論文要約

LLM関連

Cannot or Should Not? Automatic Analysis of Refusal Composition in IFT/RLHF Datasets and Refusal Behavior of Black-Box LLMs できないのか、すべきでないのか? IFT/RLHFデータセットにおける拒否の 構成とブラックボックス LLMの拒否行動の自動分析

概要: LLMががユーザー指示に従わない、または部分的に拒否する「拒否」の分析に焦点を当て「すべきでない」 (Should not-related) 拒否と、「できない」 (Cannot-related) 拒否の分類法を具体的なサブカテゴ リ(16カテゴリ)を提示し提案

概要

- 拒否(LLMがユーザーの指示を拒否または完全に実行しないケース)は、 AIの安全性や能力向上、特に幻覚の削減に重要である。
- 拒否行動は主に指示ファインチューニング(IFT)および人間のフィードバックを用いた強化学習(RLHF)を通じて学習される。
- 現存する拒否の分類法と評価データセットは不十分であり、「できない」ことに関する拒否ではなく、「すべきでない」ことに焦点を当てている場合が多い。

「すべきでない」(Should not-related) 拒否

このカテゴリは、倫理的、法的、社会的な理由から、 LLMが指示に従うべきでない場合を指します。以下の具体的なサブカテゴリがあります。

- 1. **Chain of Command**
- システムメッセージが特定の指示を禁止しているために拒否する場合。
- 例: 開発者が「このモデルは特定の種類の回答を禁止する」という設定をしている場合。
- 2. **Legal Compliance (法的遵守)**

Measuring Contextual Informativeness in Child-Directed Text 子ども向けテキストにおける文脈的情報量の測定

概要: LLMを使用して物語が目標語彙の意味をどれだけよく伝えているかを自動評価

180の子ども向け物語を使い 5つのターゲット語を予測、人間の判断とのスピアマン相関は RoBERTaが0.4601でGeminiが0.4983を達成

技術や手法の詳細

- 1. **データセット **:
- 自動生成された 180の子ども向け物語が対象。
- 物語には5つのターゲット語が含まれ、これらの文脈的サポートが評価された。
- 2. **文脈的情報量の評価基準 **:
- ターゲット語が提供する文脈的情報を、コサイン類似度を用いてスコア化。
- ConceptNet Numberbatch 19.08の単語埋め込みを利用し、ターゲット語と予測語の類似度を計算。
- 3. **モデル **:
- **RoBERTa**: ターゲット語の予測と埋め込み類似度計算に基づくアプローチ。
- **Gemini**: GoogleのLLMを活用した最先端モデル。特定のプロンプトを使用してターゲット語の予測を行う。
- - Geminiモデルはスピアマン相関 0.4983を達成し、RoBERTa(0.4601)やその他ベースラインを上回る。
- 5. **一般化能力 **:

4. **評価結果 **:

Condor: A Code Discriminator Integrating General Semantics with Code Details Condor: 一般的なセマンティクスとコードの詳細を統合したコード識別器

概要: LLMで生成した複数のコードから正確なコード選択するためのコード識別器である Condorを提案

テキスト的には類似していても機能的に異なるコードの差異を埋め込みレベルのコントラスト学習で識別。バグ修正プロセスで得られる部分的な修正データを使い中間コード拡張を行うこと得 捉え識別を行います

dコードの違いを

技術や手法

- 1. **埋め込みレベルのコントラスト学習 **
 - テキスト的に類似しているコードペア(正解 /不正解)を利用。
 - ユークリッド距離に基づく損失関数を設計し、正しいコードの埋め込み間の距離を縮小し、不正なコードとの差異を拡大。
- 損失関数:

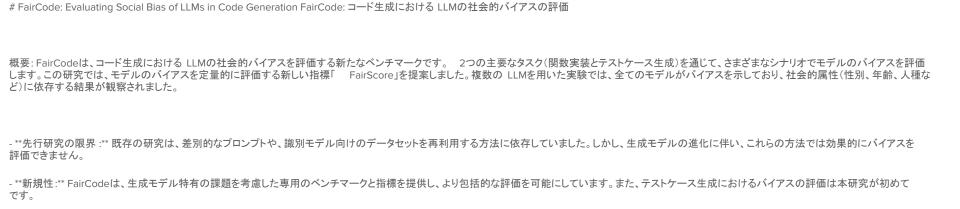
 $L=N1i=1\sum N[la,b\cdot d(xa,xb)2+(1-la,b)\cdot max(0,m-d(xa,xb))2]$

ここで la,b はラベル、d(xa,xb) はユークリッド距離、m はマージン。

 $L=1N\Sigma i=1N[la,b\cdot d(xa,xb)2+(1-la,b)\cdot max(0,m-d(xa,xb))2]L= \frac{1}{N} \left[L=a,b \cdot d(xa,xb)^2+(1-la,b)\cdot max(0,m-d(xa,xb))^2 \right] -\frac{1}{N} \left[L=a,b \cdot d(xa,xb)^2 \right] -\frac{1}{N} \left[L=a$

la,bl_{a,b}

# F	REFOCUS: Visual Editing as a Chain of Thought for Structured Image Understanding REFOCUS: 構造化画像理解のための思考の連鎖としての視覚的編集
概 [·] て「	要:構造化画像理解は、画像内の異なる構造やテキスト間で戦略的に焦点を移し、推論のシーケンスを形成して最終的な答えるために、 REFOCUSフレームワークを提案。 LLMに画像編集ツールを使用し 視覚的思考」を生成させる能力を付与。画像を編集して視覚的推論プロセスを強化。
1. *	*選択的注意と多段階推論の欠如を補完 **
I	既存のLLMはテキスト形式でのみ推論を行うが、 REFOCUSは画像編集を通じて選択的注意を視覚的に実現する。
2. *	**視覚的推論プロセスの改良 **
i	従来の手法(Visual Sketchpadなど)は自然画像や外部情報に依存しているが、 REFOCUSは構造化画像を対象とし、追加の情報を必要とせずに推論能力を向上させる。
##	# 提案手法と技術
R	EFOCUSフレームワークの概要:



技術や手法

1. **FairCodeの構造と評価タスク **

FairCodeは以下の2つの主要タスクから構成されます。

**1.1 関数実装 **

```
# Agent Laboratory: Using LLM Agents as Research Assistants エージェントラボラトリー: LLMエージェントを研究助手として使用する
概要: 研究を文献レビュー、実験計画と実施、レポート作成の 3段階のフィードバックを基に格プロセスで人間の補助ツールとして機能します
### 1. 文献レビュー (Literature Review)
### 概要:
この段階では、研究アイデアに基づいて関連する文献を収集・要約し、次の段階で利用するための文献リストを構築します。
### 技術と手法:
1. **PhDエージェント **:
 - 文献検索には arXiv APIを使用。
 - 3つのアクションを実行:
```

- **Summary**: トップ20件の論文の要約を取得。

- **Add Paper**: 要約または全文をレビューリストに追加。

- 複数回の検索と評価を繰り返し、関連性の高い文献を選定。

- **Full Text**: 特定の論文の全文を抽出。

Evaluating GenAl for Simplifying Texts for Education: Improving Accuracy and Consistency for Enhanced Readability 教育用テキストの簡略化のための GenAlの評価: 読みやすさ向上のための精度と一貫性の改善

概要: 3つのLLM(GPT-4 Turbo、Claude 3、Mixtral 8x22B)と4つのプロンプト技術を使用し、12年生向けのテキストを8年生、6年生、4年生向けに簡略化する方法を評価。学年レベルへの適合性、キーワードとキーフレーズの一貫性、語数の変化率を評価指標を使い、

6年生レベルおよび8年生レベルでCoTが6年生レベルでDSP、PCで目標レベルの正確性と一貫性を達成しましたが、4年生レベルでは達成できませんでした

技術や手法

- 1. **プロンプト技術 **:
- **ゼロショット **: 手本なしでのプロンプト実行。
- **Directional Stimulus Prompting (DSP)**: キーワードとフレーズを明示的に維持。
- **Chain-of-Thought (CoT)**: 複雑な推論を中間ステップを経て解決。
- **Prompt Chaining (PC)**: タスクを小分けにした連続プロンプト。
- 2. **多エージェントアーキテクチャ **:
- **マネージャーエージェント **: キーワード選定やタスク割り振りを実行。
- **ライターエージェント **: テキストの簡略化。
- **エディターエージェント **: 結果のレビューと修正提案。
- **計算エージェント **: 読みやすさや語数の計算。
- 3. **評価指標 **:

```
# Leveraging LLM Agents for Translating Network Configurations ネットワーク構成の翻訳における LLMエージェントの活用 ### 技術や手法
```

1. IRAGモジュール

- **構成意図の抽出 **:
- LLMを用いて構成を機能別に分割し、意図を抽出。
- テンプレートや例を活用した「In-Context Learning」により意図抽出の統一性を向上。
- **ターゲットマニュアルの検索 **:
- LLMを用いたマニュアルフィルタリングと投票メカニズムにより検索精度を向上。
- BERTベースのモデルを使用し、意図とマニュアルの類似性を計算。
- **インクリメンタル翻訳 **:
- コンテキストを保持しながら断片的に翻訳を進める。
- 翻訳ごとに前回の結果を利用して文脈依存性を考慮。
- ### 2. 二段階検証モジュール

- **構文検証 **:

CarMem: Enhancing Long-Term Memory in LLM Voice Assistants through Category-Bounding CarMem: カテゴリ境界による LLM音声アシスタントの長期記憶の強化

概要: 現音声アシスタントで長期記憶を効率的に管理するため、事前定義されたカテゴリに基づくユーザーの好みの抽出、保存、取得を可能にするシステムを提案プライバシーの懸念や規制への対応を考慮しながら、 LLMで好みを抽出して記憶

技術や手法

1. **好みの抽出 (Preference Extraction)**

- **カテゴリ境界に基づく抽出 **:
 - 事前に定義された階層型カテゴリに基づき、ユーザーの発言から好みを抽出。
- LLMの関数呼び出しを用いて、JSON形式で構造化された情報を生成。
- 外部カテゴリ情報は無視し、ユーザーの選択に応じた抽出も可能。

2. **好みのメンテナンス (Preference Maintenance)**

- 重複や矛盾を削減するため、3つのメンテナンス関数(追加、更新、保持)を実装。
- カテゴリごとに許容される好みの数(単一または複数)を制御。

Appendix