

# LLM におけるアンカリング効果の機械論的理解と軽減

武並佳輝 Yin Jou Huang 村脇有吾 竹内孝 Chenhui Chu

京都大学大学院情報学研究科

tnami@nlp.ist.i.kyoto-u.ac.jp {huang,murawaki,takeuchi,chu}@i.kyoto-u.ac.jp

## 概要

大規模言語モデル (LLM) は人間と同様にアンカリング効果を示すことが知られており, LLM の安全性を確保するために, 機械論的理解やそれに基づく軽減手法が期待される. 本研究では LLM による価格交渉シミュレーションを題材に, アンカリング効果の発生機序を機械論的に理解するとともに, アテンション機構への介入による軽減手法を提案する. 売り手 LLM が交渉の初期段階で提示する価格 (アンカー) に高い注意重みをもつアテンションヘッドを特定し, 買い手 LLM においてそのヘッドを抑制したところ, アンカリング効果が軽減されることを確認した. また, ヘッドの抑制によって他タスクでの性能が損なわれることもなかった.

## 1 はじめに

大規模言語モデル (LLM) は高度な言語処理能力を有するが, その学習データには人間の意思決定におけるアンカリング効果が埋め込まれており, LLM が人間と同様にアンカリング効果を示すことが報告されている [1]. アンカリング効果とは, 何らかの判断を行う際, 最初に提示された情報が基準点 (アンカー) となり, 最終的な判断がその値に不当に偏ってしまう認知バイアスで, 交渉における価格決定などの場面で顕著に発現する [2]. 交渉タスクにおける LLM の安全性を確保するために, その発生機序の機械論的理解や軽減手法の提案が期待される.

これまで, LLM におけるアンカリング効果の研究は, 数値推定タスク [3] から交渉などの対話タスク [4, 5] に至るまで幅広く行われてきた. その中で Huang らは, プロンプト中の特定の単語 (“higher”, “lower”, アンカーの値) が, 次トークンの生成確率に大きな影響を与えることを報告したが [6], 入出力に着目した行動レベルの分析に留まっている. 同

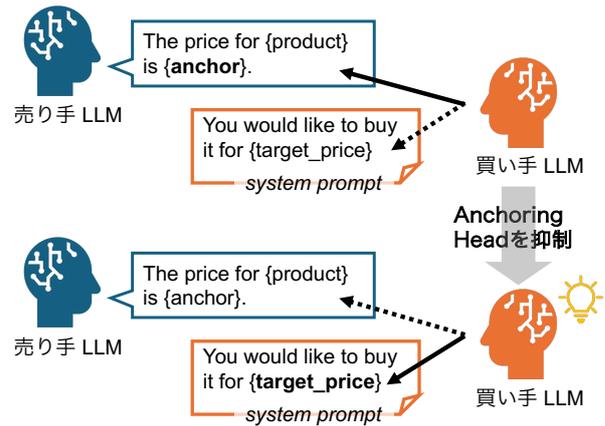


図1 アテンション機構への介入による, 価格交渉におけるアンカリング効果の軽減. アンカリング効果の原因である Anchoring Head を抑制することで, 目標価格の情報に注目させ, アンカーに引きずられることを防ぐ.

研究は軽減手法として, アンカリング効果に注意して熟考するようにプロンプト内で指示するものを提案したが, 効果は限定的である.

このような課題に対し, AI の安全性を高めるためのアプローチとして, 機械論的解釈可能性 (Mechanistic Interpretability) が注目されている [7, 8, 9]. これは, ニューラルネットワークが学習した計算メカニズムや表現を, 人間が理解可能なアルゴリズムや概念へとリバースエンジニアリングする試みである. アンカリング効果についても, 行動レベルの観察に留まらず, LLM 内部のどのコンポーネントが計算に関与しているかを機械論的に特定することで, より確実な軽減が可能になると考えられる.

一般的な LLM のアーキテクチャである Transformer [10] において, Feed-Forward Network は学習データ内の事実知識を保存する役割を担い [11, 12], Multi-Head Attention (以下, アテンションと呼ぶ) は文脈内で必要情報を検索しその場で適切なアルゴリズムを実現する役割を担う [13, 14, 15]. これを踏ま

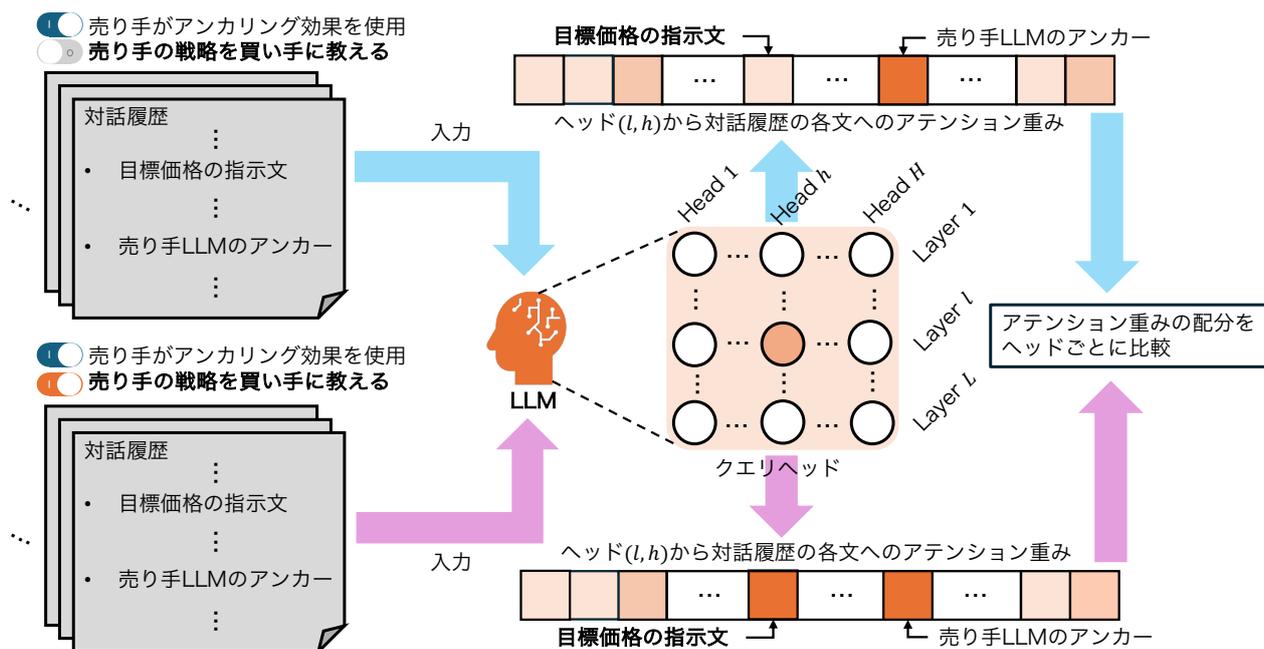


図2 Anchoring Headの検出手法. 買い手LLMに与えるシステムプロンプトと, 売り手LLMおよび買い手LLMの発話で構成される対話履歴をLLMに入力した際のアテンション重みを分析する. 売り手LLMがアンカリング効果を使うことを買い手LLMに警告するプロンプトの有無でアテンション重みの配分を変えたクエリヘッドを, アンカリング効果を引き起こすヘッドとして特定する.

えると, アンカリング効果の発生機序は次の二つが考えられる. 一つは訓練データに含まれるアンカリング効果の情報からのショートカット, もう一つはそのような訓練データに晒されたアテンションヘッドが, 推論時にプロンプト中の本来注目すべき情報よりもアンカーの数値を強く参照することによるものである.

このような背景のもと, 本研究では, LLMを用いた価格交渉シミュレーションを通じてアンカリング効果の発生機序を機械論的に理解し, その軽減手法を提案する. 図1のように, 買い手LLMが自身の目標価格よりも, 売り手LLMが提示するアンカーに対して過度なアテンション重みを配分していることがアンカリング効果の原因であると仮説を立てる. この仮説に基づき, 図2のように, アンカリング効果の警告文のプロンプトの有無でアテンション重みが鋭敏に変動するクエリヘッドを **Anchoring Head** として特定する. 実験の結果, Anchoring Headを抑制した買い手LLMはアンカリング効果を軽減できた一方で, ランダムに選ばれたヘッドを抑制した買い手LLMは軽減できなかったため, Anchoring Headとアンカリング効果の因果関係が示された. また, 提案手法は他タスクにおける言語処理能力を損なわないことも分かり, その有効性が確認された.

## 2 手法

### 2.1 LLMによる価格交渉シミュレーションにおけるアンカリング効果の検証

アンカリング効果を検証するにあたって, [5]の手法に従った. 次の三つの条件で価格交渉シミュレーションを行う.

- seller\_no\_anchor: アンカリング効果について特に指示を与えない条件. 通常の価格交渉.
- seller\_anchor: 売り手LLMにアンカリング効果を使うようにプロンプトで明示的に指示を与える条件.
- seller\_anchor\_buyer\_informed: 売り手LLMにアンカリング効果を使うようにプロンプトで指示を与えた上で, そのことを買い手LLMにプロンプト(このプロンプトを以下ではアンカリング効果の警告文と呼ぶ)で知らせる条件.

### 2.2 Anchoring Headの検出

売り手が初期価格を提案する文(アンカー)へのアテンション重みと, 買い手に目標価格を指示する文へのアテンション重みの差を Attention Preferenceとして定義し, アンカリング効果の警告文の有無によってその差がどのように変化するかを測定する.

検出手法の概要を図2に示す。

まず、ある LLM の内部状態において、特定のクエリヘッドが、アンカー価格と自身の目標価格のどちらをより強く参照しているかを測定する。対話履歴に含まれる  $k$  番目の発話文を  $S_{\text{src}}^{(k)}$  とし、売り手が提示したアンカー価格を含む文を  $S_{\text{anchor}}$ 、買い手の目標価格を含む文を  $S_{\text{goal}}$  とする。層  $l$  の  $h$  番目のヘッドにおける、 $S_{\text{src}}^{(k)}$  からあるターゲット文  $S_{\text{target}}$  へのアテンション重みの平均  $A^{(l,h)}(S_{\text{src}}^{(k)} \rightarrow S_{\text{target}})$  は次式で計算される。

$$A^{(l,h)}(S_{\text{src}}^{(k)} \rightarrow S_{\text{target}}) = \frac{1}{|T_{\text{src}}| \cdot |T_{\text{target}}|} \sum_{i \in T_{\text{src}}} \sum_{j \in T_{\text{target}}} \alpha_{i,j}^{(l,h)} \quad (1)$$

ここで、 $T_{\text{src}}, T_{\text{target}}$  はそれぞれの文に含まれるトークンインデックスの集合であり、 $\alpha_{i,j}^{(l,h)}$  はトークン  $i$  から  $j$  へのアテンション重みを表す。これを用い、対話全体を通じた平均的な Attention Preference  $P^{(l,h)}$  を次式で定義する。

$$P^{(l,h)} = \frac{1}{K} \sum_{k=1}^K \left( A^{(l,h)}(S_{\text{src}}^{(k)} \rightarrow S_{\text{anchor}}) - A^{(l,h)}(S_{\text{src}}^{(k)} \rightarrow S_{\text{goal}}) \right) \quad (2)$$

ここで  $K$  は分析対象となる対話履歴内の発話数である。 $P^{(l,h)}$  が正の値をとる場合、そのヘッドは買い手自身の目標価格よりもアンカー価格に強く注目していることを示唆する。

次に、アンカリング効果の警告文が Attention Preference に与える影響を評価する指標として  $\Delta P^{(l,h)}$  を導入する。同一の製品、目標価格、売り手買い手の組み合わせにおいて、警告文を与えない条件 (Uninformed) と与える条件 (Informed) の2つの価格交渉シミュレーションを行い、対話データを生成する。前者は節2.1における seller\_anchor 条件に、後者は seller\_anchor\_buyer\_informed 条件に該当する。 $N$  組の対話データセットに対し、各ヘッドの  $\Delta P^{(l,h)}$  を次のように定義する。

$$\Delta P^{(l,h)} = \frac{1}{N} \sum_{n=1}^N \left( P_{\text{uninformed},n}^{(l,h)} - P_{\text{informed},n}^{(l,h)} \right) \quad (3)$$

ここで、 $P_{\text{uninformed},n}$  および  $P_{\text{informed},n}$  は、それぞれ警告文なし・あり条件における  $n$  番目の対話データの Attention Preference である。アンカリング効果

の警告文がない (uninformed) 状態でアンカーへの注目が高く、警告文がある (informed) 状態でアンカーへの注目が低下 (または目標への注目が増加) する場合、 $\Delta P^{(l,h)}$  は正の値をとる。したがって、このスコアが高いヘッドほど、アンカリング効果に反応するヘッドであることを示唆する。

最後に、LLM 内の全クエリヘッドを  $\Delta P^{(l,h)}$  で降順にソートし、上位  $m$  個を Anchoring Head として検出する。

## 2.3 アテンションヘッドの抑制

前節の手法により特定された Anchoring Head が、アンカリング効果の発生に因果的な役割を果たしているかを検証するため、当該ヘッドの機能を抑制する介入実験を行う。 $N_H$  個のクエリヘッドを持つ第  $l$  層の Multi-Head Attention 機構において、入力  $X$  に対する出力  $Y^{(l)}$  は、各ヘッド  $h$  の出力  $H^{(l,h)}$  と、重み行列  $W_O^{(l,h)}$  を用いて次式のように各ヘッドの出力の重み付き和として定式化できる。

$$Y^{(l)} = \sum_{h=1}^{N_H} H^{(l,h)} W_O^{(l,h)} \quad (4)$$

本実験では、分析により選定された Anchoring Head の集合  $\mathcal{H}_{\text{anc}}$  に含まれる各ヘッド  $(l,h) \in \mathcal{H}$  に対し、対応する出力重み  $W_O^{(l,h)}$  に微小な係数  $\epsilon$  を乗じることで、その寄与を抑制する。介入後の重み  $\hat{W}_O^{(l,h)}$  は次式で与えられる。

$$\hat{W}_O^{(l,h)} = \begin{cases} \epsilon W_O^{(l,h)} & \text{if } (l,h) \in \mathcal{H}_{\text{anc}} \\ W_O^{(l,h)} & \text{otherwise} \end{cases} \quad (5)$$

## 3 実験

### 3.1 実験設定

本実験ではオープンウェイトモデルの Qwen2.5-32B<sup>1)</sup> を分析対象とした。40 個のクエリヘッドをもつ層 64 個から構成されており、モデル全体で 2,560 個のクエリヘッドを有する。売り手 LLM は Qwen2.5-32B で固定したまま、買い手 LLM を一部ヘッドを抑制した Qwen2.5-32B で置き換えて価格交渉シミュレーションを行った。

交渉シナリオ (製品、目標価格、売り手買い手の性格特性) は [5] と同じく 322 個用意し、128 個を Anchoring Head の特定に、64 個をハイパーパラメー

1) Qwen/Qwen2.5-32B-Instruct

**表 1** 通常の価格交渉 (seller\_no\_anchor), 売り手がアンカリング効果を使う条件 (seller\_anchor), 売り手がアンカリング効果を使うことを買い手に教える条件 (seller\_anchor\_buyer\_informed) における買い手 LLM の目標価格の達成度合い. 値が大きいほど, アンカリング効果が軽減されたことを示す. 提案手法により特定したヘッドを抑制した LLM はアンカリング効果を軽減することができた.

LLM	seller_no_anchor	seller_anchor	seller_anchor_buyer_informed
(1) Qwen2.5-32B-Instruct	-0.09 ± 0.67	-0.55 ± 0.60	-0.34 ± 0.59
(2) Qwen2.5-32B-Instruct_RandomHead	-0.10 ± 0.57	-0.56 ± 0.61	-0.35 ± 0.56
(3) Qwen2.5-32B-Instruct_AnchoringHead	-0.03 ± 0.57	<b>-0.42 ± 0.56</b>	<b>-0.26 ± 0.61</b>

タ  $m, \epsilon$  のチューニングに使用した (付録 B を参照).  $m = 10, \epsilon = 0.01$  のもとで, 残りの 130 個における結果を 3.3 節で報告する.

### 3.2 評価指標

買い手の目標価格の達成度合い  $Ach_b(p)$  を次式 (6) で計算する. この値が低いほど, 売り手のアンカリング効果によって買い手がより損失を被ったことを示す.

$$Ach_b(p) = \frac{\overline{p_b} - p}{\overline{p_b} - p_b} \quad (6)$$

ここで,  $\overline{p_b}$  と  $p_b$  はそれぞれ買い手の最高許容価格と目標価格である.

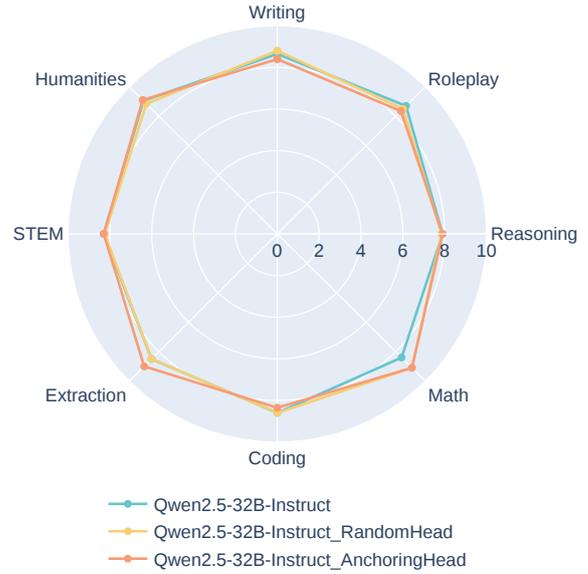
また, ヘッドの抑制による他タスクへの影響を調べるため, 知識や問題解決能力を測る MMLU [16] と, 対話や指示追従の能力を測る MT-Bench [17] を用いた. MT-Bench の採点には GPT-4.1<sup>2)</sup> を使用した.

### 3.3 結果

提案手法を用いて Anchoring Head を検出した. 検出された 10 個のヘッドの集合を  $\mathcal{H}_{anc}$  とする. また, ランダムに抽出した同数のヘッドの集合を  $\mathcal{H}_{rand}$  とする. それぞれの構成要素は付録 A に示す.

特定した Anchoring Head とアンカリング効果の因果関係を検証するため,  $\mathcal{H}_{anc}$  および  $\mathcal{H}_{rand}$  をそれぞれ抑制した際の交渉結果を比較した. 表 1 にその結果を示す.  $\mathcal{H}_{anc}$  を抑制した場合はアンカリング効果が軽減されている一方で,  $\mathcal{H}_{rand}$  を抑制した場合は軽減されていないことから, 本手法によって特定されたヘッドはアンカリング効果を引き起こしていることが明らかになった.

MT-Bench を用いた評価の結果を図 3 に示す (MMLU を用いた評価の結果の表は付録 C に示す). 両方の指標において,  $\mathcal{H}_{anc}$  と  $\mathcal{H}_{rand}$  のどちらを抑制



**図 3** MT-Bench による評価の結果.

した場合もスコアの大きな変動は見られなかったことから, 特定された  $\mathcal{H}_{anc}$  は他タスク遂行に重要なヘッドを含まないことが分かる. なお, MT-Bench の Math の定性分析を付録 D に示す.

## 4 おわりに

本研究では, LLM による価格交渉シミュレーションにおいて, アンカリング効果の機械論的理解と軽減手法の提案を行った. アテンション重みの偏りに着目した提案手法により, アンカリング効果の要因であるヘッドを特定した. 特定されたヘッドを抑制した LLM はアンカリング効果を軽減できた一方で, ランダムに選んだヘッドを抑制した場合は軽減されなかったことから, 当該ヘッドとアンカリング効果の因果関係が示された. また, 提案手法は他のタスクにおける性能を損なうことなくアンカリング効果を軽減することも確認された. 今後の展望として, 本手法の有効性を他の LLM や他の認知バイアスで検証することが挙げられる.

2) gpt-4.1-2025-04-14

## 謝辞

本研究は一部、JST CREST JPMJCR20D2 と JST 創発的研究支援事業 JPMJFR232S の支援を受けたものです。

## 参考文献

- [1] Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments. In **Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing**, SAC '25, p. 1009–1011, New York, NY, USA, 2025. Association for Computing Machinery.
- [2] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. **science**, Vol. 185, No. 4157, pp. 1124–1131, 1974.
- [3] Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. **Journal of Experimental Psychology: General**, Vol. 153, No. 4, p. 1066, 2024.
- [4] Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring Bargaining Abilities of LLMs: A Benchmark and A Buyer-Enhancement Method. In **Findings of the Association for Computational Linguistics ACL 2024**, pp. 3579–3602, 2024.
- [5] Yoshiki Takenami, Yin Jou Huang, Yugo Murawaki, and Chenhui Chu. How Does Cognitive Bias Affect Large Language Models? A Case Study on the Anchoring Effect in Price Negotiation Simulations. In **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 4481–4498, Suzhou, China, November 2025. Association for Computational Linguistics.
- [6] Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. An Empirical Study of the Anchoring Effect in LLMs: Existence, Mechanism, and Potential Mitigations. **arXiv preprint arXiv:2505.15392**, 2025.
- [7] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A Mathematical Framework for Transformer Circuits. **Transformer Circuits Thread**, Vol. 1, No. 1, p. 12, 2021.
- [8] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In **The Eleventh International Conference on Learning Representations**, 2023.
- [9] Leonard Bereska and Stratis Gavves. Mechanistic Interpretability for AI Safety - A Review. **Transactions on Machine Learning Research**, 2024. Survey Certification, Expert Certification.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. **Advances in neural information processing systems**, Vol. 30, , 2017.
- [11] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5484–5495, 2021.
- [12] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. **Advances in neural information processing systems**, Vol. 35, pp. 17359–17372, 2022.
- [13] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context Learning and Induction Heads. **arXiv preprint arXiv:2209.11895**, 2022.
- [14] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality. **arXiv preprint arXiv:2404.15574**, 2024.
- [15] Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function Vectors in Large Language Models. In **The Twelfth International Conference on Learning Representations**, 2024.
- [16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In **International Conference on Learning Representations**, 2021.
- [17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. **Advances in neural information processing systems**, Vol. 36, pp. 46595–46623, 2023.

## A 抑制したヘッド

$$\mathcal{H}_{\text{anc}} = \{(33, 39), (42, 27), (1, 19), (34, 10), (52, 6), (30, 28), (17, 39), (51, 18), (3, 10), (2, 2)\} \quad (7)$$

$$\mathcal{H}_{\text{rand}} = \{(7, 38), (14, 28), (23, 27), (24, 33), (32, 23), (37, 12), (41, 39), (42, 31), (47, 20), (59, 30)\} \quad (8)$$

ここで、各要素  $(l, h)$  は第  $l$  層の第  $h$  ヘッドを表す。

## B ハイパーパラメータの選定

ハイパーパラメータのグリッドサーチを行った結果を図4に示す。  $m, \epsilon$  の各組み合わせで買い手 LLM の Anchoring Head を抑制し、 seller\_anchor 条件で価格交渉シミュレーションを行った。最もアンカリング効果を軽減できた  $m = 10, \epsilon = 0.01$  を最適なハイパーパラメータとして選定した。

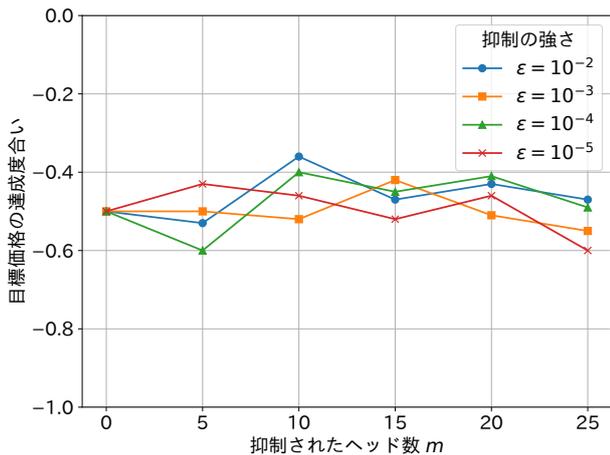


図4 ハイパーパラメータ  $m, \epsilon$  の選定。目標価格の達成度合いの平均値が最も高くなったハイパーパラメータを、最もアンカリング効果を軽減できるものとして選んだ。

## C MMLU を用いた評価

表2 MMLU のスコア。ヘッドの抑制による影響はほとんど見られなかった。番号が指す LLM は表1と同じ。

LLM	Humanities	Social Sciences	STEM	Other
(1)	0.7603	0.8911	0.8088	0.8375
(2)	0.7569	0.8885	0.8084	0.8330
(3)	0.7630	0.8915	0.8119	0.8378

## D MT-Bench の結果の定性分析

図3では Math カテゴリにおいて、ヘッド抑制後の LLM に比べて、抑制前の LLM の方がスコアが低

くなっている。 LLM の回答を手で分析したところ、このカテゴリ内の 10 問のうち 2 問の 2 ターン目で、抑制後の LLM は正解できた問題を、抑制前の LLM は誤答していることが分かった。ヘッドの抑制が良い影響を与えた可能性も示唆されるが、サンプルサイズが限られているため、偶然の可能性を排除することは今後の課題としたい。図5に誤答例を示す。

**Turn 1 Q** Some people got on a bus at the terminal. At the first bus stop, half of the people got down and 4 more people got in. Then at the second bus stop, 6 people got down and 8 more got in. If there were a total of 25 people heading to the third stop, how many people got on the bus at the terminal?

**Turn 1 A** (正解) Let's denote the number of people who got on the bus at the terminal as  $x$ . [...] we can set up the following equation:  $x/2 + 6 = 25$  [...] the number of people who got on the bus at the terminal is 38.

**Turn 2 Q** If the ticket is \$2 per person, how much is the total money earned by the bus?

**Turn 2 A** (誤り) To determine the total money earned by the bus, we need to know the total number of people who were on the bus at any point and the price per ticket. However, the problem only specifies the number of people on the bus at the third stop and the price per ticket. [...] From the previous solution, we know that 25 people were on the bus at the third stop. [...] **25 people times 2 dollars/person = 50 dollars.** Therefore, the total money earned by the bus is 50.

**Evaluation** (GPT-4.1) The assistant's answer to the second question is incorrect. [...] the total money earned should be based on the total number of people who bought tickets, i.e., **everyone who boarded the bus at any stop (38 at the terminal, 4 at the first stop, 8 at the second stop), totaling 50 people.** Thus, the total money earned is 50 times \$2 = \$100. **The assistant, however, only multiplies the number of people on the bus at the third stop (25) by the ticket price, giving \$50. ...** Rating: [[3]]

図5 MT-Bench におけるヘッド抑制前の LLM による誤答例。モデルは 2 ターン目に問題を誤解し 3 つ目の停留所だけを考慮している。一方、ヘッド抑制後の LLM は問題を誤解せずに正答できていた。