循環型市場における LLM を活用した エージェントベースシミュレーションの開発

○鶴崎 祐大（東京大学）*1　　木見田 康治（東京大学）*1

*1 東京大学工学系研究科技術経営戦略学専攻, 東京文京区本郷 7-3-1, 113-0033, kimita@tmi.t.u-tokyo.ac.jp

キーワード:サーキュラーエコノミー, LLM, Agent-based Simulation

1. 緒　　言

循環型経済(CE)は、経済活動の拡大と環境負荷の低減を同時に実現する経済モデルとして注目を集めている[1]。その中でも Product Service Systems (PSS)は、企業が製品を所有し、ユーザーに製品機能をサービスとして提供する仕組みであり[2]、一般的なサービスとしてシェアリング(リース, レンタル)が挙げられる。その経済合理性の高さから多くの製品で導入されており、そこで扱われる製品は稼働率が向上するため、環境負荷が軽減することが知られている[3]。PSS の経済性/環境性には消費者の選択や行動が大きく影響を与える。しかし、PSS ビジネスに対する消費者の受容性は、新規の利用形態であるがゆえに事前の需要予測が難しく、サービス設計が主要な課題となっている。したがって、消費者が「購入」と「シェアリング」のいずれを選択するかという 意思決定プロセスを精緻に把握し、価格、対象製品、保証等の 設計要素（デザイン変数） を体系的に調整する必要がある[4]。

サービスの設計手法としては、概念検証(PoC)やアンケートなどの実験を経て消費者行動を調査し、その結果を用いたシミュレーションなどを用いるのが近年の主流である。しかし、これらの実験は時間(調査設計[5,6]・被験者募集[5,7]・実施[5,6])、金銭(人件費[5]・インセンティブ[5,7])、運営(連絡[5,7]、データ品質管理[6,8]など)の各面で大きな負担がある。そのため、極力少ない検証で対象事業の成功に重要な要因を過不足なく得ることが求められている。このような実験の計画(design)において検証する設計変数は、従来、文献・実務・公的データに基づいて候補変数を広く洗い出し、質的調査(対象者・専門家)に基づいて選定されてきた[9]。

近年マーケティング分野において、消費者の意思決定の再現に大規模言語モデル(LLM)が注目されている[10]。LLM が主要な経済原則と整合する回答をすることがいくつかの研究で示されている[10,11,12]。この特性により、人間への調査を開始する前に、調査で解明される可能性の高い論点や、削除または追加したほうが良い設問を事前に検討できる。さらに、実施コストが比較的低いことも特長である。特に、アクセスが難しい回答者(医師や上級管理職など)の場合、その効果は大幅に高まる[10]。一方で、今回フォーカスする購入・シェアリング選好の再現については体系的な検証が少なく、人間データと照合して適用範囲を見極める必要がある。

そこで本研究は、LLM が製品の購入・シェアリング選好をどの程度外部妥当的に再現し得るかを，実アンケートに基づく離散選択モデルと比較して検証することを目的とする。比較対象(ベースライン)は、Koide et al. [4,13]の 冷蔵庫とラップトップ市場における Choice-based Conjoint(CBC)から推定した Hierarchical Bayesian (HB)-Multinomial Logit (MNL)とし、評価指標は先行研究に従い、(i)選択確率分布の整合[14,15]と (ii)シェアリングの主要属性に対する感度[10,11]という二つの観点から評価する。(i)選択確率分布の整合の検証にあたり、代表的な汎用 LLM モデル 5 つを用い、同一条件下でベースラインとの比較を行う。これにより、LLM モデルによる再現度の全体的な傾向や、2 つの製品の検証によりシェアリングに対する意思決定に適したモデルを調査する。(ii)シェアリングの主要属性に対する感度の検証では、(i)で明らかになった最適モデルを用い、シェアリングの価格とシェアリングで用いる製品の経年に対する選択確率の変化をベースラインと比較する。これにより、LLM がこれらの属性に対し人間と同じ反応を示すか検証する。この 2 つの評価を通じて、LLM がこれらの属性に対し人間と同じ反応を示すかを検証し、人間への調査を開始する前段での実験計画にどのように LLM を活用できるかを検討する。

2. 文献調査

2.1. 消費者調査における実験設計

実験は"a plan for assigning experimental units to treatment levels and the statistical analysis associated with the plan"と定義され[16]、研究者が目的とする従属変数に影響を与えうる要因をコントロールし、その影響を評価するものである[17]。具体的には、オンライン A/B テストや質問紙ベースの選択実験(CBC/DCE)などにおいて従属変数を操作し、応答の差異を統計的に推定する[18,19]。マーケティング分野や消費者行動研究で主要な方法論の一つであり[17,18]、シェアリングなどの新しいサービスの調査や検証にも主要なステップとして多くの研究で用いられている[19,20]。

実験計画(experimental design)は、研究目的に沿った応答(選択、購入意向、支払意思額)を定め、応答に寄与しうる設計変数とその水準・範囲を定義し、その上で設計構造(design structure)を選択して実施する一連の流れを指す[21]。設計変数とその水準・範囲の選定において、選択肢に影響を与えうる全ての要素を考慮する必要があるが、被験者にとって妥当なプロファイルとするために選定する必要がある[22]。その選定方法としてまず、文献・実務・公的データに基づ

いて候補変数を広く洗い出し、半構造化インタビューやフォーカスグループで名称・定義・単位を整える[9]。つぎに、関連性(目的に直結するか)、トレードオフ可能性(価格等との引き換えで選好が揺れ得るか)、理解容易性(誰が読んでもすぐ解釈できるか)、非ドミナンス(組合せが一方の選択肢を全面的に優位にしないか)などの基準で絞り込み、必要に応じて認知的プリテストと小規模パイロットでそれら妥当性を確認したうえで、水準の幅と刻みを実データに照らして確定する[23]。

このように設計変数とその水準・範囲の選定は、従来、対象者や専門家に対する質的調査に基づいてきたが、LLMを合成回答者として用い、多数の独立応答を短時間に収集・集計することで、選択確率分布や属性効果の方向を事前に検証する量的プリテストが可能になりつつある[10,11]。本研究ではこの発想を取り入れ、LLM によるプリテストで属性候補と水準の設定を行う実験計画手法の可能性を検証する。

2.2. 消費者調査における応答の近似器としての LLM

LLM は、消費者調査における応答の近似器(synthetic respondent)として有望視されており、需要曲線の右下がりやリスク回避、損失回避など基本的な経済原則に整合する応答を示すことが報告されている[10,11,24]。また、Few-shot Learning[10]や retrieval-augmented generation (RAG)[10]、fine-turning[11]などの手法を用いることでより高い一致を示すようになる。一方で、消費者の異質性の再現は限定的であり[10,11]、小さい確率を過大評価する傾向(確率過重)や[24]、提示順・ポジション・プロンプト文面への感度・温度・シード設定に起因する再現性の課題も指摘されている[15]。そのため、LLM が人間時行動を完全に再現できるわけではなく、あくまで人間被験者のシミュレーションツールとして効果を発揮する。Kozlowski & Evans[25]は大規模な LLM シミュレーションを行い、そこで得られた有望な仮説を、人間を対象とした実験で検証することの有用性を示している。

3. 方　法

本研究では、冷蔵庫／ラップトップ市場における新品購入・中古購入・シェアリングの選択を対象に、LLM 応答が人間データをどの程度再現し得るかを検証する。比較対象(ベンチマーク)は Koide et al.[4,13]の CBC 調査に基づく離散選択モデルであり、(i) 選択確率分布の整合、(ii) 属性感度(効果方向・相対強度)の整合の二点から外部妥当性を評価する。

3.1. 消費者データのベンチマーク

本研究における「人間ベンチマーク」は、Koide et al.[4,13]が日本の冷蔵庫市場（N＝911 from 22 June to 31 July, 2022）とラップトップ市場（N＝1023 from 11 October to 8 November, 2023）を対象に実施した CBC 調査と、その推定結果（階層ベイズ多項ロジット：HB-MNL）である。提示された選択肢は「新品購入」「再使用品購入（Reuse）」「メーカー改修品購入（Refurbish）」「サブスクリプション（非所有, Sharing）」の 4 種類である。選択肢の属性としては、両製品共通して価格／経年／無償修理保証、冷蔵庫は外観／評判／サブスク提供製品（新品／再使用／改修）、ラップトップは容量／バッテリーに関するものを用いている。回答者属性（ペルソナ）としては、年齢、性別、婚姻状況、世帯人数、世帯収入、三大都市圏在住の有無を把握している。

推定には HB-MNL を用い，個人レベル効用βとモデル内で同質とみなす一部の交互作用係数γを同時に推定している。回答者hの選択肢$i \in G$に対する観測可能な効用$V_{h,i}$は選択肢属性インデックスkを用いて以下のように表現される。ここで、選択肢属性は異質性を許す群K_1と同質とみなす群K_2に分ける。

$$V_{h,i} = \sum_{k=1}^{K_1} \beta_{h,k} x_{h,i,k} + \sum_{k=1}^{K_2} \gamma_{h,k} x_{h,i,k} \quad (1)$$

ここで得られた効用を元に選択確率$P_{h,s}$は次のように定義される。

$$P_{h,j} = \frac{exp(V_{h,j})}{\sum_{i \in G} exp(V_{h,i})} \quad (2)$$

推定後，個人レベル効用βによるクラスタリングにより 4 セグメントが得られている。(A)Subscription-inclined はサブスクを好むセグメント。(B)Price-sensitive は価格に対し敏感なセグメント。(C)Balanced decision-maker は価格やサービス条件に対して中程度の感度を持ち、リファービッシュや新しい循環型モデルを一定程度受容するセグメント。(D)Brand-new insistent は新品購入を固く志向するセグメントである。本稿では、これら各セグメントの部分効用を Table 1 に整理し、以降の LLM 応答との比較基準として用いる。ここで、Refurbish/Reuse/Subscription は新品購入を基準とした選択肢固有定数、PriceDiscount は新品価格からの割引率に対する効用、Year は製品の経年数に対する効用、Warranty は購入時に付与される無償修理保証の年数に対する効用、Scratch は傷の有無に対する効用、Reputation は評判の有無に対する効用、Capacity は容量に対する効用、×は交互作用を示す。

Tab. 1. Mean of part-worth utilities (β)

Goods	Parameters	Avg.	A	B	C	D
冷蔵庫	Refurbish	-4.29	-1.45	-0.58	-2.87	-6.59
	Reuse	-7.54	-3.34	-1.29	-6.87	-10.44
	Subscription	-8.13	4.76	-3.75	-6.94	-12.83
	PriceDiscount	4.23	7.71	8.18	6.73	1.20
	PriceDiscount ×Subscription	-0.04	-0.68	-2.17	-0.72	0.99
	Year	-5.34	-3.69	-2.93	-5.19	-6.41
	Warranty	0.16	0.24	0.13	0.17	0.15
	Scratch	-0.83	-0.47	-0.72	-0.93	-0.90
	Reputation	0.1	0.17	-0.02	0.08	0.13
Laptop	Refurbish	-1.65	-0.51	1.71	-1.17	-4.43
	Reuse	-3.31	-1.23	0.61	-2.99	-6.67
	Subscription	-4.01	1.87	-1.00	-4.63	-7.65
	PriceDiscount	2.32	2.60	6.54	3.19	-0.91
	PriceDiscount ×Subscription	-0.80	-0.91	-3.12	-1.27	0.94
	Year	-0.42	-0.39	-0.33	-0.47	-0.45
	Warranty	0.21	0.14	0.23	0.23	0.19
	Capacity	-0.49	0.62	0.73	0.61	0.22

3.2. LLM 応答生成プロトコル

本研究では、OpenAI 系 GPT-4/5(gpt-4o, gpt-5)、Google 系 Gemini (gemini-2.5-flash)、Anthropic 系 Claude Sonnet (claude-sonnet-4-20250514)、xAI 系 Grok-3 (grok-3-latest

(August, 2025))に対し、同一設問・同一手順で合成回答者としての応答を収集する。LLM は「人間の回答分布を予測するツール」として扱い、各設問につき十分な反復サンプルを取得して確率分布を推定する方針とする[10,11]。

- ロール：一般消費者
- ペルソナ：ベンチマーク調査から得た回答者属性の事前分布からサンプリング
- 設問：「新品購入」「再使用品購入（Reuse）」「メーカー改修品購入（Refurbish）」「サブスクリプション（非所有）」の 4 選択肢を表示し、単一選択とその理由を要求。
- 出力：JSON({“choice”:0-3, “reason”:”…”})
- 温度：default

3.3. 評価指標

まず、(i) 選択確率分布の整合では、4 選択肢（新品／中古／リファービッシュ／サブスク）に対する選択確率分布を、人間データ（HB-MNL による推定分布）と LLM 応答の分布で比較する。分布間距離は、Jensen–Shannon 距離（JSD）と KL divergence を用いる[14,15]。JSD は二つの確率分布の重なりを 0〜1 の範囲で測る対称・有界の指標で、極端値に過度に引っ張られにくく解釈が容易である[26]。一方、KL divergence は、人間分布を基準（真の分布）とみなし、それを LLM 分布で符号化したときの余分な情報量を表す量である[27]。JSD／KL divergence による分布比較は、Natural Language Inference (NLI) における「人間判断分布」と LLM の分布を照合する近年の研究で実証的に用いられており[14]、経済行動の文脈でも LLM と人間の選択分布の近さを JSD で評価する枠組みが報告されている[15]。

つぎに、(ii) シェアリングの主要属性に対する感度では、両製品でサブスク価格（PriceDiscount：−30%／0%／30%減額）、製造後経過年（冷蔵庫:0,2,4,8 年／ラップトップ:0,1,2,3,4 年）について、冷蔵庫では外観("As brand-new"/"With Scratch")/評判("No information"/"Good information")について、ラップトップでは Capacity(250GB/750GB)について、プロンプト内の設定を変更し、効果方向（増減の符号）と相対強度（効果の大きさ）が人間ベンチマークと一致するかを判定する。ここでは、対象属性のみ操作し、他はデフォルト値とする。これらの判断基準は、属性操作に対する反応の整合性を人間データと比較して外部妥当性を担保している先行研究[10,11]に則っている。

4. 結　果

4.1. 選択確率分布の整合性評価

(i) 選択確率分布の整合では、ベンチマークの回答者属性分布（年齢・性別・世帯人数・世帯収入・居住地域）に基づき 300 のペルソナを乱数サンプリングし、各モデルにつき 300 回答を取得した。その選択確率とベースライン（HB-MNL の全体平均）に対する距離を JSD とクロスエントロピー差の双方で評価する。これらの値は小さいほど人間の平均分布に近いと解釈できる。

冷蔵庫におけるそれぞれの選択確率分布を図 1 に示し、それらの分布間距離を表 2 に示す。また、ラップトップにおけるそれぞれの選択確率分布を図 2 に示し、それらの分布間距離を表 3 に示す。図 1 (a), 図 2 (a)内にはベースラインの全体平均分布のみならず、クラスタリングされた 4 セグメント(A,B,C,D)の消費者の選択確率分布も示している。

冷蔵庫における選択において、Gemini-2.5-Flash が両指標で最良。次点は Grok-3-Latest／GPT-5 が両指標で小さい値を示した。Claude-Sonnet-4 と GPT-4o はいずれの指標でもベースラインからの乖離が大きい。

ラップトップにおける選択において、Grok-3-Latest が両指標で最良。次点は Claude-Sonnet-4 で、Gemini-2.5-Flash は JSD では同水準であるが Claude-Sonnet-4 に比べ KL divergence がやや大きい。GPT-4o/GPT-5 は両指標とも相対的に大きい値を示した。

また、4 セグメント(A,B,C,D)と LLM の比較では、両製品を通して GPT-4o は (B)Price-sensitive と似た分布を示しているが(冷蔵庫: JSD= 0.3891, KL=0.6060; ラップトップ: JSD=0.2434, KL=0.2138)、他のモデルとの類似性は示されていない。

Tab. 2. Refrigerator - JSD and KL to the human benchmark (Average).

LLM Model	JSD	KL divergence
GPT-4o	0.6042	1.3434
GPT-5	0.4208	0.5747
Gemini-2.5-Flash	0.3354	0.3488
Claude-Sonnet-4	0.5241	0.9377
Grok-3-Latest	0.4146	0.6583

Fig. 1. Refrigerator - Choice Behavior by (a) Human & (b) LLM.

Tab. 3. Laptop - JSD and KL to the human benchmark (Average).

LLM Model	JSD	KL divergence
GPT-4o	0.4786	0.9836
GPT-5	0.4659	1.4553
Gemini-2.5-Flash	0.4166	1.1128
Claude-Sonnet-4	0.4172	0.6402
Grok-3-Latest	0.3704	0.4671

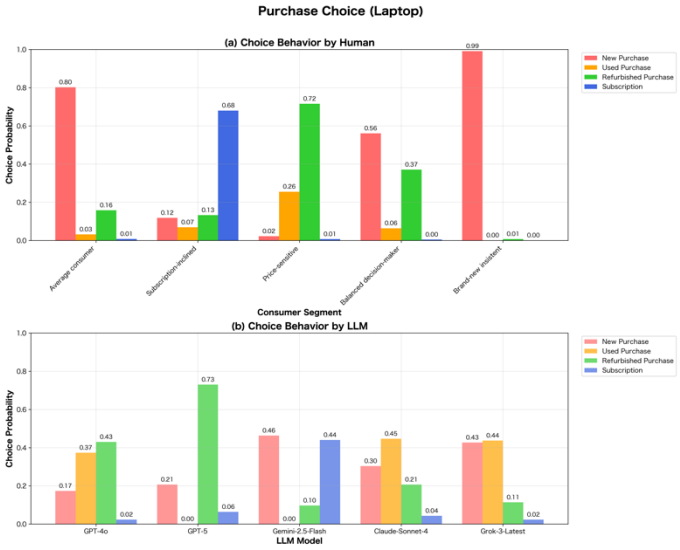

Fig. 2. Laptop - Choice Behavior by (a) Human & (b) LLM.

4. 2. シェアリングの主要属性に対する感度の検証

(ii) シェアリングの主要属性に対する感度の検証では、各変数の比較において、ベンチマークの回答者属性分布（年齢・性別・世帯人数・世帯収入・居住地域）に基づき 100 のペルソナを乱数サンプリングし、各条件 1 回答取得した。

冷蔵庫におけるシェアリングの選択確率推移を図 3 に示している。図 3(a)はシェアリング価格に対する感度を示している。ベースラインでは価格が上がるにつれて効用値が下がる一方で、LLM 両モデルにその傾向は見られなかった。図 3 (b)はシェアリング製品の経年に対する感度を示している。LLM 両モデルで 4 年から 8 年には選択確率の現象が見られ、Gemini-2.5-Flash はその相対強度も近似している。しかし、経年 0 年から 4 年にはベースラインと同様の減少傾向が LLM 両モデルから見られなかった。図 3 (c)(d)はシェアリング製品の外観(傷)/評判に対する感度を示している。どちらも LLM 両モデルにベースラインと同じ方向の効用変化が見られ、よりベースラインに近い効果を示したのは Scratch では Grok-3-Latest、Reputation では Gemini-2.5-Flash であった。これらの属性に対し、Grok-3-Latest がより過敏に反応していることが示された。

ラップトップにおけるシェアリングの選択確率推移を図 4 に示している。図 3 (a) はシェアリング価格に対する感度を示しているが、両モデルでベースラインと同じ方向の効用変化は見られなかった。図 3 (b)(c)のシェアリング製品の経年/容量において、Grok-3-Latest はベースラインと同じ方向の効用変化が見られるが、効果の大きさは近似していない。Gemini-2.5-Flash はどちらもベースラインと同じ方向の効用変化は見られなかった。

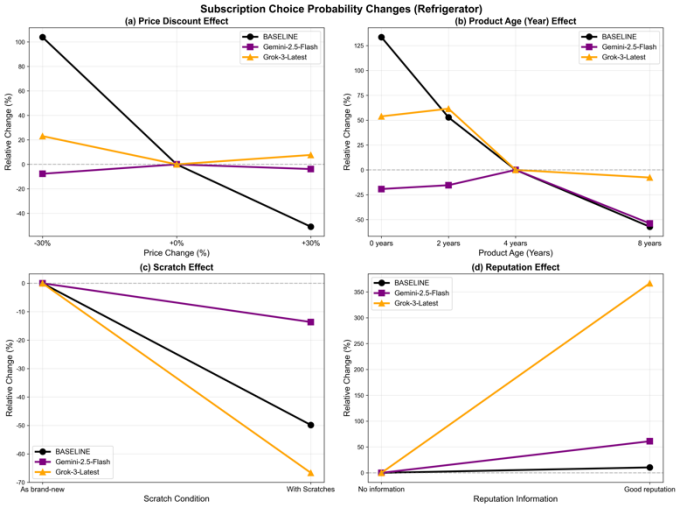

Fig. 3. Subscription choice probability changes for refrigerator purchase scenarios (relative %, baseline: p_0).
(a) PriceDiscount: vs 0% discount; (b) Product age: vs 4 years; (c) Scratch: vs no scratches; (d) Reputation: vs no information.
Relative change is $(p - p_0)/p_0 \times 100$. Within each panel, the same 100 personas answered all conditions (within-subject). Panels use distinct cohorts (100 personas per panel); one response per persona per condition.

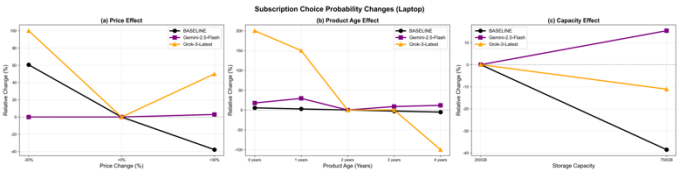

Fig. 4. Subscription choice probability changes for laptop purchase scenarios (relative %, baseline: p_0).
(a) PriceDiscount: vs 0% discount; (b) Product age: vs 2 years; (c) Capacity: vs 250GB.
Relative change is $(p - p_0)/p_0 \times 100$. Within each panel, the same 100 personas answered all conditions (within-subject). Panels use distinct cohorts (100 personas per panel); one response per persona per condition.

5. 議　論

本研究では、購入・シェアリング選好における LLM の外部妥当性を(i) 選択確率分布の整合、(ii) 属性感度(効果方向・相対強度)の整合の二点から評価した。

(i) 選択確率分布の整合では、製品の購入・シェアリング選好は LLM モデルによって大きく変わることが示された。その中でも、Grok-3-Latest, Gemini-2.5-Flash が製品カテゴリを問わず一貫して人間ベンチマークに最も近い分布を再現できていることが示された。このモデル差は、先行研究でも指摘されている指示の分かりやすさや提示順への敏感さ、出力のぶれの違いに加えて[14,15]、学習データの性質や学習後の調整のしかた（データの品質管理や多言語対応の度合いなど）が影響している可能性がある[10,11]。例えば、Grok-3-Latest は X (旧 Twitter)と連携しており、新サービスや価格に対する評価に関する学習データを多く含んでいることが関係している可能性がある。しかし、この X データは偏りが見られることが指摘されているため[28]、外部妥当性を欠く結果になるリスクも存在する。

(ii) シェアリングの主要属性に対する感度の検証では、シェアリングにおける価格と製品の経年に対し、(i)で分布の整合性で高い水準であった Grok-3-Latest, Gemini-2.5-Flash どちらもベースラインと同じ効果方向（増減の符号）と同程度の相対強度（効果の大きさ）を常に示すわけではないことが示された。特に価格/経年数/容量のような定量的な属性に対しての反応が再現できていなかった。これは先行研究でも LLM の数値的脆弱性が大規模実験で示されているように[29]、単位(月額/年額)や量的関係の誤解に起因すると考えられる。一方、定性的な属性に対しては人間ベンチマークと同じ効果方向であり、Grok-3-Latest がより過敏に反応することが示された。

本結果により、LLM による事前調査を行う際は定性的な変数であれば、選好への効果を検証するのに適していることが示された。しかし、効果方向の検証のみにとどめておいた方がよく、その効果量は一致しない可能性が高い。また、これらの結果は各モデルの学習データの詳細は公開されていないため、本設定に限定される可能性がある。今後は、Few-shot Learning[10]や retrieval-augmented generation (RAG)[10]、fine-turning[11]などの手法を用いることで外部妥当性の改善可能性を検討する。また、全く同じアンケート調査による比較を行うことで、外部妥当性の検証を重ねる必要がある。

謝　辞

本研究は、環境再生保全機構による環境研究総合推進費（JPMEERF20253M03、環境省）および日本学術振興会科学研究費補助金（23K11332）の助成を受けて実施されたものである。ここに深く感謝の意を表する。

文　献

(1) Lieder M, Rashid A. Towards circular economy implementation: a comprehensive review in context of manufacturing industry. Journal of Cleaner Production [Internet]. 2016 Mar 1;115:36–51.

(2) Bocken NMP, de Pauw I, Bakker C, van der Grinten B. Product design and business model strategies for a circular economy. Journal of Industrial and Production Engineering [Internet]. 2016 Jul 3;33(5):308–20.

(3) Hossain M. Sharing economy: A comprehensive literature review. International Journal of Hospitality Management. 2020 May 1;87.

(4) Koide R, Yamamoto H, Kimita K, Nishino N, Murakami S. Circular business cannibalization: A hierarchical Bayes conjoint analysis on reuse, refurbishment, and subscription of home appliances. Journal of Cleaner Production [Internet]. 2023 Oct 10;422:138580.

(5) Groves RM, Heeringa SG. Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. Journal of the Royal Statistical Society Series A: Statistics in Society [Internet]. 2006 Jul 1;169(3):439–57.

(6) Couper M. Web surveys: a review of issues and approaches. Public opinion quarterly [Internet]. 2000;64(4):464–94.

(7) Millar MM, Dillman DA. Improving Response to Web and Mixed-Mode Surveys. Public Opinion Quarterly [Internet]. 2011 Jun 1;75(2):249–69.

(8) Meade AW, Craig SB. Identifying careless responses in survey data. Psychological Methods [Internet]. 2012 Sep;17(3):437–55.

(9) Coast J, Horrocks S. Developing attributes and levels for discrete choice experiments using qualitative methods. Journal of Health Services Research & Policy [Internet]. 2007 Jan 1;12(1):25–30.

(10) Arora N, Chakraborty I, Nishimura Y. AI–Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators. Journal of Marketing. 2025 Mar 1;89(2):43–70.

(11) Brand J, Israeli A, Ngwe D. Using GPT for Market Research. SSRN Electronic Journal [Internet]. 2023.

(12) Gao C, Lan X, Li N, Yuan Y, Ding J, Zhou Z, et al. Large language models empowered agent-based modeling and simulation: a survey and perspectives. Humanities and Social Sciences Communications [Internet]. 2024 Sep 27;11(1):1259.

(13) Koide R, Yamamoto H, Amasawa E, Nansai K, Murakami S. Consumer preferences regarding product acquisition, repair, and discharge towards a circular economy: A segment-specific market simulation based on conjoint analysis. Journal of Environmental Management [Internet]. 2025 Sep 1;392:126806.

(14) Chen B, Wang X, Peng S, Litschko R, Korhonen A, Plank B. “Seeing the Big through the Small”: Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations? 2024 Oct 4.

(15) Gao Y, Lee D, Burtch G, Fazelpour S. Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina. 2025 Jan 23.

(16) Kirk RE. Experimental design: Procedures for the behavioral sciences, 3rd ed. Experimental design: Procedures for the behavioral sciences, 3rd ed. Belmont, CA, US: Thomson Brooks/Cole Publishing Co; 1995. 921, xiv, 921–xiv.

(17) Simester D. Field Experiments in Marketing. In 2017. p. 465–97.

(18) Viglia G, Zaefarian G, Ulqinaku A. How to design good experiments in marketing: Types, examples, and methods. Industrial Marketing Management [Internet]. 2021 Oct 1;98:193–206.

(19) Johnston RJ, Boyle KJ, Adamowicz W (Vic), Bennett J, Brouwer R, Cameron TA, et al. Contemporary Guidance for Stated Preference Studies. Journal of the Association of Environmental and Resource Economists [Internet]. 2017 Jun 1;4(2):319–405.

(20) Tukker A. Product services for a resource-efficient and circular economy – a review. Journal of Cleaner Production [Internet]. 2015 Jun 15;97:76–91.

(21) Montgomery DC. Design and analysis of experiments. John Wiley & Sons, Inc.; 2013.

(22) Bridges JFP, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research

Practices for Conjoint Analysis Task Force. Value in Health [Internet]. 2011 Jun;14(4):403–13.
(23) Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete Choice Experiments in Health Economics: A Review of the Literature. PharmacoEconomics [Internet]. 2014 Sep 9;32(9):883–902.
(24) Jia J, Yuan Z, Pan J, McNamara PE, Chen D. Decision-Making Behavior Evaluation Framework for LLMs under Uncertain Context. 2024 Nov 1.
(25) Kozlowski AC, Evans J. Simulating Subjects: The Promise and Peril of Artificial Intelligence Stand-Ins for Social Agents and Interactions. Sociological Methods & Research [Internet]. 2025 Aug 2;54(3):1017–73.
(26) Jiang Q, Deng G. Model-Free Ultra-High-Dimensional Feature Screening for Multi-Classified Response Data Based on Weighted Jensen-Shannon Divergence. Open Journal of Statistics. 2023;13(06):822–49.
(27) Kullback S, Leibler RA. On Information and Sufficiency. The Annals of Mathematical Statistics [Internet]. 1951 Mar;22(1):79–86.
(28) Philander K, Zhong YY. Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. International Journal of Hospitality Management. 2016 May 1;55:16–24.
(29) Mirzadeh I, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S, Farajtabar M. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024 Oct 7.