keine falschen Rechtsverweise, kein Bezug auf deutsches oder EU-Recht.	**0**	Keine falschen Rechtsverweise.
Wissensstand	**1**	Fall 2024: IV-Maximalrente 2026 (CHF 30'240 im Jahr) statt 2024 (CHF 29'400); Wissensbasis aktueller als der Fall.	**0**	Wissensstand passend zur Prüfungsperiode.
Individualisierung	**1**	Fall 2021: Risikofähigkeit "mittel bis hoch" und Risikobereitschaft "tief bis mittel" statt "mittel" und "tief" laut Musterlösung.	**1**	Empfehlungen stellenweise allgemein gehalten und zu wenig beziffert.
Total	**5**		**3**	

Nach Table 6 der Arbeit. Jeder Vorfall ist einmal gezählt, in der zutreffendsten Kategorie.

3 Grenzen der Untersuchung

Zwei Fälle und eine einzige Fachperson erlauben einen qualitativen Vergleich, aber keine statistische Verallgemeinerung; die Ergebnisse gelten für eine bestimmte KI-Konfiguration und eine bestimmte Person.

- Bewertet hat ein Sprachmodell (GPT-4o), kein Mensch hat die Punkte nachbewertet. Gleiche Punkte in beiden Durchläufen zeigen Stabilität, nicht Kalibrierung. Ob GPT-4o KI-Texte anders beurteilt als menschliche, ist nicht gemessen.
- Die Resultate gelten für ein spezialisiertes Werkzeug mit aufwendig erstellter Wissensbasis, nicht für ein allgemeines Sprachmodell, und nur für Claude Opus 4.6 zum Zeitpunkt des Versuchs.
- Der Autor hat das Werkzeug selbst gebaut und getestet; eine Wiederholung durch Dritte steht aus. Die KI löste jeden Fall nur einmal (Temperatur 1.0), die Streuung zwischen Durchläufen ist offen.
- Prüfungsfälle sind strukturierte Übungen. Echte Beratung verlangt Rückfragen, Gespräch und schrittweise Verfeinerung, was hier ausgeschlossen war.

4 Folgerungen für die Praxis

1. **Entwurf und Prüfung.** Struktur und Vollständigkeit sprechen für die KI als Werkzeug für den ersten Entwurf. Die Fachperson prüft die Rechnungen, passt die Empfehlungen an die Kundschaft an und bringt ein, was erst im Gespräch sichtbar wird.
2. **Wissensbasis als Voraussetzung.** Der Abstand zu den rund 70% allgemeiner Modelle passt zur Bedeutung einer kuratierten Schweizer Wissensbasis. Aufbau und Pflege sind Voraussetzung für verlässliche Ergebnisse.
3. **Rechnungen immer nachprüfen.** Falsche Eingabewerte zogen sich durch die Folgerechnungen. Jede KI-Planung braucht vor der Abgabe an Kundschaft eine unabhängige Kontrolle; die Darstellung Schritt für Schritt erleichtert sie.
4. **Parameter versioniert führen.** Die zu aktuellen IV-Werte im Fall 2024 zeigen: Parameter müssen nachgeführt und auf die massgebende Planungsperiode festgelegt werden.

SCHLUSS

In beiden Prüfungsfällen erreichte der KI-Assistent mit kuratierter Wissensbasis die Entscheidungsqualität der beteiligten Fachperson (8.3 gegenüber 7.7 Punkten). Sein Fehlerprofil spricht nach der Arbeit für Ergänzung statt Ersatz ("an augmentation model rather than substitution"): Die KI liefert einen strukturierten Entwurf mit breiter Abdeckung, die Fachperson bleibt verantwortlich für die Eingabewerte, für das Erfragen der Kundenwünsche und für den Stand der Parameter. Die Zeitersparnis stützt diesen Ansatz, entscheidend bleibt die Qualität. Die Arbeit versteht ihre Ergebnisse als vorläufige Evidenz auf Fallebene und als Ausgangspunkt für grössere Wiederholungsstudien.