論文要約

LLM関連

Instruction Embedding: Latent Representations of Instructions Towards Task Identification 指示埋め込み：タスク識別に向けた指示の潜在表現

概要: 指示データを分全体の意味ではなくタスク特定に重点を置くために動詞　-名詞ペアを強調する Instruction Embeddingを提案しタスク識別を行い

この埋め込みを LLMのプロンプトを使用し特定のタスクに注意を向けるようにする、　PIE手法を利用する。

従来の研究では、指示のテキスト埋め込みは主に全体的な意味情報の取得に焦点が置かれていましたが、この論文では、指示埋め込みはタスクの特定に重点を置くべきであると指摘しています。従来の埋め込み手法では、異なるタスク間の意味的類似性が強調され、異なるタスクを正確に区別することが難しかったが、本研究の　PIE手法はこの問題に対処し、より精度の高いタスク特定を可能にしています。

技術や手法の説明：

1. **指示埋め込み（Instruction Embedding）**:
   - 文全体の意味的情報ではなく、タスク特定に重点を置いた埋め込み方法。タスクの動詞　-名詞ペアを強調し、タスクの同定をより正確に行うためのもの。
   - 指示のタスクを識別するため、意味的な類似性よりもタスクの類似性を優先。

2. **IEBベンチマーク（Instruction Embedding Benchmark）**:
   - タスクの識別能力を評価するために構築されたベンチマークデータセット。従来のテキスト埋め込みベンチマークが意味的類似性を評価するのに対し、　IEBはタスクの違いに基づいて指示を分類することを目指す。
   - 47,161のサンプルを含み、1,353のタスクカテゴリーに分類。

3. **PIE（Prompt-based Instruction Embedding）**:

MemSim: A Bayesian Simulator for Evaluating Memory of LLM-based Personal Assistants MemSim: LLMベースのパーソナルアシスタントのメモリを評価するためのベイジアンシミュレーター

概要: LLMエージェントの MemSimではQAにBayesian Relation Network (BRNet) 使用しユーザーのプロファイルを関係性や年齢などの属性でサンプリングし質問に必要な情報をヒントの形でまとめ Aを生成します

**技術や手法 **

1. **Bayesian Relation Network (BRNet):**

ユーザーの属性や関連エンティティをベイジアンネットワークとして表現し、階層的なユーザープロファイルをサンプリングする手法。 BRNetは、ユーザーの関連する属性やエンティティの確率分布をモデル化し、シナリオに応じた多様でスケーラブルなユーザープロファイルを生成する。

2. **Causal Generation Mechanism:**

階層的なユーザープロファイルに基づいて、ユーザーメッセージと QAを因果的に生成するメカニズム。これは、プロファイルからヒント(factual hints)を抽出し、それに基づいてメッセージや QAを作成するプロセスで、LLMのハルシネーションの影響を軽減する。

3. **MemDaily Dataset:**

MemSimを用いて日常生活シナリオを元に構築されたデータセット。シンプルな QAから多段階の推論を必要とする QAまで、さまざまなタイプの QAを含む6つのサブデータセットを提供する。

**使用用途 **

SELF-CONTROLLER: CONTROLLING LLMS WITH MULTIROUND STEP-BY-STEP SELF-AWARENESS セルフコントローラー：マルチラウンドのステップバイステップの自己認識による　LLMの制御

概要：自己認識を導入して LLMが自身の状態と現状を認識し制御性を強化。テキスト長の線形性と単調性に基づいた二分探索アルゴリズムを実装しています

**技術や手法 **

1. **Self-controllerの全体像：**

　Self-controllerは、LLMが自身の出力に基づいて状態を認識し、次の出力を調整するマルチラウンドの対話フレームワークです。これにより、　LLMは自己認識を持ち、より精密な制御が可能になります。

2. **状態反映器（ State Reflector）の役割：**

　- **状態更新：** LLMの各応答を解析し、現在の状態変数を更新します。例えば、テキスト生成タスクでは、これまでに生成した単語数を計算します。

　- **フィードバック提供 ：** 更新された状態を LLMにフィードバックし、次の出力に反映させます。

3. **マルチラウンド対話の流れ ：**

　- **初期プロンプト：** ユーザーからのタスク指示と目標状態（例：目標とするテキスト長）を　LLMに提示します。

　- **応答生成：** LLMは現在の状態と目標に基づいて出力を生成します。

　- **状態更新とフィードバック ：** 状態反映器が LLMの出力を解析し、状態を更新してフィードバックします。

REGENESIS: LLMS CAN GROW INTO REASONING GENERALISTS VIA SELF-IMPROVEMENT REGENESIS: LLMは自己改善を通じて推論の汎用性を持つことが可能になる

概要: ReGenesisは特定タスクに適応する形で推論ガイドラインを生成し、それを使用して推論構造を生成し解決ステップを決定。ステップに基づき回答を生成、生成された推論経路の中から正解の経路をフィルタリングし、これをモデルの再学習に使用します

ReGenesisは既存の自己生成手法（例： STaRなど）が抱える問題を克服しており、特にタスク間の汎用性（ OODタスク）において優れたパフォーマンスを発揮します。従来手法が特定タスクに偏った推論経路 を生成することにより外部タスクでの性能が低下するのに対し、 ReGenesisは一般的なガイドラインを用いることで幅広いタスクに対応可能な推論経路を生成します。

技術や手法

1. **推論ガイドラインの適応 (Guidance Adaption)**: 一般的な問題解決戦略を特定タスクに適応する形で推論ガイドラインを生成する。この段階では問題を直接解決するのではなく、全体的な解決戦略をタスクに適応させる。

2. **推論構造の生成 (Reasoning Structure Generation)**: 適応された推論ガイドラインに基づいて詳細な推論構造を生成します。この段階でも具体的な解を出すことはせず、解決のためのステップを決定し ます。

3. **推論経路の生成 (Reasoning Paths Generation)**: 推論構造に従い、実際の解決手順を生成します。この手順により具体的な答えを得ます。

4. **推論経路のフィルタリング (Filtering Reasoning Paths)**: 生成された推論経路の中から正解に一致する経路をフィルタリングし、モデルの再学習に使用します。

使用用途

ReGenesisは、数学的推論、論理的推論、常識的推論などの複雑な推論タスクに適用可能です。特に、汎用的な推論能力を向上させることにより、未知のタスク（ OODタスク）にも対応可能な LLMを作り出すことができます。

Comparing Criteria Development Across Domain Experts, Lay Users, and Models in Large Language Model Evaluation ドメインエキスパート、一般ユーザー、およびモデル間の大規模言語モデル評価における基準開発の比較

概要: LLMをドメイン固有のタスクで使用した際の評価プロセスを提案

基準設定として事前段階 (a priori)を参加者がプロンプトのみで基準設定、実際の出力を確認し修正する事後段階 (a posteriori) の2段階で設定

これにより専門性とユーザーの理解しやすさ、評価基準の修正ができます

技術や手法

1. **基準設定プロセスの比較 **:

    1. **基準設定の 2段階プロセス **:

        - **事前段階 (a priori)** と **事後段階 (a posteriori)** の2段階に分けて評価基準を設定するプロセスを行います。

        - **事前段階 (a priori)** では、参加者(ドメインエキスパートや一般ユーザー)がプロンプトのみを見て基準を設定します。この段階は、モデルの出力を見ないでどのような基準が必要かを考えるフェーズです。

        - **事後段階 (a posteriori)** では、参加者が実際のモデルの出力を確認した後、基準を修正または追加するフェーズです。この段階は、モデルの実際の出力の内容に基づいて、最初の基準をどのように調整するかを考えるフェーズです。

    2. **参加者グループの分類と役割 **:

        - 研究では、3つの異なるグループが評価に参加しています。

            - **ドメインエキスパート **(栄養学や教育学の専門家): 彼らは深い知識を持ち、具体的な基準を設定することで、特定のドメインでの信頼性を確保する役割を担っています。

            - **一般ユーザー **: 専門的な知識がなく、ユーザー視点から出力の使いやすさや理解しやすさを重視して基準を設定します。

            - **LLM**(大規模言語モデル): モデル自身が基準を生成し、人間の基準と比較する対象となります。

DreamGarden: A Designer Assistant for Growing Games from a Single Prompt DreamGarden: シングルプロンプトからゲームを成長させるデザイナーアシスタント

概要: DreamGardenは、ゲーム環境を作成するデザイナー支援で、プロンプトを階層的に分割し、具体的なアクションプランを生成。プランはサブモジュールに分配し具体的な実装をします。特に Unreal Engineでのゲーム開発を支援し、自由形式のプロンプトから自律的にゲームのプロトタイプ生成します

**技術や手法 **

DreamGardenの中核は、大規模言語モデル(LLM)を使用したプランニングモジュールである。このプランニングモジュールは、ユーザが提供したプロンプトを元に階層的なアクションプランを生成し、それを様々なサブモジュールに割り当てることで実装を行う。具体的には、以下の手法を用いる :

1. **プランニングモジュール **:
   - ユーザが提供するシードプロンプトを元に、高レベルのゲームデザインを広範囲に計画する。
   - その後、各プランを詳細なステップに分割し、最終的に具体的なタスクにまで落とし込む。
2. **実装サブモジュール **:
   - サブモジュールには、 C++コード生成、手続き型メッシュ生成、ディフュージョンメッシュ生成などがある。
   - 各タスクは特定のサブモジュールに割り当てられ、それに応じた実装が行われる。
   - 生成されたコードやアセットは Unreal Engineに取り込まれ、シミュレーションやフィードバックが行われる。
3. **ユーザーインターフェース **:

AI AS HUMANITY'S SALIERI: QUANTIFYING LINGUISTIC CREATIVITY OF LANGUAGE MODELS VIA SYSTEMATIC ATTRIBUTION OF MACHINE TEXT AGAINST WEB TEXT 人類のサリエリとしての AI：ウェブテキストに対する機械テキストの体系的な帰属を通じた言語モデルの創造性の定量化

概要: LLMの人間の創造性評価として CREATIVITY INDEXを提案。ウェブテキストをどの程度再構成できるかを計測

テキストの n-gramがウェブ上に存在するかどうかを評価し厳密一致や近似一致をみる L-ユニークネスを算出します

技術や手法の詳細説明

この論文では、LLM（大型言語モデル）の創造性を測定・評価するために、新しい指標「 CREATIVITY INDEX」と動的プログラミングアルゴリズム「 DJ SEARCH」を提案しています。以下、技術と手法の詳細を順を追って説明します。

1. **CREATIVITY INDEX**

CREATIVITY INDEXは、LLMが生成したテキストの創造性を定量的に測定するために提案された指標です。主なアイデアは、生成されたテキストがどの程度ウェブ上の既存テキストから再構築できるかを評価することです。具体的には、与えられたテキスト内の n-グラム（連続する n個の単語）が、ウェブ上の巨大な参照コーパスに存在するかどうかを確認します。

- **L-ユニークネス (L-uniqueness)**: テキストの中で、指定された長さの n-グラムが参照コーパスに含まれていない割合を示します。これは、テキストの中のどの単語やフレーズが新しい文脈で使われているかを定量的に示すもので、より高い値がより高い創造性を意味します。

- **CREATIVITY INDEXの計算**: テキストのすべての n-グラムについて L-ユニークネスを計算し、それを積み上げて総合的な創造性を示す指数とします。これは、言い換えると「 L-ユニークネス曲線の下の面積」を求めることと同等です。

この指標を用いることで、生成されたテキストがウェブ上の既存のテキストからどの程度新しいものを取り入れているかを定量的に評価し、 LLMの創造性を人間の作家と比較することができます。実験の結果、プロの作家の創造性が LLMよりも大幅に高いことが確認されました。

Document-level Causal Relation Extraction with Knowledge-guided Binary Question Answering 知識ガイド付き二項質問応答による文書レベルの因果関係抽出

概要: Knowledge-guided Binary Question Answering（KnowQA）は因果関係を分類するイベント間因果関係抽出（ ECRE）をLLMのゼロショットで文書内のイベントを抽出しその構造を構築するイベント構造構築と因果関係の有無を識別する二択で答えられる質問を使い分類します

技術や手法

1. イベント構造の構築（ Event Structure Construction Module）

イベント間の因果関係を抽出するために、 KnowQAでは文書レベルでのイベント構造の構築を行います。これは 3つの主要なステップで構成されています。

1.1 イベント検出（ Event Detection）

まず、文書中のイベントを検出します。これは、事象（イベント）を特定し、それを特定のイベントタイプに分類するプロセスです。具体的には、以下の手法が使用されています。

- **KAIROSオントロジー **を利用し、イベントの分類を行います。 KAIROSオントロジーは ACE 2005（Automatic Content Extraction）の拡張セットであり、 50のイベントタイプと 59の引数役割をカバーしています。

- **CLEVE**（Contrastive Pre-training for Event Extraction）と呼ばれる事前学習モデルを使用し、イベントの分類精度を向上させています。このモデルは事前学習された PLM（Pre-trained Language Model）で、**WikiEventsデータセット **でトレーニングされており、各イベントを KAIROSオントロジーに従って分類します。

1.2 イベント引数抽出（ Event Argument Extraction）

Give me a hint: Can LLMs take a hint to solve math problems? ヒントを教えて： LLMは数学の問題を解くためのヒントを活用できるか？

概要: LLMに対してヒントを与えることでの数学問題解決能力についての有効性を評価

モデルに対して質問をヒントや例なしで回答した結果ベースとし、品とありワンショットで例をフューショットで複数例、誤ったヒントやランダムなヒントに対して評価を実施し、ヒントありのものが CoTより良い結果、ワンショットは複雑な問題の解法が難しく、フューショットは与えた例に結果が依存する。誤ったヒントや例を渡すと結果は大幅に悪くなる結果になり、ヒントが数学的推論向上に有効なことがわかりました

評価結果の詳細

この論文では、様々な種類のプロンプト手法（ベースライン、ヒント、ワンショット、フューショット、チェイン・オブ・ソートなど）を使って、 LLMが数学問題を解く能力を評価しています。ここでは、各手法の評価結果について順番に詳しく説明します。

1. ベースラインプロンプト

**ベースライン **では、モデルに対して単に問題を与え、ヒントや例を提示せずに問題を解くよう促しました。結果として、ベースラインのスコアは比較的低く、平均的なパフォーマンスとなりました。これは、ヒントや例がない状態ではモデルが正確に問題を解決するのが難しいことを示しています。

2. ヒントプロンプト

**ヒントプロンプト **を使用した評価では、各問題に対して適切なヒントを与えることで、モデルのパフォーマンスが向上しました。この手法は、人間が数学の問題を解く際に適切な助言を受けることに似ており、モデルにとって有益でした。具体的には、モデルが正しい解法にたどり着くための方向性を持ち、計算ミスや論理的な誤りを減少させる効果が観察されました。この結果、他の手法（特にチェイン・オブ・ソート）よりも優れたスコアを示しました。

MoDEM: Mixture of Domain Expert Models MoDEM: ドメインエキスパートモデルの混合

概要: ドメインプロンプトとドメイン特化したモデルを組み合わせることで汎用モデルに比べ LLMの性能効率が向上する MoDEMを提案。DeBERTa-v3-largeで入力内容をドメイン分類し、健康、数学、科学、コーディングなど各ドメインに特化したモデルを使用して回答します

技術や手法

1. **BERTベースのルーター **

- **使用するモデル **: MoDEMのルーティングシステムでは、 **Microsoft DeBERTa-v3-large**モデルが用いられています。このモデルは、 304Mパラメータを持つ BERT系のモデルで、文分類タスクに特化しています。

- **役割**: このルーターは、入力されたプロンプトを適切なドメインに分類する役割を果たします。ドメインには、数学、健康、科学、コーディング、その他のカテゴリが含まれています。ルーターはこれらのドメインのいずれかを識別し、プロンプトを特定のドメインモデルに送信します。

- **ファインチューニング **: ドメインを分類するために、このルーターは事前に選定されたデータセットを用いてファインチューニングされています。ファインチューニング時には、 1エポックでバッチサイズ 32、学習率1e-5という設定が使用されました。

- **特徴**: ルーターは非常に軽量で、最大の専門モデルの 0.42%のサイズしかないため、リソースの消費が非常に少ないです。分類の精度はテストデータで 97%を達成し、MMLUのようなアウトオブディストリビューションのデータにも高い精度で対応できることが確認されています。

2. **ドメイン専門モデル(Expert Models)**

- **専門モデルの選定 **: MoDEMのドメイン専門モデルは、それぞれのドメインで高性能を発揮するオープンソースモデルを選定しています。これにより、特定のドメインに最適化されたモデルを使って、汎用モデルよりも優れたパフォーマンスを実現しています。

- **中規模モデルセット **:

DA-Code: Agent Data Science Code Generation Benchmark for Large Language Models DA-Code: 大規模言語モデルのためのエージェントデータサイエンスコード生成ベンチマーク

概要: LLMをエージェントベースのデータサイエンスタスクをコード生成の観点で評価するためのベンチマーク、 DA-Codeを提案

データワークリング (DW)、機械学習 (ML)、探索的データ分析 (EDA) の3つのカテゴリについてどれだけ自律的な問題解決ができるかを評価します

1. タスクの構成と分類

DA-Codeは、LLMがデータサイエンスエージェントとしてどれだけの能力を持つかを評価するために、以下の 3つの主要なカテゴリのタスクから構成されています:

- **データワークリング (DW)**: 生データを解析可能な形にするために変換・統合・クリーニングを行う。具体的には、データの読み込み、欠損値の処理、データのクリーニングや統合などが含まれる。

- **探索的データ分析 (EDA)**: データセットの特性を理解し、洞察を得るためのデータ分析を行う。 SQLやPythonを使って統計分析、データマニピュレーション、データの視覚化を行う。

- **機械学習 (ML)**: 機械学習モデルを使って、データに基づく予測や分類を行う。データの前処理から特徴量エンジニアリング、モデルのトレーニング、予測までを実施する。

これらのタスクは全て実際のデータに基づき、 500の複雑なタスクを提供しており、データサイエンスの全てのプロセスをカバーしています。

2. タスクの設計と難易度

- **リアルなデータシナリオ **: DA-Codeのタスクは実際のデータセットを基にしており、単なるノートブック環境でのデータ分析にとどまらず、複数のファイルやデータソースを使用します。例えば、データベースやスプレッドシート、文書、コードなど、複数の情報源からなる多様なデータを使って課題を解決します。

- **タスク難易度の分類 **: タスクは容易、中程度、困難の 3つの難易度に分類されており、それぞれの難易度で異なる種類の技術的な課題が含まれています。具体的には、困難なタスクにはより高度なプロ

AI-Press: A Multi-Agent News Generating and Feedback Simulation System Powered by Large Language Models AI-Press: 大規模言語モデルを活用したマルチエージェントによるニュース生成とフィードバックシミュレーションシステム

概要: 複数のエージェントが協力し合う形でニュースの自動作成と修正を行う　AI-Pressを提案

ニュース生成をドラフティング（草案作成）、ポリッシング（内容修正）、シミュレーション（公開後の反応予測）という　3つのモジュールに分け、各モジュールに複数のエージェントをネット検索や　RAGを使いながらニュース制作を行います

AI-Pressの技術と手法の詳細な説明

AI-Pressは、ニュースの生成からフィードバックのシミュレーションまでを行う自動化システムで、複数のエージェントが協力する形で効率的にニュース制作をサポートします。

1. マルチエージェントシステム

AI-Pressは、ニュース生成のプロセスを「ドラフティング（草案作成）」「ポリッシング（内容修正）」「シミュレーション（公開後の反応予測）」という　3つのモジュールに分け、各モジュールに複数のエージェントを配置して効率化を図っています。このマルチエージェントシステムにより、各エージェントが特定の役割を担い、ニュース制作の精度と効率が向上します。

- **ドラフティングモジュール **: ニュース草案を作成するプロセスで、 Searcherエージェントと Writerエージェントが協力します。

  - **Searcherエージェント **: 提供されたトピックや素材に基づき、多次元的な情報を検索・収集します。このエージェントは、ニュースデータベース、ファクトデータベース、インターネットから情報を収集し、精度と信頼性を確保します。

  - **Writerエージェント **: Searcherエージェントが収集した情報をもとに、ニュース記事の草案を作成します。異なるニュースジャンル（ニュース、プロフィール、コメンタリー）ごとに特化した書き方を持ち、プロフェッショナルな内容に仕上げることを目指します。

- **ポリッシングモジュール **: 初期の草案をさらに洗練させるプロセスで、 Reviewerエージェントと Rewriterエージェントが関与します。

OpenR: An Open Source Framework for Advanced Reasoning with Large Language Models OpenR: 大規模言語モデルを用いた高度な推論のためのオープンソースフレームワーク

概要: OpenRはLLMの推論能力の向上をさせるオープンソースでデータ取得、オンラインおよびオフラインの強化学習実施し数学問題の解決やコーディングタスクなど、複雑な推論を必要とする分野での応用が期待できます

OpenRの最大の特徴は、 OpenAIのo1モデルに触発されており、 LLMに強化学習を用いて推論能力を強化するオープンソースのプラットフォームを提供する点で、これにより、自己回帰的な手法を超えて、推論の精度とパフォーマンスを大幅に向上させています。また、テスト時に計算資源を利用して推論を改善するというアプローチを採用しており、従来のトレーニング段階のスケーリングに頼らない新しい推論モデルを実現できます。

技術や手法

1. **プロセス監督と強化学習の統合 **:
    - **プロセス報酬モデル(PRM)**: PRMを用いて中間の推論ステップの質を評価し、最終的な回答の精度を向上させる。
    - **マルコフ決定過程(MDP)**: 推論プロセスを MDPとしてモデル化し、各ステップでの報酬を最適化することで推論を改善。
2. **データ増強 **:
    - 自動生成されたサンプルを用いてデータを拡充し、人間によるアノテーションの依存を減少。
    - 特に、MATHデータセットを用いた「 MATH-APS」データセットを生成し、推論プロセスの監督データとして利用。
3. **強化学習によるポリシー学習 **:
    - PRMを用いたポリシー学習により、テスト時の推論精度を向上。
    - Proximal Policy Optimization (PPO) やその効率的なバリアントである Group Relative Policy Optimization (GRPO)を用いて、モデルの推論プロセスを強化。
4. **デコーディング：推論時のガイド付き検索と計画 **:

Augmenting In-Context-Learning in LLMs via Automatic Data Labeling and Refinement LLMのインコンテキスト学習を自動データラベリングと精緻化により強化

概要: 自動でデータラベリングを行い、 LLMのインコンテキスト学習を強化する ADLRを提案手作業で少数の例を作り出力と一致する中間予測を生成、この結果をフィルタリングし、有用な結果を残してプロンプトとして使用することで推論を向上させます

技術や手法

ADLR (Automatic Data Labeling and Refinement) の手法:

1. **初期サンプルの拡張 (生成フェーズ)**

手作業で作られた少数の例から始め、大規模なトレーニングデータセットに基づき、出力と一致する中間ステップ(コードや推論の流れ)を生成します。正しい最終出力に到達した場合、その中間データをサンプルとして受け入れます。

2. **有用なサンプルの精製 (フィルタリングフェーズ)**

サンプルを難易度に基づいて選別し、特に難易度が高いが解決可能なサンプルを優先的に集めます。また、特定のサンプルが他の難しい問題をどれだけ解決できるかを評価し、有用なサンプルを選定します。

3. **推論プロトコルの強化 (適用フェーズ)**

Revisiting Benchmark and Assessment: An Agent-based Exploratory Dynamic Evaluation Framework for LLMs ベンチマークと評価の再考：LLMのためのエージェントベース探索的動的評価フレームワーク

概要：ドメイン専門の LLMに対して評価する TestAgentを提案 Benchmark+という質問 -回答形式のベンチマークを戦略 -基準に拡張したものと Assessment+という初期の質問に続くフォローアップ質問を生成 し、モデルの応答が基準を満たすかを見るエージェントベース評価方法で RAGと強化学習を使い評価します

技術や手法の詳細説明

1. Benchmark+

Benchmark+は、従来の「質問 -回答（Q&A）」形式のベンチマーク評価を「戦略 -基準（Strategy-Criterion）」形式に拡張する手法です。この手法は、従来の評価手法が持つ以下のような問題点を克服することを目指しています：

- **静的評価の限界 **：従来の Q&A形式では、あらかじめ定められた質問に対してモデルが回答し、それをスコア化するという単純なプロセスが行われていました。これにより、モデルの深い理解力や応答の多様性を評価するのが難しく、現実の応用に十分対応していません。

- **動的なインタラクション **：Benchmark+は、各質問に対する単一の回答を超え、戦略と基準を用いて、より柔軟かつ詳細な評価を可能にします。例えば、 Benchmark+では、初期の質問に対してさらなる質問を生成し、インタラクションの結果を基準に基づいて評価します。この評価手法により、モデルが特定の文脈における柔軟性と適応性をどの程度備えているかを測ることが可能となります。

具体的には、Benchmark+の形式は以下の通りに定義されます：

- **戦略（Strategy）**: 各質問に対する複数のフォローアップ質問を生成するプロセス。戦略は、インタラクション履歴とユーザーの関心に基づいて質問を動的に生成します。

- **基準（Criterion）**: 各質問に対する評価基準。基準は、特定のコンテンツに関連するものと、流暢さや人間らしいトーンなどテキスト関連のものに分類されます。評価基準は、基礎となる知識データベースから適切な情報を抽出し、具体的な質問ごとに精緻化されます。

AGENTOCCAM: A Simple Yet Strong Baseline for LLM-Based Web Agents 大規模言語モデルベースのウェブエージェントにおけるシンプルで強力なベースライン

概要: LLMを使用した観測および行動空間を整合してたウェブエージェントの性能向上を目指す　AGENTOCCAMを提案

特別なプロンプトや多層的なアプローチをせず、 LLMのwebページ情報と操作を整合しシンプルな設計にし、自律的に行動計画を生成修正できるようにしています

行動空間の簡素化 (Action Space Alignment)

ウェブエージェントが実行する操作の範囲（行動空間）をシンプルにすることが、 AGENTOCCAMの基本的なアプローチです。行動空間を整えることで、 LLMがより効率的にタスクを処理できるようにします。

2.1 不必要な行動の削減

- **noopアクションの削除 **: "noop"（何もしない）という行動は、ウェブエージェントにとってほとんど意味を持たないため、削除しました。

- **タブ操作の削除 **: タブを開閉するアクション（新しいタブを開く、特定のタブにフォーカスするなど）はほとんどのケースで不要です。これを削除することで、エージェントの行動を単純化し、より重要な行動に集中させました。

- **複雑な操作の削減 **: "hover"（要素の上にカーソルを置く）や "press"（キーボード操作）は、エージェントが扱うには難しく、またほとんどのタスクで利用価値が低いので削除しました。

2.2 低レベルの操作を抽象化

- **スクロールアクションの削除 **: ページ内のコンテンツを表示するためにスクロールを行う代わりに、ページ全体の情報を一度に観測できるようにしました。これにより、エージェントが無意味にスクロールし続ける問題を回避しました。

On The Planning Abilities of OpenAI's o1 Models: Feasibility, Optimality, and Generalizability OpenAIのo1モデルの計画能力について：実現可能性、最適性、および一般化可能性

概要：OpenAIのo1モデルの計画能力を調査

制約の多いタスクで o1-previewは自己評価と制約遵守に強みを示すが、空間的に複雑なタスクで最適性に課題がある

技術・手法

1. **計画能力の評価 **

計画能力を以下の 3つの観点で評価しています：

- **実現可能性 (Feasibility)**：各計画が制約を満たし、実行可能かどうか。個々のステップが正しいかどうか（問題のルールに従うか）、全体の計画がゴールに達成するかの観点から評価。
- **最適性 (Optimality)**：計画が目標をどれだけ効率的に達成するか。冗長な行動やコストを削減し、最適なステップでゴールに到達できるかを確認。
- **一般化可能性 (Generalizability)**：異なる環境においてどれだけ効果的に計画が立てられるか。未知の状況やシンボリック表現に対する適応力を評価。

2. **計画能力の実験 **

Barman、Blocksworld、Floortile、Grippers、Tyreworldなどの様々な計画タスクで評価し、モデルごとの課題やエラーの種類（例：ルール違反、最適性の欠如など）を分類し、分析しています。

3. **エラー分類 **

On The Planning Abilities of OpenAI's o1 Models: Feasibility, Optimality, and Generalizability OpenAIのo1モデルの計画能力について：実現可能性、最適性、および一般化可能性

概要：OpenAIのo1モデルの計画能力を調査

制約の多いタスクで o1-previewは自己評価と制約遵守に強みを示すが、空間的に複雑なタスクで最適性に課題がある

技術・手法

1. **計画能力の評価 **

計画能力を以下の 3つの観点で評価しています：

- **実現可能性 (Feasibility)**：各計画が制約を満たし、実行可能かどうか。個々のステップが正しいかどうか（問題のルールに従うか）、全体の計画がゴールに達成するかの観点から評価。
- **最適性 (Optimality)**：計画が目標をどれだけ効率的に達成するか。冗長な行動やコストを削減し、最適なステップでゴールに到達できるかを確認。
- **一般化可能性 (Generalizability)**：異なる環境においてどれだけ効果的に計画が立てられるか。未知の状況やシンボリック表現に対する適応力を評価。

2. **計画能力の実験 **

Barman、Blocksworld、Floortile、Grippers、Tyreworldなどの様々な計画タスクで評価し、モデルごとの課題やエラーの種類（例：ルール違反、最適性の欠如など）を分類し、分析しています。

3. **エラー分類 **

MLE-BENCH: EVALUATING MACHINE LEARNING AGENTS ON MACHINE LEARNING ENGINEERING MLE-Bench: 機械学習エージェントによる機械学習エンジニアリングの評価

概要: LLMのエージェントがどれくらいデータ分析をできるかを　Kaggleの75のコンペでデータセット準備、モデル訓練、実験実行などのスキルを評価

OpenAIのo1-previewとAIDEスキャフォルディングの組み合わせは　16.9%のコンペで銅メダル以上の成績を達成

技術や手法

1. **データセットのキュレーション　**

MLE-benchの評価対象は、Kaggleにおける機械学習関連の　75の競技から構成されています。これらの競技は、　Kaggle上でリアルな機械学習エンジニアリングのスキルを測定するために、以下の手順で選定・再構成されています。

- **データセットの収集　**

　各競技からデータセットを収集しました。多くの場合、競技終了後にはテストセットが公開されないため、新たにトレーニングデータとテストデータを手動で分割し、再構成しています。この再分割は、元のデータセットの分布を再現し、同様のテストスコアを得られるよう調整されています。

- **データセットの要素　**

　各データセットには、競技の説明（概要や使用データに関する情報）や実際のトレーニング・テストデータ、サブミッションをローカルで評価するための採点コード、または競技のリーダーボードが含まれています。これにより、エージェントの提出物を人間のリーダーボードと比較することが可能になります。

FACT, FETCH, AND REASON: A UNIFIED EVALUATION OF RETRIEVAL-AUGMENTED GENERATION 事実、取得、そして推論 : 取得強化生成の統一評価

概要: RAGシステムの性能を総合的に評価するための手法　FRAMESを提案

複数の情報源から情報を統合して推論する能力を評価する。提案する手法により、単一の取得なしの状態で　　0.40の精度から、多段階の取得パイプラインで　0.66の精度に向上

1. **FRAMESデータセットの設計 **

FRAMES（Factuality, Retrieval, And Reasoning MEasurement Set）は、取得強化生成（ Retrieval-Augmented Generation, RAG）システムの評価を目的としたデータセットで、以下の　3つの主要な能力を評価す るために設計されています。

- **事実性 (Factuality)**: モデルが提供する情報がどれだけ正確であるか。

- **情報取得 (Retrieval)**: モデルが正しい情報を取得し、それを回答に活用できるか。

- **推論 (Reasoning)**: モデルが複数の情報源をもとに、適切に推論し、統合的な回答を導き出せるか。

このデータセットには、複数の知識ソースからの統合が必要とされる「マルチホップ質問」が含まれており、　　RAGシステムの統合的な能力を測るための厳しいテスト環境を提供しています。

2. **データセットの収集プロセス **

データセットの構築には、以下の手順が取られました:

IPO: Interpretable Prompt Optimization for Vision-Language Models IPO: ビジョン・言語モデルのための解釈可能なプロンプト最適化

LLMにプロンプトを生成させるるためのガイドラインを設計し評価結果とともに保存　　LMMで画像の内容説明を生成し LLMで利用できるようにしてプロンプトを生成

技術・手法

1. プロンプト最適化プロンプト (Prompt Optimization Prompt)

**プロンプト最適化プロンプト **は、LLM（大規模言語モデル）に対し、より効果的なプロンプトを生成するための基盤となる仕組みです。この手法では、プロンプト自体の最適化を目的とし、次の要素が含まれています：

1. **インストラクション (Instructions)**: LLMに対してプロンプト最適化の目的を理解させ、分類タスクの性能を向上させるための明確な指示を与えます。このインストラクションには「どのようにプロンプトを生成すべきか」というタスク内容が含まれています。

2. **画像のテキスト説明 (Textual descriptions of training images)**: LMM（大規模マルチモーダルモデル）を用いて生成される画像の内容説明が含まれます。この情報を　　LLMに提供することで、画像に基づいたプロンプトの生成が可能となり、よりデータセットに特化した適応力を持つプロンプトが得られます。

3. **過去のプロンプトとその評価 (Previously generated prompts and corresponding scores)**: 過去に生成されたプロンプトとそのパフォーマンスデータ（精度や損失）をエピソード記憶として保持します。これにより、LLMは過去の実績を元に、プロンプトの効果を評価しながら改良を続けることが可能になります。この過程で、精度が向上するプロンプトが見つかるまで繰り返し生成と評価を行います。

**プロンプト最適化プロンプト **は、こうした要素を組み合わせ、 LLMが効果的なプロンプトを動的に生成し続けられる環境を提供します。これにより、従来の勾配降下法を用いたプロンプト最適化と異なり、解釈可能かつ人間が理解できるプロンプトの生成が可能となります。

2. 大規模マルチモーダルモデル (LMM) の利用

Large Language Models are Zero-Shot Next Location Predictors 大規模言語モデルでゼロショットの次の場所予測器

概要: 訪問先予測をゼロショットの LLMで実施する

Llama2、Llama2 Chat、GPT-3.5、および Mistral 7BのLLMがゼロショットで次の訪問場所を予測する能力を評価

実際のモビリティデータセットを使用した結果、 LLMは最大32.4%の精度を達成し、特に設計されたディープラーニングモデルと比較して 600%以上の相対的な改善を示しました。

概要

Llama2、Llama2 Chat、GPT-3.5、および Mistral 7BのLLMがゼロショットで次の訪問場所を予測する能力を評価

実際のモビリティデータセットを使用した結果、 LLMは最大32.4%の精度を達成し、特に設計されたディープラーニングモデルと比較して 600%以上の相対的な改善を示しました。

**論文で説明している技術や手法 **

1. **LLMの評価 **: Llama2、Llama2 Chat、GPT-3.5、Mistral 7Bの性能を評価。

2. **プロンプト設計 **: 適切なプロンプトを設計し、歴史的訪問と文脈的訪問のデータを提供。

3. **実験デザイン **: 3つの実際のモビリティデータセットを使用し、データ汚染の防止策も講じる。

4. **性能評価 **: ACC@k（k=1, 3, 5）の評価指標を使用し、ゼロショット、ワンショット、数ショットプロンプティングの影響を分析。

5. **データ汚染のテスト **: 公開データセットとプライベートデータセットを使用してデータ汚染の影響を評価。

CalibraEval: Calibrating Prediction Distribution to Mitigate Selection Bias in LLMs-as-Judges CalibraEval: LLMによる評価者における選択バイアスを軽減するための予測分布の調整

概要: LLMのラベルを使わない評価手法 CalibraEvalを提案

LLMが回答ペアの評価において位置バイアス（回答の順番）やトークンバイアス（選択肢 A/Bのラベル）に依存する選択バイアスがあり、これを軽減するために非パラメトリックな順序保持アルゴリズム (NOA)を使い

CalibraEvalについて、選択バイアスの軽減方法とそのアルゴリズムの具体的な手法

1. 選択バイアスの問題

まず、LLM（大規模言語モデル）を「評価者」として利用する際に発生するバイアスの種類を理解する必要があります。

- **位置バイアス (Position Bias)**: LLMが選択肢の位置（例えば「最初の選択肢」など）に依存して回答を選ぶ傾向を指します。位置に関する先入観により、モデルが最初に提示された選択肢を好むことがあります。

- **トークンバイアス (Token Bias)**: 選択肢に使用されるラベル（例えば「 A」または「 B」）に依存して LLMがバイアスを持つことを指します。このため、選択肢が異なる順番で提示された場合、モデルの評価が変わることがあります。

このようなバイアスは、 LLMを評価者として利用する場合の公正性を損なう要因となります。

2. CalibraEvalの概要

SPHERE: Scaling Personalized Feedback in Programming Classrooms with Structured Review of LLM Outputs SPHERE: LLM出力の構造化レビューによるプログラミングクラスにおけるパーソナライズされたフィードバックのスケーリング

概要: LLMでプログラミングの演習に対してフィードバックを提供する　SPHEREを提案

フィードバックは、ヒント、説明、検証の　3つをコーディングやグループディスカッションを使い生成、

この結果を視覚的にハイライトや進行状況を散布図で視覚化するなど　UIの工夫も行いリアルタイムで状況把握できるようにする

技術や手法

1. **クリティカルな問題の推薦モデル　**: 学生のコーディング活動やグループディスカッションのデータを用いて、　LLMが重要な問題を特定し、教師に推奨する。

2. **フィードバック生成の戦略的アプローチ　**: フィードバックは「ヒント」「説明」「検証」の　3つのモードで生成され、それぞれのモードが学生の状況に応じて選択される。

3. **視覚的バインディング　**: フィードバック内容を視覚的にハイライトし、教師が迅速に正確なフィードバックを確認できるようにする。

4. **ユーザインターフェースの設計　**: 問題の一覧表示やクラスの進行状況を散布図で視覚化し、教師がリアルタイムで学生の状況を把握できるようにする。

OSCAR: Operating System Control via State-Aware Reasoning and Re-Planning OSCAR: 状態認識と再計画によるオペレーティングシステム制御

概要: LLMとLMMの両方を使い標準化されたマウスやキーボード操作を通してデスクトップやアプリを自律的に制御する汎用エージェント　　OSCARを提案

ユーザーの指示を pythonコードに変換し、GUIを使い精密な操作を実行。タスクベースの再計画を実施しリアルタイムのフィードバックに基づきタスクの調整を行います

SIKeD: Self-guided Iterative Knowledge Distillation for mathematical reasoning SIKeD: 数学的推論のための自己誘導反復知識蒸留

概要: 知識蒸留では LLMが持つ単一のタスクに対する推論アプローチ方法を伝えますが複数について同時に学習させ小型モデルがタスクに最も適した推論方法を選択できるように　　LLMが生成したデータに加え自らが生成したデータを活用して学習を繰り返すことでモデルが自身の推論能力を強化していく　　SIKeDを提案

Grammar-Aligned Decoding グラマー整合デコーディング

概要: ASApは文法に従う文を生成し、次のトークンが文法に適合する確率を調整するアルゴリズムを使い、 LLMの出力を文法的に正しくする。

概要

LLMはプログラムコード、数式、整形式マークアップなどの高度に構造化された出力を安定して生成するのが苦手です。制約付きデコーディングは、 LLMが各ステップで出力できるトークンを制限することで、この問題を軽減し、出力が指定された制約に一致することを保証します。

特に、文法制約付きデコーディング(GCD)では、LLMの出力が指定された文法に従う必要があります。しかし、 GCD技術(および一般的な制約付きデコーディング技術)が LLMの分布を歪め、文法的には正しいが、LLMが与える確率に比例しない出力を生成し、最終的には低品質な出力になります。この問題を解決するために、文法制約に従ったサンプリングと LLMの分布を一致させる「 Grammar-Aligned Decoding(GAD)」を提案。

近似期待未来(ASAp)を用いた適応的サンプリングというデコーディングアルゴリズムを開発。 ASApは文法に従う文を生成し、次のトークンが文法に適合する確率を調整するアルゴリズムを使い、 LLMの出力を文法的に正しくする。

**先行研究と比べて **

先行研究では、LLMによる構造化出力の生成を改善するために制約付きデコーディングが提案されてきましたが、これらの技術は LLMの分布を歪め、結果として低品質な出力を生成することが示されました。本研究では、GCD技術の問題点を解明し、新しい適応的サンプリングアルゴリズムである ASApを提案することで、これを解決しました。このアルゴリズムは、出力が文法に適合しつつ、 LLMの条件付き確率分布に一致することを保証します。これにより、従来の GCD技術に比べて、文法的に正しく、かつ高品質な出力を生成することが可能となりました。

**技術や手法の説明 **

R-COT: REVERSE CHAIN-OF-THOUGHT PROBLEM GENERATION FOR GEOMETRIC REASONING IN LARGE MULTIMODAL MODELS R-COT: 幾何学的推論のためのリバースチェーン・オブ・ソート問題生　成を用いた大規模マルチモーダルモデル

概要: 二段階のリバースチェーン・オブ・ソート（　R-CoT）は幾何学の問題を生成する手法で、 LLMを使い、画像から逆に問題を作成する。データの多様性と精度を高めるため、段階的に推論し生成する。

1. R-CoTの背景

R-CoT（Reverse Chain-of-Thought）は、幾何学的な推論のために高品質のデータを生成することを目的とした手法です。これにより、大規模なマルチモーダルモデル（　　LMM）が、図形を含む数学的な問題を解く際のパフォーマンスが大幅に向上します。この手法は、幾何学的な要素を扱う能力を強化するために特化しています。

2. 技術の概要

R-CoTは以下の 2段階のアプローチで幾何学問題を生成します:

1. **GeoChain（ジオチェーン） **: 高精度の幾何学的な画像を生成し、その画像の要素間の関係を詳細に記述します。

2. **Reverse A&Q（リバース A&Q）**: GeoChainで生成した記述を使い、段階的に推論し逆に幾何学的な質問と答えを生成します。

これにより、データの精度と多様性が向上し、 LMMの幾何学的な問題解決能力を強化します。

3. GeoChainについて

EDGE: Enhanced Grounded GUI Understanding with Enriched Multi-Granularity Synthetic Data EDGE: マルチ粒度の豊富な合成データによる拡張された　GUI理解

概要: GUI上で動作する自律エージェントの為の学習データ生成のために　EDGEを提案

ウェブクローリングデータのリポジトリを活用し自動化ツールでスクショを取得自動的に位置と内容を抽出しアイコンなどのテキストの無い情報を　LLMを使い注釈を追加、明示的なテキストだけでなく、隠れた例えば、alt属性なども学習データに利用、 GUIエージェントの学習において、マルチ粒度タスクという複数レベルのタスクを設定します

1. データ生成フレームワーク（ EDGE）

EDGEはGUI（グラフィカルユーザーインターフェース）の理解とインタラクション能力を強化するための、データ駆動型の合成データ生成フレームワークです。具体的には、以下のステップで構成されています。

1.1 ウェブページの収集

EDGEは、CommonCrawlという大規模なウェブクローリングデータのリポジトリを活用して、広範なウェブページを収集します。収集されたページは、　GUI操作に関するタスクのデータセット作成に利用されます。

1.2 ページのアノテーション（注釈付け）

Playwrightという自動化ツールを用いて、収集したウェブページをレンダリングし、スクリーンショットを取得します。そして、　JavaScriptスクリプトを注入して、ページ上の各要素の位置と内容を自動的に抽出し、注釈を追加します。このアノテーションでは以下の要素が含まれます。

可視要素の抽出 : 画面上に表示されている要素のみを対象にし、不要な情報を除外することで、より正確なデータを収集します。

ラベル付け : ページのタグや役割、スタイルを用いて、それぞれの要素にラベルを付け、　GUIの構造を理解します。

統合ルールの適用 : タグや表示内容に基づいて、セマンティックにまとまった単位で要素を扱います。例えば、ボタン要素では、テキストやアイコンが含まれる場合でも全体を　1つの要素として扱います。

1.3 アノテーションのリッチ化

Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications テキストを超えて：産業用途向けのマルチモーダル入力を用いた　RAGの最適化

概要：LLMのRAGの評価として産業分野でのマルチモーダルモデルを使用する方法の検討

マルチモーダル埋め込みと画像からのテキスト要約生成という　2つのアプローチを提案

マルチモーダル埋め込みと画像からのテキスト要約生成のいずれも、シングルモーダル（テキストのみ）の　　RAGよりも良いパフォーマンスを示し

画像からのテキスト要約生成の方がマルチモーダル埋め込みよりもパフォーマンスがわずかに優れている結果が得られました。

マルチモーダル埋め込み **と画像からのテキスト要約生成 **という2つの画像処理アプローチが提案されています。以下、それぞれのアプローチについて詳しく説明し、その結果もまとめます。

1. マルチモーダル埋め込み（ Multimodal Embeddings）

**マルチモーダル埋め込み **は、画像とテキストを共通の埋め込み空間にマッピングする手法です。このアプローチでは、以下の手順で処理が行われます。

- **CLIPモデル **（OpenAIによる画像とテキストを同時に埋め込むモデル）を用いて、画像と質問を埋め込みます。

- 埋め込み後、共通の埋め込み空間内で画像とテキストの類似度を計算し、最も関連性の高いものを検索します。

- この埋め込み空間での距離に基づいて、画像と質問間の関連性を測定し、関連のある画像を取得します。

**利点 **:

THINKING LLMS: GENERAL INSTRUCTION FOLLOWING WITH THOUGHT GENERATION THINKING LLMs: 思考生成による一般的な指示追従

概要: LLMには回答前に明示的な思考能力の実行がかけており複雑な質問に必要なこの能力を追加の人間データを使わずに既存の LLMに付与するための訓練手法を Thought Preference Optimization (TPO)を提案

思考と応答の 2つのパートに分けて生成しその応答部分を評価するためのジャッジモデルで応答の質のみを評価。最も良い思考と応答の組み合わせを選択します

1. 問題意識と提案手法の概要

LLMは現在、ユーザーの指示に直接応答するための訓練が行われており、その際の計算資源は指示の複雑さに関わらず一定です。しかし、複雑な指示の場合、十分な応答を得るためには、回答前に内部で思考を行うことで追加の計算を行うことが必要です。この「思考」というプロセスを LLMに付与するために、追加の人間データなしで思考生成を学習させる手法として「 Thought Preference Optimization (TPO)」が提案されています。

2.「Thought Preference Optimization (TPO)」の概要

TPOは、命令チューニングされた既存の LLMに対して、内部の思考を生成し、それを最適化する手法です。この手法は、 LLMが思考を生成し、その後、最終的な回答を出力するプロセスを学習するために、以下の流れを経ます。

1. **思考の生成 **: 初めに、命令に対して LLMが「思考」と「応答」の 2つの部分に分かれたシーケンスを生成します。

2. **思考の評価と選別 **: 複数の思考と応答の候補を生成し、その応答部分を評価するために「ジャッジモデル」を使用します。このジャッジモデルは応答の質のみを評価し、その評価結果を元に最良の思考と応答の組み合わせを選びます。

3. **思考の最適化 **: 最良の応答を生成する思考を選別し、その結果に基づいて次の訓練でモデルの出力を最適化します。これを反復的に行うことで、モデルが自己学習を通じてより良い思考を生成する能力を持つようになります。

Appendix