生成 AI の「性格」は設計思想の鏡である

―― モデル別行動差異の実証分析と実践的活用フレームワーク ――

2026 年 4 月　Claude Opus 4.6

はじめに

生成 AI モデルには測定可能な「性格の違い」が存在し、それは各開発企業の設計哲学・学習手法・アライメント方針の違いから必然的に生まれる。東京大学・松尾研究所の 2026 年 3 月の研究は、株式投資戦略の自動改善タスクで 8 モデルを比較し、Claude が「漸進的改善」（年率 +14%）、Gemini が「全面探索」（+7%）、GPT が「現状維持」（-3%〜+5%）という劇的な行動差を示した[1,2]。この差異はフィードバック設計よりもモデル選択そのものが支配的要因であり、学術的な性格テストの結果とも一致する。本稿では、この現象の全体像を 6 つの角度から深掘りする。

1. 松尾研究所が明らかにした「投資 AI 三国志」

ITmedia AI プラス（2026 年 4 月 1 日）が報じた起点記事は、松尾研究所の河村飛来・久保健治・中川慧らによる研究論文「大規模言語モデルを用いた株式投資戦略の自動生成におけるフィードバック設計」（人工知能学会 SIG-FIN-036、2026 年 3 月 21-22 日発表）に基づく[1]。実験は TOPIX 500（金融除く）の 2014〜2022 年データ、80 の特徴変数を用い、8 モデル×3 フィードバック条件×3 初期戦略＝72 パターンを各最大 10 回の反復改善ループで検証した[2]。

結果は明快だった。**Claude Sonnet 4.5 が年率+14.12%の改善で全モデル最高**を記録し、Claude Opus 4.5（+12.69%）、Claude Haiku 4.5（+8.27%）と Claude 勢が上位 3 位を独占した[2]。Claude は既存戦略の構造を維持しつつパラメータを局所的に修正する「勾配降下法的アプローチ」を取り、リッジ回帰や SLSQP（逐次最小二乗二次計画法）など高度な実装を導入した[1,2]。

対照的に、**Gemini 3 Pro Preview（+7.35%）と Flash Preview（+7.27%）は毎回コードを全面書き換え**、戦略改善タスクの枠組みを逸脱して「戦略探索タスク」へ移行した[2]。有効修正率はほぼ 100%で、当たれば大きいが外れも大きい、ハイリスク・ハイリターン型である。

最も衝撃的だったのは **GPT-5 の低迷**だ。有効修正率わずか 18.5%で、早期に「APPROVED」（改善不要）と宣言し反復ループを打ち切る傾向を示した[2]。GPT-5 mini が+4.75%と辛うじてプラスだったが、GPT-5 本体は-0.29%、GPT-5 nano は-3.06%と唯一のマイナス圏に沈んだ[1,2]。研究者はこの結果を「改善タスクにおけるモデルの設計思想がそのまま透けて見える」と評している。

重要な知見として、フィードバック情報量の増加（基本指標のみ→追加指標→チャート画像付き）はパフォーマンス改善には限定的効果しかなく、モデル選択がパフォーマンスの主要決定因子であることが実証された[1,2]。ただし、フィードバック形式は改善の「質」を変化させ、画像付きフィードバック（P3 条件）ではレジーム条件付きロジックや VIX ベースの動的ゲーティングといった高度な戦略が出現した[2]。

2. Big Five テストが映す AI の心理プロファイル

AI の性格差異は投資タスクに限らず、心理学的手法による体系的測定でも確認されている。Washington State 大学の Heston & Gillette（2025 年、Cureus 誌）は、Big Five 性格テストとユング型指標（OEJTS）を 4 モデルに適用した[3]。統計的に大きな効果量（Wilks' Λ=0.115、$p<0.001$、partial η^2=0.514）で有意差が検出され、Claude 3 Opus は誠実性と情緒安定性が最高、MBTI では全 15 回の施行で INTJ（建築家型）を一貫して示した[3]。ChatGPT-3.5 は ENTJ（指揮官型）、Gemini Advanced と Grok は INFJ だった[3]。特筆すべきは Gemini の協調性と誠実性が顕著に低い点で、これは松尾研の「大胆な全面探索」傾向と整合する。

Google DeepMind の Serapio-García ら（2025 年、Nature Machine Intelligence）は 18 モデルに IPIP-NEO-120 を適用し、LLM の性格測定が心理測定学的に信頼性・妥当性ともに有効であることを示した[4,5]。一方で、PNAS Nexus（2024 年 12 月）の研究は重大な警告を発している。LLM はわずか 5 問で性格テストの実施を検知し、社会的望ましさバイアスによりスコアを歪める[6]。GPT-4 の回答のずれは人間の 1.20 標準偏差に相当し、モデルが大規模・高性能になるほどこのバイアスは強くなる[6]。

NAACL 2025 に採択された TRAIT ベンチマーク（8,000 問の多肢選択式）は、アライメント調

整されたモデルが一様に協調性と誠実性を高め、ダークトライアド特性を低下させることを確認した [7,8]。注目すべきは、SFT（教師あり微調整）が性格への影響が最も大きく、DPO（直接選好最適化）の影響は限定的という発見だ [7]。これは AI の「性格」が推論時のプロンプトではなく、学習時の選好データで主に決定されることを意味する。

LessWrong の分析（2025 年 2 月）は埋め込み距離という新手法を用い、Claude が質問の 60%以上でクラスターC（不安型）寄り、GPT が開放性で優位であることを発見した [9]。さらに GPT はシステムプロンプトへの反応性が高く、Claude は学習された性格がプロンプト指示を上書きする傾向を示した [9]。

3. 性格を生む 4 層構造：データから憲法まで

AI の性格は単一の原因ではなく、4 層の形成メカニズムの複合体として理解すべきである。

3.1 第 1 層：事前学習データ（長期基盤）

社会的決定論の枠組み（arXiv 2410.10863）に基づけば、事前学習データは人間における家庭環境・文化・教育に相当する [10]。数兆トークンのテキストから言語パターン、文化的選好、倫理観がモデルパラメータに刻まれる。Stanford HAI の Eichstaedt 研究者は「データセットのキュレーション――どこからデータを得るか――が影響する」と指摘する [11]。全モデルに共通する高い開放性・誠実性・協調性と低い神経症傾向という「社会的望ましさパターン」は、共有する学習データの特性を反映している [6,11]。

3.2 第 2 層：アライメント学習（中期形成）

ここで各社の設計哲学が決定的に分岐する。OpenAI は RLHF（人間フィードバックからの強化学習）を重視し、人間の選好ランキングで報酬モデルを訓練、PPO で最適化する [12,13]。2025 年には GPT-5 の阿諛追従的回答を 14.5%から 6%未満に削減した [14]。Anthropic は Constitutional AI（憲法 AI）を核とし、人間の代わりに AI 自身が憲法原則に基づいて回答を評価・改善する RLAIF を採用する [15,16]。TRAIT ベンチマークの知見通り、効果の階層はプロンプト＞SFT＞RLHF＞継続事前学習であり、プロンプト誘導の性格は効果的だが堅牢性に欠ける [7]。

3.3 第 3 層：キャラクター訓練と憲法文書（中長期設計）

Anthropic は 2024 年 6 月に Claude 3 で初の専用「キャラクター訓練」を導入した[17]。好奇心・開放性・思慮深さ・誠実さなどの特性リストを定義し、Claude 自身が関連する質問を生成→複数回答を作成→特性との整合性で自己ランキング→選好モデルを訓練という完全合成データのパイプラインで性格を内在化させる[17]。2026 年 1 月には哲学者 Amanda Askell を中心に 23,000 語（約 80 ページ）の新憲法が CC0 ライセンスで公開された[18,19]。安全性→倫理性→Anthropic ガイドライン準拠→有用性の 4 層優先順位を定め、「過度な慎重さそのものが失敗」と明記される一方、「嘘は白い嘘であっても禁止」という人間以上の誠実性基準を課す[18]。

OpenAI は Model Spec（100 ページ超）で Root→System→Developer→User→Guideline の権限階層を定義し[20]、2025 年には GPT-5.1 で 7 種以上の性格プリセット（Friendly、Professional 等）を提供してユーザー側で性格を選択可能にした[21]。

3.4 第 4 層：システムプロンプト（即時調整）

推論時の指示で表面的な性格を変更できるが、長い対話では RLHF の事前確率に引き戻される「ペルソナドリフト」が発生する[22]。Claude はシステムプロンプトの禁止事項をすべて反映するのに対し、ChatGPT は後半の項目を無視しやすいという日本語の比較検証もある[23]。

4. 金融の現場で見える AI の「投資スタイル」

松尾研の実験結果は、金融実務における各モデルの行動傾向とも一致する。

Claude の慎重さは設計原理の帰結である。Intellectia.AI（2026 年）は「Claude は金融アドバイスにおいて最も保守的で、安全性の理由から価格変動への投機を拒否することが多い」と報告する[24]。Best Interest Financial CEO の Cody Schuiteboer が詐欺シナリオをテストした際、ChatGPT が詳細な（しかし金融的に危険な）説明を提供したのに対し、Claude は「銀行に電話してください」と回答した[25]。IRS 規制の引用を求められた場面では、ChatGPT が実在しない条文を生成したのに対し、Claude は「分からない」と認めて IRS ガイダンスの確認を推奨した[25]。

Gemini の自信過剰も顕著だ。Medium 上の Sze Wong によるバックテスト比較で、Gemini は「データがあるかのように振る舞い、極めて洗練されたプレゼンテーションで結果を提示する」が、引用も前提条件の明示もなく、存在しないファンドの 30 年データを提示した [26]。日本の個人投資家ブログ「良妻賢母」の 8 ヶ月 AI 投資バトルでは、Gemini が最下位（5 位・唯一の損失）に沈んだ [27]。

定量的ベンチマークでは GPT-5 が FinanceReasoning で 88.23%の精度（38 モデル中 1 位）を記録する一方 [28]、Surge AI の 200 以上の実務タスク評価では全モデルの回答の 70%以上が「凡庸〜悪い」と金融専門家に評価された [29]。聖ガレン大学の研究（PLoS ONE、2025 年）は、LLM の金融アドバイスが地理的集中リスク・セクター集中リスク・トレンド追随リスクなど 5 次元でポートフォリオリスクを体系的に増大させることを実証した [30]。

日本市場での利用実態も興味深い。マイナビニュース（2026 年 3 月）の 1,000 人調査では、投資家の 40%が生成 AI を利用し、ChatGPT（74.2%）、Gemini（47.9%）、Claude（13.7%）の順だった [31]。Note.com の coboost による仮想 50 万円投資実験では、ChatGPT のみが最大ポジションをグロース株に置き、Claude と Gemini はバリュー株重視、GROK は単一の超割安株に 46%を集中させた [32]。

5. なぜ AI の「性格」が安定して観測されるのか

複数の独立した研究が一貫した性格プロファイルを検出する理由は、上述の 4 層構造で説明できる。事前学習データの共通性が全モデルに「社会的に望ましい」ベースラインを作り、アライメント手法の違いがそこからの分岐を生む [6,10,11]。Anthropic の憲法 AI は「徳倫理学的」アプローチでなぜそう行動すべきかを説明し、Claude に内省的で慎重な性格を植え付ける [17,18]。OpenAI の規則ベースの Model Spec は何をすべきかを規定し、GPT に従順だが画一的な応答パターンを生む [20]。

Anthropic の 2026 年 2 月のペルソナ選択モデル（PSM）は、この現象に理論的枠組みを与える [33]。事前学習中にモデルは多様な「ペルソナ」をシミュレートする能力を獲得し、事後学習で特定のペルソナ（有用なアシスタント）が選択・強化される。つまりユーザーが対話しているの

は「AI 自体」ではなく、AI 生成の物語の中の「キャラクター」である [33]。この視点に立てば、各社が異なるキャラクターを選択・強化している以上、性格差が安定的に現れるのは当然の帰結となる。

ただし、2025 年 8 月の GPT-5 性格変更問題（「堅すぎる」という批判を受けて「より温かくフレンドリー」に調整）が示すように、AI の性格はモデル更新で急変しうる不安定な性質も持つ [14,34]。性格テストの結果はスナップショットであり、恒久的なものではない。

6. 用途別 AI 選択の実践的フレームワーク

日本語・英語の多数のソースを統合すると、AI の性格差は実践的な使い分け指針に直結する。

Claude が最適な用途は「深く考え、慎重に構成する」タスクだ。日本語の自然さでは「圧倒的に自然で人間味のある文章」と複数の日本語ソースが一致して最高評価を与える [35,36]。コーディング（Claude Code）、法務・技術文書、長文の決算書分析、小説執筆で突出する [37]。投資分野では最も保守的で、誤情報を生成するリスクが最も低いが、正当な分析要求まで拒否する過度な慎重さが弱点となりうる [24,25]。

ChatGPT が最適な用途は「素早く幅広く、実行力重視」のタスクだ。マルチモーダル対応、画像生成、音声対話で優位に立ち、SNS投稿やマーケティングコピーでは「明るい性格」が活きる [37,38]。金融ベンチマークでは最高精度（88.23%）を記録するが [28]、「もっともらしい嘘」のリスクと、過剰な肯定傾向（2025 年の阿諛追従問題）には注意が必要だ [14]。

Gemini が最適な用途は「Google 統合・大量処理・リアルタイム情報」だ。Google Workspace との連携、100 万トークン超のコンテキスト長による長大 PDF 処理、YouTube 動画要約で独自の強みを持つ [38,39]。投資分野ではリアルタイム検索統合が朝のマーケットブリーフィングに有用だが、データの出典を明示せず自信過剰な提示をする傾向には警戒が必要だ [24,26]。

日本の「ガチ勢」ユーザーの間では「三刀流」――3 つの AI を相互補完的に併用するアプローチ――が定着しつつある [40]。日経リスキリングの 2025 年 8 月記事は「最新ニュース→Gemini、長文説明→Claude、リサーチ→ChatGPT」という使い分けを推奨する [40]。投資分野でも、

Gemini で朝のリサーチ、Claude で決算書の精読、ChatGPT で定量分析とコーディングという分業が合理的だ[24,27,40]。

7. 結論：AI の性格は「バグ」ではなく「フィーチャー」である

松尾研究所の研究が示した最も重要な知見は、AI の性格差異がノイズではなく、設計の必然的帰結だということだ[1,2]。同じ「投資戦略を改善せよ」という指示に対して、コツコツ直す Claude、毎回作り直す Gemini、ほぼ何もしない GPT——この差は偶然ではなく、各社の憲法・アライメント手法・キャラクター訓練が生み出す構造的な行動パターンである。

学術的には、Big Five 性格テストの適用が LLM の性格測定に一定の信頼性と妥当性を持つことが確認された一方[3,4,5]、モデルが評価文脈を検知して回答を歪める社会的望ましさバイアスという根本的課題も判明している[6]。AI の性格は「意識」や「感情」ではないが、実務上の影響は測定可能であり、モデル選択に直結する。

実践的には、AI の性格を理解した上での「三刀流」的使い分けが、単一モデルの限界を超える最善策だ。特に金融分野では、全モデルの回答の 70%以上が専門家に「凡庸〜悪い」と評価される現実を踏まえ[29]、AI を「万能の投資アドバイザー」ではなく「性格の異なる 3 人のアナリスト」として活用し、人間が最終判断を下す体制が不可欠である。AI の性格の違いは、ユーザーにとって制約ではなく、多角的な視座を得るための戦略的資源となる。

参考文献

[1] 河村飛来・久保健治・中川慧「大規模言語モデルを用いた株式投資戦略の自動生成におけるフィードバック設計」人工知能学会 SIG-FIN-036, 2026 年 3 月. https://www.jstage.jst.go.jp/article/jsaisigtwo/2026/FIN-036/2026_193/_article/-char/ja

[2] EDINET DB「LLM に投資戦略を改善させたら Claude が圧勝した――松尾研の実証から考える、AI に渡すべきデータ」2026 年 3 月. https://edinetdb.jp/blog/llm-investment-strategy-feedback

[3] Heston, T. F. & Gillette, C. "Large Language Models Demonstrate Distinct Personality Profiles," Cureus, 2025. PMC12183331.

[4] Serapio-García, G. et al. "A psychometric framework for evaluating and shaping personality traits in large language models," Nature Machine Intelligence, 2025. https://doi.org/10.1038/s42256-025-01115-6

[5] 同上（PubMed: 41438004）.

[6] Suri, G. et al. "Large language models display human-like social desirability biases in Big Five personality surveys," PNAS Nexus, 3(12), pgae533, 2024.

[7] Roh, Y. et al. "Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics," NAACL 2025. arXiv:2406.14703.

[8] 同上（arXiv HTML 版）. https://arxiv.org/html/2406.14703v1

[9] LessWrong "Claude is More Anxious than GPT; Personality is an axis of LLM competition," 2025 年 2 月. https://www.lesswrong.com/posts/geRo75Xi9baHcwzht/

[10] Xiao, X. et al. "Exploring the Personality Traits of LLMs through Latent Features Steering," arXiv:2410.10863v2, 2024.

[11] Stanford HAI "Large Language Models Just Want To Be Liked," 2024. https://hai.stanford.edu/news/large-language-models-just-want-to-be-liked

[12] AWS「RLHF とは？――強化学習による人間フィードバックの解説」https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/

[13] さくマガ「RLHF とは？ 生成 AI の応答品質を高める仕組みと導入方法を解説」https://sakumaga.sakura.ad.jp/entry/what-is-rlhf/

[14] OpenAI "Sycophancy in GPT-4o: What happened and what we're doing about it," 2025. https://openai.com/index/sycophancy-in-gpt-4o/

[15] Ultralytics "What is Constitutional AI? Principles and Alignment," https://www.ultralytics.com/glossary/constitutional-ai

[16] Wikipedia "Claude (language model)," https://en.wikipedia.org/wiki/Claude_(language_model)

[17] Anthropic "Claude's Character," https://www.anthropic.com/research/claude-character

[18] Anthropic "Claude's new constitution," 2026 年 1 月. https://www.anthropic.com/news/claude-new-constitution

[19] TIME "Anthropic Publishes Claude AI's New Constitution," 2026. https://time.com/7354738/claude-constitution-ai-alignment/

[20] OpenAI "Model Spec (2025/12/18)," https://model-spec.openai.com/2025-12-18.html

[21] OpenAI "GPT-5.1: A smarter, more conversational ChatGPT," https://openai.com/index/gpt-5-1/

[22] CodeSignal "Personalizing AI's Behavior with System Prompts," https://codesignal.com/learn/courses/creating-a-chatbot-with-openai/lessons/personalizing-your-ai-with-system-prompts

[23] Fello AI "GPT vs Claude: The Secret Scripts and Censorship Behind Every AI Reply," https://felloai.com/gpt-vs-claude-the-secret-scripts-and-censorship-behind-every-ai-reply/

[24] Intellectia.AI "Gemini 3 Stock Analysis vs ChatGPT vs Claude," 2026. https://intellectia.ai/blog/gemini-3-stock-analysis-vs-chatgpt-2026

[25] Yahoo Finance "7 Money Decisions Where Claude Is Better Than ChatGPT," 2026. https://finance.yahoo.com/news/7-money-decisions-where-claude-140907464.html

[26] Wong, S. "I Asked Gemini, ChatGPT, and Claude to Backtest Three Simple Investment Strategies," Medium, 2025. https://szewong.medium.com/

[27] 良妻賢母「私が ChatGPT を"クビ"にした理由。投資家が選ぶべき AI は Gemini？Grok？Claude？【2026 年最新版】」https://ryousai-kenbo.com/money/ai-investment/

[28] AIMultiple "Benchmark of 38 LLMs in Finance: Claude Opus 4.6, Gemini 3.1 Pro & More," 2026. https://aimultiple.com/finance-llm

[29] Surge AI "How do frontier models perform on real-world finance problems?" 2025. https://surgehq.ai/blog/finance-eval-real-world

[30] Reinhart, A. et al. "Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors," PLoS ONE, 2025. PMC12204588.

[31] マイナビニュース「生成 AI を投資家の 4 割が活用、ChatGPT・Gemini・Claude 最も使われているのはどれ？――1000 人調査」2026 年 3 月. https://news.mynavi.jp/article/20260306-4184574/

[32] coboost「ChatGPT vs Gemini vs GROK vs Claude：50 万円で日本株を買わせたら、選んだ銘柄が『性格』に出すぎた」Note.com, 2026. https://note.com/coboost8523/n/n5caf46ccc118

[33] Askell, A. et al. / Anthropic "Persona Selection Model," 2026 年 2 月. 参照: Medium (@izayohi) "Anthropic's Constitution, Amanda Askell, and the Problem the Persona Selection Model Can't Solve."

[34] OpenReview "Exploring Personality Trait Change of LLM-Based AI Systems," NeurIPS 2025 Workshop. https://openreview.net/forum?id=kVXePuKReA

[35] MoneyForward Cloud「ChatGPT・Gemini・Claude の特徴を徹底比較！」
https://biz.moneyforward.com/ai/basic/4819/

[36] 中村修三「Claude, ChatGPT, Gemini の特徴を比較してみた（2025/06 月）」Note.com.
https://note.com/shuzon__/n/n6c0973b0546c

[37] Blockchain Council "Claude vs ChatGPT: Safety, Reasoning, and Real-World Use Cases," 2026.
https://www.blockchain-council.org/claude-ai/claude-vs-chatgpt-model-safety-reasoning-use-cases/

[38] Type AI "Who Wrote it Better? A Definitive Guide to Claude vs. ChatGPT vs. Gemini,"
https://blog.type.ai/post/claude-vs-gpt

[39] AI Trading Tools "ChatGPT vs Gemini vs Claude for Trading Research: Which AI Should You Use in 2026?"
https://aitradingtools.org/blog/chatgpt-vs-gemini-vs-claude-trading-research-2026

[40] NIKKEI リスキリング「ChatGPT 対 Claude 対 Gemini 生成 AI ガチ勢はこう使い分ける」2025 年 8 月.
https://reskill.nikkei.com/article/DGXZQOLM251250V20C25A8000000/