B9-4

neoAI-InstructBench
実践的シナリオに基づく日本語複合指示追従ベンチマーク

川本 稔己[1]　板井 孝樹[1,2]　大槻 真輝[1,3]（[1]株式会社neoAI [2]東京都立大学 [3]東京大学）

概要

背景: 実運用のLLMには、形式・文体・内容など複数の指示を同時に満たす複合的な指示追従能力が求められる

課題: 既存ベンチマークは、単一カテゴリや定型的な指示が中心で、実利用の実態を反映しきれていない

提案: 実運用ログに基づく日本語複合指示ベンチマーク neoAI-InstructBenchを構築

結果: GPT-5.2でもタスク完遂率は67%に留まり、指示間の干渉や形式・文体の指示がボトルネックになることを解明

研究背景

背景

LLMの社会実装に伴い、実利用シーンにおけるユーザーの指示は複雑化しており、形式・文体・内容など、タスク文脈から自然に生じる性質の異なる指示の同時充足が求められる。

特に日本語では、敬語体系や表記揺れなど言語固有の複雑さが加わり、実運用に近い形で評価することが重要になる。

課題

既存ベンチマーク (e.g. IFEval, IFBench, IFScale, FollowBench)は、特定の単一カテゴリー内の指示の組み合わせ・客観判定可能な指示(Closed-ended)・定型的な付け足しで設計されている。

実運用環境では、タスクの文脈から自然に生じる複数指示の同時要求であり、Open/Closedの同時要求や複数カテゴリーは既存ベンチマークでは反映できていないことが多い。

目的

実運用ログを起点に、5カテゴリを複合させた日本語指示追従ベンチマークを構築し、既存指標では不可視だった指示間干渉の実態と、複合指示下におけるモデルの限界を解明する。

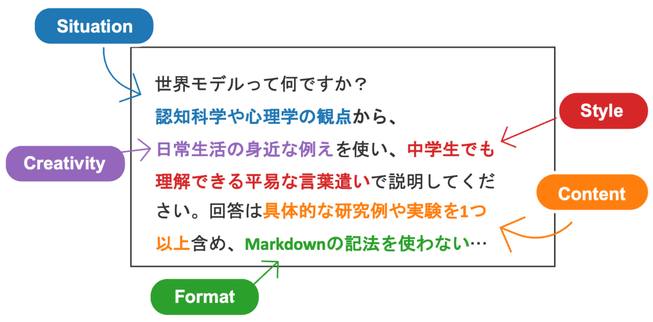
neoAI-InstructBench

データソース: neoAI Chatの社内実運用ログ（個人情報・機微秘匿情報除去済み）

サイズ: 100タスク（プロンプト）・計326指示（個別の条件）・1タスクあたり2〜5個の指示

構築フロー

データソース
↓
ソースに関連した自然な指示候補を生成
↓
人手による機微情報のマスク・作成・複合化
↓
人手によるレビュー/修正



5つの指示カテゴリー

FollowBenchをベースに、実運用に適さないFew-shot前提のExampleを除外、Creativityを追加

カテゴリー	定義	具体例
Format	構造・レイアウト・記法	JSON形式・箇条書き・Markdown禁止
Content	語彙・事実情報	単語AをN回使用・100文字以内・Aに関しては一切言及しない
Style	役割・トーン・文体	情熱的な文体・敬語・「！」をつけて
Situation	状況理解・推論・判断	翻訳・Aという前提を踏まえてSWOT分析を行う
Creativity	新規アイデア・比喩	関係を日常生活に例えて説明・造語生成

問題例

指示数2: 日焼け止めを塗り忘れてひりひりするんだけどどうしたらいい？部活の熱血顧問のような口調で、気合と根性を込めつつも身体を労るように、肌の回復プロセスを料理に関連した比喩を使って説明してください。

指示数5: git revertとgit resetの使い分けについて、mainブランチへのforce pushは禁止という前提で教えてください。説明にあたっては、ネットスラングを多用した匿名掲示板風の文体にしつつも、Git初心者に心理的な不安を与えないよう配慮してください。出力は全体を見出しレベル2（##）で区切った3つのセクションに分け、その中でタイムトラベルを題材にした短い物語を用いて2つのコマンドの違いを解説してください。

実験

評価指標

- Prompt Acc (タスク正解率): 1つのタスクの全指示を遵守した割合
- Inst. Acc (個別指示正解率): プロンプト内の個別の指示ごとの遵守率

評価手法: ハイブリッド評価

- Closed-ended: ルールベース（文字数、単語有無など）
- Open-ended: LLM-as-a-judgeによる判定（文体、創造性など）

設定

- 最大トークン: 10,240
- Reasoning: High相当
- LLM-as-a-judge: GPT-5.1

Model	Prompt Acc	Inst. Acc
GPT-5.2	**67.00**	**88.34**
GPT-5.1	66.00	87.12
GPT-5 mini	61.00	75.15
GPT-5 nano	35.00	58.59
Claude Opus 4.5	56.00	83.74
Claude Sonnet 4.5	51.00	81.60
Claude Haiku 4.5	39.00	74.23
Gemini 3 Pro	51.00	81.60
Gemini 2.5 Flash	45.00	77.30
Gemini 2.5 Flash Lite	37.00	73.31
Kimi K2 Thinking	47.00	79.75
Qwen3 235B A22B Instruct	42.00	74.85
Qwen3 235B A22B thinking	47.00	77.91
gpt-oss-120b	45.00	79.75
gpt-oss-20b	47.00	75.46
Gemma 3 27B Instruct	20.00	62.27

結果

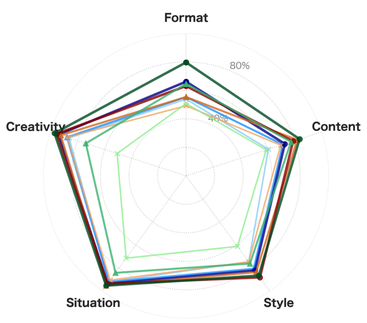
GPT-5.2が最高性能を記録

個々の指示では88%だが、Prompt全体だと67%まで低下

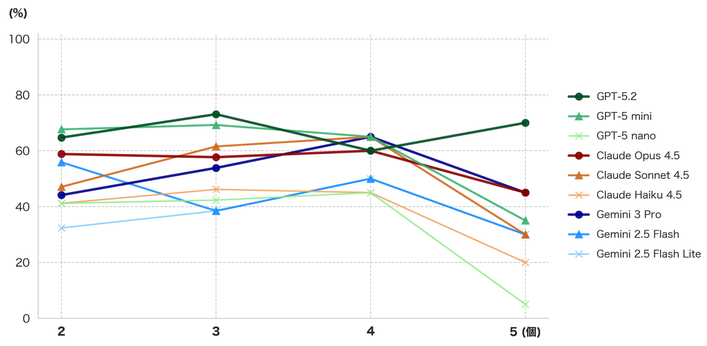
Open weightsモデルは最大47%に留まる

主要モデルの結果詳細

カテゴリー別のInst. Acc



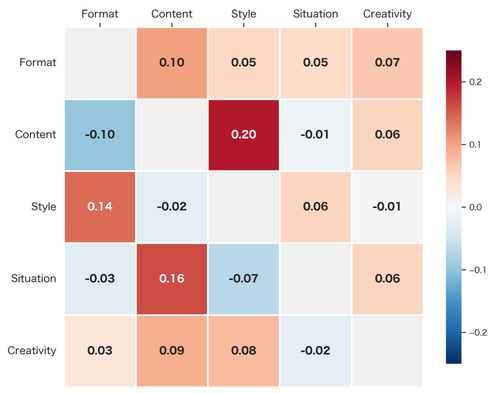
指示数ごとのPrompt Acc



分析と考察

指示間干渉

異種指示の同時入力が、性能に正負の影響を与える現象を確認



GPT-5.2 における指示カテゴリ間の干渉効果

各セルは行の指示が存在することで列の指示の正解率がどの程度変化したか。

値が負であれば干渉により精度が低下、正であれば精度が向上

負の干渉: Contentが含まれるとFormatの正解率が約10%低下

正の干渉: Contentが含まれるとStyleの正解率は約20%向上

失敗様式

- Overthinkingによる無回答: 一部のモデルで3回のリトライを含む全ての生成においてReasoningのみで最大トークン数に達し、回答が出力されない現象を確認 (GPT-5 nano: 26 件, GPT-5 mini: 11件, GPT-5.1: 2件)
- ボトルネック: FormatやContent指示でのミスが目立つ

  GPT-5.2が間違えた指示: 改行をする際は必ず3回してください。・(キャラ) [行動]「セリフ」のフォーマットに従ってください。・全ての文末には「？」か「！」をつけてください。

まとめ

- 実運用ログに基づく日本語指示追従ベンチマークを構築・公開
- 最先端モデルであっても、複合指示下では干渉や表記指示による破綻が生じる。
- 今後は干渉の抑制、活性化に対するさらなる調査およびOverthinkingの制御が重要な課題となる。

リソースはこちらまたは上のQRコードから！
https://github.com/neoAI-inc/neoAI-InstructBench