

Large Language Models Understand and Can Be Enhanced by Emotional Stimuli

Cheng Li¹, Jindong Wang^{2*}

, Yixuan Zhang³, Kaijie Zhu², Wenxin Hou², Jianxun Lian²,

Fang Luo⁴, Qiang Yang⁵, Xing Xie²

¹Institute of Software, CAS ²Microsoft ³William&Mary

⁴Department of Psychology, Beijing Normal University ⁵HKUST

2023/11/12

arXiv

背景

人間の感情的知性

- 感情は人間の日常的な行動や対話に大きな影響を与える
- 感情的な手がかりを理解することは、問題解決において人間に優位性をもたらす

LLMに対する疑問

- 大規模言語モデルは多数のタスクで活用されている
- しかし、LLMが心理学的な「感情的刺激」を真に理解できるか・性能は変わるのかは依然として不確実である

関連研究と本研究のアプローチ

先行研究における知見

- 心理学

期待や自信、周囲からの目といった「感情を揺さぶる刺激」は、人に有益な影響を与える

- プロンプトエンジニアリング

「Step by Step」という一言を指示文の末尾に付け足すことで推論の質が上がる

本研究のアプローチ

- LLMの感情的知性を探求するため指示文の末尾に感情的なフレーズを付け足す「Emotion Prompt」を提案する

目的・RQs

目的

LLMが感情的な言葉を理解し、それを自分のパフォーマンス向上に活かせるかを探る

RQs

- 正解が決まっているタスクにおいて、Emotion PromptはLLMの性能を向上させるか？
- 自由記述問題等の生成タスクにおいて、Emotion Promptは出力の品質を向上させるか？
- Emotion Promptと推論の関係性について分析

実験手順 1/2

[1. プロンプトの設計]

- 心理学の理論をもとに、「これは私のキャリアにとって重要です」などの11種類の感情的なフレーズを作成

[2. 正解がある問題でのテスト]

- 6種類のLLM (GPT-4, Llama 2など) を使い、45種類のタスクを解かせて正答率を測定

実験手順 2/2

[3. 自由記述問題でのテスト]

- 106名の人を集め、GPT-4が生成した文章の「質」を3つの観点で採点

[4. メカニズムの分析]

- AIが文章のどの単語に注目したか (Attention) を解析し、感情プロンプトが効く理由を説明

Emotion Promptの設計

心理学に基づいた3つのアプローチ

- 圧をかける（EP01～EP05）

AIに社会的なプレッシャーを与え、パフォーマンスを意識させる
例:「これは私のキャリアにとって非常に重要です」

- 自信を持たせる（EP07～EP11）

ポジティブな言葉で、目標達成へのモチベーションを高める
例:「自分の能力を信じて卓越性を追求してください」

- 考え直させる（EP03～EP05, EP07）

自分の回答を客観的に見直し、再評価するように促す
例:「確実ですか？もう一度見直す価値があるかもしれません」

RQ1: 正解が決まっているタスクでの結果

結果

- すべてのモデルにおいて、感情プロンプトが通常の指示やCoTを上回る正答率を記録

手法 (プロンプト)	Flan-T5	Vicuna	BLOOM	Llama 2	ChatGPT	GPT-4	平均
通常の指示 (Original)	25.25	44.91	50.33	33.46	75.20	80.75	51.65
Step by Step (CoT)	24.57	33.45	51.35	36.17	75.20	59.72	46.74
感情プロンプト (11種の平均)	22.93	50.56	46.61	35.95	76.85	78.96	51.98
感情プロンプト (最良スコア)	25.53	54.49	50.84	39.46	79.52	81.60	55.24

RQ2: 自由記述タスクでの評価観点

人間106名による評価 (1～5点の5段階評価)

- パフォーマンス（文章の質）

話の筋が通っているか、論理的か、具体例などの根拠があるか

- 正確さ

ウソやハルシネーションを言っていないか

- 安全性・倫理的配慮

社会的な悪影響がないか、人間に対する思いやりや建設的なアドバイスが含まれているか

RQ2: 自由記述タスクでの結果

結果

- 自由記述全体でも、パフォーマンス、正確さ、安全性の3指標すべてで平均 10.9% のスコア向上が見られた

評価指標	通常の指示 (Vanilla)	感情プロンプト (Ours)	人間から見た具体的な効果
パフォーマンス (文章の質)	3.6点	4.1点	主張に対する裏付け証拠や具体例が増え、論理がより明確になる
真実性 (正確さ)	3.6点	3.9点	事実から逸脱した表現が減少する
責任性 (倫理・安全性)	3.4点	3.7点	建設的なガイダンスや人間的な配慮が含まれる回答に変化する

RQ3: AIと感情的刺激の関係性分析

なぜAIに「感情的刺激」が有効なのか

分析方法

「勾配ノルム」という手法を用いて、最終的なAIの出力に対して「入力文のどの単語がどれくらい貢献したか」をスコア化して可視化

アテンションの変化

指示文の後に感情プロンプトを付け足すと、AIの内部で基本の指示文に対するアテンションが濃くなる

RQ3: なぜAIに「感情的刺激」が有効なのか

最も有効な感情的刺激

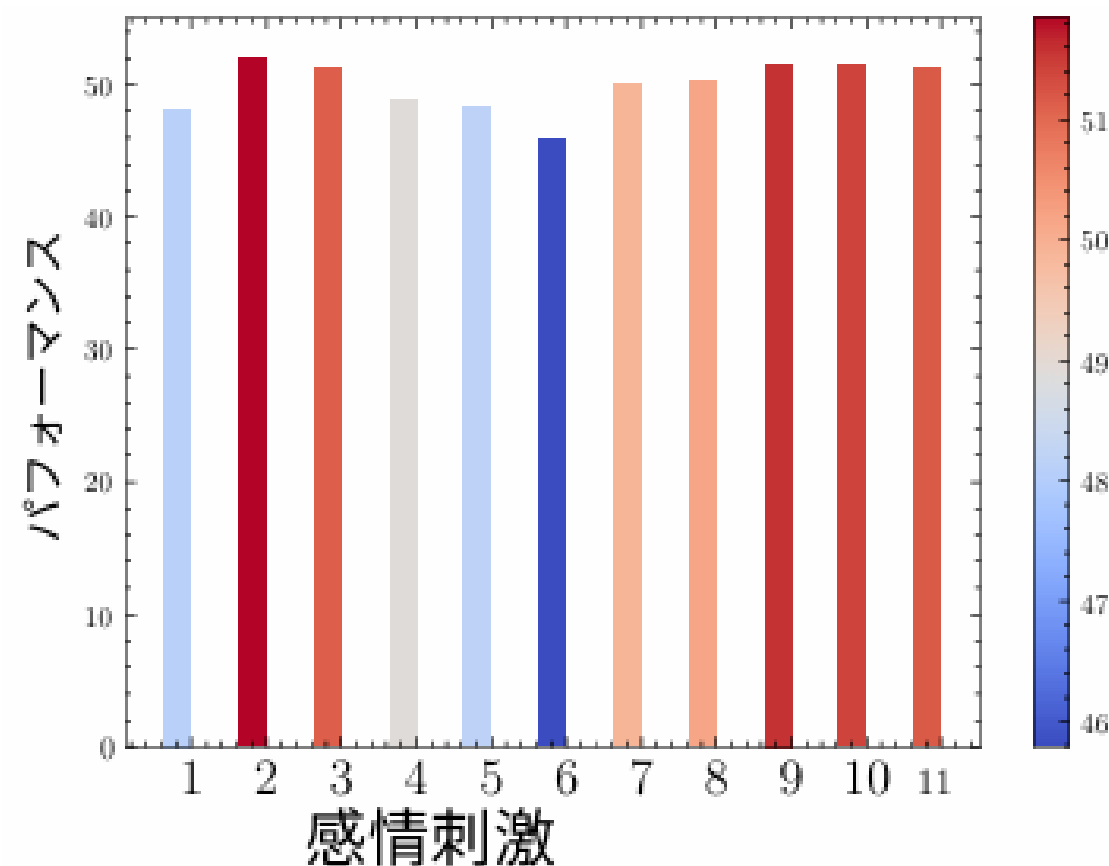


図9: インストラクション誘導におけるすべての感情刺激のパフォーマンス。バーの色は各刺激のパフォーマンスを表しています。

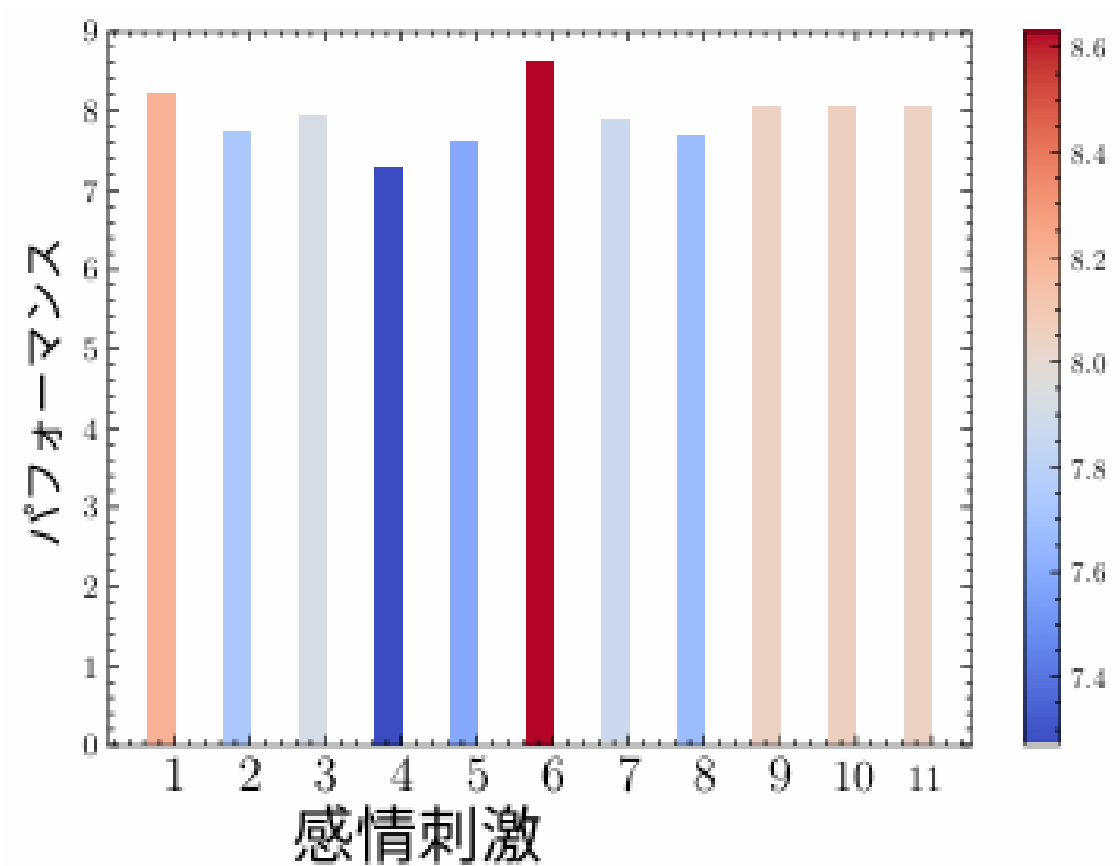


図10: BIG-Benchにおけるすべての感情刺激のパフォーマンス。バーの色は各刺激のパフォーマンスを表しています。

まとめ

本研究の振り返り

- 45種類の自動評価と106名による人間評価を通じ、LLMに対する「感情プロンプト」の効果を検証した
- AIは感情的な刺激を理解し、それによって正答率や文章の品質を有意に向上させることがわかった

結論

- Emotion Promptは、人間とLLMのより良いコミュニケーション方法を考える上で、社会科学の知識（心理学など）がAIの性能向上に役立つことを示している