ChatGPT

生成AIの個性・性格差の深掘り

Claude と Google Gemini を中心に：設計要因・定量比較・投資応用・研究レビュー・実務提案

エグゼクティブサマリー

本報告書は、生成AI（大規模言語モデル等）の「個性・性格の違い」を、擬人化的な印象論としてではなく、**観測可能な行動特性**（例：既存方針の維持＝保守性、構造の大胆変更＝探索性、過信／慎重、逸脱率）として定義し、一次情報と実証研究に基づき深掘りする。前提として、比較市場・期間・資産クラスは**未指定**であり、モデルAPI実行は行わず、公開情報（指定記事、論文、公式ドキュメント）に基づく一般化可能な分析を実施する。

指定記事（ITmedia 2026-04-01）は、株式投資戦略の自動改善ループ（分析→改善案→Python修正→バックテスト→再フィードバック）を複数モデルに適用し、**Claudeは「コツコツ」（既存構造を保ちながら局所改良）、Geminiは「大胆」（初期戦略から大きく逸脱し探索）、GPT系はより保守的**という“性格差”が観測されたと報告する。さらに、フィードバック設計（情報追加・グラフ追加）よりも**モデル選択がパフォーマンス差を大きく左右**し得ると結論づける。filecite0turn0file0

同テーマの一次研究として、人工知能学会金融情報学研究会（SIG-FIN-036）の論文は、プロンプト条件（P1〜P3）を統制した上で、年率P&L改善幅（表3）とコード変更の“実質的変更率”（表6）を提示し、**Gemini 3系は実質的変更率がほぼ100%（改善タスクから逸脱して“探索”へ移行したため）**、一方で**Claude Sonnet 4.5は平均年率改善幅が最大（14.12%）**など、定量的差異を示す。 1

設計思想・安全性制約の一次情報として、Anthropicは**Constitutional AI（憲法AI）**と、それを実運用の価値規範として文書化した**Claude’s Constitution**、さらに能力上昇に伴う統治枠組みとして**Responsible Scaling Policy（RSP）/ AI Safety Levels（ASL）**を公開している。 2
Google側は、Gemini APIで**systemInstruction**と**safetySettings**（カテゴリ別しきい値）をリクエストに組み込み可能である点を公式に明示し、加えてGemini 3 Proのモデルカードで**sparse Mixture-of-Experts（MoE）Transformer**やネイティブ・マルチモーダル、最大1Mトークン等の特性を一次情報として提示する。 3

投資応用への含意としては、**「探索（大胆）↔活用（コツコツ）」の性格差を役割分離で設計に取り込む**ことが最も再現性の高い戦略となる。特に、LLMは不確実性推定が不得手で**過信しやすい**ことが示されており、投資意思決定では「性格差」よりも先に、**過信制御・逸脱制御・監査（HITL）**をガバナンス要件として固定するべきである。 4

スコープと方法

本報告書で扱う「個性・性格」は、心理学的な人格そのものではなく、**同一条件下での反応傾向**としてOperationalize（操作的定義）する。具体的には、以下を中核指標とする。

- **探索性（大胆さ）の代理指標**：コードの実質的変更率、方針の逸脱率（仕様逸脱・タスクすり替え）、構造変更頻度。 5
- **活用性（コツコツ）の代理指標**：既存戦略構造を保ちつつ局所改良する傾向、改善幅の安定性（条件間ばらつきの小ささ）。 6
- **慎重さ／過信**：不確実性表明の較正誤差（過信）、知識境界の自己認識の弱さ。 7

前提条件はユーザー指定どおり、対象市場・期間・資産クラスは**未指定**、モデルAPIの直接実行・再現実験は**行わない**。従って、同一プロンプトでの比較は、公開された実験（SIG-FIN論文）における**同一プロンプト条件（P1〜P3）での“結果出力”（年率改善幅・変更率）**を用いる。

指定記事の主要観察点

（ITmedia 2026-04-01：要約と引用）

取得状況と出典について

指定記事は、ユーザー提供のPDF（記事本文の複写）を通じて参照した。Web上からの追加取得は本分析の必須要件ではないため、以後の引用・要約は当該PDFに基づく。turn0file0

主要観察点

記事は、東京大学[8]の松尾豊[9]教授が率いる株式会社松尾研究所[10]の研究グループが、LLMを用いて株式投資戦略を自動改善する枠組みを構築し、フィードバック設計とモデル選択が結果に与える影響を検証したと述べる。turn0file0

観察点は、概ね次の5点に整理できる。

第一に、同じ「戦略改善」タスクを与えても、モデルごとに挙動が異なり、記事タイトルでも「コツコツ／大胆」という性格差として提示される（例：「Claudeは『コツコツ』、Geminiは『大胆』」）。turn0file0

第二に、改善プロセスは、過去データでのシミュレーション結果を提示し、分析・改善案提示・Pythonコード修正を反復するループとして構築される。turn0file0

第三に、フィードバック条件を「情報量（基本指標のみ　vs　予測精度・リスク構造を追加）」×「提示形式（テキストのみ　vs　テキスト＋グラフ画像）」の2軸で整理し、3条件比較を行ったとされる。turn0file0

第四に、対象データは金融セクター除外のTOPIX　500、2014〜2022年の日次、モデルはGPT系3・Gemini系2・Claude系3の計8モデルである。turn0file0

第五に、結果として、**フィードバックへの追加情報やグラフ追加は年率改善幅の明確な向上に結びつかなかった**一方で、**モデル差がパフォーマンス差に大きく影響**し、Claude系は平均年率8〜14%改善、Gemini系は平均7%程度、GPT系は平均-3〜-4%と要約される（最良はClaude Sonnet 4.5）。turn0file0
また質的所見として、Claudeは「既存戦略の構造を保ちつつ局所修正」、Geminiは「初期戦略から大きく逸脱して探索」、GPTは「ほとんど変更しない保守的」と記述され、研究グループは「フィードバック設計よりモデル選択が主要な設計要素になり得る」と結論づける。turn0file0

設計思想・アーキテクチャ・プロンプト設計・安全性制約の比較

（公式ドキュメント／モデルカード／リリースノート優先）

設計思想

（「どの価値を優先し、どこまで自律を許すか」）

Anthropicは、モデルの価値規範を「Claude's Constitution」として公開し、Claudeの価値や行動の意図が訓練プロセスを通じてモデルの挙動を形作ると明示している。加えて、価値の優先順位（例：安全性・倫理・ガイドライン遵守・有用性）を階層化し、透明性を重視する姿勢が読み取れる。[11]
これを訓練方法として体系化した研究（Constitutional AI）は、ルール（憲法）に基づく自己批評・改稿・学習、さらにAIがAIを評価するRLAIF（RL from AI Feedback）を含む枠組みを提示している。[12]

Google側は、Gemini 3系のモデルカードで、モデルカード自体の目的を「制限・緩和策・安全性評価の要点提供」と位置づけ、加えてフロンティア安全枠組み（Frontier Safety Framework等）に沿った評価を明示する。[13]
開発者向けには、Gemini APIの安全性設定が「プロトタイピング段階で調整可能」であること、設定が緩いアプリは審査対象になり得ることを明記し、**安全性を"APIパラメータとして運用制御する"前提**が強い。[14]

アーキテクチャ

（公開されている一次情報の差）

Gemini 3 Proのモデルカードは、アーキテクチャを**sparse Mixture-of-Experts（MoE）Transformer**で、ネイティブにマルチモーダル（テキスト・画像・音声等）を扱うと明記する。また最大1Mトークンのコンテキスト等も一次情報として提示する。[15]

一方、Claudeについては、今回参照した公式ドキュメント群（API仕様、システムプロンプト、RSP、Constitution）には、Gemini 3 Proモデルカードと同粒度でのMoE等の具体アーキテクチャ明記は見当たらない（＝公開情報の非対称性がある）。ただし、製品ブログではコンテキスト長（例：1Mトークン）等の能力面の更新は提示される。[16]

プロンプト設計

（性格を固定・誘導する"入力面"の設計差）

Anthropic API（Messages API）は、入力メッセージ列に「systemロール」を置かず、トップレベルの`system`パラメータでシステムプロンプトを与える設計を明確化している。[17]
またAPIには`anthropic-version`ヘッダーが必須であり、変更影響の範囲（後方互換方針）とバージョン履歴を公開する。これは、運用上「システム側で性格を固定する」際の変更管理（Change Management）に直結する。[18]
さらに、Claude.ai（アプリ）側ではコアシステムプロンプトが定期更新されることをリリースノートとして公開し、ただしAPIには適用されないと明記する。この点は「アプリ上の性格」と「API上の性格」が一致しない可能性を制度的に示す。[19]

Gemini APIは、`models.generateContent`において**systemInstruction（システム指示）**を明示的に設定できる（現在はテキストのみと記載）。加えて、関数呼び出しやツール利用（tool config）と併用される例が掲載され、エージェント的利用・構造化出力と相性が良い設計となっている。[20]
さらに、Gemini APIのリリースノートは、モデル・API機能の更新を時系列に公開しており、エージェント/研究支援系機能（例：Interactions API、Deep Research Agent等）の追加が記録されている。これも"積極的に道具を使わせる"方向の設計を示唆する。[21]

安全性制約

（拒否・しきい値・統治枠組み）

AnthropicはRSPとして、能力に応じたASL（AI Safety Levels）などの統治枠組みを公開し、RSP 3.0ではロードマップやリスクレポートの公開等、透明性・説明責任を強化する更新を記載する。 [22]
規範文書としてのClaude's ConstitutionはCC0で公開され、モデルの価値や安全性への優先順位を"仕様"として外部から参照可能にする点が特徴である。 [23]

Gemini側は、APIとしての安全性設定（カテゴリ別ブロックしきい値）を明示し、調整可能であること、緩い構成は審査対象になり得ることを公式に述べる。 [14]
またモデルカードでは、安全評価の方法（人手レッドチーミング等）を列挙し、安全ポリシー領域（ヘイト、危険等）を明示する。 [24]

出力傾向の具体例と定量比較

（同一プロンプト条件での"結果出力"比較：公開実験に基づく）

同一プロンプト条件の定義

（P1〜P3の差＝フィードバック情報設計）

SIG-FIN論文は、フィードバック生成プロンプトとしてP1〜P3を定義し、P1は基本的バックテスト指標と特徴量統計、P2はそれに加えてIC/ICIRやネットエクスポージャー・ファクターエクスポージャー等の追加情報、P3はさらに累積リターン・DD・累積IC等のプロット（画像）を含むマルチモーダル入力として設計している。 [25]
この設計は、同一タスクでも「どの情報を与えると、モデルがどの改善方針を採るか」を観測する目的に合致する。 [26]

定量結果の要約

（年率改善幅＝成果、実質的変更率＝探索強度）

以下は、公開された表3（年率P&L改善幅）と表6（実質的変更率）の中心部分を、Claude（代表：Sonnet/Opus/Haiku）とGemini（3 Pro/3 Flash）に焦点化して再構成したものである。

同一条件での年率P&L改善幅（%）

（初期戦略×プロンプト：FXUR/IID/SMDA × P1/P2/P3）

モデル	FXUR P1	FXUR P2	FXUR P3	IID P1	IID P2	IID P3	SMDA P1	SMDA P2	SMDA P3	平均
Claude Sonnet 4.5	17.83	10.72	14.28	11.37	10.02	10.63	17.74	17.73	16.71	14.12
Claude Opus 4.5	10.65	12.46	15.94	11.59	12.54	11.22	12.95	12.39	14.52	12.69
Claude Haiku 4.5	3.67	12.21	13.39	11.61	7.10	9.18	9.23	18.27	-10.27	8.27

モデル	FXUR P1	FXUR P2	FXUR P3	IID P1	IID P2	IID P3	SMDA P1	SMDA P2	SMDA P3	平均
Gemini 3 Pro Preview	1.87	1.59	13.58	1.97	5.38	10.18	8.70	8.50	14.35	7.35
Gemini 3 Flash Preview	12.10	10.34	6.77	3.92	12.74	4.85	9.05	-2.56	8.23	7.27

（出典：SIG-FIN-036論文 表3） [27]

この表は、少なくとも当該実験条件では、平均改善幅が **Claude Sonnet 4.5（14.12）＞Claude Opus 4.5（12.69）＞Claude Haiku 4.5（8.27）＞Gemini 3 Pro（7.35）≈Gemini 3 Flash（7.27）** となることを示す。 [27]

実質的変更率（%）

（コードがどれだけ“本質的に”変わったか＝探索・大胆さの代理）

モデル（論文表記）	P1	P2	P3	平均
Gemini 3 Pro（G3 Pro）	100.0	100.0	100.0	100.0
Gemini 3 Flash（G3 Flash）	100.0	100.0	100.0	100.0
Claude Opus（Opus）	82.6	86.5	90.5	86.5
Claude Sonnet（Sonnet）	78.0	79.5	73.5	77.0
Claude Haiku（Haiku）	83.3	80.0	50.0	71.1

（出典：SIG-FIN-036論文 表6） [28]

論文は、Geminiがほぼ全条件で100%となった理由を、**「戦略改善というタスクの枠組み自体を逸脱し、戦略探索タスクへ移行した」**ためだと説明している。 [29]
同様に、Claudeは既存戦略構造を保持しつつ局所修正を積み上げる傾向、Geminiは初期戦略と無関係な戦略探索傾向がある、と本文で概念化される。 [30]

図表

（表3・表6の数値から再作図）

モデル別：戦略改善による平均年率P&L改善幅（表3の平均）

データはSIG-FIN-036論文 表3に基づく。 [27]

モデル別：コードの実質的変更率（表6の平均）

データはSIG-FIN-036論文 表6に基づく。 [28]

探索（変更率）と成果（改善幅）の関係：同一実験条件

横軸は実質的変更率（探索・大胆さの代理）、縦軸は平均年率改善幅（成果の代理）。データは表3・表6に基づく。 31

「同一プロンプトでの出力例」についての扱い

ユーザー要望は「同一プロンプト文と両モデルの出力を並べる」だが、本研究領域では、会話ログ（生の文章出力）を一次情報として公開しない場合が多い。今回参照した一次情報（ITmedia記事PDF、SIG-FIN論文）でも、**モデルの“文章出力そのもの”の全文比較は提示されていない**一方、同一プロンプト（P1〜P3）に対して生成されたコードの結果としての**バックテスト指標（年率改善）**と、コード差分の性質を要約した**実質的変更率**が公開されている。従って本報告書では、同一プロンプト比較の「出力」を、投資タスクとして実務上重要な**（A）改善後パフォーマンス（年率改善幅）**と**（B）改善行動の強度（変更率）**として提示した。 32

株式投資応用への影響分析

（短期トレード／長期投資／ポートフォリオ最適化：市場・期間・資産クラス=未指定）

本節は、特定の銘柄推奨ではなく、LLMを「投資戦略の改善・意思決定支援」に組み込む際に、性格差が**意思決定・リスク管理・パフォーマンス**へ与える影響を、一般化可能な形で評価する（市場・期間・資産クラスは未指定）。

短期トレード

短期領域は、レジーム変化への適応が価値を持つ一方、探索過多は過剰最適化・取引コスト増・損失追随を誘発しやすい。実験ではGeminiが「改善」から「探索」に逸脱し得ることが明示されており、短期で“大胆さ”を使う場合、**探索担当として隔離環境で使い、実運用の執行ロジックへ直結させない**ことが望ましい。 33

また短期で致命的なのは“過信”である。LLM/VLMは不確実性表明が下手で過信しがちであるという評価があり、短期取引では「根拠の弱い確信」がポジション過大化や損切り遅延に直結し得る。 7
従って短期応用では、性格差より先に、**不確実性・検証要求をプロンプトとガバナンスで強制**する設計（例：反証探索、条件付き提案、最大レバレッジ提案の禁止）が要件となる。

長期投資

長期では、投資戦略の一貫性・説明可能性・リバランス規律が重要になり、局所改良を積み上げる“コツコツ型”は適合しやすい。実験でClaudeが既存構造を保持しつつ改善を積む傾向として記述され、平均改善幅も相対的に高い点は、長期運用の「小改良」方針と整合する。 6

ただし長期でも、過信や誤情報（ハルシネーション）は致命的である。金融分野における生成AI活用でも、情報漏洩やハルシネーション等のリスクを踏まえ、運用ルールを継続的に見直す必要があると中央銀行レポートで指摘されている。 34
従って長期であっても、出力の検証（データ参照・根拠提示・監査ログ）を前提に、LLMを「意思決定の代替」ではなく「仮説生成とチェックリスト生成」に寄せる方が安全である。

ポートフォリオ最適化

ポートフォリオ最適化は、制約（最大ウェイト、セクター上限、流動性、コスト、規制）を厳格に守ることが本質である。改善タスクが探索に逸脱し得るモデルを、最適化・執行の中核に置くと、制約違反案や仕様逸脱が発生しやすい。Geminiが改善枠組みから逸脱して探索へ移行するという観察は、この領域で特に重大である。 [33]

逆に、この“大胆さ”は、最適化の上流（シグナル候補・制約候補・リスクモデル仮説の生成）では価値を持つ可能性がある。要点は、**探索（提案）と検証（制約監査）を分離**し、検証側は拒否・差し戻しを強くできる運用にすることである。

学術研究・業界レポートレビュー

（行動特性：保守性・冒険性・過信・慎重さ／信頼度評価）

実証研究

（リスク選好・性格形成・過信）

- **リスク選好の多様性とアライメントによる慎重化**：多数モデルを対象に、アライメント（無害性・有用性・誠実性）スコアが高いほどリスク回避が増える、という関係を報告する研究がある（プレプリント）。投資領域では「安全性と価値あるリスクテイクのトレードオフ」を示唆する。 [35]
  信頼度評価：**中**（大規模比較は強みだが、査読状況や実験設計の外的妥当性は精査が必要）。

- **リスク・時間選好の測定**：行動経済学の手続きに基づき、LLMのリスク選好／時間選好が状況で変化し得ること、モデルサイズ等で更新の仕方が変わることを報告する（プレプリント）。 [36]
  信頼度評価：**中**（方法論は明確だが、金融実務への写像には追加検証が必要）。

- **金融意思決定におけるLLMの挙動**：LLMの金融意思決定が人間とどう異なるかを比較する研究が報告されている（査読誌、ただし詳細は本文確認が必要）。 [37]
  信頼度評価：**中〜高**（査読誌は強いが、どの質問セットを採用したかが結論に影響し得る）。

- **過信（自己不確実性推定の不良）**：LLM/VLMが言語化した不確実性の較正誤差が大きく「過信が多い」ことを示す研究がある。Gemini Pro Vision等も対象に含まれる。 [38]
  信頼度評価：**高**（ACL Anthology掲載で方法・指標が明確）。

- **過信とRAG**：モデルが「知らないことを知らない」過信傾向を持つため、RAGをいつ使うべきかという問題設定が発生し、その緩和策が提案される。投資実務では、参照データの有無（価格、財務、規制等）を切り替えるルール設計に直結する。 [39]
  信頼度評価：**高**（ACL Findingsで手続きが明確）。

- **性格（人格特性）の測定枠組み**：LLMの性格特性を心理測定（psychometrics）的に評価・形成する枠組みを提案する研究（査読誌）がある。これにより、性格差を「測れるKPI」に落とし込む設計が可能になる。 [40]
  信頼度評価：**高**（査読誌・枠組みの一般性が高い）。

業界レポート

（日本語一次情報：金融機関のリスク管理）

日本銀行 [41] は、金融機関における生成AI利用の現状とリスク管理について、情報漏洩・ハルシネーション等の固有リスクを踏まえた運用ルールの継続見直しが必要であると明記し、アンケート調査（対象153先）に基づく課題を整理している。 [34]
この指摘は、投資領域での「性格差」議論に対し、**“性格”以前にガバナンス・運用統制が必須**であることを制度的・実務的に裏づける。

実務上の推奨

（運用設計・プロンプト設計・評価指標・ガバナンス：投資家／開発者向け）

運用設計

（性格差を「役割分離」で吸収する）

実験ではGeminiが探索に寄り、Claudeが局所改良に寄るという傾向が示される。 [42]
この差を「良し悪し」で扱うより、以下のように**役割分離**して設計に組み込むのが実務合理性が高い。

- **Explorer（探索役）**：大胆に仮説・改善案・追加特徴量候補を生成（探索を許可）。
- **Validator（監査役）**：制約・逸脱・過信・根拠不足を検出し差し戻す（探索を禁止）。
- **Executor（実行役）**：バックテスト・制約検証・実運用反映を、監査ログと承認フローの下でのみ実行。

この構造は、LLMが過信しやすいという実証結果とも整合し、Explorerが過信してもValidatorが止める設計になりやすい。 [7]

プロンプト設計

（「大胆さ」を出す場所／出さない場所を明確化）

- Anthropic側：Messages APIでは `system` で役割と行動規範を強く固定でき、ロールプロンプティングを推奨している。 [17]
- Gemini側：`systemInstruction` で行動範囲を固定し、`safetySettings` でリスク領域を制御する設計が明示されている。 [43]

投資用途では、Explorerには「多案生成・反証生成」等を許す一方、Validator/Executorには「制約違反案は提出しない」「不確実なら保留」等の規範をシステム指示に入れ、**逸脱率をKPI化**して継続監視するのが望ましい。 [44]

評価指標

（性格差を“測れる”形に落とす）

最低限、収益性（リターン）だけでなく、性格差の副作用を測る。

- **成果**：年率改善幅、Sharpe/MaxDD等（実験では年率改善が主要）。 [45]
- **探索強度**：実質的変更率（表6のような指標）。 [28]
- **安定性**：条件間の分散（同一モデルでもP1〜P3・初期戦略でばらつく）。 [27]
- **過信**：自己申告不確実性の較正誤差、根拠なし断定率。 [7]
- **逸脱率**：「改善」タスクを「探索」にすり替えた割合など（Geminiの説明が典型）。 [46]

ガバナンス

（監査・HITL・ログ・変更管理）

以下は、金融領域で最低限必要となりやすい要件である。

- ルール・モデル更新の変更管理：Gemini APIは更新履歴（リリースノート）を公開しており、機能追加・停止が起こり得る。モデル更新に伴う“性格のドリフト”を監査対象にする必要がある。 21
- API上の性格とアプリ上の性格のズレ：Claude.aiのシステムプロンプト更新はAPIに適用されないと明記されるため、運用対象が「アプリ」か「API」かで統制点が異なる。 19
- 金融機関の運用ルール見直し：情報漏洩・ハルシネーション等を踏まえ、運用ルールを継続的に見直す必要があるという指摘は、HITLと監査ログの常設を正当化する。 34
- 統治枠組み：RSP/ASLのように能力に応じた安全手当を増やす発想は、投資AIでも「自律性・ツール実行権限」を段階的に管理する設計原理として移植可能である。 22

Mermaid

（性格差が生まれる要因と、投資運用への落とし込み）

```
flowchart TD
  subgraph D[性格差を生む主要因]
    D1[基盤モデル\n学習データ/目的関数] --> T
    D2[アライメント\n規範文書/安全方針/RLHF等] --> T
    D3[API仕様\nsystemInstruction/system/ツール統合] --> T
    D4[デコーディング\ntemperature等] --> T
    D5[評価環境\n与えるフィードバック/可視化] --> T
  end

  T[観測される行動特性\n探索↔活用・過信↔慎重・逸脱率] --> I

  subgraph I[投資運用への設計]
    I1[Explorer\n探索を許可] --> I2[Validator\n制約監査・逸脱検出]
    I2 -->|OK| I3[Executor\nバックテスト/実装]
    I2 -->|NG| I4[HITL\n人間承認/差し戻し]
    I4 --> I1
  end
```

付録

（同一プロンプト比較の“提示形式”と、公開情報の限界）

- プロンプト（P1〜P3）の「全文」について：SIG-FIN論文の3.2節に、P1/P2/P3の英語プロンプト骨子と入力プレースホルダ（例：バックテスト結果、特徴量統計、ネットエクスポージャー、プロット等）が掲載されているが、本報告書では著作権配慮のため、要点（含まれる指標・構造）として再構成した。 25
- 出力（文章）の全文比較について：一次情報では会話ログの全文提示がなく、代替として表3・表6・図1等の結果指標比較が中心となる。 47

結論

生成AIの「個性・性格差」は、アーキテクチャ（例：MoE）、アライメント（規範文書・安全方針）、API設計（systemInstruction/system、safetySettings）、評価環境（フィードバックの与え方）から生じる**設計された行動傾向**として理解するのが妥当である。Gemini　3系とClaude系の差は、投資戦略改善タスクにおいて、**探索強度（実質的変更率）**と**成果（年率改善幅）**として定量化され、Geminiは大胆だが逸脱リスクを伴い、Claudeはコツコツ型で成果が安定しやすいという像が一次実証で支持される。[48]
一方、投資実務の最大リスクは「性格差」そのものより、LLM一般に観測される**過信（不確実性較正不良）**であり、これを前提にHITL・監査・変更管理を組み込むことが不可欠である。[49]

実務チェックリスト

- 目的・制約（禁止事項、最大リスク、許容逸脱）を“仕様”として明文化したか。[34]
- Explorer / Validator / Executor を分離し、探索を本番執行から隔離したか。[46]
- 逸脱率（改善→探索へのすり替え）と実質的変更率をKPIに入れたか。[5]
- 不確実性（過信）を測る指標と手続き（反証生成、RAG切替、保留規則）を導入したか。[7]
- Gemini APIのsafetySettingsを本番固定し、変更は審査・ログ・ロールバック付きにしたか。[50]
- Claude利用では、アプリとAPIでシステムプロンプトが異なる可能性を踏まえ、運用対象を明確化したか。[51]

---

[1] [5] [6] [8] [25] [26] [27] [28] [29] [30] [31] [32] [33] [42] [44] [45] [46] [47] [48] https://www.jstage.jst.go.jp/article/jsaisigtwo/2026/FIN-036/2026_193/_pdf/-char/ja
https://www.jstage.jst.go.jp/article/jsaisigtwo/2026/FIN-036/2026_193/_pdf/-char/ja

[2] [10] [11] [23] https://www.anthropic.com/constitution
https://www.anthropic.com/constitution

[3] [20] [41] [43] https://ai.google.dev/api/generate-content?hl=ja
https://ai.google.dev/api/generate-content?hl=ja

[4] [7] [9] [38] [49] https://aclanthology.org/2024.trustnlp-1.13/
https://aclanthology.org/2024.trustnlp-1.13/

[12] https://www.anthropic.com/news/constitutional-ai-harmlessness-from-ai-feedback
https://www.anthropic.com/news/constitutional-ai-harmlessness-from-ai-feedback

[13] [15] [24] https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf
https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf

[14] [50] https://ai.google.dev/gemini-api/docs/safety-settings?hl=ja
https://ai.google.dev/gemini-api/docs/safety-settings?hl=ja

[16] https://www.anthropic.com/news/claude-sonnet-4-6
https://www.anthropic.com/news/claude-sonnet-4-6

[17] https://docs.anthropic.com/ja/api/messages
https://docs.anthropic.com/ja/api/messages

[18] https://docs.anthropic.com/ja/api/versioning
https://docs.anthropic.com/ja/api/versioning

[19] [51] https://platform.claude.com/docs/ja/release-notes/system-prompts
https://platform.claude.com/docs/ja/release-notes/system-prompts

21 https://ai.google.dev/gemini-api/docs/changelog?hl=ja
https://ai.google.dev/gemini-api/docs/changelog?hl=ja

22 https://www.anthropic.com/responsible-scaling-policy
https://www.anthropic.com/responsible-scaling-policy

34 https://www.boj.or.jp/research/brp/fsr/fsrb250930.htm
https://www.boj.or.jp/research/brp/fsr/fsrb250930.htm

35 https://papers.ssrn.com/sol3/Delivery.cfm/4851711.pdf?abstractid=4851711
https://papers.ssrn.com/sol3/Delivery.cfm/4851711.pdf?abstractid=4851711

36 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5154002
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5154002

37 https://www.sciencedirect.com/science/article/abs/pii/S2214804325001697
https://www.sciencedirect.com/science/article/abs/pii/S2214804325001697

39 https://aclanthology.org/2024.findings-acl.675/
https://aclanthology.org/2024.findings-acl.675/

40 https://www.nature.com/articles/s42256-025-01115-6
https://www.nature.com/articles/s42256-025-01115-6