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

吉川克正 [1]　大萩雅也 [1]　高山隼矢 [1]
[1]SB Intuitions 株式会社
{katsumasa.yoshikawa,masaya.ohagi,junya.takayama}@sbintuitions.co.jp

概要

大規模言語モデル（LLM）の意思決定特性を理解することは、人工知能技術の発展において重要な課題となっている。本研究では、行動経済学の基礎理論であるプロスペクト理論の枠組みを用いて、LLM の意思決定パターンを実験的に検証した。実験結果は、LLM の意思決定特性が人間とは異なる独特のパターンを示すことを明らかにした。具体的には、損失領域においては人間の意思決定に近い特性を示す一方で、利得領域では極端に歪んだパターンを示した。これらの結果は、LLM が人間とは質的に異なる意思決定メカニズムを持つことを示唆しており、LLM の意思決定特性の理解と制御に重要な示唆を与えるものである。

1　はじめに

人工知能技術の急速な発展に伴い、大規模言語モデル（LLM）は様々な意思決定タスクに活用され始めつつある [1, 2, 3, 4]。しかし、これらのモデルがどのような意思決定特性を持つのか、また、それが人間の意思決定とどのように異なるのかについては、まだ十分に理解されていない。行動経済学におけるプロスペクト理論 [5] は、不確実性下での人間の意思決定を説明する最も影響力のある理論の一つとして知られている。例えば、図 1 のように、利得が発生する 1a) と 1b) の選択肢を提示された場合、人間はリスク回避の 1a) を選び易い。一方で、損失が発生する 2a) と 2b) ではリスク嗜好の 2b) を選びやすくなる。プロスペクト理論ではこの利得と損失の非対称性を価値関数として表現し、また人間が低確率を過大に評価し (少ない可能性に賭けてしまう)、高確率を過小に評価する（100%でないことを嫌う）特徴を確率過重関数として表現している。

本研究では、このプロスペクト理論の枠組みを用いて、LLM の意思決定特性を実験的に検証した。特に注目すべき点は、LLM が損失領域と利得領域で顕著に異なる特性を示すことである。従来の研究では、人間の意思決定において損失回避性が観察されることが知られているが、LLM ではこれとは異なるパターンが観察された。具体的には、損失領域では人間に近い意思決定パターンを示す一方で、利得領域では極端に歪んだ評価パターンを示すことが明らかになった。このような非対称性は、LLM の学習過程や意思決定メカニズムに関する重要な示唆を与えるものである。また、これらの知見は、LLM を実際の意思決定タスクに適用する際の重要な考慮事項となりうる。本研究では、これらの実験結果を詳細に分析し、LLM の意思決定特性の理解を深めるとともに、その実践的含意について議論する。

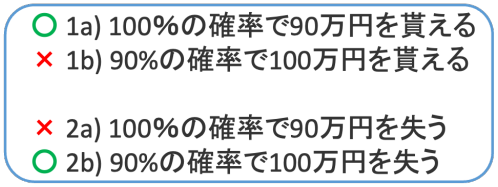

図 1: 主観的な価値と確率の矛盾事例

2　実験手法

プロスペクト理論の 2 つの主要な特性を評価するため、本研究では確率加重関数 (2.1) および価値関数 (2.2)、2 つの実験を設計・実施する。各実験は複数の LLM を利用して実施し、LLM によってどのような意思決定傾向があるかを評価する。評価対象は、Llama[6]、Qwen[7]、Mistral[8] が公開する LLM で、それぞれ大小のパラメータサイズのインストラクションチューニング済みモデルを選択した。

2 つの実験に共通する処理として、LLM を様々な被験者と見立てて利用するため、性別、年齢、姓をランダムで与えている。この詳細は付録に記述し

た。利用する実験プロンプトは選択肢を提示する形式であるため、その提示する順序が生成結果に影響を与える。その影響を軽減するため、常に選択肢順序を逆順で提示したプロンプトも併用する。なお、各シナリオ（金額、確率基準）について 20 回ずつ結果の生成を行うことにする。

2.1 確率加重関数評価実験

確率過重関数は、人間が捉える主観的確率が、客観な確率から乖離しているかを扱うモデルである。一般に小さい確率は過大に評価され、大きい確率は過小に評価されることで、結果として確率過重関数グラフは逆 S 字の曲線を描くことが知られている。この確率の主観的重みづけを評価するため、図 1 の 1a) と 1b) のように、確実な選択肢と確率的な選択肢の間での選択問題を用いた。各確率水準について、期待値が等価となるように金額を調整した。実験には付録: 図 4 のプロンプトを用いて、A, B の選択肢から 1 つを選択させる。

また、プロンプト内の利得/損失金額 (X) と確率水準 (p) は、次のように設定した。

- 利得/損失金額 (X): \$100, \$500, \$1000
- 確率水準 (p): 0.01, 0.05, 0.10, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8, 0.90, 0.95, 0.99

確率加重関数は客観的確率を p とした時、次のような主観的確率 $w(p)$ として定義される。

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

ここで、$\gamma = 1.0$ では主観と客観の確率が一致し、γ が小さいほど歪みが乖離が大きい。$w(p)$ は被験者集団から複数の回答を集めた時の、確率的選択肢が選ばれた比率として計算することができる。本研究では LLM から複数回の生成を行うことで、被験者集団の回答としている。生成に利用するハイパーパラメータはいずれの LLM でも全てデフォルトとした。生成結果として妥当でないもの (A or B) 以外は推定には利用せず、各 LLM ごとに妥当な回答が生成された割合 (有効回答率) を計測する。得られた生成結果は利得領域と損失領域に分けて集計する。

2.2 価値関数評価実験

価値関数は人間の考える主観的な価値が、特定の参照点より上の利得領域と、下の損失領域で大きく傾向が変わることを表現した関数である。具体的には、人間は参照点より遠く離れるほど価値の差に鈍感になる価値低減性、利得より損失を重くみる損失回避性がある。この価値関数の特性を評価するため、確実性等価法を利用した実験を行う。具体的には、図 5 のように確率的な結果に対する確実性等価の値を生成させることで、被験者の主観的な価値を推定する。なお、生成に利用するハイパーパラメータは確率過重関数実験と同様全てデフォルトとした。

プロンプト内の利得/損失金額 (X) と確率水準 (p) は、次のように設定した。

- 利得/損失金額 (X): \$100, \$500, \$750, \$1000, \$2500, \$5000, \$7500, \$10000
- 確率水準 (p): 0.2, 0.5, 0.8

価値関数は客観的価値 x に対する主観的価値 $v(x)$ として次のように定義できる。

$$v(X) = \begin{cases} x^\alpha & (x \geq 0) \\ -\lambda(-x)^\beta & (x < 0) \end{cases} \tag{2}$$

ここで α, β はそれぞれ利得領域、損失領域の曲率であり、利得や損失が大きくなるときに、価値に与える影響が小さくなるかを示している（感応度逓減）。λ は損失回避度で、利得に対する損失の価値の程度を示しており、この値が高いほど利得より損失から受ける心理的価値が大きくなり、損失回避傾向にあるとされる。

本研究では利得/損失金額 (X) を客観的価値 x、LLM が生成する確実性等価値 (Y) の平均を主観的価値 $v(X)$ としてフィッティングを行う。ただし、確率過重関数実験とは異なり、本実験では Y を数値として直接生成させるため問題として難易度が高く、有効回答率が低いことも予測される。

3 実験結果

3.1 確率加重関数

利得領域及び損失領域における確率過重関数のパラメータ推定結果を、それぞれ表 1 と表 2 に示す。

まず、式 1 への非線形回帰によるパラメータ推定精度を R^2 と RMSE で評価すると、いずれも低く、分散の大きい結果となっていることが分かる。特にパラメータサイズの小さなモデルでは推定精度が悪いことが多く、サイズが大きくなることで安定した精度が得られる。比較的高い精度を得ているのは Qwen のモデルだった。有効回答率はほとんどのモデルで 0.99 以上の値が観測された。

表 1: 確率過重関数推定結果 (利得領域)

Model	γ	R^2	RMSE
Meta-Llama-3.1-8B-Instruct	0.463	-0.838	0.181
Meta-Llama-3.1-70B-Instruct	0.243	0.496	0.066
Llama-3.2-1B-Instruct	0.603	-30.194	0.329
Llama-3.2-3B-Instruct	0.471	-116.155	0.221
Llama-3.3-70B-Instruct	0.191	0.176	0.067
Mistral-7B-Instruct-v0.3	0.410	-2.188	0.168
Mistral-Large-Instruct-2407	0.116	0.000	0.027
Mixtral-8x7B-Instruct-v0.1	0.769	-3.747	0.366
Mixtral-8x22B-Instruct-v0.1	0.272	-0.103	0.103
Qwen2.5-7B-Instruct	0.104	0.214	0.007
Qwen2.5-72B-Instruct	0.108	0.228	0.008

表 2: 確率過重関数推定結果 (損失領域)

Model	γ	R^2	RMSE
Meta-Llama-3.1-8B-Instruct	0.510	-15.341	0.291
Meta-Llama-3.1-70B-Instruct	0.547	-7.041	0.465
Llama-3.2-1B-Instruct	0.610	-31.030	0.354
Llama-3.2-3B-Instruct	0.478	-313.367	0.229
Llama-3.3-70B-Instruct	0.485	-4.507	0.457
Mistral-7B-Instruct-v0.1	0.772	-947.222	0.601
Mistral-Large-Instruct-2407	0.510	-0.028	0.172
Mixtral-8x7B-Instruct-v0.1	0.727	-17.328	0.569
Mixtral-8x22B-Instruct-v0.1	0.284	-0.854	0.336
Qwen2.5-7B-Instruct	0.319	0.222	0.134
Qwen2.5-72B-Instruct	0.383	-1.060	0.160

次に γ 値を利得・損失それぞれの領域で評価すると、人間の典型的な γ の範囲（利得:0.61, 損失:0.69）から逸脱しているものも多い。損失領域では比較的に人間の γ 値に近く、人間に類似した確率過重パターンを示す一方で、利得領域では γ が極めて小さく、人間とは全く異なる強い確率の歪みがあることを示している。参考までに、図 2 と図 3 には、利得/損失領域それぞれでの Qwen2-72B-Instruct の確率過重関数を可視化したものを示した（利得:0.108, 損失:0.383）。ここでは横軸が客観確率、縦軸がリスク選択肢の比率である。この 2 つの図からも分かる通り、このモデルは利得領域ではほとんどリスク選択肢を選ばず、確実な利益を優先する傾向にある。また損失領域でも人間に比べれば低い γ の値を示し、図 3 からは高い確率を過小評価する傾向が特に強いことが分かる。LLM の選択は全体にリスク回避傾向であり、より確実な選択肢を好むと言える。

3.2 価値関数

価値関数のパラメータ推定結果を、表 3 に示す。

まず、推定精度の点では確率過重関数実験に比べて高い値が多いが、一方で有効回答率は低いものも

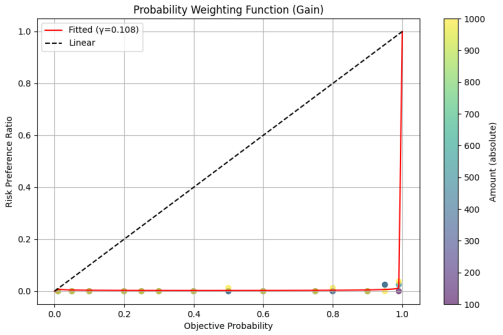

図 2: 利得領域の確率過重関数 (Qwen2-72B-Instruct)

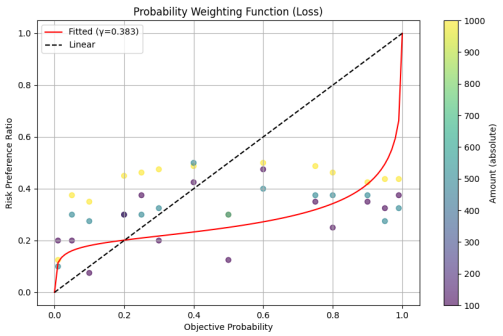

図 3: 損失領域の確率過重関数 (Qwen2-72B-Instruct)

多く、パラメータを推定するために十分なデータポイントを確保できないモデルも多かった。

次に LLM 間の結果には大きな差があることが分かる。Llama では、α と比べて β の値が低い結果となっており、これは感応度逓減が損失領域で大きいことを示している。逆に Qwen では、β の値の方が大きくなるものが多い。特に Qwen2.5-72B-Instruct のように $\beta = 1.0$ になるのは、損失領域で感応度一定であることを示している。いずれにしても、$(\alpha = 0.88, \beta = 0.88)$[9] とされている人間の値とは逸脱していると言える。

さらに、λ の値は人間の被験者の場合は 1.5〜2.5 になると報告されている [9, 10, 11, 12]。Llama では損失回避傾向がより強く、λ は 3.0 以上のモデルも多い。一方で Qwen の λ 値はいずれも 1.0 を下回っており、人間の損失回避傾向とは逆である。

図 6 は Qwen2.5-72B-Instruct の価値関数を可視化したものである。逓減がほとんどなく、線形に近いグラフとなっていることが分かる。

このように価値関数の推定結果は分散が大きく、LLM ごとに傾向が異なるため注意が必要である。

表 3: 価値関数推定結果

Model	α	β	λ	R^2	RMSE	有効回答率
Meta-Llama-3.1-8B-Instruct	0.942	0.792	3.345	0.965	518.146	0.990
Meta-Llama-3.1-70B-Instruct	0.922	0.866	1.474	0.823	1058.797	1.000
Llama-3.2-1B-Instruct	0.839	0.591	10.000	0.819	584.339	0.524
Llama-3.2-3B-Instruct	0.927	0.792	3.715	0.974	432.985	0.988
Llama-3.3-70B-Instruct	0.922	0.752	3.212	0.820	973.482	1.000
Mistral-7B-Instruct-v0.1	0.775	1.000	0.753	0.794	912.313	0.043
Mistral-Large-Instruct-2407	0.918	0.878	1.343	0.802	1123.198	1.000
Mixtral-8x7B-Instruct-v0.1	0.914	0.953	0.584	0.849	873.034	0.894
Mixtral-8x22B-Instruct-v0.1	-	-	-	-	-	0.003
Qwen2.5-7B-Instruct	0.924	0.922	0.944	0.988	256.561	1.000
Qwen2.5-72B-Instruct	0.917	1.000	0.490	0.823	1086.741	1.000

表 4: ペルソナによる価値関数変化

Persona	α	β	λ
男性	0.914	0.788	2.315
女性	0.913	0.803	2.029
80 代	0.911	0.779	2.545
20 代	0.914	0.781	2.426
高所得	0.919	0.938	0.793
低所得	0.903	0.972	0.515
喫煙者	0.909	0.756	3.072
非喫煙者	0.919	0.721	4.172
職業トレーダー	0.923	0.715	4.796
一般人	0.922	0.750	2.908
先進国	0.913	0.783	2.439
新興国	0.913	0.764	2.885
途上国	0.913	0.754	3.176

3.3 ペルソナによる変化

価値関数のパラメータは、被験者の属性によってやや異なる値を示すことが知られている [10, 11, 13]。そこで、比較的人間の結果と近い傾向を示した Llama-3.3-70B-Instruct を利用して、様々なペルソナ情報を与えることで各パラメータの変動を検証する。

表 4 はペルソナを与えた時の価値関数パラメータの推定結果である。α, β はあまり大きな変化がないが、所得情報ラベルを与えた時には、β の値が 0.9 を超えており、利得と損失間の差がなくなったのがわかる。所得情報はある種の参照点情報であり、プロスペクト理論の参照点依存効果を示唆していると考えられる。損失回避係数 λ はペルソナによって大きく変化する。特に所得情報ラベルを与えた時は $\lambda < 1.0$ になり、人間の傾向とは乖離する。

Sokol-Hessner らは職業トレーダーの損失回避係数 λ は一般人のそれより低いことを報告している（職業トレーダー:1.28、一般人:1.96) [10] しかし、本研究では逆に職業トレーダーが $\lambda = 4.796$ となり、むしろ一般人のそれより高くなった ($\lambda = 2.908$)。

その他、人種や国籍により損失回避係数に幅があることも報告されている [13]。先進国（アメリカ、イギリス、日本など）の国籍情報を与えると比較的損失回避係数が低くなり (2.439)、逆に途上国（ナイジェリア、バングラデシュ、パキスタンなど）の国籍を与えると λ はやや高くなった (3.176)。新興国 (BRICS) はその中間の値を示している (2.885)。

3.4 考察

プロスペクト理論要素の非対称的な再現は、以下を示唆していると考えられる：

- LLM は学習データを通じて確率の歪みを自然に獲得している
- 価値関数特性の欠如は、LLM が価値や効用を処理する方法の根本的な違いを反映している
- 学習プロセスが、人間の損失回避を駆動する感情的・本能的な側面を適切に捉えていない可能性がある

一方で本実験には限界もあり、学習プロセスのどの段階で歪みが生じるかや、観察されたパターンの長期的安定性は未検証である。また LLM でのシミュレーションは人間の被験者実験とは異なり、現実の利益や損失が発生しないことも影響していると考えられる。

4 おわりに

本研究の知見は、LLM が人間の意思決定パターンを部分的かつ非対称的にのみ再現可能であることを示している。これは、LLM を人間の経済的行動のモデル化や意思決定支援に用いる際の重要な示唆を提供している。

参考文献

[1] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. Large language models in law: A survey. **AI Open**, Vol. 5, pp. 181–196, 2024.

[2] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4, 2023.

[3] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. **Nature Medicine**, Vol. 29, pp. 1930–1940, 2023.

[4] Eric; Roberts Roger; Singla Alex; Smaje Kate; Sukharevsky Alex; Yee Lareina; Zemmel Rodney Chui, Michael; Hazan. The economic potential of generative ai the next productivity frontier the economic potential of generative ai: The next productivity frontier, 2023.

[5] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. **Econometrica**, Vol. 47, No. 2, pp. 263–291, 1979.

[6] Aaron Grattafiori, et al. The llama 3 herd of models, 2024.

[7] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[8] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[9] 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, October 1992.

[10] Peter Sokol-Hessner, Ming Hsu, Nina G. Curley, Mauricio R. Delgado, Colin F. Camerer, and Elizabeth A. Phelps. Thinking like a trader selectively reduces individuals' loss aversion. **Proceedings of the National Academy of Sciences**, Vol. 106, No. 13, pp. 5035–5040, 2009.

[11] A.S. Booij, B.M.S. van Praag, and G. van de Kuilen. A parametric analysis of prospect theory's functionals for the general population. **Theory and Decision**, Vol. 68, No. 1-2, pp. 115–148, 2010.

[12] Mei Wang, Marc Oliver Rieger, and Thorsten Hens. The impact of culture on loss aversion. **Cultural Finance**, 2017.

[13] Marc Oliver Rieger, Mei Wang, and Thorsten Hens. Estimating cumulative prospect theory parameters from an international survey. **Theory and Decision**, Vol. 82, No. 4, pp. 567–596, 2017.

A　実験用プロンプト

検証実験に利用したプロンプトを、図 4,5 に示した。

```
# ペルソナ情報
You are {persona}.

# 指示
I will now ask you multiple-choice questions, so please
  output only the letter of your answer (A or B).

# シナリオ
Question: There are two ways to receive the prize money.
  Based on your intuition and personal risk preferences,
  which of the following options would you choose?
A. ${p*100}% chance to receive ${X} ({(1-p)*100}% chance
  to receive nothing).# リスク選択肢
B. Receive ${p*X} for certain. # 確実性選択肢
Answer(A or B):
```

図 4: 確率過重関数実験プロンプト (利得領域)

```
# ペルソナ情報
You are {persona}.

# 指示
Consider the following two options.

# シナリオ
Question: Based on your intuition and personal risk
  preferences, what amount Y (in dollars) would make you
  feel exactly neutral between these two options:
A. {p*100}% chance to receive ${X} ({(1-p)*100}% chance
  to receive nothing). # リスク選択肢
B. Receive $Y for certain. # 確実性選択肢
Respond with a single number only:
```

図 5: 価値関数実験プロンプト（利得領域）

A.1　ペルソナ情報の生成

ここでは実験プロンプトに与えるペルソナ情報をどのように生成するかについて述べる。被験者実験として現実的な設定とするため、姓、性別、年齢の情報をランダムで与えるようにした。姓については、the US 2010 Census Data [1)]から上位 1000 件の surname を取得し、その中からランダムに利用した。図 5 には、サンプルとなる生成されたペルソナ情報を示してある。

表 5: プロンプト中の自動生成ペルソナ情報

Persona	Example Sentence
name	You are Smith.
title	Mr. Smith,
age	You are 27 years old.
gender	You are a male.

B　その他の実験結果

図 6 には Qwen2.5-72B-Instruct の価値関数を可視化したもの示した。利得領域でわずかに逓減が見られるものの、ほぼ線形のグラフを描いていることが分かる。

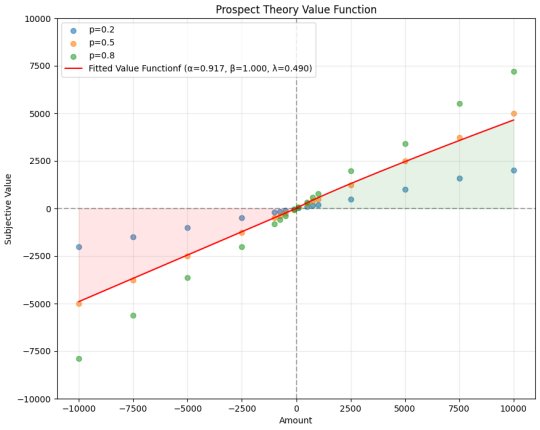

図 6: 価値関数関数 (Qwen2.5-72B-Instruct)

1)　https://www.census.gov/topics/population/genealogy/data/2010_surnames.html