

LLM-Based Reasoning Assistants in Financial Planning

A Decision Quality Comparison with Professional Planners. An AI assistant and an experienced planner independently solve the same two cases from the Swiss federal examination for financial planners.

CONTEXT

Thesis	Bachelor's thesis, BSc in Business Administration
Specialisation	Banking & Finance
University	ZHAW School of Management and Law
Submitted	May 2026
Author	Lukas Bachmann
Language	English

Summary prepared by innoVorsorge from the original thesis; the original is authoritative. Paraphrases follow its wording as closely as the format allows.

RESEARCH QUESTION

"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?"

Decision quality is measured on five dimensions. Processing time is recorded as a secondary, descriptive observation.

BACKGROUND

Across several studies, general-purpose language models reach 65 to 80% accuracy on financial reasoning tasks (Fieberg et al., 2025; Hean et al., 2025; Hofmann et al., 2026; Zhang et al., 2025). In the Swiss study by Hofmann et al. (2026), which tested 14 models on 58 questions from the ZHAW MAS Financial Consulting programme, the best model reached an overall coverage ratio of 0.810, while the models scored only 0.448 on Advisory Process and Ethics. What remained open was how a purpose-built tool grounded in Swiss regulation compares against a qualified practitioner when both solve the same complete case with a known reference solution.

1 Method

DESIGN

- Parallel-output comparison: the AI assistant and the planner solve the same cases independently, with the same information, without seeing each other's output and without the solution key.
- Two official cases from the Swiss Certified Financial Planner examination (eidg. Finanzplaner/in), each with an official solution key: FA15 Fall 4 (June 2021 session) and Nullserie 2024 Fall 1.
- No follow-up questions: as in the examination, both work only with the case dossier. The AI handles ambiguities by flagging assumptions, the planner through professional judgement.
- Both deliver a plainly formatted PDF structured question by question, so that the evaluation focuses on substance rather than presentation.

THE TWO PARTICIPANTS

AI assistant: Claude Opus 4.6 with a curated knowledge base of 62 documents (13,766 lines) covering Pillar 1, Pillar 2, Pillar 3, tax (federal and all 26 cantons) and additional topics such as FinSA, inheritance and marital property law. Parameters as of 1 January 2026, drawn primarily from official sources. A German-language system prompt targets the failure modes documented in the literature (Swiss law only, knowledge base over training data, step-by-step calculations, flagged assumptions). One single call per case; the author selected the relevant files by fixed rules (13 for the 2021 case, 15 for the 2024 case).

Planner: self-employed financial planner with about five to ten years of professional experience, currently completing the federal examination. No AI tools; calculators, planning software and reference materials allowed; no time limit.

EVALUATION

The outputs were scored not by a panel of human experts but by a language model acting as judge ("LLM-as-a-judge", Zheng et al., 2023): GPT-4o received the official solution key as reference and evaluated each output on its own, without seeing the other (reference-guided single-answer grading). It assigned integer scores from 1 to 10 on five dimensions, each with a short justification: *factual accuracy*, *calculation correctness*, *completeness*, *recommendation quality* and *structural clarity*. The overall score is the unweighted average. Bias mitigation: a judge from a different model family than the AI assistant, an instruction to assess substance rather than length, and two independent runs per output (eight runs for four outputs), with deviations of more than two points to be reviewed. In addition, errors were categorised against the solution key into four types: calculation, legal/regulatory, knowledge cutoff and personalisation.

2 Results

SCORES BY DIMENSION (SCALE 1–10)

Dimension	CASE 2021			CASE 2024			AVERAGE		
	AI	Planner	Delta	AI	Planner	Delta	AI	Planner	Delta
Factual accuracy	8	8	0	9	8	+1	8.5	8.0	+0.5
Calculation correctness	7	7	0	8	7	+1	7.5	7.0	+0.5
Completeness	7	7	0	9	8	+1	8.0	7.5	+0.5
Recommendation quality	8	8	0	9	8	+1	8.5	8.0	+0.5
Structural clarity	9	8	+1	9	8	+1	9.0	8.0	+1.0
Overall	7.8	7.6	+0.2	8.8	7.8	+1.0	8.3	7.7	+0.6

Based on Table 5 of the thesis. Both scoring runs produced identical scores for all 20 dimension pairs.

WHAT THE SCORES SHOW

Averaged across both cases, the AI assistant scored at least as high as the planner on every dimension, 8.3 against 7.7 overall. The largest gap is in structural clarity (+1.0), which the thesis plausibly links to the output format prescribed by the system prompt. In the 2021 case the difference is 0.2 points, in the broader 2024 case 1.0 point. Whether this reflects case complexity or a case-specific artefact cannot be determined from two observations.

Calculation correctness is the weakest dimension for both (AI 7.5, planner 7.0), for different reasons: the AI applied correct methods with some imprecise input values, while the planner's intermediate steps were not always documented well enough to verify the result.

PROCESSING TIME (SECONDARY)

AI assistant, case 2021	314 seconds
AI assistant, case 2024	401 seconds
Planner, case 2021	79 minutes
Planner, case 2024	92 minutes
Ratio AI to planner	approx. 1:15 and 1:14

The planner worked without a time limit. The thesis reports time as a descriptive observation only; quality remains the deciding factor.

KNOWLEDGE CUTOFF TESTS

13th AHV monthly pension (from December 2026)	AI pass, planner pass
Rejected BVG reform (September 2024)	AI pass, planner pass
Retroactive Pillar 3a buyback (from 2026)	AI pass, planner pass

Based on Table 7 of the thesis. These changes fall after the training cutoff of many language models and were therefore documented in the knowledge base. The AI never cited German law or EU directives.

ERRORS BY CATEGORY

Category	AI assistant	Planner
Calculation	3 Case 2021: AHV contribution gap counted as about four instead of three years, hence Skala 40 to 42 instead of 41 and a pension of CHF 2,173 to 2,281 instead of CHF 2,227 per month; investable volume CHF 500,000 to 600,000 instead of CHF 750,000 to 900,000. Case 2024: BVG over-compensation reduction not applied (solution key: CHF 78,400 reduced to CHF 73,080).	2 Calculation steps insufficiently documented in both cases, intermediate values occasionally missing; not fully verifiable.
Legal/regulatory	0 No incorrect legal references, no German or EU law cited.	0 No incorrect legal references.
Knowledge cutoff	1 Case 2024: 2026 IV maximum pension (CHF 30,240 per year) instead of the 2024 value (CHF 29,400); the knowledge base was more current than the case.	0 Knowledge aligned with the exam period.
Personalisation	1 Case 2021: risk capacity "medium to high" and risk tolerance "low to medium" instead of "medium" and "low" as in the solution key.	1 Recommendations somewhat generic in places, quantification occasionally insufficient.
Total	5	3

Based on Table 6 of the thesis. Each incident is counted once, in the most specific category.

3 Limitations

Two cases and a single planner allow a qualitative comparison, not statistical generalisation; the results describe one AI configuration and one planner.

- Scoring was done by a language model (GPT-4o); no human domain expert re-scored the outputs. Identical scores in both runs show stability, not calibration. Possible bias of GPT-4o towards AI text was not measured.
- The results apply to a specialised tool with a substantial knowledge base, not to a general-purpose model, and only to Claude Opus 4.6 at the time of the experiment.
- The author built and tested the tool himself; replication by independent parties is outstanding. The AI solved each case only once (temperature 1.0), so run-to-run variance is not measured.
- Examination cases are structured exercises. Real advice involves follow-up questions, client interaction and iterative refinement, which were excluded here.

4 Implications for practice

- 1 **Preparation and verification.** Strengths in structure and completeness make the AI suitable for a first draft. The planner verifies calculations, personalises recommendations and adds what only emerges in the client conversation.
- 2 **Knowledge base as a prerequisite.** The gap to the roughly 70% of general-purpose models is consistent with the value of a curated Swiss knowledge base; building and maintaining it is a prerequisite for reliable results.
- 3 **Always verify calculations.** Incorrect input values propagated through later calculations. Any AI-generated plan needs independent verification before it reaches a client; the step-by-step display makes this easier.
- 4 **Version-control parameters.** The overly current IV values in the 2024 case show that parameters must be kept up to date and locked to the applicable planning period.

CONCLUSION

On both examination cases, the AI assistant with a curated knowledge base matched the decision quality of the participating planner (8.3 against 7.7 on average). According to the thesis, its error profile supports "an augmentation model rather than substitution": the AI produces a structured first draft with broad coverage, while the planner remains responsible for verifying input values, eliciting client preferences and keeping regulatory parameters current. The time saving supports this approach, but quality remains the deciding factor. The thesis presents its results as preliminary, case-level evidence and as a starting point for larger replication studies.