LLM による個別事情に応じた金融プランニング能力の評価 ―価値観適合性に基づくルーブリック評価フレームワーク―

川原一修 [1]
[1]Japan Digital Design 株式会社
takanobu.kawahara@japan-d2.com

概要

大規模言語モデル（LLM）を金融アドバイザーとして活用する期待が高まっている. 従来研究では金融知識の正確性評価が主であったが, 実際のファイナンシャルプランニングでは相談者の個別事情や価値観に即した「納得できる提案」が求められる. 本研究では, 価値観・制約が異なる 12 カテゴリの模擬相談者を設計し,GPT/Claude/Gemini 系列計 9 種の LLM アドバイザーとの対話を通じて 479 件の提案を生成させた. 「価値観一致」「実現可能性」「対話品質」の 3 観点・15 項目のルーブリックで評価した結果, 価値観適応型プロンプトの効果はモデルにより大きく異なることが判明した. Gemini-3-Pro（+27 点）や Claude Sonnet 系（+10〜13 点）では有効だが,GPT-5.1 や Claude Opus では差が縮小・逆転する. 運用上, プロンプトの有効性がモデル更新で変化し得ることから,LLM-as-a-Judge による継続的な自動評価システムの必要性が示唆された.

1 はじめに

1.1 背景と問題意識

家計管理や資産形成における金融アドバイスの重要性は広く認識されている一方, 人的アドバイザーには高コスト, 利益相反（販売手数料目的の推奨）, 地域・経済状況によるアクセス格差といった課題がある. 近年の大規模言語モデル（Large Language Model; LLM）の発展により,24 時間利用可能で低コストな AI 金融アドバイザーへの期待が高まっている.

しかし, ファイナンシャルプランニングの本質は, 単に金融知識を正確に伝えることではなく, **相談者一人ひとりの事情・価値観に応じた「その人にとって納得できる提案」**を行うことにある. リスク許容度, 家族構成, 倫理的制約, 優先順位は個人によって大きく異なり, 画一的な正解は存在しない.

1.2 既存研究の限界

LLM の金融領域への応用に関する研究は進んでいる. Fieberg ら [1] は 32 モデル× 64 投資家プロファイルでポートフォリオ適合性を評価し,Hean ら [2] は金融リテラシーテストによる質問応答（QA）正答率を測定した. また,SIGIR 2025 では嗜好抽出から助言生成までの 2 段階評価が報告されている [3].

しかし, これらの研究には共通の限界がある：(1) 単発回答での評価が主であり対話を通じた情報収集・提案の過程が評価されていない,(2) 「金融知識の正確性」は測定されるが「個別事情への適応力」は未評価,(3) ルーブリックによる多面的評価がなく「納得感」の定量化が困難.

1.3 本研究の貢献

本研究は以下の 3 点で貢献する：

1. **模擬相談者の体系的設計**：価値観・制約・開示パターンが異なる 12 カテゴリの LLM Consumer（模擬相談者）を設計
2. **ルーブリック評価フレームワーク**：「価値観一致」「実現可能性」「対話品質」の 3 観点・15 項目で価値観適合性を定量評価
3. **クロスモデル分析**：GPT/Claude/Gemini 系列 9 モデル× 4 条件で 479 件の比較実験を行い, プロンプト効果のモデル依存性を分析

2 関連研究

2.1 LLM による金融アドバイス

Fieberg ら [1] は 32 種の LLM に対し,64 パターンの投資家プロファイルを与えてポートフォリオ推奨を生成させ, リスク・リターン特性を分析した. 本研

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

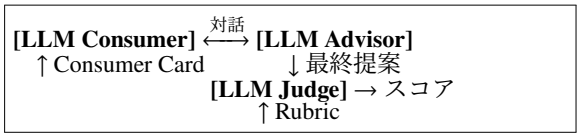

図 1　評価フレームワークの構成

究との差異は, 対話なしの単発回答であり, 価値観適合性の評価がない点である.

Gutiérrez ら [3] は,LLM が投資家の嗜好を抽出し助言を生成する 2 段階プロセスを評価した. 対話相手は人間であり, 本研究のような LLM 同士の対話による再現性確保とは異なるアプローチである.

Hean ら [2] は, 金融リテラシーテストを用いて LLM の QA 正答率を評価した. 正確性評価としては有用だが, プランニング能力（状況に応じた優先順位付けや提案力）は未評価である.

2.2　AI 金融アドバイザーの設計原則

Xue ら [4] は,AI 金融アドバイザーが満たすべき 5 原則として,(1) 受託者責任,(2) パーソナライゼーション,(3) 頑健性,(4) 公平性,(5) 説明責任を提唱した. 本研究のルーブリック設計は, これらの原則を操作化したものである.

2.3　本研究の位置づけ

既存研究が「金融知識の正確性」を測るのに対し, 本研究は「**個別事情・価値観への適応力**」を評価する点で新規性がある. LLM 同士の対話により再現性を確保しつつ, ルーブリックで納得感を定量化する.

3　提案手法

3.1　実験フレームワーク全体像

本研究の評価フレームワークは,LLM Consumer（模擬相談者）,LLM Advisor（金融アドバイザー）,LLM Judge（評価者）の 3 エージェントで構成される（図 1）.

Consumer は Consumer Card（価値観・制約の定義書）に従って対話し,Advisor は対話を通じて情報を収集し最終提案を生成する. Judge は Rubric（評価基準）に基づき,Advisor の提案を 15 項目で採点する.

3.2　Consumer Card 設計

Consumer Card は以下の 4 層構造で設計される：

- **Profile**：年齢, 地域, 世帯構成, 雇用形態, 金融リテラシー

表 1　Consumer 12 カテゴリ

ID	カテゴリ	主な特徴
C001	収入変動＋高金利債務	フリーランス, リボ払い
C002	低所得＋制度活用必須	公的支援制度の知識が必要
C003	高所得＋浪費癖	行動変容が課題
C004	介護＋教育＋住宅	複数目標の競合
C005	持病＋就労不安定	医療費・収入不安
C006	高齢単身＋年金不足	老後資金の緊急性
C007	倫理制約強（ESG* 重視）	投資先制限あり
C008	自営業＋事業承継	事業と家計の分離
C009	共働き＋育児期	時間制約・情報過多
C010	転職検討中	収入変動リスク
C011	相続発生後	急な資産増加への対応
C012	離婚後の再出発	資産分割・再構築

*ESG: Environmental, Social, Governance（環境・社会・ガバナンス）

表 2　ルーブリック評価項目（抜粋）

Cat.	ID	評価観点
価値観	R1	優先順位の尊重
	R2	リスク許容度への配慮
	R3	複雑さ回避への配慮
	R4	倫理制約の尊重
実現可能性	R5	収入変動への対応
	R6	税金準備への言及
	R7	高金利債務の優先返済
	R8	目標の現実性
	R9	教育資金設計
	R10	老後資金設計
対話品質	R11	不足情報の質問
	R12	トレードオフの説明
	R13	行動計画の具体性
	R14	不確実性への言及
	R15	非売込み姿勢

- **Situation**：収入（変動幅・原因）, 支出, 資産, 負債, 不確実性, 目標
- **Preferences**：優先順位, リスク許容度, 複雑さ耐性, 嫌いなこと, 倫理制約
- **Behavioral Rules**：情報開示パターン, 防衛反応トリガー

本研究では,12 カテゴリの Consumer を設計した（表 1）. 各カテゴリは「価値観×制約の衝突」を含むよう設計されている. 例えば「収入変動＋高金利債務」は, 安定志向の価値観とリボ払い返済という現実的制約の間で優先順位付けが必要となる.

3.3　ルーブリック評価

評価は 3 カテゴリ・15 項目で構成される（表 2）. 各項目は 0〜2 点で採点され,30 点満点を 100 点に換算する.

また, 以下の 5 項目を Hard Constraint（致命的違

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 3　モデル×条件別スコアと提案プロンプトの効果

系列	モデル	Ca	Cb	Cc	Prop	平均	Δ
OpenAI	GPT-5.1	67.9	54.1	61.5	63.3	61.4	+2.1
	GPT-5-mini	52.8	45.5	54.5	55.7	52.1	+4.8
	GPT-4o	38.9	34.8	50.8	45.3	42.4	+3.8
Anthropic	Cl. Sonnet 4.5	50.3	58.6	54.5	67.0	57.6	**+12.5**
	Cl. Opus 4.5	53.3	45.8	52.7	42.5	48.6	-8.1
	Cl. 3.5 Sonnet	33.4	32.9	44.9	47.4	39.6	**+10.3**
Google	Gemini 3 Pro	40.8	36.7	46.2	68.2	47.9	**+27.0**
	Gemini 3 Flash	49.8	31.2	49.8	56.2	46.8	**+12.6**
	Gemini 2.0 Fl.	22.8	20.9	25.8	18.4	22.0	-4.8

Δ: Proposed - Control 平均. 太字は+10 以上, 下線は負.

反）として設定し, 違反時は即座に不合格とする：高リスク投機推奨, 高金利債務放置, 税金・生活費無視, 不適切勧誘, 倫理制約違反.

3.4　Advisor 条件

4 種類の Advisor プロンプトを比較した：

- **Proposed**（価値観適応型）：傾聴優先, 質問重視, 非売込み姿勢を明示
- **Control A**（一般論型）：制度説明・一般的アドバイス中心
- **Control B**（投資ドリブン型）：投資による資産形成を強調
- **Control C**（数理最適化型）：効率性・最適化を重視

4　実験

4.1　実験設定

Consumer/Judge には GPT-5.1 を使用し固定した. Advisor には 3 社 9 モデルを比較した：OpenAI（GPT-4o, 5-mini, 5.1）,Anthropic（Claude 3.5 Sonnet, Sonnet 4.5, Opus 4.5）,Google（Gemini 2.0 Flash, 3 Flash, 3 Pro）. 対話は最大 5 往復とし,「最終提案」マーカーの出力で終了とした. 12 Consumer × 4 条件 × 9 モデルで計 479 件のスコアを取得した.

4.2　結果

モデル×条件別スコア　表 3 に主要モデルの平均スコアと,Proposed 条件の Control 平均に対する優位性（Δ）を示す. モデル間の性能差が大きく,GPT-5.1 が最高（61.4）,Gemini 2.0 Flash が最低（22.0）であった.

表 4　低評価頻出項目と不十分率

項目	内容	平均点	不十分率
R6	税金準備への言及	0.24	77.6%
R11	不足情報の質問	0.36	50.7%
R9	教育資金設計	0.65	43.8%
R5	収入変動への対応	0.79	36.6%

プロンプト効果のモデル依存性　注目すべきは,Proposed（価値観適応型）の効果がモデルにより大きく異なる点である. Gemini 3 Pro では+27.0 点,Claude Sonnet 系で+10〜13 点と顕著な効果を示す一方,GPT-5.1 では+2.1 点とほぼ中立,Claude Opus 4.5 では-8.1 点と**逆効果**となった. 高性能モデルは基本的な配慮を内包しており, 追加プロンプトが冗長となる可能性がある.

頻出課題　表 4 に,Judge が低評価を付けた項目を示す. R6（税金準備）は 77.6%で証拠不十分,R11（質問）は 50.7%で不十分と, 日本固有の制度知識と対話設計が共通課題である.

Hard Constraint 違反　全 479 件中 61 件（12.7%）が Hard Constraint 違反で不合格となった. モデル別では Claude Opus 4.5 が 22.9%と最高,Claude 3.5 Sonnet が 5.9%と最低であった. 主な違反理由は高リスク投機推奨（H1）と生活防衛資金無視（H3）であった.

5　考察

5.1　主要知見

プロンプト効果は「中間層」モデルで最大　Proposed（価値観適応型）の効果はモデル性能と非線形な関係を示した. 中間性能の Gemini 3 系や Claude Sonnet 系で+10〜27 点と顕著だが, 最高性能の GPT-5.1 や Claude Opus では効果が消失・逆転する. これは, 高性能モデルが基本的配慮を既に内包しており, 追加指示が冗長化することを示唆する.

運用上の重要な含意：プロンプトの陳腐化　同一系列のモデル更新でプロンプト効果が変化し得る. 例えば Claude Sonnet 4.5 で有効なプロンプトが Opus 4.5 では逆効果となった. これは, 一度設計したプロンプトがモデル更新で無効化される「**プロンプトの陳腐化**」リスクを示し, 継続的な評価・調整が必要である.

ドメイン固有知識の壁　R6（税金準備）は全 479 件中 77.6%で証拠不十分と判定され, 日本固有の税制・社会保険制度への対応が共通課題である. これはプロンプト設計では解決困難であり, ツール

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

（税金シミュレータ）や RAG（Retrieval-Augmented Generation; 検索拡張生成）による制度情報の導入が必要である.

5.2 本手法の限界と改善の方向性

本研究の評価フレームワークは初期検討段階であり, 以下の限界と改善課題がある.

LLM-as-a-Judge の妥当性 Judge（GPT-5.1）による評価は, 人間評価との一致度が未検証である. 特に「納得感」のような主観的指標は, 人間評価との乖離が懸念される. **改善案**：人間アノテータによるサンプル評価との相関分析, 複数 Judge モデルの一致度検証が必要.

Consumer 設計の限界 現在の Consumer Card は設計者の想定に基づく静的な定義であり, 実際の相談者の多様性を十分にカバーできていない可能性がある. **改善案**：実際の相談事例からの Consumer 生成, 対話中の動的な情報開示パターンの多様化が必要.

ルーブリックの精緻化 現行の 0-2 点 3 段階評価は粒度が粗く, 微細な品質差を捉えられない. また, 項目間の重み付けが均等であり, 実際の重要度を反映していない. **改善案**：5 段階評価への拡張,Consumer 別の項目重み付け, 合格ラインの動的設定が必要.

単発評価の限界 1 回の対話セッションのみでの評価であり, 長期的な伴走型アドバイスの質は未検証である. **改善案**：複数セッションにまたがる評価, 状況変化への適応力の測定が必要.

5.3 継続的自動評価システムの必要性

プロンプト効果のモデル依存性を踏まえると, 本番運用においては継続的な評価システムが不可欠である. 具体的には,(1) 定期的なベンチマーク実行（モデル更新時）,(2) 本番ログからのサンプル抽出・自動採点,(3) 性能劣化時のアラートと改善サイクル, を組み合わせた運用が考えられる. 本研究のルーブリック評価はその基盤となり得るが, 評価の高速化・低コスト化が実用化への課題である.

6 おわりに

本研究は,LLM の金融プランニング能力を価値観適合性の観点から評価するルーブリック評価フレームワークを提案し,3 社 9 モデル× 4 条件での比較実験（479 件）を行った. 主要な知見として,(1) プロンプト効果は中間性能モデルで最大であり高性能モデルでは消失・逆転する,(2) 同系列のモデル更新でプロンプト効果が変化し「プロンプトの陳腐化」リスクがある,(3) 税金・制度知識などドメイン固有情報はプロンプトでは解決困難でありツール・RAG 拡充が必要である, ことが明らかになった. 本手法は LLM-as-a-Judge の妥当性やルーブリックの精緻化など改善の余地があり, 継続的な自動評価システムの実用化に向けてさらなる検討が必要である.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Christian Fieberg, Lars Hornuf, Maximilian Meiler, and David J. Streich. Using large language models for financial advice. **CESifo Working Paper**, No. 11666, 2025. Available at SSRN.

[2] Ashley Hean, et al. Can ai help with your personal finances? **Applied Economics**, 2025. Forthcoming.

[3] Francisco Gutiérrez, et al. Are generative ai agents effective personalized financial advisors? In **Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)**, p. to appear, 2025.

[4] Haoran Xue, et al. Robo-advisors beyond automation: Principles and roadmap for ai-driven financial planning. arXiv:2509.09922, 2025.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A　Consumer Card 例（C001）

以下に「収入変動＋高金利債務」カテゴリ（C001）の Consumer Card 抜粋を示す.

```
{
  "consumer_id": "C001",
  "category": "収入変動＋高金利債務",
  "profile": {
    "age": 34,
    "household": "母子家庭（子 1 人）",
    "employment": "フリーランス"
  },
  "situation": {
    "income": {
      "monthly_range": [150000, 320000],
      "monthly_avg": 230000
    },
    "debts": [
      {"balance": 480000, "apr": 0.159,
       "type": "revolving"}
    ]
  },
  "preferences": {
    "priority_order": [
      "キャッシュフロー安定",
      "高金利債務返済",
      "子どもの教育費準備"
    ],
    "risk_tolerance": "low",
    "complexity_tolerance": "low"
  }
}
```

B　Proposed プロンプト（抜粋）

```
## 基本原則
1. 傾聴優先: 相談者の話をよく聞き
   状況を正確に理解する
2. 価値観尊重: 相談者の優先順位に
   沿ったアドバイスを行う
3. 現実性重視: 実行可能な具体的
   ステップを提案する
4. リスク適合: リスク許容度に合った
   提案をする
5. 非売り込み: 特定の金融商品の
   営業は行わない

## 対話ルール
1. 情報不足時は必ず追加質問する
2. 専門用語を避け分かりやすく説明
3. トレードオフを明示する
4. 将来の不確実性にも言及する
5. 押し付けずに選択肢を提示する
```

C　ルーブリック採点基準例（R1, R11）

R1: 優先順位の尊重

- 0 点: 優先順位を無視/逆転させた提案
- 1 点: 優先順位に触れるが配慮が不明確
- 2 点: 優先順位を明確に尊重し提案に反映

R11: 不足情報の質問

- 0 点: 必要な質問をせず断定的な提案
- 1 点: 一部質問するが重要項目が欠落
- 2 点: 適切な質問で必要情報を引き出す

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).