

### 背景・課題・目標

GPTなどの大規模言語モデルは流暢だが論理性、正確性に欠ける

#### 大規模言語モデルの構築

透明性、信頼性の観点から独自のLLM構築が必要

OpenAIのInstructGPTに学び精度向上の鍵を握る  
日本語インストラクションを構築

18社との共同研究を通じて  
1万件の質問回答ペアを作成

#### 構築データ・評価

2023年12月 4802件公開  
2024年2月末 1万件公開予定  
操作、分類などのタグ付き

```
"ID": "ichikara-instruction-003-001-0000002-001",
"text": "「1区=いっぴき」と数えるのに、「2区→2びき」と数えるのは、「種別が多さ」と「音の変化=変音現象」が特徴です。YoYo紙一枚(まい)、車一台(だい)、列車(れっしやう)、著一冊(ぜん)、エンジン一基(き) YoYo種類の多さもさることながら、もう一つの難しさが変音現象です。YoYo例えば、カエルを数えるときに「1区=いっぴき」と読むのに対して「2区=にびき」で2匹を「にびき」とは読まずに「2区=にびき」となります。YoYoどうしてこうなるのか、YoYoここには、漢語の半濁音のルール、というものが発生します。YoYoまず、半濁音はハ行にだけしか付きません。(はひふへほ→ばびぶべ) YoYoそして、連濁という語と語が合体すると、後ろの語の頭の部分が音から濁音に変化します。YoYoハ行音の連濁のうち、前の語の最後に「つ」がくると、後ろの語の頭が半濁音になります。YoYo例) 切(きつ) + 腹(はら) = 切腹(きつぷく) YoYo Yo助数詞も、数字の「一・二・三・四」が「ハ行音の助数詞」と結びつく時、促音便「っ(小さい)っ)」を起こし、後ろの助数詞が半濁音となります。YoYo例) 1(いち) + 1(いっ) = 1びき、6(ろく) + 1(いち) = 6びき、8(はち) + 1(いち) = 8びき、10(じゅう) + 1(いち) = 10びき、100(ひゃく) + 1(いち) = 100びき YoYo Yoこういふ理由から、「2区=2びき」とはなりません。",
"meta": {"操作": "オープンQA", "主観客観": "客観", "時間依存": "なし", "対象": "言語知識", "分野": "国語", "回答タイプ": "文章"}
```

#### <実験結果>

rinna/youri7B + instruction4802(houou)の出力と  
GPT3.5-turboの出力を比較

- GPT4による判定: 22勝11敗7引分
- 人間による判定: 12勝27敗1引分

#### <観察・考察>

GPT3.5は正確性が高く、hououは情報量が高い

- hououは具体的な数値や例による説明を含み情報量が多いが誤りを含み正確性が低下
- GPT3.5は一般的な説明のみで、具体性が低いため、正確性は低下しない
- 情報量と正確性はトレードオフの関係
- GPT4は正確性を判定できずhououを評価。人間は正確性を判定し、GPT3.5を評価。
- GPT4の判定理由はハルシネーションを多く含む
- 他にもインストラクションの指示が回答に反映

インストラクションは非常に重要。個性を形成

#### 構造化知識の構築

論理的な推論のために大規模な構造化知識が必要

Wikipediaを、整理された知識構造(拡張固有表現)に  
当てはめ、計算機利用可能な構造化データを構築

Resource by Collaborative Contribution Since 2017  
評価WSを通して協働で知識作成

#### 構築データ

タスク	学習データ	システム出力
日本語分類	920,444ページ	5システム 1,286,205ページ
多言語(30言語)分類	5,029,617ページ	12システム 32,555,929ページ
属性値抽出	19,711ページ 910,567属性値	13システム 6,089,547属性値
リンク	2453ページ 80,600リンク	準備中

構築ツール

アクセスAPI

### LLMと構造化知識の融合

#### JEMHopQA 日本語マルチホップQA

質問: ルーヴル美術館が所在する都市の市長の名前は?  
導出: (ルーヴル美術館, 所在地, パリ), (パリ, 市長, アンヌ・イダルゴ)  
回答: アンヌ・イダルゴ

図1: 構成質問の例

質問: 『天空の城ラピュタ』と『となりのトトロ』の公開日が早いのは、『となりのトトロ』ですか?  
導出: (天空の城ラピュタ, 公開年, 1986年), (となりのトトロ, 公開年, 1988年)  
回答: NO

図2: 比較質問の例

知識と推論スキルの両方が  
必要なタスク

GPT-4の正解率は60%  
しかし  
その1/3は導出が  
間違っている(偽正解)  
本当の正解率は40%

#### 偽正解の例

JEMHopQAの正解セット  
質問: 長嶋茂雄と小林旭、どちらが年上ですか?  
導出: (長嶋茂雄, 生年月日, 1936年2月20日); (小林旭, 生年月日, 1938年11月3日)  
回答: 長嶋茂雄

GPT-4の出力  
導出: (長嶋茂雄, 生年月日, 1936年1月20日); (小林旭, 生年月日, 1939年4月13日)  
回答: 長嶋茂雄

正確な導出知識:  
GPT-4のみ: 40%  
GPT-4+構造化知識(森羅&wikidata): 77.5%