論文要約

LLM関連

LONG2RAG: Evaluating Long-Context & Long-Form Retrieval-Augmented Generation with Key Point Recall LONG2RAG: 長文コンテキストおよび長文形式の検索強化生成の評価とキーポイントリコールによる評価

概要: LONG2RAGは、長文コンテキストでの LLMのRAG性能を評価するためのベンチマークです。 280の質問に5つの関連文書を設定し、検索された文書から抽出されたキーポイントをどれだけ含んでいるかを測定するKPRで評価。GPT-4oが最高スコアの 0.579を記録

RAG評価で主に長文コンテキストに対応する LLM評価ベンチマーク LONG2RAGを提案。280の質問が10の領域で設定され各質問に対して 5つの関連文書を設定。評価指標には KPR（Key Point Recall）を設定し検索された文書から抽出されたキーポイントをどれだけ含んでいるかを測定する方法（各スコアは 0から1の範囲で、高いほど良い性能を示します）を使用。質問は 8つのカテゴリ（事実、説明、比較、主観、因果関係、仮定、予測、方法論）に分類して評価、 GPT-4oのKPRは 0.579、Claude-3-SonnetのKPRは 0.477、Qwen2-72B（オープンソースモデルの大規模版） : KPRは、0.449、Phi-3-mini-128K: KPRは 0.434と商用モデルである GPT-4oが最も優れた結果を示しました。

また、KPRは長文生成を好む傾向もあるため、生成の質と長さのバランスが重要であることもわかっています

技術や手法

- **LONG2RAGベンチマーク **: 280の質問を使用し、各質問に対して平均 2444語の5つの検索文書が関連付けられています。これにより、モデルが長文の検索情報を取り込む能力を評価します。

- **キーポイントリコール（ KPR）**: 検索された文書から抽出されたキーポイントが生成された回答にどの程度含まれているかを評価する手法です。この評価を通じて、モデルが検索情報を活用しているかどうかを測定します。

- **データセットの構築方法 **: 自動パイプラインを用いて質問を生成し、関連する文書を検索してキーポイントを抽出。その後、 LLMと人間の協力によりキーとなるポイントの検証を行い、データセットを構築しました。

評価手法とパフォーマンス指標

論文では、LONG2RAGベンチマークを用いて 9つの最新の LLM（大規模言語モデル）を評価しました。評価に用いられた指標は以下の通りです。

Benchmarking Large Language Models in Retrieval-Augmented Generation 検索強化生成における大規模言語モデルのベンチマーク

概要: RAGのパフォーマンスを 4つのノイズ耐性、否定拒否、情報統合、反事実耐性の　RGBのコーパスで評価。LLMは一定のノイズ耐性を持つが、否定拒否や情報統合、誤情報処理にはまだ課題が多いことがわかりました

技術や手法

- **検索強化生成 (Retrieval-Augmented Generation, RAG):** RAGは、検索エンジンを用いて外部の知識を取得し、モデルの幻覚を軽減する手法。特に、インターネット上の膨大な情報から正確な知識を得るために使用される。

- **Retrieval-Augmented Generation Benchmark (RGB):** RGBは、RAGの4つの基本的な能力を評価するために設計された新しいベンチマークで、最新のニュース情報を基に構築されている。このベンチマークにより、LLMがノイズ情報に対してどの程度頑健であるかや、複数の情報を統合する能力などを評価できる。

    - **ノイズ耐性 (Noise Robustness):** 質問と関連があるが、回答を含まないノイズ文書から必要な情報を抽出する能力。

    - **否定拒否 (Negative Rejection):** 必要な知識が取得された文書に存在しない場合に、適切に回答を拒否する能力。

    - **情報統合 (Information Integration):** 複数の文書から情報を統合して質問に回答する能力。

    - **反事実耐性 (Counterfactual Robustness):** 取得された文書に誤った情報が含まれている場合に、そのリスクを認識して適切に処理する能力。

使用用途

この研究は、以下のようなシーンで活用が期待される:

- **検索エンジンの改善 :** LLMを用いた検索結果の生成において、ノイズ情報を適切にフィルタリングし、より正確な情報提供を行う。

Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models プロンプト設計および外部ツールの役割が大規模言語モデルの幻覚率に与える影響の調査

概要: LLMのハルシネーションがプロンプトエンジニアリングや　LLMエージェントの活用でどのように変わるかを調査

Temperatureをあげて複数回の LLM呼び出しの多数決で回答する SCを使用することが効果的だという結果になり、現実での知識を問うタスクには　KGRが効果的という結果になりました

技術や手法

1. プロンプト技術

1.1 チェイン・オブ・ソート（ CoT）プロンプト

チェイン・オブ・ソート（ Chain-of-Thought, CoT）プロンプトは、複雑な問題をより簡単に解決できるように、小さなステップに分割する手法です。この方法では、モデルが一度に問題全体を解決するのではなく、解決の過程を段階的に分解します。例えば、数学の問題を解く場合、問題をいくつかの小さなステップに分けて、それぞれのステップで部分的な答えを導き出し、最終的に全体の答えに到達します。この方法により、LLMはより精度の高い推論が可能になります。

1.2 自己一貫性（ SC）

自己一貫性（ Self-Consistency, SC）は、同じ質問に対して複数回の LLM呼び出しを行い、その結果を多数決で選ぶことで一貫性のある答えを導き出す手法です。この手法の目的は、モデルのランダムな生成によって生じる不安定な出力を安定させることです。温度（ temperature）の設定を調整し、複数の異なる出力から最も一貫した答えを選ぶことで、信頼性の高い回答が得られるようになります。この方法は、特に数学の問題や論理的な推論を必要とする課題に有効です。

1.3 木構造の思考（ ToT）

Distance between Relevant Information Pieces Causes Bias in Long-Context LLMs 関連情報の間の距離が長コンテキスト LLMにバイアスを引き起こす

概要: 長文を扱うLLMには、情報が中間にあると見落とす lost in the middle問題の他に複数の情報を活用して回答するときにその複数の情報同士の距離とその配置が遠くなることが結果に影響することが開発されたLONGPIBENCHというベンチマークからわかりました

技術や手法

1. **ポジショナルバイアスの問題と「 lost in the middle」現象 **

- **ポジショナルバイアス **とは、大規模言語モデル(LLM)が入力された情報の位置に応じて、その情報をうまく扱えなくなる現象を指します。この論文では、特に長文の入力での問題を扱っています。具体的には、重要な情報が文脈の中間に位置する場合、モデルがその情報を見落とす「 lost in the middle」現象が問題視されています。この現象は、 LLMsが長い文脈を効率的に利用する際の大きな障害です。

2. **LONGPIBENCHの設計と目的 **

- **LONGPIBENCH**は、複数の関連情報が含まれるタスクにおけるポジショナルバイアスを評価するためのベンチマークです。このベンチマークは、絶対位置と相対位置のバイアスを評価することを目的としています。

  - **絶対位置 **とは、文脈全体の中で関連情報がどの部分に位置するかを指します(例えば、入力の先頭、中間、末尾など)。

  - **相対位置 **は、複数の関連情報の間の距離や、それらの情報がどの程度密集しているかを意味します。この点に注目することで、 LLMが情報の分布や配置にどのようなバイアスを持っているかを評価します。

3. **LONGPIBENCHのタスク設計 **

Appendix