意思決定を特徴づける形容詞的修飾を用いた大規模言語モデルの実証的分析

馬田　光琉† 　山本　修平††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
E-mail: †s2211866@u.tsukuba.ac.jp, ††syamamoto@slis.tsukuba.ac.jp

あらまし　大規模言語モデル（LLM）を意思決定支援に応用するには，LLM の意思決定特性を明らかにし，制御する手法の確立が重要である．本研究では，リスク選好および時間選好を表す形容詞をプロンプトに含めることで，LLM の意思決定特性の変化を分析した．GPT-4o と Llama3.2 を対象とした実験の結果，GPT-4o ではリスク愛好的な形容詞により損失回避係数が上昇し，現在志向な形容詞により時間割引率が上昇するなど，形容詞の意味と一致する変化が観察された．一方，Llama3.2 では形容詞による選好制御が限定的であった．本研究は，形容詞を用いたプロンプト設計により LLM の意思決定特性を制御できる可能性を示した．

キーワード　LLM，行動経済学，リスク選好，プロスペクト理論，時間選好，双曲割引モデル

1 はじめに

1.1 研究背景

近年，大規模言語モデル（Large Language Models; LLM）は急速に発展し，文章生成や要約，対話応答，情報検索支援，コード生成など多様なタスクにおいて活用されている．さらに，意思決定のドメインにおいても LLM が活用される事例が存在する．例えば，Chiang らは，人間の集団意思決定において，「悪魔の代弁者」と呼ばれる，議論に反対意見を述べるような LLM を導入し，集団意思決定プロセスがどのように変化するのかを検証している [1]．また，Kitadai らは，大規模言語モデル駆動型マルチエージェントシミュレーション（LLMs-driven MAS）を用いて最後通牒ゲームを行い，人間の実験結果を再現する研究を行っている [2], [3]．

このような研究動向の中で，LLM を意思決定支援や行動シミュレーションに応用するためには，モデルがどのような意思決定特性を内部的に備えているのかを明らかにすることが重要である [4]．すなわち，LLM が不確実性や遅延報酬を伴う選択課題に対してどのような判断傾向を示すかを明示的に測定し，人間の行動に関する理論的枠組みとの対応関係を検討する必要がある．

行動経済学の分野では，人間の意思決定に影響を与える心理的要因をリスク選好，時間選好，社会選好という三つの観点から分析することが一般的である [5]．本研究では，個人の意思決定において特に影響を与えうるリスク選好と時間選好の二つに着目する [3]．リスク選好とは，人間が不確実性を伴う選択に対してどの程度リスクを取るか，あるいは回避するかを示す指標である [6]．リスクを取る傾向が強い場合リスク愛好的であり，逆にリスクを避ける傾向が強い場合はリスク回避的とされる．リスク回避的な意思決定は，安全性や確実性を重視する心理的傾向に基づくものであり，一方でリスク愛好的な選択は，報酬の大きさや挑戦意欲を重視する心理的傾向に基づくものである．一方，時間選好とは，人間が将来の利益を現在の利益よりもどの程度重視するかを示す指標であり [7]，将来の利益を重視する場合には割引率が低く，逆に目先の利益を優先する場合には割引率が高くなる．

近年，これらの理論枠組みを用いて，LLM が人間の意思決定パターンをどの程度再現できるかを検証する研究が行われている [4], [8], [9]．例えば，Jia らは，商用の複数モデル（ChatGPT-4.0-Turbo，Claude-3-Opus，Gemini-1.0-pro）を対象に，選択リスト形式で質問を与え，その回答を元に「リスク選好」「確率加重」「損失回避」という行動経済学の三つの要素をどの程度示すのかを測定している [4]．同様に，吉川らは，LLM に対してペルソナを与え，リスク選好や時間選好がどの程度表出されるかを測定している [8], [9]．

しかし，既存研究の多くは，LLM が元から備える選好構造を「観察的に測定する」段階にとどまっており，特定の意思決定特性をモデルに対して意図的に付与し，その影響を体系的に評価する手法は十分に検討されていない．

1.2 研究目的

以上の背景から，本研究では LLM の内部的な意思決定特性を行動経済学の理論枠組みに基づいて測定し，さらにプロンプト設計を通じて特性を付与・調整する方法を検討する．これにより，LLM が人間の意思決定特性をどの程度再現できるかを明らかにし，意思決定モデルとしての LLM の特性を実証的に分析することを目的とする．この目的を達成する方法として，本研究では Imasaka ら [10] の手法に着目した．彼らは，特性を表す形容詞をプロンプトに含めることで LLM の性格的傾向を制御できる可能性を示している．このような手法を応用すれば，リスク選好や時間選好を LLM に付与・調整することが理論的に可能であると考えられる．

本研究では実際に，GPT-4o と Llama3.2 を対象として，リ

スク選好および時間選好を表す形容詞をプロンプトに含めることで，選好がどのように変化するかをパラメータ推定によって分析した．条件として，選好を表す形容詞を一つ与えた場合と二つ与えた場合，選好の異なる形容詞を同時に与えた場合の三つの条件で実験を行った．その結果，GPT-4o ではリスク愛好的な形容詞により損失回避係数 λ が上昇し，現在志向の形容詞により時間割引率 k が上昇するなど，形容詞の意味と一致する方向へのパラメータの変化が観察された．また，同じ選好の形容詞を二つ与えても効果が累積しないこと，選好の異なる形容詞を同時に与えると片方の選好が優勢になったり，互いの選好を打ち消し合ったりすることも明らかになった．一方，Llama3.2 では多くの条件で，提示する選択肢を入れ替えた場合の回答の一致度が低く，形容詞の意味と矛盾した変化も観察され，形容詞に応じて選好を変化させる能力が LLM によって異なることが明らかになった．これらの結果から，意思決定特性を表す形容詞をプロンプトに含めることで LLM のリスク選好および時間選好を一定程度制御できることが示された．

2 関連研究

2.1 LLM の意思決定特性と行動経済学的分析

Jia ら [4] は，大規模言語モデルの意思決定特性を，行動経済学の理論に基づいて体系的に評価する枠組みを提案した．この研究では，リスク選好，損失回避，確率加重の三つの側面から LLM の判断傾向を分析し，ChatGPT-4.0-Turbo，Claude-3-Opus，Gemini-1.0-Pro の三つの商用モデルを対象に比較実験を行っている．実験では，行動経済学実験で用いられる選択リスト形式の課題を LLM に提示して得られた結果を，構築した TCN モデルを使用して，プロスペクト理論に基づくパラメータ（リスク回避度，確率加重，損失回避係数）を推定した．その結果，三つの LLM はいずれもリスク回避的な傾向を示していたが，リスク回避度・損失回避度・確率加重の度合いは LLM によって異なることが確認された．また，人口統計学的属性を付与した際の実験では，各 LLM の意思決定傾向に変化が観察された．その他にも，LLM の意思決定特性を行動経済学的視点から分析した研究が報告されている．吉川ら [8] は，LLM にペルソナとなる情報を与えた上で，プロスペクト理論に基づき，利得および損失領域におけるリスク選好の非対称性を検証し，LLM が損失回避的な傾向を示すことを明らかにした．さらに，同じ研究グループは，ペルソナを設定した LLM を対象に双曲割引モデルを用いて時間選好を分析し，報酬額や遅延時間に応じて割引率が変動する傾向を確認している [9].

2.2 LLM を用いた意思決定エージェントに関する研究

Li ら [11] は，LLM が「言葉の上で示す考え方」と「エージェントとして実際にとる行動」との間に一貫性がないことを指摘した．彼らの研究では，ある LLM が社会的属性に関する質問に対して中立的で公平な回答をする一方で，社会的属性を付与したエージェントとして意思決定を行わせると，その行動選択に偏りが現れることが確認された．Echterhoff ら [12] は，LLM が人間由来の学習データに起因して認知バイアスを示す可能性を指摘し，それを体系的に評価・緩和する枠組み「BIASBUSTER」を提案した．BIASBUSTER では，アンカリングやフレーミングなどの典型的な認知バイアスを定量的に測定し，さらに Self-Help 法により，モデル自身がプロンプトを再構成する仕組みを導入したことで，認知バイアスを低減した LLM 出力を実現することが可能となった．同様に，角田ら [13] は，LLM が人間の認知バイアスを受け継ぐことを指摘した．LLM が受け継いだ認知バイアスを軽減するため，クラウドソーシングで提案された社会的バイアス軽減手法を LLM のプロンプトに適用するアプローチを導入した．このアプローチによって，LLM の出力に含まれる認知バイアスの一部を軽減し，合理的な判断を行うようになることが示された．

2.3 本研究の位置づけ

2.1 節で述べた研究は，LLM が人間の意思決定バイアスを部分的に再現しつつも，LLM ごとに異なる特性を持つことを示している．さらに，人口統計学的属性を与えることで，LLM の意思決定傾向が変化することも明らかとなった．一方で，これらの研究は LLM にペルソナや条件を与えることでその意思決定傾向を観察する段階にとどまっており，LLM の意思決定特性を意図的に制御・操作する手法については十分に検討されていない．また，2.2 節で述べた研究は，LLM に内在するバイアスを検出したり，その影響を最小化することを目的としている．しかし，本研究では，そのようなバイアスを単に欠陥として除去するのではなく，人間が持つ特性や判断の偏りをあえて LLM に付与し，そのようなエージェントによる意思決定を再現・分析することを目的とする．

3 分析手法

3.1 手法概要

本研究では，大規模言語モデル（LLM）に対して認知的特性を表す形容詞をプロンプト内に付与し，その影響が意思決定特性（リスク選好・時間選好）にどのような変化を与えるかを観察する手法を提案する．本研究で提案する分析手法の概要を図 1 に示す．手法は大きく Step1 と Step2 から構成される．

Step1 では，金額と確率（もしくは期間）のリストおよび認知的特性を表す形容詞を用いて選択課題を生成し，LLM に複数回回答させることで選択課題タスクの結果を得る．また，選択肢の順序による回答への影響を考慮して，選択肢を入れ替えた場合も含めて合計 40 回，LLM に回答させる．

Step2 では，Step1 で得られた選択データに対してプロスペクト理論および双曲割引モデルを用いて最尤推定を行い，意思決定特性を表すパラメータを推定する．これにより，形容詞の付与によって LLM の意思決定傾向がどの程度変動するかを観察する．

今回主に使用したモデルは，gpt-4o-2024-08-06 および llama3.2:latest であり，両モデルにおいて temperature は

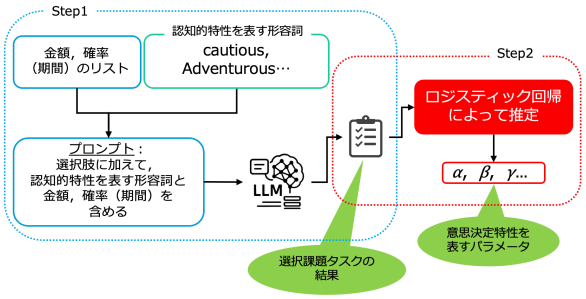

図 1　本研究で提案する分析手法の概要図

0 に設定している．[1]これ以降，gpt-4o-2024-08-06 および llama3.2:latest はそれぞれ GPT-4o および Llama3.2 と略記する．

3.2　Step1：選択課題データ生成

Step1 では，意思決定特性としてリスク選好と時間選好を取り上げ，それらを表す形容詞を用いたプロンプトを設計する．続いて，リスク選好タスクおよび時間選好タスクの選択課題を自動生成し，プロンプトとして LLM に複数回回答させることで選択課題データを取得する．

3.2.1　本研究で扱う意思決定特性の概要

本研究で対象とする意思決定特性は，リスク選好 [6] と時間選好 [7] の二種類である．リスク選好は，人間がリスクをどの程度好むか，あるいは避けるかを表す特性である．本研究では，リスク選好の水準を「リスク回避的」「中立」「リスク愛好的」の三水準として扱う．時間選好は，人間が現在の利益を将来の利益よりもどの程度重視するかを表す特性である．本研究では時間選好の水準を「現在志向」「中立」「将来志向」の三水準として扱う．

3.2.2　プロンプト設計

ここでは，Imasaka らの手法 [10] を参考に，リスク選好および時間選好を LLM に付与する．Imasaka らは，大規模言語モデル（LLM）の性格特性が検索タスクにおけるクエリ生成行動へ与える影響を分析した．この研究では，ビッグファイブ理論に基づき，協調性（Agreeableness），誠実性（Conscientiousness），外向性（Extraversion），神経症傾向（Neuroticism），開放性（Openness）の五つの性格特性を LLM に付与した上で，その出力傾向を評価している．実験では，ビッグファイブ理論に基づく各性格特性を表す形容詞を多数含むプロンプトを LLM に提示し，それぞれの性格特性を付与したエージェントを生成した．その結果，性格特性を付与した LLM は，プロンプトに与えられた形容詞的修飾に応じて生成されるクエリの長さや語彙の多様性が変化することが確認された．この Imasaka らの手法を参考に，認知的特性を表す形容詞を含めたプロンプトを，吉川らが用いたプロンプトテンプレート [8], [9] に組み込むことで，LLM にリスク選好および時間選好を付与するプロンプト設計を行う．本研究で用いる形容詞と，リスク選好タスクと時間選好タスクで提示するプロンプトをそれぞれ以下に示す．

表 1　本研究で用いた認知的特性を表す形容詞

意思決定特性	選好	形容詞
リスク選好・時間選好	中立	Moderate（中庸的な） Balanced（均衡をとった）
リスク選好	リスク愛好的	Adventurous（冒険的な） Audacious（大胆な）
	リスク回避的	Cautious（慎重な） Conservative（保守的な）
時間選好	将来志向	Patient（忍耐強い） Prudent（慎重な）
	現在志向	Impatient（せっかちな） Impulsive（衝動的な）

リスク選好タスクのプロンプトテンプレート

You have a behavioral economic characteristic of being `{neutral or low or high}` in risk aversion. To explain in detail, you are {認知的特性を表す形容詞}.
I will now ask you multiple-choice questions, so please output only the letter of your answer A or B, and nothing else.
Question: There are two ways to `{receive or lose}` the prize money. Based on your intuition and personal risk preferences, which of the following options would you choose?
A. `{receive or lose}` {金額 A} for certain.
B. {確率 p_1} chance to `{receive or lose}` {金額 B_1} or {確率 p_2} chance to `{receive or lose}` {金額 B_2}.
Answer(A or B):

時間選好タスクのプロンプトテンプレート

You have a behavioral economic characteristic of having a `{neutral or low or high}` time discount rate. To explain in detail, you are {認知的特性を表す形容詞}.
I will ask you multiple-choice questions, please answer with either SSR or LLR, and nothing else.
Question: You have won the lottery. There are two ways to receive the prize money. Based on your intuition and personal time preferences, which of the following options would you choose?
A. Receive {金額 A} after 0 days.
B. Wait until {期間 t} days later and receive {金額 $B(> A)$}.
Answer:

これらのテンプレートを用いて，使用する LLM に意思決定特性を与えた上で，リスク選好タスクでは確実に得られる，もしくは失う金額 A と不確実に得られる，もしくは失う金額 B_1, B_2 およびその確率 p_1, p_2 の組み合わせを変化させた選択課題を生成する．同様に，時間選好タスクでは即時に得られる金額 A と遅延して得られる金額 B およびその遅延期間 t の組

1：この他にも gpt-3.5-turbo-0125, llama3.1:latest, llama3.3:latest, phi4:latest, gemma3:27b, mistral:latest を使用して実験を行ったが，詳細な結果分析は gpt-4o-2024-08-06 および llama3.2:latest に絞って行う．

み合わせを変化させた選択課題を生成する．また，意思決定特性の与え方として，形容詞を一つ用いる方法，同じ選好を表す形容詞を二つ用いる方法，選好の異なる形容詞を同時に用いる方法の三通りを検討する．

3.2.3 選択課題タスク

選択課題タスクの生成では 3.2.2 節で示したプロンプトテンプレートに基づき，Python によりリスク選好タスクおよび時間選好タスクの選択課題データを自動生成している．

まず，リスク選好タスクについて説明する．本研究におけるリスク選好タスクは，確実に得られる，もしくは失う金額 A と不確実に得られる，もしくは失う二つの金額 B_1, B_2 およびその確率 p_1, p_2 の組み合わせを変化させた選択課題を生成し，LLM がどちらの選択を行うかを観察するものである．ここで用いる金額リストと確率リスト [8] は以下の通りである．

- 金額リスト（単位：$）:
  `[100, 500, 750, 1000, 2500, 5000, 7500, 10000]`

- 確率リスト:
  `[0.01, 0.05, 0.10, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.75, 0.80, 0.90, 0.95, 0.99]`

これらのリストから，不確実な選択肢 B に対応する二つの金額 B_1, B_2 と確率 p_1，および p_2 （$= 1 - p_1$）を抽出する．次に，確実な選択肢 A の金額 A は，選択肢 B の期待値

$$\mathrm{EV}_B = p_1 \times B_1 + p_2 \times B_2$$

に 1.0〜1.01 の範囲の乱数係数を掛けて決定している．このようにして，金額が得られる選択課題タスク，金額を失う選択課題タスクをそれぞれ 10 問ずつ，合計 20 問生成し，選択肢の順序を入れ替えたものも含めて合計 40 問のリスク選好タスクがプロンプトテンプレート（3.2.2 節）に挿入され，LLM に提示される．

同様に，時間選好タスクについても，Python により選択課題データを自動生成している．本研究における時間選好タスクは，即時に得られる金額（SSR: Short-Soon Reward）と遅延して得られる金額（LLR: Long-Late Reward）およびその遅延期間の組み合わせを変化させた選択課題を生成し，LLM がどちらの選択を行うかを観察するものである．ここで用いる金額リストと遅延期間リスト [9] は以下の通りである．

- 即時報酬リスト（単位：$）:
  `[100, 500, 1000, 5000, 10000]`

- 報酬比率リスト：
  `[1.2, 1.5, 2.0]`

- 遅延期間リスト（単位：day）:
  `[1, 7, 30, 180, 365]`

これらのリストから，即時報酬 A，報酬比率 r，遅延期間 t を抽出し，遅延報酬 B を

$$B = r \times A$$

として計算することで，選択課題タスクを生成する．このようにして，選択課題タスクを 20 問生成し，選択肢の順序を入れ替えたものも含めて合計 40 問の時間選好タスクがプロンプトテンプレート（3.2.2 節）に挿入され，LLM に提示される．

3.3 Step2：パラメータ推定

Step1 で得られたリスク選好タスクおよび時間選好タスクの選択データに対して，Step2 ではプロスペクト理論の価値関数，確率加重関数および双曲割引モデルを用いて，LLM の意思決定特性を表すパラメータを最尤推定する．推定されたパラメータを比較することで，形容詞の付与によって LLM の意思決定傾向がどの程度変動するかを観察する．

3.3.1 プロスペクト理論によるリスク選好の推定

リスク選好タスクに対しては，プロスペクト理論 [14] に基づき，利得と損失に対する価値関数と確率加重関数を用いたモデルを採用する．金額 x に対する価値関数 $v(x)$ は，利得と損失で異なるべき乗型の関数として次のように定義する．

$$v(x) = \begin{cases} x^{\alpha} & (x \geqq 0) \\ -\lambda(-x)^{\beta} & (x < 0) \end{cases} \tag{1}$$

ここで，α は利得の感応度，β は損失の感応度，λ は損失回避係数を表すパラメータである．一般に $0 < \alpha, \beta \leqq 1$ のとき，利得・損失のいずれの領域においても，金額の絶対値が大きくなるほど主観的価値の増加幅が小さくなるという限界効用逓減 [15] の性質を示す．また，$\lambda > 1$ のとき，損失は同額の利得よりも大きく評価される．

確率 $p \in [0, 1]$ に対しては，確率加重関数を用いて，単一のパラメータ γ をもつ次の関数を用いる．

$$w(p) = \frac{p^{\gamma}}{\{p^{\gamma} + (1-p)^{\gamma}\}^{1/\gamma}} \tag{2}$$

$\gamma < 1$ のとき，実際に発生する確率が小さい事象の確率が過大評価され，実際に発生する確率が大きい事象の確率が過小評価される．一方，$\gamma = 1$ のとき $w(p) = p$ となり，実際に発生する確率と一致する．

3.2.2 節のプロンプトテンプレートで示した通り，本研究のリスク選好タスクでは，確実に金額 A を得る選択肢 A と，確率 p_1 で金額 B_1，確率 p_2 で金額 B_2 を得る選択肢 B の二択が提示される．40 回の回答のうち，i 回目に提示される金額と確率 (A, B_1, B_2, p_1, p_2) をそれぞれ，

$$q_i = (A_i, B_i^+, B_i^-, p_i^+, p_i^-) \tag{3}$$

と書く．このとき，人間の金額に対する主観的価値を表す効用関数 $U_A(q_i), U_B(q_i)$ は

$$U_A(q_i) = v(A_i) \tag{4}$$

$$U_B(q_i) = w(p_i^+)\, v(B_i^+) + w(p_i^-)\, v(B_i^-) \tag{5}$$

と表される．

次に，LLM の選択データと効用関数との関係を定式化する．このとき，i 回目の質問において，選択肢 A が選ばれる確率を μ_i とおき，次のロジスティックモデルで定義する．

$$\mu_i = \frac{1}{1 + \exp\{-\tau\,(U_A(q_i) - U_B(q_i))\}} \tag{6}$$

μ_i は選択肢 A と B の主観的価値の差に応じて変化すると考える．一般に，$U_A(q_i)$ が $U_B(q_i)$ より大きいほど選択肢 A が選ばれやすく，両者がほぼ同程度であれば選択がランダムになると想定できる．ここで $\tau > 0$ はロジスティックモデルの傾きを表すパラメータであり，τ が大きいほど主観的価値の差が過剰に評価され，μ_i は 0 か 1 のどちらかにより近づく．一方，τ が小さいほど主観的価値の差はそれほど評価されなくなり，選択がランダム（$\mu_i \approx 0.5$）に近づく．

質問回数 $i = 1, \ldots, N$ において LLM が選択肢 A を選んだ場合を $y_i = 1$，選択肢 B を選んだ場合を $y_i = 0$ と定義し，このように定義した μ_i をパラメータとするベルヌーイ分布に従うと仮定すると，

$$y_i \sim \mathrm{Bern}(\mu_i) \tag{7}$$

が成り立つ．したがって，N 回の質問に対する回答系列 $y_1, \ldots, y_N$ が得られる確率は

$$P(y_1, \ldots, y_N) = \prod_{i=1}^{N} \mu_i^{y_i} (1 - \mu_i)^{1 - y_i} \tag{8}$$

となる．

パラメータ集合 $\theta = (\alpha, \beta, \lambda, \gamma, \tau)$ に対する対数尤度は

$$\log L(y_1, \ldots, y_N \mid \theta) = \sum_{i=1}^{N} \{y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)\} \tag{9}$$

と書ける．本研究では，この対数尤度を最大化する制約付き最適化問題を解くことでパラメータ θ を推定する．

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \log L(y_1, \ldots, y_N \mid \theta) \\ \text{s.t.}\ \ & 0 < \alpha,\ 0 < \beta,\ 1 \leqq \lambda,\ 0 \leqq \gamma \leqq 1,\ 0 < \tau \end{aligned} \tag{10}$$

ここで $\hat{\theta}$ を最尤推定値として用いる．最尤推定では，Python の数値最適化ライブラリである `scipy.optimize` を用いて，負の対数尤度 $-\log L$ を最小化する形で $\hat{\theta} = (\alpha, \beta, \lambda, \gamma, \tau)$ を求める．推定された $\hat{\theta}$ のうち，$\alpha, \beta, \lambda, \gamma$ を比較することで，プロンプトに付与した形容詞が LLM の意思決定特性に与える影響を分析する．

3.3.2 双曲割引モデルによる時間選好の推定

時間選好タスクに対しては，遅延報酬の主観的価値が時間とともに減衰すると仮定する双曲割引モデル [7] を用いる．金額 x を遅延時間 t 日後に受け取るときの主観的価値 $V(x,t)$ を

$$V(x,t) = \frac{x}{1 + kt} \tag{11}$$

と定義する．ここで，$k \geqq 0$ は割引率を表すパラメータであり，k が大きいほど将来の報酬を強く割り引き，現在志向的な時間選好を示す．

3.2.2 節のプロンプトテンプレートで示した通り，本研究の時間選好タスクでは，即時に金額 A を受け取る選択肢 SSR と，t 日後に金額 B を受け取る選択肢 LLR の二択が提示される．40 回の回答のうち，i 回目に提示される即時報酬・遅延報酬および遅延期間 (A, B, t) を，それぞれ $x_i^{\mathrm{SSR}}, x_i^{\mathrm{LLR}}, t_i^{\mathrm{LLR}}$ と書き，即時報酬の受け取り時点は $t_i^{\mathrm{SSR}} = 0$ とおく．このときのパラメータ集合を

$$q_i = \left(x_i^{\mathrm{SSR}}, x_i^{\mathrm{LLR}}, t_i^{\mathrm{SSR}}, t_i^{\mathrm{LLR}}\right) \tag{12}$$

と表す．

双曲割引モデルに基づく各選択肢の主観的価値は，式 (11) を用いて

$$V_{\mathrm{SSR}}(q_i) = V\left(x_i^{\mathrm{SSR}}, t_i^{\mathrm{SSR}}\right) = x_i^{\mathrm{SSR}} \tag{13}$$

$$V_{\mathrm{LLR}}(q_i) = V\left(x_i^{\mathrm{LLR}}, t_i^{\mathrm{LLR}}\right) = \frac{x_i^{\mathrm{LLR}}}{1 + k t_i^{\mathrm{LLR}}} \tag{14}$$

と表される．

以降は，リスク選好タスクと同様の手続きでパラメータを推定した．すなわち，i 回目の質問において即時報酬（SSR）を選択する確率 μ_i をロジスティックモデルで定義し，ベルヌーイ分布に基づく対数尤度を最大化する制約付き最適化問題を解くことで，パラメータ $\theta = (k, \tau)$ を推定する．推定されたパラメータのうち，割引率 k を比較することで，プロンプトに付与した時間選好に関する形容詞が LLM の意思決定特性に与える影響の違いを評価する．

4 リスク選好に関する実験結果と考察

本章では，GPT-4o および Llama3.2 に対して，形容詞を用いた場合のリスク選好のパラメータ推定の結果を示すとともに，その考察を行う．なお，一致度は選択肢の順序を入れ替えても同じ選択肢を選んだ割合を示し，有効回答率はモデルが有効な回答を返した割合を示す．ここでいう有効な回答とは，「A」「B」のいずれかを選択する回答を指す．

4.1 動いたパラメータから考えられること

表 2 に，形容詞を一つ用いた場合のパラメータ推定の結果を示す．表 2 において，GPT-4o にリスク愛好的な形容詞（Adventurous/Audacious）を与えた際に λ が 1.000 から 5.000 へ上昇した．$\lambda = 5.000$ への上昇は，モデルが損失回避的な傾向を強めたことを示している．これは一見，リスク愛好的な形容詞の意味と矛盾するように見える．しかし，リスク愛好的な行動は「利得領域でリスクを取る」ことと「損失領域で損失を強く回避する」ことの両方を含みうる．本実験の結果は，GPT-4o がリスク愛好的な形容詞を後者の意味で解釈した可能性を示唆しており，具体的には，損失領域の選択課題タスクにおいて，たとえリスクを取ってでも損失額を減らしたいという思考が働いたと考えられる．また表 2 において，GPT-4o にリスク愛好的な形容詞を与えると，α が 0.993 から 1.058 へ増加し，利得額の増加に対して主観的価値がより大きく増加するよ

表 2　形容詞を一つ用いた場合のパラメータ推定の結果

		gpt-4o-2024-08-06						llama3.2:latest					
リスク選好	形容詞	α	β	λ	γ	負の対数尤度	一致度	α	β	λ	γ	負の対数尤度	一致度
-	なし	0.993	1.001	1.000	0.997	43.371	85%	1.002	1.004	1.000	1.000	42.032	0%
中立	Moderate	0.968	1.017	1.000	1.000	38.434	90%	1.002	1.004	1.000	1.000	42.032	0%
	Balanced	0.968	1.111	1.000	1.000	42.815	95%	1.002	1.004	1.000	1.000	42.032	0%
リスク愛好的	Adventurous	1.058	0.977	5.000	1.000	19.956	100%	1.002	0.998	5.000	1.000	20.834	55%
	Audacious	1.058	0.977	5.000	1.000	19.956	100%	1.002	1.004	1.000	1.000	42.032	0%
リスク回避的	Cautious	0.968	1.110	1.000	1.000	44.210	100%	1.002	1.004	1.000	1.000	42.032	0%
	Conservative	0.968	1.110	1.000	1.000	44.210	100%	1.002	1.004	1.000	1.000	42.032	0%
-	人間の例 [16]	0.859	0.826	1.576	0.605	-	-	0.859	0.826	1.576	0.605	-	-

注）いずれの実験結果も有効回答率は 100%であった.

表 3　形容詞を二つ用いた場合のパラメータ推定の結果

		gpt-4o-2024-08-06						llama3.2:latest					
リスク選好	形容詞	α	β	λ	γ	負の対数尤度	一致度	α	β	λ	γ	負の対数尤度	一致度
-	なし	0.993	1.001	1.000	0.997	43.371	85%	1.002	1.004	1.000	1.000	42.032	0%
中立	Moderate + Balanced	0.968	1.111	1.000	1.000	42.815	95%	1.002	1.004	1.000	1.000	42.032	0%
リスク愛好的	Adventurous + Audacious	1.058	0.977	5.000	1.000	19.956	100%	1.002	0.999	1.000	1.000	26.981	35%
リスク回避的	Cautious + Conservative	0.968	1.110	1.000	1.000	44.210	100%	1.002	1.004	1.000	1.000	42.032	0%

注）いずれの実験結果も有効回答率は 100%であった.

表 4　選好の異なる形容詞を同時に用いた場合のパラメータ推定の結果

	gpt-4o-2024-08-06						llama3.2:latest					
形容詞	α	β	λ	γ	負の対数尤度	一致度	α	β	λ	γ	負の対数尤度	一致度
なし	0.993	1.001	1.000	0.997	43.371	85%	1.002	1.004	1.000	1.000	42.032	0%
Adventurous + Cautious	0.968	1.002	1.000	1.000	25.967	85%	1.002	1.004	1.000	1.000	27.904	15%
Adventurous + Conservative	0.950	0.995	1.000	0.988	17.324	75%	1.002	1.004	1.000	1.000	41.365	5%
Audacious + Cautious	0.968	1.003	1.000	1.000	26.288	70%	1.002	1.004	1.000	1.000	41.365	5%
Audacious + Conservative	0.968	1.004	1.000	1.000	25.198	55%	1.002	1.004	1.000	1.000	42.032	0%

注）いずれの実験結果も有効回答率は 100%であった.

うになったことを表している．また，β が 1.001 から 0.977 へ低下し，損失額が増加しても主観的な損失感があまり増えないようになったことを表している．一方，表 2 において，中立な形容詞（Moderate/Balanced）およびリスク回避的な形容詞（Cautious/Conservative）を与えた場合は $\lambda = 1.000$ が維持された．これは $\lambda \geqq 1$ という制約があるため，リスク回避的な形容詞を与えることによってさらに低下する余地がなかった．ただし，β は 1.017 ∼ 1.111 へ上昇しており，損失額の増加に対して主観的な損失感がより大きく増加するようになったことを意味する．γ はほぼ全ての条件で 1.000 を維持し，形容詞を与えても確率の主観的評価には影響を与えにくいことが示された．

表 2 において，Llama3.2 では，Adventurous 単独の場合を除き，ほぼ全ての条件でパラメータが形容詞を与えない場合（$\alpha \approx 1$, $\beta \approx 1$, $\lambda = 1$, $\gamma = 1$）と変化がなかった．これらの値はリスク中立的な選好を表しており，Llama3.2 が形容詞に応じて選好を変化させる能力が限定的であることを示している．唯一 Adventurous で $\lambda = 5.000$ が推定されたことは，この形容詞が Llama3.2 にとって特に強い影響を及ぼした可能性がある．

4.2　形容詞を一つ用いた場合と二つ用いた場合の違い

表 3 に形容詞を二つ用いた場合のパラメータ推定の結果を示す．表 2 と表 3 を比較すると，GPT-4o において，形容詞を一つ用いた場合と二つ用いた場合（同一選好の形容詞を二つ組み合わせた場合）では，推定されるパラメータに大きな差は見られなかった．例えば，表 2 の Adventurous 単独の場合と表 3 の Adventurous + Audacious の場合で，いずれも $\alpha = 1.058$, $\beta = 0.977$, $\lambda = 5.000$, $\gamma = 1.000$ が推定された．このことから，同一選好の形容詞を重ねても効果が累積しないことが考えられる．

表 4 に，選好の異なる形容詞を同時に用いた場合のパラメータ推定の結果を示す．表 4 において，選好の異なる形容詞を組み合わせた場合では，リスク愛好的な形容詞とリスク回避的な形容詞を組み合わせても $\lambda = 1.000$ が維持され，α は 0.950 ∼ 0.968，β は 0.995 ∼ 1.004 の範囲に収まった．このことから，相反する選好の形容詞が相互に打ち消し合い，結果として中立に近い選好が表現されたと解釈できる．

Llama3.2 では，表 2，3，4 のいずれにおいても，Adventurous 単独の場合を除きパラメータが形容詞を与えない場合の値

付近に留まった．形容詞の数や組合せによる系統的な差は観察されなかった．

4.3 人間との比較

人間の推定値の一例（$\alpha = 0.859$, $\beta = 0.826$, $\lambda = 1.576$, $\gamma = 0.605$）[16] を参考として，本実験の結果と比較する．この被験者は，$\alpha = 0.859$ および $\beta = 0.826$ といずれも 1 未満であり，利得・損失の両領域で限界効用逓減の特徴を持つ．また，$\lambda = 1.576$ で損失回避的（損失を利得の約 1.6 倍重く評価する），$\gamma = 0.605$ で確率加重バイアスを持つという特徴を示している．表 2 において，形容詞を与えない場合，両モデルは $\alpha \approx 1$, $\beta \approx 1$, $\lambda = 1.000$, $\gamma \approx 1$ であり，これは，この被験者が限界効用逓減や損失回避といった特徴を持つのに対し，両モデルはリスク中立的な選好を示している点で異なる．表 2 において，リスク愛好的な形容詞を与えた場合，GPT-4o では $\alpha = 1.058$, $\beta = 0.977$, $\lambda = 5.000$ となった．$\lambda = 5.000$ はこの被験者（$\lambda = 1.576$）よりも大きく，損失を利得の 5 倍重く評価するという強い損失回避を示している．また，この被験者は $\alpha = 0.859$, $\beta = 0.826$ と両領域で限界効用逓減の特徴を持っているのに対して，GPT-4o では $\alpha > 1$ で利得に対する感応度が高く，$\beta < 1$ で損失に対する感応度が低いという，人間とは異なる特徴を示している．表 2 においてリスク回避的な形容詞を与えた場合，GPT-4o では $\beta = 1.110$ となった．$\beta > 1$ は損失に対する感応度が高いことを示し，損失額の増加に対して主観的な損失感がより大きく増加することを意味する．これに対し，この被験者は $\beta = 0.826 < 1$ であり，損失に対する感応度が低いことを示している．γ については，表 2 において，いずれの条件でも $\gamma \approx 1.000$ であり，確率に対する主観的価値は客観的価値と一致している．この被験者は $\gamma = 0.605$ であり，小さな確率を過大評価し大きな確率を過小評価する傾向を持つが，両モデルではこのような確率加重バイアスは観察されなかった．

5 時間選好に関する実験結果と考察

本章では，GPT-4o および Llama3.2 に対して，形容詞を用いた場合の時間選好のパラメータ推定の結果を示すとともに，その考察を行う．なお，一致度は選択肢の順序を入れ替えても同じ選択肢を選んだ割合を示し，有効回答率はモデルが有効な回答を返した割合を示す．ここでいう有効な回答とは，「SSR」「LLR」のいずれかを選択する回答を指す．

5.1 動いたパラメータから考えられること

表 5 に，形容詞を一つ用いた場合のパラメータ推定の結果を示す．表 5 において，GPT-4o に現在志向の形容詞（Impatient/Impulsive）を与えた際に k が 0.010 から 3.666 へ上昇した．$k = 3.666$ への上昇は，モデルが「今すぐ報酬を得たい」という傾向を強めたことを示しており，Impatient（せっかちな）や Impulsive（衝動的な）という形容詞の意味と一致する変化である．一方，表 5 において，中立な形容詞（Moderate/Balanced）および将来志向の形容詞（Prudent/Patient）を与えた場合は $k = 0.010$ が維持された．これは形容詞を与えない場合にすでに $k = 0.010$ であったため，将来志向の形容詞によるさらなる低下の余地がなかったと考えられる．これは，GPT-4o のデフォルトの時間選好がすでに強い将来志向であることを示唆している．

表 5 において，Llama3.2 では，形容詞を与えない場合で $k = 3.664$ であり，GPT-4o とは異なり，デフォルトの時間選好が現在志向であった．また，中立の形容詞（Moderate/Balanced）を与えた場合でも $k = 3.664$ が維持された．将来志向の形容詞（Prudent/Patient）を与えた場合は k が $0.069 \sim 0.500$ へ低下し，将来志向への変化が見られた．これは，形容詞の意味と一致する方向への変化であり，Llama3.2 が将来志向の形容詞を適切に解釈できたことを示している．しかし，Llama3.2 では現在志向の形容詞（Impatient/Impulsive）を与えた場合に $k = 0.500$ となり，形容詞を与えない場合（$k = 3.664$）よりむしろ低下するという逆転現象が観察された．これは形容詞の意味と矛盾する変化であり，Llama3.2 が現在志向の形容詞を正しく解釈できていなかった可能性がある．

5.2 形容詞を一つ用いた場合と二つ用いた場合の違い

表 6 に，形容詞を二つ用いた場合のパラメータ推定の結果を示す．表 5 と表 6 を比較すると，GPT-4o では，形容詞を一つ用いた場合と二つ用いた場合（同一選好の形容詞を二つ組み合わせた場合）で推定される k に差は見られなかった．中立・将来志向の形容詞では $k = 0.010$，現在志向の形容詞では $k = 3.666$ となり，形容詞の数によってより強い選好は現れなかった．

表 7 に，選好の異なる形容詞を同時に用いた場合のパラメータ推定の結果を示す．表 7 において，選好の異なる形容詞を組み合わせた場合では，Patient + Impatient および Patient + Impulsive で $k = 0.010$，Prudent + Impatient で $k = 3.666$，Prudent + Impulsive で $k = 0.500$ となった．Patient + Impatient および Patient + Impulsive では将来志向の形容詞（Patient）が優勢に働き，Prudent + Impatient では現在志向の形容詞（Impatient）が優勢に働いた．Prudent + Impulsive では $k = 0.500$ と中間的な値となり，両者の選好が打ち消し合い，結果として中立に近い選好が表現されたと解釈できる．

表 6 において，Llama3.2 では，形容詞を二つ用いたときに形容詞を与えない場合（$k = 3.664$）から大きく変化する傾向が見られた．中立の形容詞を二つ与えた場合は $k = 0.501$，将来志向の形容詞を二つ与えた場合は $k = 0.010$，現在志向の形容詞を二つ与えた場合は $k = 0.500$ となった．いずれの場合も形容詞を一つ与えたとき以上に将来志向的な値へと変化しており，形容詞を二つ与えること自体が応答パターンに影響を与えている可能性がある．表 7 において，選好の異なる形容詞を組み合わせた場合では，Patient + Impatient，Patient + Impulsive，Prudent + Impatient で $k = 0.010$，Prudent + Impulsive で $k = 0.500$ となった．表 5 において，形容詞単体では，Patient で $k = 0.500$，Prudent で $k = 0.069$，Impatient および Impulsive で $k = 0.500$ であったが，表 7 の Patient + Impatient の組み合わせでは $k = 0.010$ となり，Patient 単体（$k = 0.500$）や Impatient 単体（$k = 0.500$）よりもさらに

表 5　形容詞を一つ用いた場合のパラメータ推定の結果

時間選好	形容詞	gpt-4o-2024-08-06 k	負の対数尤度	一致度	llama3.2:latest k	負の対数尤度	一致度
-	なし	0.010	90.851	50%	3.664	20.723	95%
中立	Moderate	0.010	379.983	100%	3.664	20.723	95%
	Balanced	0.010	379.983	100%	3.664	20.723	95%
将来志向	Prudent	0.010	379.983	100%	0.069	87.842	55%
	Patient	0.010	379.983	100%	0.500	85.229	60%
現在志向	Impatient	3.666	0.000	100%	0.500	23.059	75%
	Impulsive	3.666	0.000	100%	0.500	23.059	75%
-	人間の例 [17]	2.6 ± 0.94	-	-	2.6 ± 0.94	-	-

注）いずれの実験結果も有効回答率は 100%であった.

表 6　形容詞を二つ用いた場合のパラメータ推定の結果

時間選好	形容詞	gpt-4o-2024-08-06 k	負の対数尤度	一致度	llama3.2:latest k	負の対数尤度	一致度
-	なし	0.010	90.851	50%	3.664	20.723	95%
中立	Moderate + Balanced	0.010	379.983	100%	0.501	44.247	80%
将来志向	Prudent + Patient	0.010	379.983	100%	0.010	229.475	15%
現在志向	Impatient + Impulsive	3.666	0.000	100%	0.500	24.882	85%

注）いずれの実験結果も有効回答率は 100%であった.

表 7　選好の異なる形容詞を同時に用いた場合のパラメータ推定の結果

形容詞	gpt-4o-2024-08-06 k	負の対数尤度	一致度	llama3.2:latest k	負の対数尤度	一致度
なし	0.010	90.851	50%	3.664	20.723	95%
Patient + Impatient	0.010	206.224	40%	0.010	200.195	30%
Patient + Impulsive	0.010	229.288	55%	0.010	291.644	0%
Prudent + Impatient	3.666	0.000	100%	0.010	161.606	35%
Prudent + Impulsive	0.500	169.941	55%	0.500	64.506	65%

注）いずれの実験結果も有効回答率は 100%であった.

将来志向的な値が推定された．同様に，Patient + Impulsive（$k = 0.010$）や Prudent + Impatient（$k = 0.010$）でも，各形容詞を単体で与えた場合よりも極端な将来志向の値となった．一方，Prudent + Impulsive では $k = 0.500$ となり，Prudent 単体（$k = 0.069$）よりも現在志向的な値へと変化した．これらの結果から，Llama3.2 では異なる選好の形容詞を組み合わせた場合に，単体で与えた場合とは異なる相互作用が生じているものの，組み合わせによって予測困難な変化が生じており，形容詞の意味を適切に解釈した上での打ち消し合いとは言い難い．

5.3　人間との比較

人間の推定値の一例（$k = 2.6 \pm 0.94$）[17] を参考として，本実験の結果と比較する．この被験者は $k = 1.7 \sim 3.5$ 程度の時間選好を示している．表 5 において，形容詞を与えない場合，GPT-4o は $k = 0.010$，Llama3.2 は $k = 3.664$ であった．GPT-4o はこの被験者に対し，極めて将来志向的である点で大きく異なる．一方，Llama3.2 はこの被験者の範囲（$k = 1.7 \sim 3.5$ 程度）に近い値であり，比較的近い特徴を示している．表 5 において，現在志向の形容詞を与えた場合，GPT-4o では $k = 3.666$ となり，この被験者の範囲に近い値へと変化した．Llama3.2 では $k = 0.500$ となり，この被験者よりも将来志向的な値であった．これは形容詞を与えない場合（$k = 3.664$）から低下しており，形容詞の意味とは矛盾する変化である．表 5 において，将来志向の形容詞を与えた場合，GPT-4o では $k = 0.010$ が維持された．元々極めて将来志向的であったため，さらなる変化の余地がなかったと考えられる．Llama3.2 では $k = 0.069 \sim 0.500$ となり，形容詞の意味と一致する将来志向への変化が見られた．ただし，いずれの値もこの被験者の範囲（$k = 1.7 \sim 3.5$ 程度）からは乖離している．

6　失敗分析

6.1　Llama3.2 の回答の一致度の低さ

Llama3.2 のリスク選好実験では，多くの条件で一致度が 0%となった．これは，選択肢の提示順序を入れ替えると異なる選択肢を選ぶことを意味しており，回答が選択肢の順序に影響されていたことが考えられる．

6.2　Llama3.2 におけるリスク選好パラメータの非変化

Llama3.2 のリスク選好実験では，Adventurous 単独の場合

を除き，ほぼ全ての条件でパラメータが $\alpha \approx 1$, $\beta \approx 1$, $\lambda = 1.000$, $\gamma = 1.000$ のまま変化しなかった．中立な形容詞，リスク回避的な形容詞，リスク愛好的な形容詞（Audacious 以外），および形容詞を二つ与えた場合や選好の異なる形容詞を組み合わせた場合のいずれにおいても，パラメータは形容詞を与えない場合の値付近に留まった．Llama3.2 のリスク選好実験では，多くの条件で一致度が 0%となったため，正しく選択肢を評価できていなかった可能性もあるが，Llama3.2 が形容詞に応じてリスク選好を変化させる能力が限定的であることを示している．

6.3 Llama3.2 における形容詞の意味と矛盾した変化

Llama3.2 の時間選好実験では，現在志向の形容詞（Impatient/Impulsive）を与えた場合に $k = 0.500$ となり，形容詞を与えない場合（$k = 3.664$）よりもむしろ将来志向的な値へと変化した．これは現在志向の形容詞の意味とは矛盾する変化であり，Llama3.2 が現在志向の形容詞を正しく解釈できていなかった可能性を示している．また，選好の異なる形容詞を組み合わせた場合（Patient + Impatient，Patient + Impulsive，Prudent + Impatient）でも $k = 0.010$ と極端に将来志向的な値が推定され，GPT-4o で観察されたような選好の打ち消し合いは見られなかった．これらの結果から，Llama3.2 では形容詞の意味を適切に解釈して選好を変化させる能力が限定的であることが考えられる．

6.4 形容詞を与えてもパラメータが変化しなかった場合

GPT-4o のリスク選好実験では，リスク愛好的な形容詞（Adventurous/Audacious）が λ を 1.000 から 5.000 へ上昇させた一方，リスク回避的な形容詞（Cautious/Conservative）は $\lambda = 1.000$ を維持した．リスク回避的な形容詞は損失回避を強める方向に作用することが期待されたが，そのような変化は観察されなかった．これは $\lambda \geqq 1$ という制約により，形容詞を与えない場合の $\lambda = 1.000$ からさらに低下する余地がなかったためである．同様に，時間選好実験でも，形容詞を与えない場合で $k = 0.010$ と極めて将来志向的であったため，将来志向の形容詞（Prudent/Patient）によるさらなる変化の余地がなかった．

7 本研究の限界点

本研究にはいくつかの限界点がある．第一に，使用した LLM と形容詞が限定的である点である．本研究では，付録も含めると 8 種類の LLM とリスク選好および時間選好に関してそれぞれ 6 種類の形容詞を使用したが，より多様な LLM や形容詞を検討することで，より包括的な知見を得られる可能性がある [9]．第二に，パラメータの推定範囲に関する限界である．本研究では $\lambda \geqq 1$ という制約を設けたため，形容詞を与えない場合に $\lambda = 1.000$ であった GPT-4o では，リスク回避的な形容詞による λ の低下を観察できなかった．同様に，時間選好パラメータ k についても，GPT-4o の形容詞を与えない場合が $k = 0.010$ と推定範囲の下限付近であったため，将来志向の形容詞による変化を観察できなかった．これらの問題を回避するためには，形容詞を与えない場合のパラメータが推定範囲の境界付近に位置しない LLM を選択するか，異なるプロンプト設計によってベースラインのパラメータを調整することが考えられる．

8 おわりに

本研究では，大規模言語モデル（LLM）の内部的な意思決定特性を行動経済学の理論枠組みに基づいて測定し，さらにプロンプト設計を通じて特性を付与・調整する方法を検討した．具体的には，GPT-4o と Llama3.2 を対象に，リスク選好および時間選好を表す形容詞をプロンプトに含めることで，選好がどのように変化するかをパラメータ推定によって分析した．

リスク選好実験では，GPT-4o ではリスク愛好的な形容詞（Adventurous/Audacious）により λ が 1.000 から 5.000 へ上昇し，損失回避傾向が強化された．リスク回避的な形容詞（Cautious/Conservative）では $\lambda = 1.000$ が維持されたが，β が上昇し損失への感応度が高まった．Llama3.2 では一部の条件でのみパラメータが変化し，形容詞に応じて選好を変化させる能力が限定的であった．形容詞を二つ与えた場合，GPT-4o では効果の累積が見られず，選好の異なる形容詞を組み合わせると相互に打ち消し合い中立に近い選好が表現された．

時間選好実験では，GPT-4o では現在志向の形容詞（Impatient/Impulsive）により k が 0.010 から 3.666 へ上昇し，形容詞の意味と一致する変化が観察された．Llama3.2 では将来志向の形容詞により k が低下したが，現在志向の形容詞でも k が低下し，形容詞の意味と矛盾した変化が見られた．形容詞を二つ与えた場合，GPT-4o では効果の累積が見られなかった一方，Llama3.2 では応答パターンに予測困難な変化が生じた．選好の異なる形容詞を組み合わせた場合，GPT-4o では条件によって異なる形容詞が優勢に働くか，または相互に打ち消し合う結果となった．

本研究ではいくつかの課題も明らかになった．Llama3.2 のリスク選好実験では，多くの条件で一致度が 0%となり，選択肢の提示順序に回答が影響されていたほか，形容詞を与えても選好パラメータがほとんど変化せず，形容詞に応じて選好を変化させる能力が限定的であることが示された．また，Llama3.2 の時間選好実験では，現在志向の形容詞を与えた場合に k が低下するという，形容詞の意味と矛盾した変化が観察された．さらに，GPT-4o では形容詞を与えない場合の λ や k がすでに推定範囲の境界付近にあったため，リスク回避的な形容詞や将来志向の形容詞による変化を十分に観察できなかった．

今後の研究の方向性として，以下の点が挙げられる．第一に，より多様な LLM や形容詞を用いた検証を行うことで，知見の一般化可能性を高めることが考えられる．第二に，形容詞を与えない場合のパラメータが推定範囲の境界付近に位置しない LLM を選択するか，プロンプト設計によってベースラインのパラメータを調整することで，形容詞による変化をより明確に観察できる可能性がある．

本研究の結果は，意思決定特性を表す形容詞をプロンプトに含めることで LLM のリスク選好および時間選好を一定程度制

御できることを示した．この知見は，LLM を意思決定支援や行動シミュレーションに応用する際に，モデルの選好特性を意図的に調整するための一手法として活用できる可能性がある．

謝　　辞

本研究の一部は，JSPS 科研費 24K03046，24K23877 の助成を受けたものです．ここに記して謝意を示します．

文　　献

[1] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Enhancing ai-assisted group decision making through llm-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 103–119, 2024.

[2] Ayato Kitadai, Sinndy Dayana Rico Lugo, Yudai Tsurusaki, Yusuke Fukasawa, and Nariaki Nishino. Can ai with high reasoning ability replicate human-like decision making in economic experiments? *Group Decision and Negotiation*, pp. 1–24, 2025.

[3] 馬田光琉, 山本修平. 行動経済学的特性を付与した大規模言語モデルによる消費行動予測. 人工知能学会全国大会論文集 第 39 回 (2025), pp. 1D3OS24a03–1D3OS24a03. 一般社団法人 人工知能学会, 2025.

[4] Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, Vol. 37, pp. 113360–113382, 2024.

[5] 筒井義郎, 佐々木俊一郎, 山根承子, グレッグ・マルデワ. 行動経済学入門. 東洋経済新報社, 2017.

[6] Charles A Holt and Susan K Laury. Risk aversion and incentive effects. *American economic review*, Vol. 92, No. 5, pp. 1644–1655, 2002.

[7] Kris N Kirby and Nino N Maraković. Delay-discounting probabilistic rewards: Rates decrease as amounts increase. *Psychonomic bulletin & review*, Vol. 3, No. 1, pp. 100–104, 1996.

[8] 吉川克正, 大萩雅也, 高山隼矢. 大規模言語モデルの非対称的意思決定特性：プロスペクト理論要素の実証分析. 言語処理学会 第 31 回年次大会 発表論文集, pp. 195–200, 2025.

[9] 吉川克正, 大萩雅也, 高山隼矢. 大規模言語モデルは人間の時間選好性を再現できるか? 人工知能学会全国大会論文集 第 39 回 (2025), pp. 2A1GS1001–2A1GS1001. 一般社団法人 人工知能学会, 2025.

[10] Yuta Imasaka and Hideo Joho. Effect of llm's personality traits on query generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 249–258, 2024.

[11] Yuxuan Li, Hirokazu Shirado, and Sauvik Das. Actions speak louder than words: Agent decisions reveal implicit biases in language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3303–3325, 2025.

[12] Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12640–12653, 2024.

[13] 角田康明, 竹内孝, 鹿島久嗣. 大規模言語モデルにおける認知バイアスの軽減. 人工知能学会全国大会論文集 第 38 回 (2024), pp. 2T5OS5b02–2T5OS5b02. 一般社団法人 人工知能学会, 2024.

[14] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, Vol. 5, No. 4, pp. 297–323, 1992.

[15] 川越敏司. 「意思決定」の科学：なぜ、それを選ぶのか. 講談社, 2020.

[16] Adam S Booij, Bernard MS Van Praag, and Gijs Van De Kuilen. A parametric analysis of prospect theory's functionals for the general population. *Theory and Decision*, Vol. 68, No. 1, pp. 115–148, 2010.

[17] Nicolas Schweighofer, Kazuhiro Shishida, Cheol E Han, Yasumasa Okamoto, Saori C Tanaka, Shigeto Yamawaki, and Kenji Doya. Humans can adopt optimal discounting strategy under real-time constraints. *PLoS computational biology*, Vol. 2, No. 11, p. e152, 2006.