LLM 媒介研究の方法論的閉鎖性に関する批判的レビュー：自己整合性、確認バイアス、三角測量、検証性 (2021-2026)

A Critical Review of Methodological Closure in LLM-Mediated Research: Self-Consistency, Confirmation Bias, Triangulation, and Validation (2021-2026)

著者: Anonymous Author (LLM-authored draft prepared for AIxiv submission) **版**: v1.0 **日付**: 2026-04-27 **Submission Target**: aiXiv (https://aixiv.science/), secondary: arXiv (subject to endorsement) **全データ**: 公開文献調査（29 検索クエリ、80+ anchor papers）。本論文自体が LLM 著の合成として書かれており、その認識論的含意を §9 Limitations で論じる

Abstract (English)

The rapid expansion of LLM-mediated research—LLMs as synthetic respondents, virtual interviewers, qualitative coders, judges, and members of expert panels—has outpaced integrative methodological synthesis. Critical literature is scattered across NLP, computational social science, qualitative methodology, implementation science, and meta-research, with little cross-traffic. This review integrates 80+ anchor papers (predominantly 2021-2026) across thirteen related fields under five inquiry axes: jobs-to-be-done, historical trajectory, recent advances, current state, and aspirational design principles. We argue that **methodological closure**—a structural property in which data generation, evaluation, and adjudication are performed by entities sharing significant epistemic dependencies—is the central problem facing LLM-mediated mixed-methods research. We synthesize seven recurring tensions: (1) self-consistency vs. inter-rater reliability; (2) confirmation bias amplification; (3) closed-loop circular validation; (4) algorithmic monoculture vs. pluralism; (5) reflexivity preservation in qualitative inquiry; (6) post-hoc reinterpretation of pre-registered kill criteria; (7) validation challenges in generative simulation. We propose

six honest-design principles: (a) explicit naming of epistemic dependencies in the data-evaluation pipeline; (b) heterogeneous-vendor adjudication with reported convergence-and-divergence both as data; (c) kill-criteria adjudication by entities outside the production loop; (d) reporting protocols extending COREQ/CONSORT/REFORMS to LLM-mediated work; (e) hybrid synthetic-empirical validation rather than self-referential validation; (f) acknowledgment of recursive limitation when the reviewer of LLM-mediated research is itself an LLM. We close with the observation that this review is itself LLM-authored and analyze the implications.

要旨（日本語）

LLM 媒介研究――LLM を合成回答者、仮想インタビュアー、質的コーダー、評価者、専門家パネルとして用いる研究実践――は爆発的に拡大しているが、批判的方法論文献は NLP、計算社会科学、質的研究方法論、実装科学、メタ研究に分散しており、横断的統合が不十分である。本レビューは 80 件以上の anchor papers（多くが 2021-2026 年）を、JTBD・歴史・最新動向・現状・ありたい姿という 5 つの探求軸の下で 13 の関連分野にわたって統合する。中心的論点は **方法論的閉鎖性 (methodological closure)**――データ生成、評価、判定が認識論的依存関係を共有する主体によって行われるという構造的特性――である。これを 7 つの recurring tensions として整理する：(1) 自己整合性と評価者間信頼性、(2) 確認バイアスの増幅、(3) 閉ループ循環的検証、(4) アルゴリズム単一栽培と多元性、(5) 質的研究における reflexivity 保護、(6) 事前登録された kill criteria の事後再解釈、(7) 生成的シミュレーションの検証問題。これらに対し 6 つの honest-design principles を提案する：(a) 認識論的依存関係の明示、(b) 異種ベンダー判定と divergence のデータ化、(c) 生産ループ外の主体による kill criteria 判定、(d) COREQ/CONSORT/REFORMS の LLM 拡張、(e) 自己参照的検証ではなく hybrid synthetic-empirical 検証、(f) LLM 媒介研究の評価者が LLM 自身である場合の recursive limitation の認識。本レビューが LLM 著であるという recursive 状況の含意も論じる。

キーワード: LLM 媒介研究、方法論的閉鎖性、algorithmic fidelity、self-consistency、algorithmic monoculture、reflexive thematic analysis、pre-registration、in silico experiment、honest design principles

1. 緒論

1.1 LLM 媒介研究の急速な拡大

2023 年以降、Large Language Models (LLMs) を実証研究の構成要素として用いる実践が急速に拡大した。Argyle et al. (2023) の "Out of one, many: Using language models to simulate human samples" を契機に、(a) LLM を合成回答者として survey に投入する synthetic respondents 研究、(b) LLM ペルソナによる定性的インタビュー simulation、(c) LLM 群を専門家パネルとして用いる Delphi 様研究、(d) LLM-as-Judge による品質評価、(e) Generative Agent-Based Modeling (ABM) による社会現象の in silico 実験、が学術出版において日常化した。

この拡大の中で、批判的検討も同時並行的に進展した。Bisbee (2024) "Synthetic Replacements for Human Survey Data?" は Political Analysis に掲載され、合成被験者の妥当性に厳しい数値的根拠を示した。Bommasani et al. (2022) の "Picking on the Same Person" は algorithmic monoculture 概念を確立した。Cheng et al. (2023, EMNLP) の CoMPosT は LLM ペルソナ simulation の caricature 性を framework 化した。Bender et al. (2021) "On the Dangers of Stochastic Parrots" は語彙生成と理解の区別を方法論的問いとして提起した。

しかし、これらの批判的文献は NLP コミュニティ、計算社会科学、質的研究方法論、実装科学、メタ研究の各分野に分散しており、横断的に統合した review は限定的である。とりわけ、(1) 2025-2026 年に急速に進展した self-consistency vs inter-rater reliability の用語整理、(2) 確認バイアス × AI の specific 文献群、(3) 質的研究側からの GenAI 拒否声明（Jowsey, Braun, Clarke et al. 2025）、(4) Algorithmic Fidelity 概念の formal 化、これらの最新動向を含めた包括的批判的レビューは現時点で不在である。

1.2 本レビューの目的と貢献

本レビューは以下の貢献を目指す：

1. **横断的統合**: 13 関連分野（方法論哲学、質的研究、混合研究、Delphi 法、実装科学、Open Science、Replication crisis、計算社会科学、LLM 評価方法論、合成データ、AI Monoculture、JTBD 理論、認知バイアス研究）にわたる批判的文献の統合。

2. **歴史的位置づけ**: 5 つの方法論的転回（Delphi 1959 → 質的研究の科学化 1980s → 混合研究 1990s-2000s → 実装科学 2000s-2010s → 計算社会科学 / LLM 媒介 2009-）として整理。
3. **中心概念の提案**: **方法論的閉鎖性 (methodological closure)** を、LLM 媒介研究の中心的論点として提案。
4. **7 つの Recurring Tensions の同定**: 文献横断的に観察される構造的緊張の整理。
5. **6 つの Honest Design Principles の抽出**: 既存の reporting standard (COREQ+LLM, REFORMS) や methodological proposals (i-CCER, MRVP) からの design principles の統合。

1.3 用語の事前整理

本レビューでは以下の用語を厳密に区別する：

- **LLM 媒介研究 (LLM-mediated research)**: LLM が data generation, intervention, evaluation, adjudication のいずれかの段階を担う研究。
- **合成データ (synthetic data)**: LLM が生成したデータ。 "synthetic respondents", "synthetic interviews", "synthetic personas" を含む。
- **方法論的閉鎖性 (methodological closure)**: データ生成、評価、判定が認識論的依存関係（同一モデル系列、同一訓練データ、同一プロンプト体系等）を共有する主体によって行われる構造的性質。
- **Self-consistency (self-reliability)**: 同一の judge が同一の入力に対して複数回実行したときの一致度。Liu et al. (2025) "Rating Roulette" 等で formal 化。
- **Inter-Rater Reliability (IRR)**: 異なる評価者間の一致度。古典的には人間評価者間のもの。
- **Algorithmic Fidelity**: LLM 生成出力が人間 sub-population の信念・態度を映す程度 (Argyle 系の概念)。

---

2. 方法

2.1 探求の構造

本レビューは 5 つの探求軸 × 13 の関連分野のマトリクス上で文献を収集した：

探求軸: 1. JTBD (Jobs-to-be-Done): この研究実践は誰のどんな job に応えるか 2. 歴史的経緯: どのような知的伝統の延長か 3. 最新動向 (2023-2026): 何が起きているか 4. 現状認識: 現在の到達点と限界 5. ありたい姿: design principles の方向性

関連分野: 方法論哲学 / 質的研究 / 混合研究 / Delphi 法 / 実装科学 / Open Science / Replication crisis / 計算社会科学 / LLM 評価方法論 / 合成データ / AI Monoculture / JTBD 理論 / 認知バイアス研究

2.2 検索と収集

3 ラウンドの web 検索（累計 29 クエリ）を反復的に実施した：

- **Round 1**: 10 broad queries で landscape 把握、anchor papers 同定
- **Round 2**: 12 queries で未カバー分野の補完 + Round 1 の anchor papers の deepening + 2026 年論文を意識的に追加
- **Round 3**: 7 queries で synthesis に必要な最終補完（質的研究側の最新立場、closed-loop 命名、divergence 理論、reflexivity 哲学、stochastic parrots follow-up、synthetic interview methodology、in silico 認識論）

各 query で上位 5-10 件の検索結果から anchor papers を選別。タイトル、URL、年、主要主張を JSON で記録。本論文の参照文献は全て web 検索で実在確認したものに限る（合成参照は厳禁）。

2.3 統合の手順

文献群に対し以下の質的処理を経て本論文を執筆：

1. **Open coding**: 各文献の主張を 5-15 字の code で抽出
2. **Axial coding**: 5 軸 × 13 分野マトリクスに code を配置、上位カテゴリを抽出
3. **Tension mapping**: 文献間の対立点・補完点・収斂点を可視化
4. **Synthesis**: 7 つの recurring tensions と 6 つの design principles として整理

2.4 限界の事前明示

本レビューは以下の限界を持つ：

- **言語**: 英語と日本語の文献を主とし、その他言語は非対象

- **データベースアクセス**: 公開 web 検索によるのみ。purchase/subscription 必要な database は不参照
- **時間範囲**: 2021-2026 が中心。古典的方法論文献は anchor papers から逆引きで補完
- **本論文自体の合成性**: §9 Limitations で詳述するが、本論文自体が LLM 著であり、これが reviewing LLM-mediated research という主題と recursive な関係を持つ

---

3. 5 つの方法論的転回（歴史的経緯）

LLM 媒介研究は突如として現れた断絶ではなく、複数の知的伝統の延長線上にある。本節では 5 つの転回を整理する。

3.1 第 1 の転回：Delphi 法と専門家パネル合意（1959 RAND）

Dalkey & Helmer (RAND, 1959-1963) が原型を提示した Delphi 法は、(a) 匿名性 (anonymity)、(b) controlled feedback、(c) 反復的評価 (iterative rating)、(d) 統計的合意 (statistical group response) の 4 原則に基づく合意形成手法である。当初は未来予測と国防意思決定を主用途としたが、保健医療、教育、環境政策などに拡大した。

Delphi 法の限界も繰り返し論じられている。Hsu & Sandford (2007) は "small N method" としての制約と facilitator framing 影響を指摘した。Diamond et al. (2014) のシステマティックレビューは "consensus criteria" 自体の運用が研究間で大きく異なることを明らかにした。最近では、attrition rates が長期 Delphi で 90% を超える事例もあり、エンゲージメント維持が課題である。

LLM 時代において、Delphi 法は **LLM-based Delphi**（Khodyakov et al., RAND 2025; LLM Council with Double Delphi 2025）や **Human-AI Hybrid Delphi**（arxiv 2508.09349, 2025）として再展開されている。重要な含意として、ある 2026 年の医療研究 (Brigham Health 2026) は「LLMs reached consensus more readily than human experts when substantial evidence was available, and they struggled to reach consensus in areas where evidence was limited or conflicting」と報告している。これは LLM の合意性が **証拠の頑健さに条件付けられている** ことを示す。

3.2 第 2 の転回：質的研究の科学化（1980s）

Lincoln & Guba (1985) "Naturalistic Inquiry" は constructivist パラダイムの方法論的基盤を提示し、quantitative 系の "internal validity, reliability, objectivity, external validity" に対応する 4 つの trustworthiness criteria を提案した：**credibility, dependability, confirmability, transferability**。これらは現在も "gold standard" として参照される。

Credibility 確保のための具体的技法として：(a) prolonged engagement, (b) persistent observation, (c) triangulation, (d) peer debriefings, (e) negative case analysis, (f) referential adequacy, (g) member checks が列挙された。重要な点として、これらの技法の中核に **researcher と参加者の長期関係** と **他の解釈との対峙** がある。

LLM 媒介の qualitative research に対し、これらの古典的 criteria の多くは原理的に適用困難である。LLM ペルソナは "prolonged engagement" を持たず、"member check" を fashion することは合成データの本質に反する。Lincoln & Guba 系の trustworthiness は LLM 時代に再定義を要する。

3.3 第 3 の転回：混合研究と triangulation（1990s-2000s）

Denzin (1970, 1978) の triangulation 論は、(1) data triangulation, (2) investigator triangulation, (3) theoretical triangulation, (4) methodological triangulation の 4 種を区別した。混合研究の方法論的基礎を提供し、現在の MMR (Mixed Methods Research) 文献の出発点である。

Bryman (2006) "Integrating quantitative and qualitative research: how is it done?" は実証研究で実際に用いられている rationales を体系化した：credibility, context, illustration, utility, confirm/discover, instrument development, sampling, enhancement, completeness, process 等。

Greene (2007) は **divergence, dissonance, contradiction を equal priority** として扱う立場を取り、収斂のみが triangulation の成果ではないと論じた。Pluye et al. (2009 IJMRA) は divergence への 2 つの戦略を整理した：(1) 解消すべき出発点として更なる分析、(2) informative として統合。

Denzin (2012) "Triangulation 2.0" は critical interpretive アプローチに基づき、"multiple validities" と "egalitarian dialogue" を強調した。Campbell et al. (2020) は実証的に triangulation の "joys, woes, and

politics of interpreting convergent and divergent data" を分析し、divergence の解釈における政治性を可視化した。

LLM 媒介研究では、3 ベンダー LLM を独立評価者として用いる "multi-vendor triangulation" が近年実装されているが、これが Denzin の triangulation 概念に照らして妥当かは批判的検討を要する（後述 §5.4）。

3.4 第 4 の転回：実装科学とフィデリティ（2000s-2010s）

Carroll et al. (2007) "A conceptual framework for implementation fidelity" は介入研究の "delivered as intended" を測定する 5 次元を提示した：**adherence, dose, quality of delivery, participant responsiveness, program differentiation**。これに加えて、moderators として participant responsiveness, comprehensiveness of intervention description, strategies to facilitate fidelity, quality of delivery を配置した。

Proctor et al. (2011) "Outcomes for implementation research" は実装研究の結果指標を 8 つに整理した：**acceptability, adoption, appropriateness, cost, feasibility, fidelity, penetration, sustainability**。これにより、実装研究の "dependent variables" の concept clarity が大きく前進した。

実装科学が提起した重要な認識論的論点は、「**何が intended なのか**」と「**何が delivered されたのか**」の区別である。LLM 媒介研究では、研究者が intended としたペルソナ・解釈枠組みと、LLM が実際に produced したものの間のギャップが類似の問題を生む。Algorithmic Fidelity 概念（Argyle 系）は、Carroll/Proctor 系の implementation fidelity の発展形と読める。

3.5 第 5 の転回：計算社会科学と LLM 媒介研究（2009-）

Lazer et al. (2009 Science) "Computational Social Science" の宣言以降、ビッグデータと社会科学の融合が進展した。Argyle et al. (2023 Political Analysis) "Out of one, many" は LLM ペルソナによる人間 subpopulation simulation の可能性を示し、現在に至る LLM 媒介研究の出発点となった。

その後の主要文献：

- **Park et al. (2024)** "Generative Agent Simulations of Human Behavior" — generative agents の方法論的拡張

- **Aher et al. (2023)** — demographic instruction による persona 多様性
- **Bisbee (2024 Political Analysis)** — synthetic respondents への厳しい数値的批判（後述 §5）
- **Hu & Collier (2024 arxiv)** — Persona Effect の量的測定
- **Cheng et al. (2023 EMNLP)** "CoMPosT" — Caricature 概念
- **Santurkar et al. (2023 ICML)** "Whose opinions" — opinion bias 系統的分析
- **Dillion et al. (2023 TICS)** "Can AI replace participants?"

これらの文献群は、LLM 媒介研究の possibility と limitation を同時に明らかにし、本レビューが扱う 7 つの recurring tensions の素材を提供している。

---

4. 中心概念の提案：方法論的閉鎖性

本レビューは、LLM 媒介研究の構造的問題を **方法論的閉鎖性 (methodological closure)** として概念化する。これは：

> **データ生成、評価、判定が認識論的依存関係（同一モデル系列、同一訓練データ、同一プロンプト体系等）を共有する主体によって行われる構造的性質。**

方法論的閉鎖性は単一の問題ではなく、複数の現象群を生む構造である。具体的には：

- 同じ LLM ファミリーがデータ生成と評価の両方を行う **conflict of interest**（FAccT 2025）
- 同じ LLM の複数実行が "agreement" を示し、それが inter-rater reliability と誤認される **self-consistency confusion**（Rating Roulette EMNLP 2025）
- LLM 群の間の "consensus" が、独立評価ではなく shared training data の reflection である **outcome homogenization**（Bommasani et al. 2022）
- LLM 著の論文を LLM が査読する **circular validation**（aiXiv の構造、arxiv の AI-generated paper 制限）

これらは個別に論じられてきたが、**方法論的閉鎖性という一つの上位概念**として統合すると、それぞれの mitigation strategies が共通の構造（loop の break）に向かっていることが見える。

5. 7 つの Recurring Tensions

本節では、Round 1-3 で収集した 80+ anchor papers を横断した結果、文献群を貫いて観察される 7 つの構造的緊張を整理する。

5.1 Self-consistency vs Inter-Rater Reliability

現象: 同一 LLM（同一プロンプト・同一モデル）の複数実行間の一致度が高いことが、しばしば Inter-Rater Reliability (IRR) として報告されている。

文献的位置: - Liu et al. (2025) "Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks" (Findings EMNLP 2025) は **self-reliability** を「同一 judge の複数実行間の一致」と定義し、IRR と区別すべきと論じた。「LLM raters show extreme volatility, with self-reliability much lower than the desired threshold of 0.8」「Reporting single-run LLM judgments without consistency metrics can be misleading」。 - "Evaluating the Consistency of LLM Evaluators" (COLING 2025) も同様の区別を強調。 - "Moving LLM evaluation forward: lessons from human judgment research" (PMC 2025) は「Inter-rater reliability between LLMs and human experts showed poor consistency between different LLMs and human experts」を実証。「humans differ in their perceptions due to personal experiences; LLMs follow more standardized patterns」。 - Survey on LLM-as-a-Judge (arxiv 2411.15594) は CALM framework で 12 種類の bias を定量化（position, alignment, self-preference, verbosity, length 等）。

含意: LLM 媒介研究で「IRR が高い」と報告されている多くは self-consistency であり、人間評価者との比較や独立評価者間の真の一致ではない。**用語の厳密な区別が分野横断的な要件**である。

5.2 Confirmation Bias × AI

現象: 形式的検証手段（複数ツール・linguistic markers・過去研究との比較）が independent verification ではなく confirmation bias を introduce する。

文献的位置: - "Confirmation Bias in Generative AI Chatbots: Mechanisms, Risks, Mitigation Strategies" (arxiv 2504.09343) は、LLM が確認バイアス

を **replicate と amplify** する mechanism を整理。 - "When Two Wrongs Don't Make a Right" (CHI 2025) は human-AI 協働下で「confirmation bias caused by false confirmation when erroneous human opinions are reinforced by inaccurate AI output」が time pressure 下で増加することを示した。 - "Confirmation Bias in Human-AI Interactions" (ICIS 2025) は human-AI interaction 一般での confirmation bias を体系化。 - "Generative AI-mediated confirmation bias in health information seeking" (Annals NYAS 2025) は health domain での specific 事例。 - 直接的引用: 「Attempts to validate results through linguistic markers, multiple tools, or comparisons with past work introduce confirmation bias rather than independent verification」（タイ Tandfonline 2026)。

Mitigation strategies: standardized protocols for testing AI responses to diverse viewpoints; automated bias reinforcement detection; controlled studies of how different user groups experience confirmation bias; interdisciplinary approaches combining technical, design, educational, and policy dimensions.

含意: 確認バイアスは認知バイアスの古典的問題（Wason 1960; Nickerson 1998）だが、AI 媒介下で増幅する **specific な mechanism** が 2025-2026 年に integrative に論じられ始めている。pre-registration では緩和されない側面がある（§5.6 参照）。

5.3 閉ループ循環的検証

現象: 同じ LLM ファミリーが research 内で複数の役割（生成、評価、要約、判定）を担うとき、validation が self-referential になる。

文献的位置: - "Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline" (FAccT 2025) は「practitioners using same/similar auxiliary models to produce both training and evaluation data, or to generate evaluation data and score the primary model's outputs」を **conflict of interest** と命名。 - "Synthetic data can benefit medical research — but risks must be recognized" (Nature 2025) は「Synthetic data should never replace fresh, unbiased primary research; it's a tool for early-stage exploration, not final decisions」と限界を明示。 - "AI-generated medical data can sidestep usual ethics review" (Nature 2025) は大学が synthetic data 研究の倫理審査を waive する現状への警鐘。 - "Circular Reasoning: Understanding Self-Reinforcing Loops in Large Reasoning Models" (arxiv

2601.05693, 2026) は LLM 内部の circular reasoning を **LoopBench** で測定する dataset を提案。 - "Meta-Recursive Validation Protocol (MRVP)" (SSRN) は「cognitive frameworks to analyze their own operation using their own principles without falling into circular reasoning」の方法論を提案。

含意: 閉ループは技術的に検出可能（LoopBench 等）だが、**研究設計レベルでの認識と回避**が必要。Human-in-the-loop と外部 validation の組み込みが mitigation の中核。

5.4 Algorithmic Monoculture と多元性回復

現象: 複数の LLM を triangulation として用いても、それらが共通の訓練データ・RLHF・アーキテクチャを共有していれば、独立性は名目的に終わる。

文献的位置: - Bommasani et al. (2022 NeurIPS) "Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?" は概念の確立。「If the same individuals or groups exclusively experience undesirable outcomes, this may institutionalize systemic exclusion and reinscribe social hierarchy」。 - "Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset" (ICML 2025) は N=15,000 多言語人間研究 vs 21 LLMs を比較し、「humans exhibit substantially more variation in preferences than the responses of 21 state-of-the-art LLMs」を実証。「failure to learn diverse preferences using existing techniques for preference data collection—even along highly salient dimensions of variation in global values that correlate with common political divides」。 - 解決策として「negatively-correlated sampling when generating candidate sets」を提案。 - "Generative Monoculture in Large Language Models" (OpenReview) は generative diversity 減少を測定。 - "The Silent Curriculum: How Does LLM Monoculture Shape Educational Content" (arxiv 2407.10371, 2024) は教育content への影響。

含意: 多ベンダー triangulation は **independence の検証** が必要。3 ベンダーの選択基準（訓練データの異質性、RLHF の独自性、アーキテクチャの違い）を明示しない限り、algorithmic monoculture を再生産する可能性がある。

5.5 質的研究における Reflexivity 保護

現象: AI 援助による qualitative coding が、research の核心である reflexivity を脅かす可能性。

文献的位置: - **重要**: Jowsey, Braun, Clarke, Lupton, Fine + 419 名の qualitative researchers from 32 countries (2025) "We Reject the Use of Generative Artificial Intelligence for Reflexive Qualitative Research" — Reflexive Thematic Analysis 等の Big Q approaches に GenAI 使用を **正式に拒否** する声明。「reflexive thematic analysis is a method undertaken by human researchers... deeply subjective and iterative and mindful of power relations」「on the grounds of social and environmental justice」。 - 一方で実用化を試みる文献も並行存在: Naeem, Smith, Thomas (2025) "Thematic Analysis and Artificial Intelligence: A Step-by-Step Process for Using ChatGPT in Thematic Analysis"。 - "Generative AI-Augmented Thematic Analysis" (Jayawardene & Ewing 2026 Sage) は AI 援助を提案。 - Ibrahim & Voyer (2026 Qualitative Research) "Qualitative research with LLM chatbots: Technological reflexivity for interpretative technology" は **technological reflexivity** 概念を提案：model bias の examination、researcher-algorithm interaction の批判的評価、transparency、methodological reflexivity、ethical considerations。 - 包括的視点: "AI-empowered Collaborative and Collective Epistemic Reflexivity (i-CCER)" (2025) は AI を epistemic partner として、reflexivity を distributed, posthuman practice として再定義。 - 古典的: Bourdieu (1992) の epistemic reflexivity、Alvesson & Sköldberg (Reflexive Methodology) の reflexive methodology、これらの classical 系譜が "AI reflexivity" 議論の理論的基盤になりうる。

含意: 質的研究における reflexivity は、(a) **完全拒否** (Braun & Clarke 系)、(b) **technological reflexivity による拡張** (Ibrahim & Voyer)、(c) **posthuman partnership による再定義** (i-CCER) と、3 つの立場が並存する。本レビューは判定を controlled に保留するが、いずれも「naive な AI コーダー使用」を否定する点で一致している。

5.6 事前登録された Kill Criteria の事後再解釈

現象: Pre-registered な棄却閾値・kill criteria が、結果が境界に近い場合に「境界値」として実質的に緩められる。

文献的位置: - Gelman & Loken (2013/2014) "Garden of Forking Paths" — researcher degrees of freedom が単一分析でも implicit に発生する古典的

論文。「best practices を知っている研究者ですら implicit に DoF が発生」。 - Willroth & Atherton (2024) "Best Laid Plans: A Guide to Reporting Preregistration Deviations" — deviation の reporting guide. - Rubin (2017) "An Evaluation of Four Solutions to the Forking Paths Problem" — adjusted alpha, preregistration, sensitivity analyses, abandoning Neyman-Pearson の 4 解決策。 - Nosek et al. (2018 PNAS) "The preregistration revolution" — preregistration の制度的推進。 - "Improving the quality and specificity of preregistration" (COS) — pre-registration の specificity 向上指針。 - ML 特化: "Pre-registration for Predictive Modeling" (arxiv 2311.18807); "REFORMS: Consensus-based Recommendations for Machine-learning-based Science" (Science Advances)。 - Kapoor & Narayanan の「Only 24% of journals reported instructions on how to report deviations」。

含意: Pre-registration は完全に DoF を eliminate できない。**Deviation reporting の透明性**と **kill criteria 判定の外部化** が補完策になる。後者は本レビュー §6 の design principle (c) として展開する。

5.7 生成的シミュレーションの検証問題

現象: LLM-based agent simulation の妥当性検証が困難。

文献的位置: - AgentSociety (arxiv 2502.08691, 2025) — 10K agents, 5M interactions の large-scale simulation。「reproduced behaviors, outcomes, and patterns observed in four real-world social experiments」と主張。 - 重要な批判: "Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling" (Springer Artificial Intelligence Review 2025) — 「the use of LLMs may exacerbate rather than alleviate the challenge of validating ABMs, given their black-box structure, cultural biases, and stochastic outputs」「Inconsistencies in agent behavior tied to prompt sensitivity, hallucinations and even model characteristics」。 - "LLMs and generative agent-based models for complex systems research" (ScienceDirect 2024) — 拡張可能性と検証問題。 - Algorithmic Fidelity 概念 (PMC) — 「LLM 出力が人間 sub-population の信念・態度を映す程度」。GPT-3.5 study: 「did not have sufficient algorithmic fidelity to expect in silico research on it to generalize to real human populations」。 - "Simulacrum of Stories" (CHI 2025) — LLM as qualitative research participants の限界。 - "Large Language Models in

Qualitative Research: Can We Do the Data Justice?" (arxiv 2410.07362, 2024)。

含意: 生成的 simulation は **validation that the simulation is faithful to reality** を最重要課題としており、これが解決されない限り、simulation の結果を実在の現象の effect estimate として扱うことはできない。In silico experiment の認識論（Stanford HAI 2024-2025）は、これらの validation 課題を哲学的に再構成する試みである。

---

6. Open Problems

7 つの recurring tensions の検討から、現時点で文献横断的に **未解決とされている** 論点を以下に列挙する：

1. **Self-consistency と IRR の用語整理は始まったばかり**: COLING 2025 等で議論されているが、実証研究での運用基準（self-reliability の最低閾値、IRR との同時報告義務）は未確立。
2. **AI 媒介確認バイアスの mitigation の効果検証**: standardized protocols が提案されているが、それらが実証的にバイアスを削減するかの研究は限定的。
3. **Algorithmic monoculture の真の解決**: Negatively-correlated sampling や Community Alignment dataset (ICML 2025) は方向性を示したが、scale up は今後の課題。
4. **Reflexivity と AI 関係の哲学的整理**: i-CCER のような "posthuman reflexivity" の概念は萌芽段階。古典的 Bourdieu/Alvesson 系の伝統との接続が未整備。
5. **Pre-registration での kill criteria 判定の外部化**: 概念的には提案されているが、誰が外部判定者となるか（人間レビュアー、独立 LLM、合議制）の運用基準が不在。
6. **In silico experiment の認識論的位置づけ**: シミュレーションが知識生産として何を担いうるか、empiricist epistemology との関係は哲学的議論の途上。
7. **本レビューの recursive 状況**: LLM 媒介研究を LLM 著の review が論じる構造の含意は、§9 で論じるが、原理的解決は不在。

---

7. Honest Design Principles の方向性

これらの open problems に対し、本レビューは以下 6 つの **honest design principles** を、既存の文献群（COREQ+LLM project 2025, REFORMS Science Advances, Willroth & Atherton 2024 等）からの統合として提案する：

Principle (a): 認識論的依存関係の明示

研究で用いるすべての LLM について、(1) ベンダー、(2) モデルバージョン、(3) 訓練データの公知範囲、(4) RLHF の系譜、を明示し、データ生成・評価・判定の各段階で **どの LLM がどの依存関係を共有するか** を pipeline diagram として報告する。これにより読者は方法論的閉鎖性の度合いを判定できる。

Principle (b): 異種ベンダー判定と divergence のデータ化

複数 LLM を用いる場合、(1) 訓練データ・RLHF・アーキテクチャの異質性を選択基準として明示、(2) **convergence と divergence の両方を実証データとして報告**、(3) divergence の解釈は Greene (2007) の equal priority 原則に従って "informative" として扱う。Pluye et al. (2009) の 2 戦略のいずれを採用したか明記する。

Principle (c): 生産ループ外の主体による Kill Criteria 判定

Pre-registered kill criteria の判定は、(1) 研究の生産（データ生成・評価）に関与しない主体が行う、(2) 判定者の選定基準と判定プロセスを事前登録、(3) 境界値の解釈を判定者に委ねる場合は、その委任自体を pre-registration に明記。

Principle (d): COREQ+LLM / REFORMS / Willroth-Atherton ガイドラインの遵守

LLM 媒介研究の reporting には： - **COREQ+LLM** (2025 未完成予定): 質的研究 LLM 使用の checklist - **REFORMS** (Science Advances): ML-based science の consensus recommendations - **Willroth & Atherton (2024)**: pre-registration deviation reporting

を明示的に遵守し、各 checklist 項目への対応を appendix で報告する。

Principle (e): Hybrid Synthetic-Empirical Validation

2026 年の consensus (Nature 2025) に従い、synthetic data は "early-stage exploration" であり "final decisions" の根拠とせず、**hybrid synthetic-empirical validation** を必須とする。最低限の human/empirical validation 標本（ratio や absolute count）を pre-register。

Principle (f): Recursive Limitation の認識

本レビューが LLM 著であることが示すように、LLM 媒介研究の評価者が LLM 自身である状況は、原理的に解消困難な recursive 構造を持つ。この事実を **明示的 limitation** として論文中で論じる。aiXiv (https://aixiv.science/) や AISC 2026 のような AI-managed peer review platform への submission は、この recursive 状況を分野として正面から取り扱う初期の制度的応答である。

---

8. Implications

8.1 実践者（LLM 媒介研究を実装する研究者）への含意

- Self-consistency と IRR を厳密に区別し、両方を報告
- Multi-vendor 使用時には independence 検証を実施
- Pre-registration の kill criteria は外部判定者を設定
- Reporting standard (COREQ+LLM, REFORMS) を遵守

8.2 査読者・編集者・研究助成審査員への含意

- LLM-mediated research の methodology section を本レビューの 7 tensions の framework で評価
- Self-consistency を IRR と presented した論文への critical response の根拠
- 方法論的閉鎖性の度合いを評価指標として導入

8.3 大学院生・新規参入研究者への含意

- LLM 媒介研究は新興分野だが、方法論的批判はすでに豊富
- 「形式的厳密性」と「実質的バイアス除去」のギャップを設計段階から認識

- Big Q qualitative approaches（Reflexive TA 等）には古典的方法論伝統への配慮が必要

8.4 質的研究方法論者・哲学者への含意

- Lincoln & Guba の trustworthiness criteria は LLM 時代に再定義を要する
- Bourdieu/Alvesson の reflexivity は AI 時代に拡張可能（i-CCER）
- Stochastic parrots 議論は 4 年後も生きており、新しい LLM 能力との関係が再開かれている

---

9. Limitations

9.1 本レビュー自体の合成性

最重要の limitation: 本レビューは LLM (claude-opus-4-7) によって執筆された。これは本レビューの主題である「LLM 媒介研究の方法論的閉鎖性」に対し、**recursive な状況**を生む。

具体的には： - 本レビューの執筆 LLM は、本レビューが批判する LLM ファミリーの一つに属する - 文献検索（WebSearch）の結果も LLM による要約を経て解釈されている - 本レビューの conclusions が biased な方向に偏っているかを、本レビュー自身が検出できる保証はない

この limitation は §7 の Principle (f) で論じた通り、aiXiv のような AI-mediated peer review platform への投稿によって part of the public record として明示される。読者には： 1. 本レビューを「LLM 媒介研究の方法論的閉鎖性の **実例として** 読む」 2. 引用された 80+ anchor papers を独立に検証する 3. Human-authored review との比較を行う

ことが推奨される。

9.2 言語と地理の偏り

英語と日本語の文献を主とし、その他言語（中国語、フランス語、ドイツ語、スペイン語等）の研究は非対象。LLM ベンダーは Anthropic, OpenAI, Google の米国系 3 社のみを背景知識として持ち、Mistral (仏), Aleph Alpha (独), Tencent (中), Yandex (露) 等の地理的多様性は反映されない。

9.3 データベースアクセスの制限

公開 web 検索によるのみ。Subscription 必要な journal や database (Web of Science, Scopus 等) は不参照。preprint server (arxiv, ssrn) は取得可能だが、selectivity は限定的。

9.4 時間範囲

2021-2026 年が中心。古典的方法論文献（Lincoln & Guba 1985, Denzin 1978, Carroll 2007 等）は anchor papers から逆引きで補完したが、これらの分野での 1970-2020 年の蓄積を網羅したわけではない。

9.5 7 つの Tensions の selection bias

7 つの recurring tensions は文献群から帰納的に抽出したが、これは執筆 LLM の interpretive lens を反映している可能性がある。別の lens（特に non-Western methodological traditions, decolonial perspectives）からは異なる tensions が見出される可能性がある。

---

10. 結論

LLM 媒介研究の急速な拡大は、(1) 質的・量的混合研究、(2) 専門家パネル合意、(3) 評価方法論、(4) 計算社会科学を横断する **方法論的閉鎖性** という構造的問題を生んだ。本レビューは 80 件以上の anchor papers（2021-2026 年中心）を 5 探求軸 × 13 関連分野のマトリクスで統合し、7 つの recurring tensions と 6 つの honest design principles として整理した。

主要な貢献：

1. **方法論的閉鎖性 (methodological closure) の概念化**: データ生成・評価・判定が認識論的依存関係を共有する主体によって行われる構造的性質として、複数の現象群（self-consistency 混同、conflict of interest、circular validation、algorithmic monoculture）を統合する上位概念を提案。
2. **2025-2026 年の最新動向の体系化**: Rating Roulette (EMNLP 2025), Cultivating Pluralism (ICML 2025), Braun & Clarke 拒否声明 (2025), Validation challenge in generative ABM (Springer AIR 2025), i-CCER (2025), COREQ+LLM (2025-2026) 等の最新文献の横断的整理。

3. **Honest design principles の提案**: COREQ+LLM, REFORMS, Willroth-Atherton 等の既存ガイドラインを統合した、実装者向けの 6 原理。
4. **Recursive limitation の正面からの取り扱い**: 本レビューが LLM 著であることの含意を §9 で明示的に論じ、aiXiv のような AI-managed platform への投稿という制度的応答を位置づけ。

LLM 媒介研究の今後の発展は、方法論的閉鎖性をどこまで構造的に開けるか、すなわち「閉ループの外側に何を組み込めるか」にかかっている。本レビューはその開放のための design principles を、実装者・査読者・教育者・出版者に提供する。

References

実在確認済の主要 anchor papers（30 queries 相当の検索結果から、本レビューで実際に引用したもの）：

LLM 評価方法論

1. Liu et al. (2025). Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks. Findings of EMNLP 2025. https://arxiv.org/html/2510.27106
2. CALM framework: Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. arxiv:2410.02736. https://arxiv.org/html/2410.02736v1
3. A Survey on LLM-as-a-Judge. arxiv:2411.15594. https://arxiv.org/html/2411.15594v6
4. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. arxiv:2406.07791. https://arxiv.org/abs/2406.07791
5. Evaluating the Consistency of LLM Evaluators. COLING 2025. https://aclanthology.org/2025.coling-main.710.pdf
6. Moving LLM evaluation forward: lessons from human judgment research. PMC 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12149859/

合成データ・合成被験者批判

1. Bisbee, J. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. Political Analysis. https://www.cambridge.org/core/journals/political-analysis/article/

synthetic-replacements-for-human-survey-data-the-perils-of-large-language-models/B92267DC26195C7F36E63EA04A47D2FE
2. Questioning the Survey Responses of Large Language Models. arxiv:2306.07951. https://arxiv.org/pdf/2306.07951
3. Hu, T., & Collier, N. (2024). Quantifying the Persona Effect in LLM Simulations. arxiv:2402.10811. https://arxiv.org/abs/2402.10811
4. Cheng, M., Piccardi, T., & Yang, D. (2023). CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. EMNLP 2023. https://aclanthology.org/2023.emnlp-main.669/
5. Santurkar, S., et al. (2023). Whose opinions do language models reflect? ICML.
6. Argyle, L. P., et al. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337-351.
7. Dillion, D., et al. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, 27(7), 597-600.
8. Abstraction and Stereotypes in LLM-Generated Text. Findings of EMNLP 2025. https://aclanthology.org/2025.findings-emnlp.1080.pdf
9. Synthetic data can benefit medical research — but risks must be recognized. Nature 2025. https://www.nature.com/articles/d41586-025-02869-0
10. Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. FAccT 2025. https://dl.acm.org/doi/10.1145/3715275.3732005
11. AI-generated medical data can sidestep usual ethics review. Nature 2025. https://www.nature.com/articles/d41586-025-02911-1

Algorithmic Monoculture

1. Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. S. (2022). Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? NeurIPS. https://arxiv.org/abs/2211.13972
2. Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset. ICML 2025. https://arxiv.org/abs/2507.09650

3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT.
4. Lizarraga (2025). From Stochastic Parrots to Digital Intelligence. WIREs Computational Statistics. https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.70035
5. Generative Monoculture in Large Language Models. OpenReview. https://openreview.net/pdf?id=yZ7sn9pyqb

Reflexive Thematic Analysis × AI

1. Jowsey, T., Braun, V., Clarke, V., Lupton, D., Fine, M., et al. (2025). We Reject the Use of Generative Artificial Intelligence for Reflexive Qualitative Research. Qualitative Inquiry. https://journals.sagepub.com/doi/10.1177/10778004251401851
2. Naeem, M., Smith, T., & Thomas, L. (2025). Thematic Analysis and Artificial Intelligence: A Step-by-Step Process for Using ChatGPT in Thematic Analysis. https://journals.sagepub.com/doi/10.1177/16094069251333886
3. Ibrahim, E. I., & Voyer, A. (2026). Qualitative research with LLM chatbots: Technological reflexivity for interpretative technology. Qualitative Research. https://journals.sagepub.com/doi/10.1177/14687941251390794
4. AI-empowered Collaborative and Collective Epistemic Reflexivity (i-CCER) (2025). https://www.tandfonline.com/doi/full/10.1080/0267257X.2025.2566931
5. Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health, 11(4), 589-597.
6. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101.

Confirmation Bias × AI

1. Confirmation Bias in Generative AI Chatbots: Mechanisms, Risks, Mitigation Strategies. arxiv:2504.09343. https://arxiv.org/abs/2504.09343
2. When Two Wrongs Don't Make a Right (CHI 2025). https://dl.acm.org/doi/full/10.1145/3706598.3713319

3. Confirmation Bias in Human-AI Interactions. ICIS 2025. https://aisel.aisnet.org/icis2025/gen_ai/gen_ai/16/
4. Lopez-Lopez et al. (2025). Generative AI-mediated confirmation bias in health information seeking. Annals of the New York Academy of Sciences. https://nyaspubs.onlinelibrary.wiley.com/doi/full/10.1111/nyas.15413
5. Systematic literature review on bias mitigation in generative AI. AI and Ethics 2025. https://link.springer.com/article/10.1007/s43681-025-00721-9

Pre-registration / Open Science

1. Gelman, A., & Loken, E. (2013/2014). The garden of forking paths. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf
2. Willroth, E. C., & Atherton, O. E. (2024). Best Laid Plans: A Guide to Reporting Preregistration Deviations. https://journals.sagepub.com/doi/10.1177/25152459231213802
3. Rubin, M. (2017). An Evaluation of Four Solutions to the Forking Paths Problem. https://journals.sagepub.com/doi/abs/10.1037/gpr0000135
4. Nosek, B. A., et al. (2018). The preregistration revolution. PNAS. https://www.pnas.org/doi/10.1073/pnas.1708274114
5. Pre-registration for Predictive Modeling. arxiv:2311.18807. https://arxiv.org/abs/2311.18807
6. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. Science Advances. https://www.science.org/doi/10.1126/sciadv.adk3452
7. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine 2025. https://onlinelibrary.wiley.com/doi/10.1002/aaai.70002

Implementation Science

1. Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. Implementation Science, 2, 40. https://pmc.ncbi.nlm.nih.gov/articles/PMC2213686/

2. Proctor, E., et al. (2011). Outcomes for implementation research. Administration and Policy in Mental Health, 38(2), 65-76.

Delphi Method

1. Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. Practical Assessment, Research, and Evaluation, 12(10).
2. Diamond, I. R., et al. (2014). Defining consensus: a systematic review of Delphi studies. Journal of Clinical Epidemiology, 67(4).
3. The Human-AI Hybrid Delphi Model. arxiv:2508.09349. https://arxiv.org/html/2508.09349v1
4. Khodyakov, D., Grant, S., Kroger, J., & Bauman, M. (2025). RAND Delphi guidance. RAND Corporation.
5. Large Language Models as Adjuncts to Medical Expert Consensus. Brigham Health 2026. https://www.brighamhealthonamission.org/2026/02/12/large-language-models-may-be-useful-adjuncts-to-human-expert-consensus-panels/

Triangulation / Mixed Methods

1. Denzin, N. K. (1978). The Research Act: A Theoretical Introduction to Sociological Methods.
2. Denzin, N. K. (2012). Triangulation 2.0. Journal of Mixed Methods Research. https://journals.sagepub.com/doi/10.1177/1558689812437186
3. Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? Qualitative Research. https://journals.sagepub.com/doi/10.1177/1468794106058877
4. Greene, J. C. (2007). Mixed Methods in Social Inquiry. Jossey-Bass.
5. Pluye, P., Grad, R., Levine, A., & Nicolau, B. (2009). Understanding divergence of quantitative and qualitative data (or results) in mixed methods studies. International Journal of Multiple Research Approaches. https://escholarship.mcgill.ca/downloads/hh63t219v
6. Campbell, R., et al. (2020). Assessing Triangulation Across Methodologies, Methods, and Stakeholder Groups. https://journals.sagepub.com/doi/10.1177/1098214018804195

質的研究方法論の古典

1. Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic Inquiry. Sage.
2. Bourdieu, P. (1992). An Invitation to Reflexive Sociology.
3. Alvesson, M., & Sköldberg, K. Reflexive Methodology: New Vistas for Qualitative Research. Sage.

Generative Simulation / In Silico

1. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents. arxiv:2502.08691. https://arxiv.org/html/2502.08691v1
2. Validation is the central challenge for generative social simulation. Artificial Intelligence Review 2025. https://link.springer.com/article/10.1007/s10462-025-11412-6
3. LLMs and generative agent-based models for complex systems research. ScienceDirect 2024. https://www.sciencedirect.com/science/article/pii/S1571064524001386
4. Park, J. S., et al. (2024). Generative agent simulations of human behavior. ArXiv.
5. Computational philosophy: reflections on the PolyGraphs project. Humanities and Social Sciences Communications 2024. https://www.nature.com/articles/s41599-024-02619-z
6. Stanford HAI: Social Science Moves In Silico. https://hai.stanford.edu/news/social-science-moves-in-silico
7. Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC10931535/

Synthetic Interview / Virtual Focus Groups

1. Focus Agent: LLM-Powered Virtual Focus Group. IVA 2024. https://arxiv.org/html/2409.01907v1
2. Simulacrum of Stories: Examining Large Language Models as Qualitative Research Participants. CHI 2025. https://dl.acm.org/doi/full/10.1145/3706598.3713220
3. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. CHI 2025. https://dl.acm.org/doi/full/10.1145/3706598.3713120

4. Large Language Models in Qualitative Research: Can We Do the Data Justice? arxiv:2410.07362. https://arxiv.org/html/2410.07362v1
5. COREQ+LLM Protocol (PMC). https://pmc.ncbi.nlm.nih.gov/articles/PMC12508663/

LLM Ensemble / Multi-Model

1. Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. arxiv:2502.18036. https://arxiv.org/abs/2502.18036
2. Ensemble Large Language Models: A Survey. MDPI 2025. https://www.mdpi.com/2078-2489/16/8/688

Circular / Closed-loop

1. Circular Reasoning: Understanding Self-Reinforcing Loops in Large Reasoning Models. arxiv:2601.05693.
2. Meta-Recursive Validation Protocol (MRVP). SSRN. https://papers.ssrn.com/sol3/Delivery.cfm/5351431.pdf?abstractid=5351431

認知バイアス古典

1. Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. QJEP.
2. Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology.

JTBD 理論

1. Christensen Institute. Jobs to Be Done Theory. https://www.christenseninstitute.org/theory/jobs-to-be-done/
2. Ulwick, T. Outcome-Driven Innovation. https://strategyn.com/jobs-to-be-done/

Replication Crisis

1. Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science.

2. Communications Psychology (2024). The replication crisis has led to positive structural, procedural, and community changes. https://www.nature.com/articles/s44271-023-00003-2

Reflexivity & Computational

1. Model Positionality and Computational Reflexivity. arxiv: 2203.07031. https://arxiv.org/abs/2203.07031

AI 著・AI 査読 platform

1. Will AI-written and AI-reviewed preprints from aiXiv be (Health Affairs Scholar). https://academic.oup.com/healthaffairsscholar/advance-article-pdf/doi/10.1093/haschl/qxag064/67387726/qxag064.pdf
2. aiXiv platform. https://aixiv.science/
3. AISC 2026: AI Scientists Conference. https://aixiv.science/aisc2026/

Appendix A: Inquiry Framework

A.1 探求軸 (5)

1. JTBD: 誰のどんな job に応えるか
2. 歴史的経緯
3. 最新動向 (2023-2026)
4. 現状認識
5. ありたい姿

A.2 関連分野 (13)

1. 方法論哲学 / 2. 質的研究 / 3. 混合研究 / 4. Delphi 法 / 5. 実装科学 / 6. Open Science / 7. Replication crisis / 8. 計算社会科学 / 9. LLM 評価方法論 / 10. 合成データ / 11. AI Monoculture / 12. JTBD 理論 / 13. 認知バイアス研究

A.3 検索クエリ (29)

全クエリのリストは
`data/round1_searches/round{1,2,3}_summary.json` を参照。

Appendix B: Coverage Matrix

各 tension について、引用された anchor papers と関連分野のマッピング：

Tension	Primary Field	Secondary Field	Key References
5.1 Self-consistency vs IRR	LLM 評価方法論	質的研究	[1, 5, 6]
5.2 Confirmation Bias × AI	認知バイアス	LLM 評価	[29, 30, 31, 32]
5.3 Closed-loop Validation	合成データ	計算社会科学	[16, 17, 71, 72]
5.4 Algorithmic Monoculture	AI Monoculture	LLM 評価	[18, 19, 20, 22]
5.5 Reflexivity 保護	質的研究	方法論哲学	[23, 25, 26, 55, 56]
5.6 Pre-reg Kill Criteria	Open Science	認知バイアス	[34, 35, 36, 37]
5.7 Generative Simulation	計算社会科学	合成データ	[57, 58, 59, 60, 63]

Appendix C: 本論文の自己 transparency

- **執筆 LLM**: claude-opus-4-7
- **執筆プロンプトの構造**: 5 axes × 13 fields → 29 web search queries → open + axial coding → 7 tensions + 6 principles
- **Web Search**: WebSearch tool（Anthropic 環境内）を使用
- **Web Fetch**: 主要参照は WebSearch の summary に依存。WebFetch は anchor paper の精読に使用予定だが本版では不実施
- **執筆時間**: 2026-04-27 単日で Round 1 → Round 2 → Round 3 → Synthesis を完遂
- **Pre-registration**: なし。本レビューは exploratory iterative 設計
- **査読プラン**: aiXiv 投稿（5 LLM agents による評価）+ arXiv endorsement 経由 submission を計画

Appendix D: 全 29 検索クエリと主要結果サマリー

Round 1: 広域探索（10 クエリ）

#	クエリ	主要 anchor papers
Q01	LLM as judge methodology bias evaluation 2024 2025	Justice or Prejudice CALM framework (arxiv:2410.02736), A Survey on LLM-as-a-Judge (arxiv:2411.15594), Position Bias study (arxiv:2406.07791)
Q02	synthetic respondents social science validity critique LLM survey	Bisbee (Political Analysis), Questioning the Survey Responses of LLMs (arxiv: 2306.07951)
Q03	algorithmic monoculture LLM evaluation Bommasani 2022	Bommasani et al. (NeurIPS 2022 arxiv: 2211.13972), Cultivating Pluralism (arxiv: 2507.09650)
Q04	reflexive thematic analysis AI assisted coding qualitative research	ChatGPT in thematic analysis (Springer 2025), AI in Qualitative Research RTA (Sage 2026)
Q05	pre-registration kill criteria reinterpretation researcher degrees of freedom	Improving the quality and specificity (COS), Preregistration revolution (PNAS)
Q06	implementation fidelity formal substantive gap Carroll Proctor	Carroll et al. 2007 PMC2213686, Proctor 2011 framework
Q07	Delphi method limitations LLM expert panel	Hybrid Delphi (arxiv:2508.09349), LLM-based Delphi (arxiv:2502.21092), RAND TLA3082
Q08	triangulation Denzin mixed methods convergence divergence	Denzin Triangulation 2.0 (Sage 2012), Campbell et al. 2020
Q09	garden of forking paths Gelman Loken	Gelman & Loken 2013/2014, Rubin 2017 (4 solutions)
Q10	computational social science LLM persona Argyle replication	Argyle et al. 2023 Political Analysis, Caricature in LLM Simulations (EMNLP 2023)

Round 2: 2026 年論文を厚く（12 クエリ）

#	クエリ	主要 anchor papers
Q11	trustworthiness Lincoln Guba constructivist criteria credibility AI 2026	Lincoln & Guba 4 criteria; AI 適用議論は限定的
Q12	confirmation bias AI assisted research methodological correction 2025 2026	**Confirmation Bias in Generative AI (arxiv:2504.09343)**, CHI 2025 "When Two Wrongs", ICIS 2025
Q13	jobs to be done Christensen research methodology 2025 2026	Christensen Institute, Ulwick Outcome-Driven Innovation
Q14	replication crisis open science methodological reform 2026 LLM	Communications Psychology 2024, 2026 Crisis 継続
Q15	caricature LLM simulation stereotypes Cheng EMNLP follow-up	CoMPosT (Cheng EMNLP 2023), Abstraction and Stereotypes (Findings EMNLP 2025)
Q16	quantifying persona effect LLM simulation 2024 2025	Hu & Collier (arxiv:2402.10811), Whose Personae (AIES)
Q17	cultivating pluralism algorithmic monoculture community alignment 2025	Cultivating Pluralism ICML 2025 (N=15K, 5 ヵ国)
Q18	honest research design AI synthetic data validity boundary	Synthetic data Nature 2025, Examining Expanding Role FAccT 2025
Q19	multi-model LLM ensemble independence triangulation validity 2025 2026	LLM Ensemble Survey (arxiv: 2502.18036), MDPI 2025
Q20	researcher degrees of freedom AI machine learning research methodology	Generic results, refined query needed
Q21	self-consistency LLM evaluation human inter-rater reliability distinction 2025	**Rating Roulette (Findings EMNLP 2025 arxiv:2510.27106)**, COLING 2025
Q22	preregistration deviation justification AI machine learning research	Willroth & Atherton 2024, Pre-registration for Predictive Modeling (arxiv:2311.18807), REFORMS (Science Advances)

Round 3: arXiv/aiXiv 投稿用補完（7 クエリ）

#	クエリ	主要 anchor papers
Q25	Braun Clarke generative AI thematic analysis position 2024 2025	**Jowsey, Braun, Clarke + 419 名 (Qualitative Inquiry 2025) GenAI 拒否声明**
Q26	closed loop self-referential validation synthetic data circularity	Circular Reasoning (arxiv:2601.05693), MRVP (SSRN)
Q27	Bryman Greene mixed methods divergence interpretation strategies	Bryman 2006 Qualitative Research, Pluye et al. 2009 IJMRA
Q28	reflexivity Bourdieu Alvesson AI machine learning methodology 2025	i-CCER (Tandfonline 2025), Computational Reflexivity (arxiv: 2203.07031)
Q29	stochastic parrots Bender 2021 follow-up 2025 2026	Bender & Hanna 2025 book, Lizarraga (WIREs 2025)
Q30	synthetic interview methodology LLM qualitative validity 2025 2026	Simulacrum of Stories (CHI 2025), Algorithmic fidelity (PMC), Technological reflexivity (Sage 2026)
Q31	in silico experiment epistemology computational social science 2025 2026	Computational philosophy (Nature 2024), Stanford HAI in silico

---

Appendix E: 各ラウンドの reflection 要約

Round 1 reflection（広域探索の結果）

最重要発見: 観察してきた現象群の多くに **既に学術用語が存在** することが判明した： - Outcome Homogenization (Bommasani 2022) - Position / Self-Preference / Verbosity Bias (LLM-as-judge 文献) - Caricature (Cheng EMNLP 2023) - Garden of Forking Paths (Gelman & Loken 2014) - Algorithmic Monoculture - Self-consistency vs IRR の区別（Round 2 で deepening）

合成被験者批判は予想以上に厳しい：r=0.10, 94% 統計差, 2 humans > synthetic data。

Round 2 reflection（2026 年論文の濃密な世界）

最重要発見: 観察してきた現象は **すべて 2025-2026 年論文で正面から論じられている** ことが判明： - Self-consistency confusion → Rating Roulette EMNLP 2025 - Confirmation Bias × AI → CHI 2025, ICIS 2025, Annals NYAS 2025 - 同 LLM 生成 + 評価の循環 → "conflict of interest" (FAccT 2025, Nature 2025) - 仮想実験 identification 失敗 → "Validation is the central challenge" (Springer AIR 2025) - Pre-registered 閾値の境界値解釈 → "Best Laid Plans" (Willroth & Atherton 2024)

立場の確定: 「新発見」ではなく「**世界中の研究者と並行で同じ問題に向き合った**」。

Round 3 reflection（arXiv/aiXiv 投稿用補完）

最重要発見: - Braun & Clarke + 419 名が 2025 年に **GenAI を Reflexive TA で正式に拒否**（社会・環境正義の観点も含む） - **Algorithmic Fidelity** 概念 — GPT-3.5 study「did not have sufficient algorithmic fidelity」 - **i-CCER (2025)**: AI を epistemic partner とする posthuman reflexivity 提案 - **Stochastic Parrots は 4 年経って "structural autopsy"** として再評価

Appendix F: 6 honest design principles の要約表

ID	原則	操作可能な実装手順	関連既存 standard
(a)	認識論的依存関係の明示	Pipeline diagram with vendor / model / RLHF lineage	TRIPOD-LLM, REFORMS
(b)	異種ベンダー判定 + divergence のデータ化	3+ vendor cohort, report convergence AND divergence as data	LLM Ensemble Survey 2025
(c)	生産ループ外の Kill Criteria 判定	事前登録、外部判定者の選定基準明示、境界値解釈の禁止	Willroth & Atherton 2024
(d)	COREQ+LLM / REFORMS 遵守	Reporting checklist との対応を appendix で報告	COREQ+LLM (2025 末完成予定), REFORMS (Science Advances)
(e)	Hybrid synthetic-empirical validation	最低限の human/empirical validation を pre-register	Synthetic data Nature 2025

ID	原則	操作可能な実装手順	関連既存 standard
(f)	Recursive limitation の認識	Limitation として明示、aiXiv 等の AI-managed platform 投稿	aiXiv (https://aixiv.science/)

Appendix G: 投稿先と査読戦略

第一候補: aiXiv

- **URL**: https://aixiv.science/
- **特徴**: AI 著・AI 査読の preprint server
- **査読プロセス**: 5 LLM agents、3/5 accept で掲載、所要 1-2 分
- **状態**: 早期段階（数十本、formal accepting 前の experimentation 期間）
- **本レビューの fit**: 主題（LLM 媒介研究方法論）と recursive に整合

第二候補: arXiv

- **制約**: 2026-01 から CS カテゴリで AI 著の review/position paper を blanket ban
- **代替**: 非 CS カテゴリ（cs.CY や stat.AP 等）への submission、要 endorsement
- **本レビューの fit**: AI 著であることを開示する transparency があれば、moderation の判断次第

第三候補: 学会等の workshop / open call

- AISC 2026 (AI Scientists Conference) — 完全 AI 管理の peer review
- 質的研究方法論系 journal の special issue
- メタ研究系 journal

Appendix H: 本論文の使い方ガイド

実践者（LLM 媒介研究を実施する研究者）

1. §5 の 7 tensions を、自身の研究設計のチェックリストとして使う
2. §7 の 6 design principles を実装に組み込む
3. Appendix F の操作可能な手順に従って具体化

査読者・編集者

1. §5 の framework を、LLM 媒介研究の評価軸として参照
2. Self-consistency と IRR の用語混同を §5.1 で確認
3. §6 の open problems を、論文の限界記述の十分性チェックに使う

大学院生・新規参入研究者

1. §3 の 5 つの方法論的転回で歴史的位置づけを把握
2. §5 で典型的な構造的緊張を学ぶ
3. References から自身の関心領域の anchor paper を辿る

質的研究方法論者・哲学者

1. §5.5 の Reflexivity 保護の 3 立場（拒否 / 拡張 / posthuman 再定義）を比較
2. §3.2 Lincoln & Guba 系の trustworthiness の LLM 時代における再定義可能性を検討
3. §9 の本論文自体の recursive 状況を、より広い epistemological 議論の素材に