EMNLP 2024 Paper Abstracts

Sentiment Analysis, Stylistic Analysis, and Argument Mining

1. UniGen: Universal Domain Generalization for Sentiment Classification via Zero-shot Dataset Generation

This paper introduces UniGen, a method for universal domain generalization in sentiment classification through zero-shot dataset generation. The proposed technique allows for the generation of datasets applicable to any domain sharing the same label space, improving the efficacy of small task-specific models while maintaining low parameter size.

https://arxiv.org/abs/2405.01022

608. Diversity Over Size: On the Effect of Sample and Topic Sizes for Topic-Dependent Argument Mining Datasets

This paper investigates the effects of dataset composition on argument mining in few- and zero-shot settings, demonstrating that significant reductions in training sample sizes can still yield high performance. Additionally, it introduces a new dataset for future benchmarking in the field of argument mining.

https://arxiv.org/abs/2205.11472

611. Dynamic Multi-granularity Attribution Network for Aspect-based Sentiment Analysis

This paper introduces the Extensible Multi-Granularity Fusion (EMGF) network for Aspect-based Sentiment Analysis (ABSA), which integrates various linguistic and structural features to enhance sentiment evaluation. Experimental results demonstrate EMGF's superior performance compared to existing ABSA methods on standard datasets.

https://arxiv.org/abs/2402.07787

1139. MASIVE: Open-Ended Affective State Identification in English and Spanish

This paper introduces MASIVE, a dataset aimed at identifying a wide range of affective states in English and Spanish from Reddit posts. It demonstrates that fine-tuning multilingual models on this dataset improves performance on emotion identification tasks and highlights the importance of native-written data for effective emotion analysis.

https://arxiv.org/abs/2407.12196

1142. Flee the Flaw: Annotating the Underlying Logic of Fallacious Arguments Through Templates and Slot-filling

This paper presents a novel approach to annotating logical fallacies in arguments using explainable templates and a slot-filling technique. It finds that existing language models are ineffective at detecting these fallacies, leading to the creation of a publicly available dataset to support further research.

http://arxiv.org/abs/2406.12402v1

1171. I love pineapple on pizza != I hate pineapple on pizza: Stance-Aware Sentence Transformers for Opinion Mining

162. Message Passing on Semantic-Anchor-Graphs for Fine-grained Emotion Representation Learning and Classification

1079. An Empirical Analysis of the Writing Styles of Persona-Assigned LLMs

170. DGLF: A Dual Graph-based Learning Framework for Multi-modal Sarcasm Detection

207. D2R: Dual-Branch Dynamic Routing Network for Multimodal Sentiment Detection

216. External Knowledge-Driven Argument Mining: Leveraging Attention-Enhanced Multi-Network Models

954. Media Attitude Detection via Framing Analysis with Events and their Relations

1041. Semantics and Sentiment: Cross-lingual Variations in Emoji Use

Resources and Evaluation

2. Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation

This paper introduces a method for cleansing noisy datasets by using large language models (LLMs) for data annotation, specifically enhancing the Multi-News dataset. It provides an efficient alternative to human annotators, demonstrating the utility of LLMs in improving dataset quality for downstream tasks, such as multi-document summarization.

https://arxiv.org/abs/2404.09682

83. DecorateLM: Data Engineering through Corpus Rating, Tagging, and Editing with Language Models

This paper presents DecorateLM, a data engineering method that improves the quality of the pretraining corpus for large language models through data rating, tagging, and editing. By refining a massive corpus of 100 billion tokens, this approach demonstrates a significant boost in model performance, showcasing the importance of quality data in training language models.

https://arxiv.org/abs/2410.05639

121. GLaPE: Gold Label-agnostic Prompt Evaluation for Large Language Models

This paper proposes GLaPE, a gold label-agnostic method for evaluating and optimizing prompts for large language models (LLMs). GLaPE provides reliable prompt evaluations based on self-consistency, allowing for effective prompt optimization without relying on manual annotations or gold labels.

https://arxiv.org/abs/2402.02408

175. How Hard is this Test Set? NLI Characterization by Exploiting Training Dynamics

This paper addresses the evaluation of Natural Language Inference (NLI) models by proposing a method that automatically creates a challenging test set, categorizing it into difficulty levels based on training dynamics. By minimizing spurious correlations in popular NLI datasets, the research offers a more authentic evaluation of model performance and implications for various natural language understanding applications.

https://arxiv.org/abs/2410.03429

210. A User-Centric Multi-Intent Benchmark for Evaluating Large Language Models

This paper introduces a user-centric benchmark for evaluating large language models (LLMs) based on real-world user intent, which was derived from a study involving 712 participants. It demonstrates correlations between benchmark scores and human preferences, validating its effectiveness in guiding users in selecting suitable LLM services.

https://arxiv.org/abs/2404.13940

248. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models

Prometheus 2 is introduced as an advanced open-source language model designed for evaluating other language models, addressing major shortcomings of existing evaluators by aligning closely with human and proprietary model assessments. The new model supports direct assessment and pairwise ranking formats based on user-defined criteria, demonstrating superior correlation and agreement with human judges on various benchmarks.

https://arxiv.org/abs/2405.01535

285. Quality Matters: Evaluating Synthetic Data for Tool-Using LLMs

The paper evaluates the quality of synthetic data used for training large language models (LLMs) in utilizing external tools, highlighting the importance of data reliability. It presents two assessment approaches and demonstrates that training on high-quality data significantly enhances model performance.

https://arxiv.org/abs/2409.16341

322. Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA

This paper introduces Loong, a benchmark for assessing long-context capabilities of large language models through realistic multi-document question answering tasks. It emphasizes the importance of relevant document inclusion in evaluating model performance, as well as identifying the limitations of current long-context language models.

https://arxiv.org/abs/2406.17419

398. RepEval: Effective Text Evaluation with LLM Representation

This paper introduces RepEval, a novel metric for automatic text evaluation leveraging LLM representations, which is adaptable across various tasks and requires minimal sample pairs for construction. RepEval demonstrates higher effectiveness and correlation with human judgments compared to traditional and previous LLM-based evaluation methods across multiple datasets.

https://arxiv.org/abs/2404.19563

596. Data, Data Everywhere: A Guide for Pretraining Dataset Construction

This paper presents a systematic study of the entire pipeline involved in constructing pretraining datasets for language models. The authors identify effective techniques and categorize prevalent data sources to provide actionable guidance for improving dataset quality, thereby enhancing model performance on downstream tasks.

http://www.arxiv.org/abs/2407.06380

658. QGEval: Benchmarking Multi-dimensional Evaluation for Question Generation

QGEval introduces a benchmark for evaluating question generation based on multiple criteria, addressing the shortcomings of current evaluation methods. The study finds that generated questions often lack clarity and answer consistency, and existing automatic evaluation metrics do not align with human judgments.

https://arxiv.org/abs/2406.05707

764. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations

This paper presents a systematic review of the challenges and limitations in evaluating large language models (LLMs), highlighting inconsistencies in evaluation setups that affect findings and interpretations. It offers recommendations to improve the reliability and robustness of LLM evaluations to ensure their proper deployment in real-world applications.

https://arxiv.org/abs/2407.04069

891. Themis: A Reference-free NLG Evaluation Language Model with Flexibility and Interpretability

This paper presents Themis, a large language model specifically designed for natural language generation (NLG) evaluation, which improves on existing methods by allowing for reference-free and interpretable assessments. The authors also introduce a new evaluation corpus, NLG-Eval, containing human and GPT-4 annotations to enhance the evaluation of NLG tasks.

https://arxiv.org/abs/2406.18365

896. Leveraging Large Language Models for NLG Evaluation: Advances and Challenges

This paper discusses the use of Large Language Models (LLMs) for the evaluation of Natural Language Generation (NLG), introducing a formal taxonomy for LLM-based evaluation metrics. It also highlights the challenges in this area, including issues of bias and robustness, aiming to provide a comprehensive understanding of current methodologies and promote better evaluation practices.

https://arxiv.org/abs/2401.07103

949. Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation

The paper introduces FLAMe, a family of Foundational Large Autorater Models designed to improve the evaluation of outputs from large language models (LLMs). FLAMe outperforms existing popular models in a variety of quality assessment tasks with a focus on efficiency and reduced bias in evaluation.

https://arxiv.org/abs/2407.10817

1118. CodeJudge: Evaluating Code Generation with Large Language Models

This paper introduces CodeJudge, a framework designed to evaluate the semantic correctness of code generated by Large Language Models (LLMs) without relying on test cases. Experimental results indicate that CodeJudge significantly outperforms current evaluation methods, demonstrating the efficacy of using LLMs for code evaluation across multiple programming languages.

https://arxiv.org/abs/2410.02184

480. CliMedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models in Clinical Scenarios

CliMedBench is a large-scale benchmark designed to evaluate the medical capabilities of Large Language Models (LLMs) in clinical scenarios, consisting of 33,735 real-world derived questions across 14 expert-guided scenarios. Findings indicate that Chinese medical LLMs struggle with reasoning and consistency, while certain general-domain LLMs show promise in medical applications but are limited by input capacity.

http://arxiv.org/abs/2410.03502

536. Precise Model Benchmarking with Only a Few Observations

This paper proposes an empirical Bayes (EB) estimator to improve the precision of estimating a large language model's performance on specific topics within question-answering datasets. The EB method balances direct and regression estimates, leading to lower mean squared error and more reliable confidence intervals compared to traditional approaches.

https://arxiv.org/abs/2410.05222

40. Tokenization Is More Than Compression

This paper investigates the role of tokenization in natural language processing and challenges the notion that fewer tokens necessarily lead to improved performance. It introduces a new tokenizer, PathPiece, and provides insights into the design decisions that enhance the effectiveness of tokenization methods.

https://arxiv.org/abs/2402.18376

269. Forgetting Curve: A Reliable Method for Evaluating Memorization Capability for Long-Context Models

This paper critiques existing evaluations on the memorization capability of long-context language models and introduces the 'forgetting curve' method as a reliable measurement tool. The forgetting curve method demonstrates robustness across various models and contexts, highlighting significant differences in memorization performance among different architectures.

https://arxiv.org/abs/2410.04727

423. Do Text-to-Vis Benchmarks Test Real Use of Visualizations?

This paper evaluates the effectiveness of existing benchmarks for testing the capabilities of large language models to generate visualization code and highlights a significant misalignment with real-world user practices. The findings underscore the necessity for new benchmarks that accurately reflect users' visualization needs, guiding future data creation efforts.

https://arxiv.org/abs/2407.19726

452. LawBench: Benchmarking Legal Knowledge of Large Language Models

The paper presents LawBench, an evaluation benchmark designed to assess the legal knowledge capabilities of large language models (LLMs) across three cognitive levels: memorization, understanding, and application. The study evaluates 51 LLMs on various legal tasks, revealing that while fine-tuning on legal text improves performance, there is still a need for further advancements for LLMs to reliably perform legal tasks.

https://arxiv.org/abs/2309.16289

647. SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading

The paper introduces SciEx, a benchmark developed to evaluate large language models (LLMs) on university-level science exam questions, both from English and German sources. It emphasizes the challenges in assessing LLM performance on freeform answers and explores the potential of LLMs to grade their peers' responses with high correlation to expert grading.

https://arxiv.org/abs/2406.10421

702. SUPER: Evaluating Agents on Setting Up and Executing Tasks from Research Repositories

SUPER introduces a benchmark designed to evaluate the capability of Large Language Models (LLMs) to autonomously reproduce results from machine learning and natural language processing research repositories. The benchmark highlights the challenges researchers face and aims to provide a resource to measure progress, showing that current models struggle with a significant number of tasks.

https://arxiv.org/abs/2409.07440

843. AKEW: Assessing Knowledge Editing in the Wild

This paper introduces AKEW, a practical benchmark for assessing knowledge editing in language models, encompassing various knowledge update settings. The authors illustrate the gap between existing methods and real-world knowledge editing scenarios through extensive experiments, highlighting the need for practical evaluation metrics.

https://arxiv.org/abs/2402.18909

946. Mathador-LM: A Dynamic Benchmark for Mathematical Reasoning on Large Language Models

Mathador-LM introduces a dynamic benchmark for evaluating the mathematical reasoning abilities of large language models (LLMs) through a game-based approach. The results indicate that contemporary models perform poorly compared to average third graders on this benchmark, highlighting issues with existing mathematical reasoning evaluations.

https://arxiv.org/abs/2406.12572

519. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization

This paper introduces APPLS, a meta-evaluation testbed developed to assess metrics for Plain Language Summarization (PLS). It identifies key criteria for evaluating PLS and finds that current metrics are insufficient to fully capture the quality across all criteria, recommending a suite of methods instead.

https://arxiv.org/abs/2305.14341

65. DocHieNet: A Large and Diverse Dataset for Document Hierarchy Parsing

180. BC-Prover: Backward Chaining Prover for Formal Theorem Proving

779. Rethinking the Evaluation of In-Context Learning for LLMs

1085. Re-Evaluating Evaluation for Multilingual Summarization

796. More DWUGs: Extending and Evaluating Word Usage Graph Datasets in Multiple Languages

Summarization

3. FIZZ: Factual Inconsistency Detection by Zoom-in Summary and Zoom-out Document

This paper presents a new method called FIZZ for detecting factual inconsistencies in abstractive summarization systems, utilizing a granular approach to evaluate facts. The proposed system demonstrates significant improvements in effectiveness and interpretability compared to existing methods.

https://arxiv.org/abs/2404.11184

552. Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems

This paper introduces the "Summary of a Haystack" (SummHay) task, aimed at evaluating the performance of long-context LLMs and RAG systems when tasked with summarization in specific domains. The findings show that current systems, even with relevant document signals, struggle to meet human performance levels in summarization quality and citation accuracy.

https://arxiv.org/abs/2407.01370

603. GlobeSumm: A Challenging Benchmark Towards Unifying Multi-lingual, Cross-lingual and Multi-document News Summarization

This paper presents GlobeSumm, a benchmark that aims to unify multilingual, cross-lingual, and multi-document news summarization into a new task called MCMS. It introduces the GLOBESUMM dataset and highlights challenges such as conflicts between news reports, along with experimental validation of the dataset's quality.

http://arxiv.org/abs/2410.04087

728. Detecting Errors through Ensembling Prompts (DEEP): An End-to-End LLM Framework for Detecting Factual Errors

This paper introduces DEEP, a framework designed to detect factual errors in text summarization created by large language models using an ensemble of prompts. It achieves state-of-the-art accuracy on various benchmarks without relying on fine-tuning or impractical thresholding techniques.

https://arxiv.org/abs/2406.13009

838. Learning to Rank Salient Content for Query-focused Summarization

This study explores the integration of Learning-to-Rank (LTR) with Query-focused Summarization (QFS) to improve summary relevance by prioritizing content. The proposed model outperforms state-of-the-art approaches on the QMSum benchmark, demonstrating enhanced understanding and relevance while maintaining fluency in generated summaries.

https://arxiv.org/abs/2411.00324

1048. Model-based Preference Optimization in Abstractive Summarization without Human Feedback

This paper introduces a new approach called Model-based Preference Optimization (MPO) for improving abstractive summarization without the need for human feedback. The method allows the model to refine its summarization abilities by utilizing internally generated preference datasets, resulting in higher quality summaries based on experimental evaluations.

https://arxiv.org/abs/2409.18618

1106. Towards Enhancing Coherence in Extractive Summarization: Dataset and Experiments with LLMs

This paper addresses the issue of coherence in extractive summarization, proposing a new human-annotated dataset that incorporates user intent to improve summary readability. Experiments with large language models indicate that the proposed dataset significantly enhances the coherence of summaries, as evidenced by a notable performance improvement in evaluation metrics.

https://arxiv.org/abs/2407.04855

1120. SYNFAC-EDIT: Synthetic Imitation Edit Feedback for Factual Alignment in Clinical Summarization

This paper presents SYNFAC-EDIT, a novel approach that uses large language models as synthetic experts to generate feedback for enhancing factual alignment in clinical summarization tasks. By leveraging expert-level edit feedback from advanced models, the study aims to bridge the gap in factual accuracy for less powerful models in the medical domain.

https://arxiv.org/abs/2402.13919

557. STORYSUMM: Evaluating Faithfulness in Story Summarization

This paper introduces STORYSUMM, a dataset for evaluating the faithfulness of LLM-generated summaries of short stories, accompanied by localized faithfulness labels and error explanations. It highlights the shortcomings of existing human annotation protocols and automatic metrics, advocating for a diverse range of methods to establish ground truth in summarization evaluation.

http://www.arxiv.org/abs/2407.06501

923. Can We Trust the Performance Evaluation of Uncertainty Estimation Methods in Text Summarization?

This paper discusses the reliability of uncertainty estimation methods used in text summarization, an essential task in natural language generation. It presents a new benchmark that evaluates these methods against multiple summary quality metrics to highlight the need for comprehensive assessment in risk-critical applications.

https://arxiv.org/abs/2406.17274

984. Knowledge Planning in Large Language Models for Domain-Aligned Counseling Summarization

The paper presents a novel framework called PIECE that enhances Large Language Models' capabilities in generating concise counseling summaries by incorporating domain-specific knowledge. By employing a planning engine to structure dialogue and filter knowledge, PIECE shows significant improvements in summary quality over multiple benchmarks, including expert evaluations.

https://arxiv.org/abs/2409.14907

1078. Mitigating the Impact of Reference Quality on Evaluation of Summarization Systems with Reference-Free Metrics

This paper presents a novel reference-free metric for evaluating abstractive summarization systems that correlates well with human relevance assessments and is cost-effective to compute. It also demonstrates that incorporating this metric can enhance the robustness of standard reference-based evaluation metrics in scenarios with low-quality references.

https://arxiv.org/abs/2410.10867

935. Are Large Language Models In-Context Personalized Summarizers? Get an iCOPERNICUS Test Done!

The paper discusses the limitations of large language models in personalized summarization, emphasizing the need for In-Context Personalization Learning (ICPL) to account for user preferences. It introduces the iCOPERNICUS framework to evaluate the personalization capabilities of these models using the EGISES measure, revealing that many state-of-the-art LLMs show degraded performance under richer prompt contexts.

https://arxiv.org/abs/2410.00149

Interpretability and Analysis of Models for NLP

18. Uncertainty in Language Models: Assessment through Rank-Calibration

This paper introduces a framework called Rank-Calibration to assess the uncertainty and confidence measures of language models, which often generate incorrect or hallucinated responses. The study highlights the importance of quantifying uncertainty in LMs, demonstrating the framework's broad applicability and interpretability through empirical evidence.

https://arxiv.org/abs/2404.03163

36. Evaluating Readability and Faithfulness of Concept-based Explanations

This paper introduces a formal framework for evaluating concept-based explanations from large language models, focusing on the measures of faithfulness and readability. The proposed methods involve quantifying these measures through perturbation techniques and meta-evaluation, making the results generalizable to other explanatory tasks.

https://arxiv.org/abs/2404.18533

84. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps

This paper introduces Lookback Lens, a simple model that detects contextual hallucinations in large language models by analyzing attention weights between the context and generated tokens. It demonstrates that this detection method effectively reduces hallucinations across various tasks and model sizes without needing retraining.

https://arxiv.org/abs/2407.07071

173. Calibrating the Confidence of Large Language Models by Eliciting Fidelity

This paper focuses on calibrating the confidence of large language models post-alignment to address their tendency to be overconfident. It presents a novel method for estimating model confidence and introduces new metrics for evaluating this calibration performance.

http://arxiv.org/abs/2404.02655v2

181. From Insights to Actions: The Impact of Interpretability and Analysis Research on NLP

This paper examines the impact of interpretability and analysis (IA) research in NLP, focusing on its citations and the perceived importance among the NLP community. The authors argue for the need to enhance IA work's actionable insights to increase its influence on NLP advancements.

https://arxiv.org/abs/2406.12618

191. Neuron-Level Knowledge Attribution in Large Language Models

The paper presents a static method to identify significant neurons in large language models for better understanding their mechanisms, outperforming existing attribution techniques. It introduces the concepts of 'value neurons' and 'query neurons' and analyzes knowledge across various network layers, aiding future research in knowledge editing.

https://arxiv.org/abs/2312.12141

280. An Unsupervised Approach to Achieve Supervised-Level Explainability in Healthcare Records

This study proposes an unsupervised method to generate explanations for electronic healthcare records, aiming to enhance model transparency without relying on costly human annotations. The authors introduce a new explanation method, AttInGrad, demonstrating its effectiveness over existing methods while achieving results comparable to supervised approaches.

https://arxiv.org/abs/2406.08958

299. LUQ: Long-text Uncertainty Quantification for LLMs

This paper addresses the challenge of uncertainty quantification (UQ) in large language models (LLMs), particularly focusing on their generation of long texts, which current UQ methods struggle with. The authors introduce a new approach called LUQ, which outperforms existing methods in predicting the factuality of LLM outputs and propose an ensemble method to further improve the accuracy of these responses.

https://arxiv.org/abs/2403.20279

343. SaySelf: Teaching LLMs to Express Confidence with Self-Reflective Rationales

The paper introduces SaySelf, a framework that enhances large language models (LLMs) by teaching them to express accurate fine-grained confidence estimates and generate self-reflective rationales about their uncertainties. Through a combination of supervised fine-tuning and reinforcement learning, SaySelf aims to reduce confidence calibration errors while maintaining task performance and generating useful reminders about knowledge limitations.

https://arxiv.org/abs/2405.20974

376. Discovering Knowledge-Critical Subnetworks in Pretrained Language Models

This paper investigates the presence of knowledge-critical subnetworks within pretrained language models, which can suppress specific memorized knowledge when removed, maintaining most of the model's original abilities. The authors propose a multi-objective differentiable masking scheme for identifying and manipulating these subnetworks, demonstrating significant sparsity and targeted knowledge suppression in GPT2 variants.

https://arxiv.org/abs/2310.03084

420. Unveiling Factual Recall Behaviors of Large Language Models through Knowledge Neurons

This paper examines how Large Language Models (LLMs) recall factual knowledge during reasoning tasks, revealing their tendency to follow alternative pathways instead of utilizing critical associations. The study shows that enhancing factual recall improves reasoning performance and that Chain-of-Thought prompting can aid in retrieving factual knowledge effectively.

https://arxiv.org/abs/2408.03247

432. XplainLLM: A Knowledge-Augmented Dataset for Reliable Grounded Explanations in LLMs

XplainLLM is a dataset designed to improve the transparency and reliability of Large Language Models (LLMs) by providing grounded explanations of their reasoning processes. This framework utilizes knowledge graphs and graph attention networks, aiming to reduce hallucinations and enhance the trustworthiness of LLM outputs.

https://arxiv.org/abs/2311.08614

564. Detecting Subtle Differences between Human and Model Languages Using Spectrum of Relative Likelihood

This study introduces a new approach to distinguish between human and model-generated texts using the spectrum of relative likelihood values. It proposes a detection methodology that demonstrates state-of-the-art performance on short text detection and offers insights into the subtle differences between human and generative model languages, drawing from psycholinguistics.

https://arxiv.org/abs/2406.19874

615. Self-AMPLIFY: Improving Small Language Models with Self Post Hoc Explanations

Self-AMPLIFY introduces a method to automatically generate rationales from post hoc explanation methods to enhance the performance of small language models (SLMs). By applying this technique across multiple datasets, Self-AMPLIFY shows significant accuracy improvements for SLMs by leveraging in-context learning without the need for human-annotated rationales.

https://arxiv.org/abs/2402.12038

637. ToolBeHonest: A Multi-level Hallucination Diagnostic Benchmark for Tool-Augmented Large Language Models

This paper presents ToolBH, a comprehensive diagnostic benchmark designed to assess hallucination issues in tool-augmented large language models (LLMs) through a multi-level diagnostic process. The benchmark highlights significant challenges faced by advanced LLMs, particularly in the area of task solvability, indicating that model errors are closely linked to their response strategies and training data.

https://arxiv.org/abs/2406.20015

692. Latent Concept-based Explanation of NLP Models

This paper presents the Latent Concept Attribution method (LACOAT), which aims to improve the interpretation of predictions made by deep learning models in NLP by generating context-based explanations that reflect the latent concepts of words. By mapping salient input words into the training latent space, LACOAT provides more informative insights than traditional methods that rely solely on the discrete nature of words.

http://arxiv.org/abs/2404.12545v3

699. Towards Interpretable Sequence Continuation: Analyzing Shared Circuits in Large Language Models

This paper investigates the interpretability of transformer models by analyzing shared circuits in large language models while predicting sequence continuations. The findings highlight that semantically related sequences utilize shared subgraphs, contributing to improved predictions and safer model editing processes.

https://arxiv.org/abs/2311.04131

734. FAC^{2}E: Better Understanding Large Language Model Capabilities by Dissociating Language and Cognition

This paper introduces FAC^{2}E, a framework designed to evaluate large language models by separating language capabilities from cognitive skills. The framework allows for a more detailed diagnosis of LLMs by examining their reasoning processes and identifying areas of improvement in knowledge utilization.

https://arxiv.org/abs/2403.00126

782. Enhancing Training Data Attribution for Large Language Models with Fitting Error Consideration

This paper presents a novel approach to training data attribution for large language models, tackling challenges in interpreting model outcomes and improving data protection. The proposed method, Debias and Denoise Attribution, significantly enhances the performance of influence functions by mitigating fitting errors and demonstrates strong generality across various model architectures.

https://arxiv.org/abs/2410.01285

795. The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis

This paper presents a comprehensive survey on in-context learning (ICL) in large language models, focusing on both theoretical and empirical aspects. It emphasizes the importance of interpreting ICL capabilities for better model utilization and for addressing risks like bias and toxicity in AI systems.

https://arxiv.org/abs/2311.00237

930. Interpreting Context Look-ups in Transformers: Investigating Attention-MLP Interactions

This study focuses on the interactions between attention mechanisms and multi-layer perceptrons (MLPs) within large language models, particularly in the context of next-token prediction. It proposes a methodology to analyze how attention heads correspond to token-predicting neurons, aiming to elucidate the workings of LLMs and enhance our understanding of their generative processes.

https://arxiv.org/abs/2402.15055

965. Information Flow Routes: Automatically Interpreting Language Models at Scale

This paper presents a method for automatically interpreting language models by constructing information flow routes represented as graphs, highlighting important nodes and edges with efficient attribution. The approach allows for the exploration of model behavior across different predictions and domains without the need for human-designed templates.

https://arxiv.org/abs/2403.00824

1111. Step-by-Step Reasoning to Solve Grid Puzzles: Where do LLMs Falter?

This paper evaluates the reasoning capabilities of large language models (LLMs) in solving grid puzzles by developing a new evaluation dataset called GridPuzzle and proposing an error taxonomy to analyze reasoning chains. It highlights that current prompting methods do not improve performance on these puzzles and emphasizes the need for a better understanding of model errors to advance LLM reasoning abilities.

https://arxiv.org/abs/2407.14790

1115. Revealing Personality Traits: A New Benchmark Dataset for Explainable Personality Recognition on Dialogues

This paper introduces Explainable Personality Recognition, which reveals the reasoning process behind recognizing personality traits in dialogues. The authors present the Chain-of-Personality-Evidence (CoPE) framework and a new dataset, PersonalityEvd, for training models to recognize personality traits and their supporting evidence.

https://arxiv.org/abs/2409.19723

142. Backward Lens: Projecting Language Model Gradients into the Vocabulary Space

This paper introduces a method for projecting the gradients of language models back into the vocabulary space to understand how information is learned and recalled. The authors demonstrate how these gradients can be represented as a low-rank linear combination of inputs from both the forward and backward passes of the model.

https://arxiv.org/abs/2402.12865

152. Knowledge Verification to Nip Hallucination in the Bud

This paper addresses the issue of hallucination in large language models (LLMs) by proposing Knowledge Consistent Alignment (KCA), which verifies and minimizes inconsistencies between external knowledge and intrinsic knowledge within LLMs. The results demonstrate KCA's effectiveness in reducing hallucinations across six different benchmarks, confirming its value in knowledge verification processes.

https://arxiv.org/abs/2401.10768

232. Estimating Knowledge in Large Language Models Without Generating a Single Token

This paper investigates a method for evaluating knowledge in large language models (LLMs) without generating text, focusing on predicting their ability to answer questions and the factuality of their responses based on internal computations. A simple probe called KEEN is proposed, which effectively correlates with existing evaluation metrics and offers insights into the model's knowledge and its changes post fine-tuning.

https://arxiv.org/abs/2406.12673

483. Perceptions of Linguistic Uncertainty by Language Models and Humans

This study investigates how language models interpret expressions of uncertainty compared to human responses, revealing that models can often map these expressions to numerical probabilities in a manner similar to humans. However, the models demonstrate a sensitivity to the truth value of statements that suggests they can be biased by prior knowledge, highlighting challenges for human-AI alignment.

https://arxiv.org/abs/2407.15814

486. Knowledge Conflicts for LLMs: A Survey

This paper surveys knowledge conflicts affecting large language models (LLMs), focusing on three conflict categories: context-memory, inter-context, and intra-memory. The study emphasizes the importance of understanding these conflicts to improve the trustworthiness and performance of LLMs in practical applications.

https://arxiv.org/abs/2403.08319

547. If CLIP Could Talk: Understanding Vision-Language Model Representations Through Their Preferred Concept Descriptions

This paper introduces Extract and Explore (EX2), a method to analyze and understand the representations of Vision-Language Models (VLMs) by examining the preferred textual descriptions for concepts. The study finds that VLMs heavily rely on non-visual attributes and that different VLMs prioritize different attributes in their representations, challenging the assumption that visual models predominantly use visual information.

http://arxiv.org/abs/2403.16442v1

781. Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries

This paper investigates how large language models (LLMs) process multi-hop queries and finds that information extraction occurs at different model layers. It introduces a 'back-patching' method to improve the accuracy of responses by addressing the limitations in the later layers of the model during multi-step reasoning tasks.

https://arxiv.org/abs/2406.12775

1012. On the Universal Truthfulness Hyperplane Inside LLMs

This paper investigates the existence of a universal truthfulness hyperplane within large language models (LLMs) that can distinguish between factually correct and incorrect outputs. The research indicates that increasing the diversity of training datasets significantly improves the model's performance across various tasks and domains, supporting the idea of a fundamental factual awareness within LLMs.

https://arxiv.org/abs/2407.08582

1089. Discovering Biases in Information Retrieval Models Using Relevance Thesaurus as Global Explanation

This paper introduces a novel method for globally explaining neural relevance models in information retrieval using a relevance thesaurus, which captures semantically relevant queries and document term pairs. The method particularly highlights existing biases within the models, such as brand name bias, affirming the importance of understanding model behavior on unseen query-document pairs.

http://arxiv.org/abs/2410.03584

4. Prompts have evil twins

This paper reveals the existence of 'evil twins', which are unintelligible prompts that can replace natural-language prompts while eliciting similar responses from language models. These obfuscated prompts can be transferred between models and are found by solving a maximum-likelihood problem, indicating potential broader applications.

https://arxiv.org/abs/2311.07064

92. Can Large Language Models Always Solve Easy Problems if They Can Solve Harder Ones?

This paper investigates the inconsistency in large language models (LLMs), particularly their tendency to solve harder problems while failing at easier ones. The authors introduce the ConsisEval benchmark and a consistency score metric to analyze these inconsistencies across various models, concluding that while more capable models show better consistency, exceptions occur.

https://arxiv.org/abs/2406.12809

155. Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models

This paper analyzes the phenomenon of false premise hallucinations in large language models (LLMs), identifying how specific attention heads contribute to this issue. It proposes a novel mitigation method, FAITH, which constrains these attention heads during inference, significantly improving the model's performance.

https://arxiv.org/abs/2402.19103

169. Understanding Higher-Order Correlations Among Semantic Components in Embeddings

This paper explores the limitations of Independent Component Analysis (ICA) in interpreting semantic components of embeddings, particularly focusing on the issue of non-independencies among components. It introduces a method to quantify these non-independencies using higher-order correlations, revealing their impact on semantic associations and visualizing the structure through a maximum spanning tree.

https://arxiv.org/abs/2409.19919

172. Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving

This paper presents a neuro-symbolic framework, named Explanation-Refiner, which integrates Large Language Models with Theorem Provers to enhance the validity of natural language explanations for Natural Language Inference models. The framework aims to automate the verification and refinement of explanations, providing formal guarantees on their logical correctness and facilitating improvements through feedback mechanisms.

https://arxiv.org/abs/2405.01379

193. Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis

This paper investigates the arithmetic mechanisms in large language models by identifying specific attention heads dedicated to distinct arithmetic operations. Additionally, it introduces the Comparative Neuron Analysis method to reveal an internal logic chain that aids in arithmetic tasks and addresses model editing for gender bias reduction.

https://arxiv.org/abs/2409.14144

246. When Reasoning Meets Information Aggregation: A Case Study with Sports Narratives

This paper investigates the importance of information aggregation in the reasoning capabilities of large language models (LLMs) using sports narratives, specifically NBA basketball data. The findings highlight that many LLMs struggle with accurately aggregating information due to narrative complexity and frequent scoring patterns, leading to significant errors in reasoning tasks.

https://arxiv.org/abs/2406.12084

348. Do Large Language Models Know How Much They Know?

This paper investigates the self-knowledge of large language models (LLMs), focusing on their ability to identify unanswerable questions. It introduces an automated approach to evaluate uncertainty in LLMs' responses and presents a dataset, SelfAware, to facilitate this evaluation, highlighting both LLMs' intrinsic self-knowledge and the gap compared to human performance.

https://arxiv.org/abs/2305.18153

411. FRoG: Evaluating Fuzzy Reasoning of Generalized Quantifiers in LLMs

This paper presents FRoG, a benchmark for evaluating fuzzy reasoning capabilities in large language models through real-world mathematical word problems involving generalized quantifiers. The findings indicate significant challenges in fuzzy reasoning for LLMs, with no consistent improvement from existing reasoning enhancement methods and a notable inverse scaling effect in LLM performance.

https://arxiv.org/abs/2407.01046

440. Unlocking the Future: Exploring Look-Ahead Planning Mechanistic Interpretability in Large Language Models

This paper investigates the look-ahead planning mechanisms in large language models (LLMs) by examining the information flow and internal representations involved in planning. The findings reveal how decision-making is encoded within the model, shedding light on the planning capabilities of LLMs and paving the way for future research in this area.

https://arxiv.org/abs/2406.16033

443. Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words?

This paper investigates how large language models (LLMs) can express their internal uncertainty in natural language responses, particularly when faced with conflicting answers. The authors propose a metric to evaluate whether LLMs appropriately hedge their responses based on their confidence levels, revealing a significant gap in LLMs' ability to communicate their uncertainty accurately, indicating a need for better alignment to enhance trustworthiness.

http://arxiv.org/abs/2405.16908v2

500. Beyond Label Attention: Transparency in Language Models for Automated Medical Coding via Dictionary Learning

This paper presents a method for improving the interpretability of language models used for automated medical coding by employing dictionary learning. By extracting understandable representations from dense embeddings, the approach enhances the transparency of model predictions even for medically irrelevant tokens.

https://arxiv.org/abs/2411.00173

543. Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs

This paper investigates how large language models (LLMs) process tokens, finding that less common words are often split into multiple semantically unrelated tokens, leading to an 'erasure' of information in early layers. It proposes a method to explore the implicit vocabulary of LLMs by examining token representation differences across layers, specifically for Llama-2-7b and Llama-3-8B.

https://arxiv.org/abs/2406.20086

569. Atomic Inference for NLI with Generated Facts as Atoms

This paper introduces atomic inference for Natural Language Inference (NLI) using LLM-generated facts, ensuring model predictions are interpretable and faithful. By decomposing NLI premises into facts and employing a two-stage generation and training process, the proposed method significantly improves performance over traditional approaches.

https://arxiv.org/abs/2305.13214

574. When Parts are Greater Than Sums: Individual LLM Components Can Outperform Full Models

This paper investigates in-context learning by analyzing the contributions of different components of large language models (LLMs). It presents a method for component reweighting that enhances classification accuracy by adjusting component activations based on limited labeled examples.

https://arxiv.org/abs/2406.13131

586. Belief Revision: The Adaptability of Large Language Models Reasoning

This paper introduces Belief-R, a new dataset to evaluate the ability of language models to revise beliefs when faced with new evidence. The study reveals that although some models can update their beliefs, they struggle in scenarios where no updates are needed, indicating a significant challenge in their adaptiveness to changing information.

https://arxiv.org/abs/2406.19764

590. LLMs Are Prone to Fallacies in Causal Inference

The paper investigates the ability of LLMs to infer causal relations from relations in text, distinguishing between memorized causal facts and those that are inferred. It finds that LLMs can be misled into inferring causality from entity mention order and struggle with counterfactual relations, suggesting limitations in their understanding of causality.

https://arxiv.org/abs/2406.12158

627. A Multi-Perspective Analysis of Memorization in Large Language Models

This paper discusses the phenomenon of memorization in large language models (LLMs), exploring its causes and the dynamics of both memorized and unmemorized content. Through experiments and embedding analyses, the research reveals relationships between model size, context size, and memorization, providing insights into predicting memorization behavior.

https://arxiv.org/abs/2405.11577

649. Fishing for Magikarp: Automatically Detecting Under-trained Tokens in Large Language Models

This paper addresses the issue of 'glitch tokens' in Large Language Models (LLMs), which are tokens that exist in the tokenizer but are under-represented in the training data. It introduces methods for detecting these problematic tokens, highlighting their prevalence and proposing solutions to enhance the efficiency and safety of LLMs.

https://arxiv.org/abs/2405.05417

650. Reasoning or a Semblance of it? A Diagnostic Study of Transitive Reasoning in LLMs

This paper examines the transitive reasoning capabilities of two Large Language Models (LLMs), LLaMA 2 and Flan-T5, to determine if they engage in genuine logical reasoning or rely on implicit cues for answering compositional questions. The results show that while both models utilize certain cues, Flan-T5 demonstrates more consistency and resilience across manipulations, suggesting a response based more on an understanding developed through fine-tuning.

https://arxiv.org/abs/2410.20200

700. Why Does New Knowledge Create Messy Ripple Effects in LLMs?

This paper investigates the causes behind messy ripple effects that occur during knowledge editing in language models, proposing GradSim as an effective indicator for identifying when updated knowledge influences related information. The authors demonstrate a strong correlation between ripple effect performance and GradSim, providing insights into failure cases associated with low GradSim values.

https://arxiv.org/abs/2407.12828

737. Learning Personalized Alignment for Evaluating Open-ended Text Generation

The paper presents PerSE, an interpretable evaluation framework that enhances the assessment of open-ended text generation by aligning it with diverse human values and preferences. It demonstrates significant improvements in correlation with human judgments compared to existing models, particularly in evaluating personalized content generation.

https://arxiv.org/abs/2310.03304

780. Cluster-Norm for Unsupervised Probing of Knowledge

This paper introduces a cluster normalization method aimed at improving unsupervised probing techniques for extracting knowledge from language models while minimizing misleading influences from unrelated dataset features. Although it does not fully resolve the distinction between real and simulated knowledge, the proposed method enhances the accuracy of knowledge identification in language models.

https://arxiv.org/abs/2407.18712

789. First Heuristic Then Rational: Dynamic Use of Heuristics in Language Model Reasoning

The paper investigates the dynamic use of heuristics in multi-step reasoning by language models (LMs), showing that they initially favor heuristic methods like lexical overlap before switching to more rationale-based methods as they approach their final answers. This finding suggests that LMs can adaptively combine different reasoning strategies depending on their position in the reasoning process.

https://arxiv.org/abs/2406.16078

854. Exploring the Role of Reasoning Structures for Constructing Proofs in Multi-Step Natural Language Reasoning with Large Language Models

This paper investigates how large language models can effectively utilize structured intermediate proof steps to enhance their multi-step reasoning abilities and explainability. The study identifies and analyzes the benefits of structure-aware demonstration and pruning techniques in improving model performance on complex reasoning tasks.

https://arxiv.org/html/2410.08436

866. CONTESTS: a Framework for Consistency Testing of Span Probabilities in Language Models

The paper introduces ConTestS, a framework designed to assess the consistency of span probabilities in language models through statistical testing methods. Experiments reveal that autoregressive models display larger inconsistencies compared to Masked Language Models, which suggests important implications for interpreting language model outputs and for selecting decoding strategies.

https://arxiv.org/abs/2409.19984

911. Finding Blind Spots in Evaluator LLMs with Interpretable Checklists

This paper investigates the effectiveness of large language models (LLMs) as evaluators for text generation tasks by examining their ability to assess various critical capabilities. The proposed framework, FBI, reveals significant shortcomings in LLM evaluators, highlighting the unreliable nature of these models and advocating for careful implementation in practical applications.

https://arxiv.org/abs/2406.13439

1007. Calibrating Language Models with Adaptive Temperature Scaling

This paper presents Adaptive Temperature Scaling (ATS), a post-hoc method aiming to enhance the calibration of large language models (LLMs) after fine-tuning with reinforcement learning from human feedback. ATS predicts a temperature parameter for each token prediction, improving calibration by 10-50% on various natural language evaluation benchmarks without hindering performance gains from RLHF.

https://arxiv.org/abs/2409.19817

1182. Connecting the Dots: Evaluating Abstract Reasoning Capabilities of LLMs Using the New York Times Connections Word Game

This paper evaluates the abstract reasoning capabilities of large language models (LLMs) using the New York Times Connections word game, finding that LLMs struggle significantly compared to human players, especially regarding certain knowledge types. It establishes a taxonomy of knowledge necessary for solving the game, highlighting the limitations of LLMs in areas beyond semantic relations.

https://arxiv.org/abs/2406.11012

150. Collaborative Performance Prediction for Large Language Models

This paper introduces Collaborative Performance Prediction (CPP), a framework that improves the accuracy of predicting large language models' performance across various tasks by utilizing historical data and design factors. It addresses limitations of existing scaling laws by emphasizing similarities across different model families.

https://arxiv.org/abs/2407.01300

368. Detection and Measurement of Syntactic Templates in Generated Text

This paper analyzes syntactic templates in text generated by language models, demonstrating that they produce templated text more frequently than human-authored texts. The authors argue that these templates, which tend to be present in pre-training data, are useful for evaluating model constructions and analyzing style memorization.

https://arxiv.org/abs/2407.00211

582. Encourage or Inhibit Monosemanticity? Revisit Monosemanticity from a Feature Decorrelation Perspective

This paper examines the concept of monosemanticity in large language models and its effects on model performance through a feature decorrelation approach. It argues that encouraging monosemanticity can enhance model capacity and proposes a regularization technique that improves representation diversity and preference alignment.

http://arxiv.org/abs/2406.17969v2

1077. CaT-Bench: Benchmarking Language Model Understanding of Causal and Temporal Dependencies in Plans

This paper presents CaT-Bench, a benchmark for evaluating the understanding of causal and temporal dependencies in instructional texts by language models. The findings indicate significant limitations in LLMs' ability to reason about step orders in plans, with a maximum F1 score of 0.73, highlighting the need for improvement in their reasoning capabilities.

http://arxiv.org/abs/2406.15823v1

1160. Multi-LogiEval: Towards Evaluating Multi-Step Logical Reasoning Ability of Large Language Models

The paper introduces Multi-LogiEval, a comprehensive dataset aimed at evaluating the multi-step logical reasoning abilities of large language models (LLMs) using diverse inference rules. The findings highlight a significant decline in performance as reasoning depth increases, suggesting a critical area for future research in enhancing LLM logical reasoning capabilities.

https://arxiv.org/abs/2406.17169

166. Evaluating Large Language Models via Linguistic Profiling

288. Investigating How Large Language Models Leverage Internal Knowledge to Perform Complex Reasoning

403. Does Large Language Model Contain Task-Specific Neurons?

1032. Lost in Tokenization: How to Measure Word Surprisal From LM Token Probabilities

116. Embedding and Gradient Say Wrong: A White-Box Method for Hallucination Detection

225. Interpretability-based Tailored Knowledge Editing in Transformers

379. Verifiable, Debuggable, and Repairable Commonsense Logical Reasoning via LLM-based Theory Resolution

428. Rethinking the Reversal Curse of LLMs: a Prescription from Human Knowledge Reversal

524. Toward Compositional Behavior in Neural Models: A Survey of Current Views

584. Reasoning Robustness of LLMs to Adversarial Typographical Errors

664. Rationalizing Transformer Predictions via End-To-End Differentiable Self-Training

740. Null-Shot Prompting: Rethinking Prompting Large Language Models With Hallucination

810. Interpretable Composition Attribution Enhancement for Visio-linguistic Compositional Understanding

825. RepMatch: Quantifying Cross-Instance Similarities in Representation Space

837. HalluMeasure: Fine-grained Hallucination Measurement Using Chain-of-Thought Reasoning

885. Text Fluoroscopy: Detecting LLM-Generated Text through Intrinsic Features

1051. Do Explanations Help or Hurt? Saliency Maps vs Natural Language Explanations in a Clinical Decision-Support Setting

1227. The Death and Life of Great Prompts: Analyzing the Evolution of LLM Prompts from the Structural Perspective

1246. SimLLM: Detecting Sentences Generated by Large Language Models Using Similarity between the Generation and its Re-generation

Low-resource Methods for NLP

5. Table Question Answering for Low-resourced Indic Languages

This paper presents a fully automatic large-scale data generation process for Table Question Answering (TableQA) specifically targeting low-resource Indic languages, such as Bengali and Hindi. The proposed method enables the training of TableQA models that outperform existing state-of-the-art models and establishes new frameworks for scalable data generation in NLP for low-resource contexts.

https://arxiv.org/abs/2410.03576

28. Clustering and Ranking: Diversity-preserved Instruction Selection through Expert-aligned Quality Estimation

This paper introduces Clustering and Ranking (CaR), a method designed to efficiently select high-quality instruction tuning data while preserving dataset diversity. The experiments demonstrate that CaR not only reduces resource allocation but also significantly improves model performance compared to existing data selection methods.

https://arxiv.org/abs/2402.18191

78. Model Balancing Helps Low-data Training and Fine-tuning

The paper discusses the challenges of low-data training and fine-tuning for pre-trained models, particularly in the context of natural language processing and scientific machine learning. It introduces TempBalance, a layer-wise learning rate scheduler that improves model performance by balancing training quality across layers, especially as the amount of available tuning data decreases.

https://arxiv.org/abs/2410.12178

109. An Effective Deployment of Diffusion LM for Data Augmentation in Low-Resource Sentiment Classification

This paper presents DiffusionCLS, a method that uses a diffusion language model to enhance data augmentation for sentiment classification tasks, particularly in low-resource scenarios. The approach ensures the preservation of important sentiment-bearing tokens while providing enough variability in the generated data to improve model performance.

http://arxiv.org/abs/2409.03203v2

168. KB-Plugin: A Plug-and-play Framework for Large Language Models to Induce Programs over Low-resourced Knowledge Bases

KB-Plugin is a framework that employs self-supervised learning to help large language models (LLMs) induce programs over low-resourced knowledge bases (KBs) without requiring extensive parallel data. The framework leverages rich-resourced KBs for training while achieving comparable performance to state-of-the-art methods with much smaller LLMs on various knowledge question answering datasets.

https://arxiv.org/abs/2402.01619

186. Cross-lingual Transfer for Automatic Question Generation by Learning Interrogative Structures in Target Languages

This paper presents a cross-lingual transfer method for automatic question generation (QG) that enables question creation in low-resource languages using models trained on English datasets. By learning interrogative structures from a small set of examples, the method shows improved performance over existing baselines and generates useful synthetic data for multilingual question-answering tasks.

http://arxiv.org/abs/2410.03197

220. Incubating Text Classifiers Following User Instruction with Nothing but LLM

This paper introduces a framework called Incubator, which generates text classification data based on user-defined class definitions without requiring human annotations. The proposed method leverages large language models to handle complex class relationships and has shown promising results compared to existing baselines across various classification tasks.

https://arxiv.org/abs/2404.10877

239. Teaching LLMs to Abstain across Languages via Multilingual Feedback

This paper proposes a method for teaching multilingual language models (LLMs) to abstain from making incorrect assertions in under-resourced languages by using multilingual feedback to identify knowledge gaps. The research demonstrates that utilizing multilingual feedback significantly improves LLM performance in low-resource languages while highlighting the importance of cultural factors in language modeling.

https://arxiv.org/abs/2406.15948

251. Voices Unheard: NLP Resources and Models for Yorùbá Regional Dialects

This paper presents the YORULECT corpus, a high-quality parallel text and speech dataset for four regional dialects of the Yorùbá language, aimed at addressing the lack of NLP resources for these dialects. The study highlights the performance disparities in various NLP tasks between the standard Yorùbá and its dialects, while demonstrating that dialect-adaptive finetuning can help narrow this gap.

https://arxiv.org/abs/2406.19564

261. Multiple Sources are Better Than One: Incorporating External Knowledge in Low-Resource Glossing

This paper tackles the issue of data scarcity in automatic glossing for low-resource languages by integrating various sources of linguistic expertise. The proposed method improves word-level accuracy significantly, especially for the most under-resourced language, Gitksan, demonstrating effective use of external knowledge in language processing for low-resource settings.

https://arxiv.org/abs/2406.11085

350. SciPrompt: Knowledge-Augmented Prompting for Fine-Grained Categorization of Scientific Topics

The paper introduces SciPrompt, a framework that enhances prompt-based fine-tuning for scientific text classification by automatically retrieving relevant label terms, particularly in low-resource settings. It showcases improved performance in classifying fine-grained scientific topics, especially in few and zero-shot scenarios, surpassing existing methods.

http://arxiv.org/abs/2410.01946v1

371. Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts

The paper presents Personalized Pieces (Per-Pcs), a framework for efficiently personalizing large language models (LLMs) through collaborative efforts while ensuring user privacy. By breaking parameter-efficient fine-tuning (PEFT) methods into shareable pieces and allowing users to assemble them, Per-Pcs enhances model personalization with reduced resource demands and improved scalability.

https://arxiv.org/abs/2406.10471

482. CSSL: Contrastive Self-Supervised Learning for Dependency Parsing on Relatively Free Word Ordered and Morphologically Rich Low Resource Languages

This paper presents a contrastive self-supervised learning method for enhancing dependency parsing in morphologically rich languages with relatively free word order. The proposed approach improves parsing performance significantly, demonstrating the potential of modifying graph-based parsing architectures to handle word order variations effectively.

http://arxiv.org/abs/2410.06944

499. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents

This paper presents MiniCheck, an efficient fact-checking model for large language model outputs that achieves GPT-4-level performance at a significantly reduced computational cost. The authors create synthetic training data to train their model, which leads to improved accuracy in evaluating the factual accuracy of LLM generations across various tasks.

https://arxiv.org/abs/2404.10774

529. Less is More: Parameter-Efficient Selection of Intermediate Tasks for Transfer Learning

This paper introduces Embedding Space Maps (ESMs) as a lightweight approach to select intermediate tasks for transfer learning, enhancing model performance with reduced resource usage. Through a comprehensive study of 12,000 source-target pairs, ESMs demonstrate significant efficiency improvements while maintaining high selection accuracy.

https://arxiv.org/abs/2410.15148

669. Self-Training for Sample-Efficient Active Learning for Text Classification with Pre-Trained Language Models

This paper explores the use of self-training to enhance sample efficiency in active learning for text classification with pre-trained language models. It introduces HAST, a new self-training strategy that demonstrates improved performance in classification while using significantly less labeled data.

https://arxiv.org/abs/2406.09206

719. Target-Aware Language Modeling via Granular Data Sampling

This paper proposes a data sampling method using low-dimensional n-gram features to enhance language model pretraining for specific domains without reducing performance on general tasks. The approach achieves high correlation with downstream task performance while allowing efficient pretraining with significantly reduced data volume, showing effective results across multiple benchmarks.

https://arxiv.org/abs/2409.14705

746. Adaptive Query Rewriting: Aligning Rewriters through Marginal Probability of Conversational Answers

This paper presents AdaQR, a novel framework for adaptive query rewriting in open-domain conversational question answering, that leverages limited annotations for training. The method improves the generalization and adaptability of rewriting models without the need for extensive retrieval or rewrite annotations.

https://arxiv.org/abs/2406.10991

753. Zero-Shot Cross-Lingual NER Using Phonemic Representations for Low-Resource Languages

This paper presents a novel zero-shot cross-lingual named entity recognition (NER) approach using phonemic representations for low-resource languages, addressing the challenges of the lack of prior knowledge for these languages. The proposed method shows significant improvements over baseline models, particularly for languages with non-Latin scripts, achieving the highest average F1 score in low-resource context.

https://arxiv.org/abs/2406.16030

792. Cross-lingual Back-Parsing: Utterance Synthesis from Meaning Representation for Zero-Resource Semantic Parsing

The paper proposes Cross-Lingual Back-Parsing (CBP), a method to enhance zero-resource semantic parsing by synthesizing target language utterances from source meaning representations. Experiments demonstrate that CBP significantly improves performance in cross-lingual semantic parsing tasks by utilizing labeled data in a source language and monolingual corpora.

https://arxiv.org/abs/2410.00513

841. Open-world Multi-label Text Classification with Extremely Weak Supervision

This paper presents a novel approach, X-MLClass, for open-world multi-label text classification using extremely weak supervision. By leveraging user descriptions and a large language model, the method iteratively refines label spaces and achieves significant improvements in classification accuracy and label coverage across various datasets.

http://arxiv.org/abs/2407.05609

1071. SynthesizRR: Generating Diverse Datasets with Retrieval Augmentation

The paper introduces SynthesizRR, a method for generating diverse datasets through the use of retrieval augmentation, which aims to enhance dataset synthesis by adding variety to the examples generated by large language models. The empirical study demonstrates that SynthesizRR improves lexical and semantic diversity, making the generated datasets more similar to human-written text and enhancing the performance of the distilled models.

https://arxiv.org/abs/2405.10040

1132. DEFT-UCS: Data Efficient Fine-Tuning for Pre-Trained Language Models via Unsupervised Core-Set Selection

This paper introduces DEFT-UCS, a framework for data-efficient fine-tuning of pre-trained language models through unsupervised core-set selection. The framework demonstrates highly accurate performance using significantly less data than state-of-the-art models for text-editing tasks.

https://arxiv.org/abs/2310.16776

1162. Task Oriented In-Domain Data Augmentation

The paper introduces TRAIT, a framework for task-oriented in-domain data augmentation to enhance large language models' performance on specialized domains by addressing the scarcity of in-domain data and creating task-aware synthetic passages. TRAIT demonstrates an average improvement of 8% in the advertisement domain and 7.5% in the math domain through better data selection and synthetic passage generation.

https://arxiv.org/abs/2406.16694

334. Do Not Worry if You Do Not Have Data: Building Pretrained Language Models Using Translationese

The paper investigates using Translationese as synthetic data for pre-training language models to address data scarcity in non-English languages. It demonstrates that models pre-trained on this synthetic data produce performance results comparable to those trained on clean data while also releasing a new dataset, IndicMonoDoc, to support further research.

https://arxiv.org/abs/2403.13638

1217. T-FREE: Tokenizer-Free Generative LLMs via Sparse Representations for Memory-Efficient Embeddings

T-FREE is a novel approach that eliminates the need for tokenizers in Large Language Models by directly embedding words through sparse representations, significantly reducing computational overhead. This method enhances performance for underrepresented languages and achieves over 85% parameter reduction, while still providing competitive results in downstream tasks.

https://arxiv.org/abs/2406.19223

46. Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models

This paper explores parameter-efficient fine-tuning (PEFT) for sparse-architecture Large Language Models (LLMs) utilizing the Mixture-of-Experts (MoE) architecture. It introduces Expert-Specialized Fine-Tuning (ESFT) which improves tuning efficiency by selecting task-relevant experts while freezing others, significantly enhancing model performance without full parameter fine-tuning.

http://arxiv.org/abs/2407.01906

74. Effective Demonstration Annotation for In-Context Learning via Language Model-Based Determinantal Point Process

This paper presents a novel selective annotation mechanism for in-context learning (ICL) that enhances few-shot learning by focusing on optimizing the selection of examples based on uncertainty and diversity. The proposed Language Model-based Determinant Point Process (LM-DPP) demonstrates effectiveness across various language models in refining input-output pair mappings for different datasets.

https://arxiv.org/abs/2408.02103

176. Zero-shot Cross-Lingual Transfer for Synthetic Data Generation in Grammatical Error Detection

This paper explores grammatical error detection (GED) in low-resource languages by generating synthetic error corpora using zero-shot cross-lingual transfer from multilingual models. The proposed method achieves performance improvements over existing annotation-free GED methods by first fine-tuning on synthetic data and then on human-annotated corpora.

https://arxiv.org/abs/2407.11854

219. Text Grafting: Near-Distribution Weak Supervision for Minority Classes in Text Classification

This paper introduces 'text grafting', a novel framework for generating near-distribution weak supervision for minority classes in text classification. By combining the strengths of language models with template mining from raw corpora, it enhances the classification accuracy for underrepresented categories.

https://arxiv.org/abs/2406.11115

497. Jellyfish: Instruction-Tuning Local Large Language Models for Data Preprocessing

This paper investigates the application of large language models (LLMs) in data preprocessing, improving the data mining pipeline by customizing and securing the data transformation process. It introduces Jellyfish, a set of LLMs tuned for various data preprocessing tasks that demonstrate competitiveness with GPT-3.5 and enhanced reasoning capabilities while ensuring data confidentiality.

https://arxiv.org/abs/2312.01678

809. Small Agent Can Also Rock! Empowering Small Language Models as Hallucination Detector

The paper introduces HaluAgent, an agent framework that empowers smaller language models to detect various types of hallucinations effectively. By fine-tuning on synthesized detection trajectories, HaluAgent achieves comparable performance to larger models like GPT-4 using limited training data.

https://arxiv.org/abs/2406.11277

942. LLM-based Code-Switched Text Generation for Grammatical Error Correction

This paper presents a study on code-switching (CSW) in natural language processing, focusing on grammatical error correction (GEC) for multilingual texts. It introduces a synthetic dataset for CSW GEC and demonstrates improvements over existing models in correcting grammatical errors in both monolingual and code-switched contexts for ESL learners.

https://arxiv.org/abs/2410.10349

970. Context-Aware Adapter Tuning for Few-Shot Relation Learning in Knowledge Graphs

The paper presents RelAdapter, a context-aware adapter designed for few-shot relation learning in knowledge graphs to address the challenges of predicting novel relations with limited training data. It enhances the adaptation process within meta-learning by using a lightweight module and incorporating contextual information about target relations, demonstrating superior performance in experiments.

https://arxiv.org/abs/2410.09123

1044. Evaluating Large Language Models along Dimensions of Language Variation: A Systematik Invesdigatiom uv Cross-lingual Generalization

This paper investigates the performance degradation of large language models on closely-related languages and dialects, focusing on understanding the linguistic distances that contribute to this decline. It presents a framework using Bayesian noise processes to model phonological, morphological, and lexical distances, providing insights into mitigating performance degradation in low-resource languages.

https://arxiv.org/abs/2406.13718

1127. Back to School: Translation Using Grammar Books

This paper explores improving machine translation for low-resource languages by incorporating grammar books into the prompts used with GPT-4. It demonstrates that using available linguistic resources can enhance translation quality across diverse, under-represented languages.

https://arxiv.org/html/2410.15263v1

1173. ArMeme: Propagandistic Content in Arabic Memes

This paper addresses the development of a new Arabic memes dataset annotated for propagandistic content, highlighting the importance of identifying misleading information in digital memes. It provides a resource aimed at stakeholders such as social media platforms and policymakers, emphasizing the need for computational tools to detect such content, especially for low-resource languages.

https://arxiv.org/abs/2406.03916

1211. Casablanca: Data and Models for Multidialectal Arabic Speech Recognition

This paper presents Casablanca, a community-driven project aimed at addressing the technological divide by collecting and transcribing a large-scale dataset for multiple Arabic dialects. The dataset includes annotations for transcription, gender, dialect, and code-switching and is intended to enhance speech recognition systems for underrepresented languages.

http://arxiv.org/abs/2410.04527

426. Multi-Dialect Vietnamese: Task, Dataset, Baseline Models and Challenges

This paper introduces the Vietnamese Multi-Dialect (ViMD) dataset, which includes 102.56 hours of audio capturing 63 provincial dialects across Vietnam, addressing the lack of fine-grained classification for dialects. It evaluates the dataset by fine-tuning pre-trained models for dialect identification and speech recognition, highlighting the geographical influences on dialects and the effectiveness of current speech recognition methods.

https://arxiv.org/abs/2410.03458

1145. Can LLM Generate Culturally Relevant Commonsense QA Data? Case Study in Indonesian and Sundanese

This study explores the use of Large Language Models (LLMs) for generating culturally relevant commonsense question answering (QA) datasets specifically for Indonesian and Sundanese languages. Through various generation methods, the researchers produced a substantial dataset and discovered that while LLMs can create adequate questions, they lack the cultural depth and face challenges related to fluency in lower-resource languages.

https://arxiv.org/abs/2402.17302

830. Unsupervised Named Entity Disambiguation for Low Resource Domains

145. Reusing Transferable Weight Increments for Low-resource Style Generation

462. Language-to-Code Translation with a Single Labeled Example

851. Joint Pre-Encoding Representation and Structure Embedding for Efficient and Low-Resource Knowledge Graph Completion

136. Towards Low-Resource Harmful Meme Detection with LMM Agents

638. TEMA: Token Embeddings Mapping for Enriching Low-Resource Language Models

1199. RAR: Retrieval Augmented Retrieval for Code Generation in Low Resource Languages

293. Academics Can Contribute to Domain-Specialized Language Models

1181. Generalizing Clinical De-identification Models by Privacy-safe Data Augmentation using GPT-4

1185. CoverICL: Selective Annotation for In-Context Learning via Active Graph Coverage

567. ARM: An Alignment-and-Replacement Module for Chinese Spelling Check Based on LLMs

Multimodality and Language Grounding to Vision, Robotics and Beyond

6. ImageInWords: Unlocking Hyper-Detailed Image Descriptions

The paper introduces ImageInWords (IIW), a human-in-the-loop framework designed to generate hyper-detailed image descriptions that address issues related to completeness and visual inaccuracies in existing models. Extensive evaluations show significant improvements in description quality, compositional reasoning, and visual fidelity when using IIW-generated data compared to previous methods.

https://arxiv.org/abs/2405.02793

8. When LLMs Meets Acoustic Landmarks: An Efficient Approach to Integrate Speech into Large Language Models for Depression Detection

This paper proposes an innovative method to integrate acoustic speech information with Large Language Models (LLMs) for detecting depression, addressing a significant limitation of LLMs which rely solely on text. By using Acoustic Landmarks to enrich the analysis of speech patterns alongside text, the approach demonstrates state-of-the-art results on the DAIC-WOZ dataset for multimodal depression detection.

https://arxiv.org/abs/2402.13276

89. UniFashion: A Unified Vision-Language Model for Multimodal Fashion Retrieval and Generation

The paper presents UniFashion, a unified vision-language model aimed at addressing multimodal retrieval and generation specifically in the fashion domain. It successfully integrates image and text generation with retrieval tasks, outperforming prior models and showcasing the synergy between these multimodal tasks.

https://arxiv.org/abs/2408.11305

114. Enhancing Advanced Visual Reasoning Ability of Large Language Models

This paper introduces Complex Visual Reasoning Large Language Models (CVR-LLM) to enhance visual reasoning capabilities by integrating the strengths of Vision-Language Models (VLMs) and Large Language Models (LLMs). The CVR-LLM utilizes a novel multi-modal in-context learning approach and Chain-of-Comparison technique to achieve state-of-the-art performance on complex visual reasoning tasks.

https://arxiv.org/abs/2409.13980

124. Towards Difficulty-Agnostic Efficient Transfer Learning for Vision-Language Models

This paper analyzes the varying transfer difficulty of downstream tasks for vision-language models (VLMs) and introduces an adaptive ensemble method that employs vision prompts and text adapters. The proposed method enhances the performance of VLMs across different task difficulties, demonstrating effectiveness especially in unseen tasks.

https://arxiv.org/html/2311.15569v2

182. Dual Modalities of Text: Visual and Textual Generative Pre-Training

This paper introduces a novel pre-training framework for pixel-based autoregressive language models, utilizing a dual-modality training approach that incorporates both visual and textual data. The findings indicate that blending visual and textual information significantly enhances the performance of these models in language understanding tasks, highlighting the potential for further research in this area.

https://arxiv.org/html/2404.10710v2

218. MPT: Multimodal Prompt Tuning for Zero-shot Instruction Learning

This paper presents a novel Multimodal Prompt Tuning (M^{2}PT) method to enhance zero-shot instruction learning capabilities of Multimodal Large Language Models (MLLMs). The proposed approach efficiently integrates visual and textual prompts during finetuning, demonstrating superior performance across various multimodal evaluation datasets.

https://arxiv.org/abs/2409.15657

237. Teaching Embodied Reinforcement Learning Agents: Informativeness and Diversity of Language Use

This paper investigates the role of diverse and informative language inputs in enhancing the learning process of embodied reinforcement learning agents. The results indicate that incorporating rich language use significantly improves the generalization and adaptation capabilities of these agents in open-world tasks.

http://arxiv.org/abs/2410.24218v1

254. VIMI: Grounding Video Generation through Multi-modal Instruction

This paper introduces VIMI, a multimodal conditional video generation framework that addresses the limitations of existing text-to-video models by incorporating a large-scale multimodal prompt dataset. VIMI employs a two-stage training process leading to contextually rich video generation, demonstrating state-of-the-art results on the UCF101 benchmark.

http://www.arxiv.org/abs/2407.06304

265. World to Code: Multi-modal Data Generation via Self-Instructed Compositional Captioning and Filtering

The paper presents a multi-modal data generation pipeline called World to Code (W2C) that generates aligned visual and textual data in Python code format. It demonstrates improved performance in visual question answering and grounding tasks by leveraging Vision-Language Models (VLMs) for extracting cross-modal information.

http://arxiv.org/abs/2409.20424

276. RWKV-CLIP: A Robust Vision-Language Representation Learner

This paper introduces RWKV-CLIP, a robust model for vision-language representation learning that addresses data noise and enhances image-text data quality through a diverse description generation framework leveraging Large Language Models. Comprehensive experiments showcase its state-of-the-art performance on various vision-language tasks, highlighting its efficient training and inference capabilities.

https://arxiv.org/abs/2406.06973

284. From the Least to the Most: Building a Plug-and-Play Visual Reasoner via Data Synthesis

This paper presents a new approach for multi-step reasoning in vision-language models (VLMs) using a least-to-most visual reasoning paradigm. It discusses a data synthesis method for producing high-quality reasoning datasets, which significantly enhances the reasoning capabilities of existing VLMs through a plug-and-play visual reasoner.

https://arxiv.org/abs/2406.19934

291. Concept-skill Transferability-based Data Selection for Large Vision-Language Models

This paper presents COINCIDE, a data selection technique designed to improve the efficiency of instruction tuning for Large Vision-Language Models by selecting a diverse and transferably beneficial subset of training data. Using only a fraction of the data, COINCIDE demonstrates superior performance and reduced training time compared to several strong baselines across multiple datasets.

https://arxiv.org/abs/2406.10995

305. How Does the Textual Information Affect the Retrieval of Multimodal In-Context Learning?

This paper investigates how textual information impacts the retrieval of in-context examples in multimodal large language models (MLLMs), emphasizing the bias toward visual data. The authors introduce a novel supervised MLLM-retriever, MSIER, that enhances multimodal in-context learning efficiency through neural network-based example selection, validated by extensive testing across multiple tasks.

https://arxiv.org/abs/2404.12866

311. Words Worth a Thousand Pictures: Measuring and Understanding Perceptual Variability in Text-to-Image Generation

This paper investigates perceptual variability in text-to-image generation using diffusion models and introduces W1KP, a novel measure of variability. The study reveals how different prompts influence image reusability and variability, providing insights based on linguistics and evaluation metrics.

https://arxiv.org/abs/2406.08482

342. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

The paper introduces Video-LLaVA, a unified visual representation model that enhances multi-modal understanding by aligning the feature spaces of images and videos before projection into a language model. This method has shown to improve performance across several benchmarks compared to models focused solely on images or videos.

http://arxiv.org/abs/2311.10122v3

356. Finer: Investigating and Enhancing Fine-Grained Visual Concept Recognition in Large Vision Language Models

This paper investigates the performance of Large Vision-Language Models (LVLMs) in fine-grained visual categorization, revealing significant deficiencies in classification accuracy and explanatory capabilities. It introduces a benchmark, Finer, aimed at evaluating the fine-grained visual understanding of LVLMs and improving their explainability.

https://arxiv.org/abs/2402.16315

358. VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment

This paper presents VLFeedback, a large-scale AI feedback dataset aimed at improving the alignment of large vision-language models (LVLMs) through AI-generated human-like feedback. The dataset comprises over 82K multi-modal instructions and reveals that fine-tuning with this feedback significantly enhances performance metrics on various tasks.

https://arxiv.org/abs/2410.09421

361. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities

This paper presents GAMA, a novel Large Audio-Language Model designed to understand non-verbal speech and non-speech sounds, integrating an LLM with various audio representations. The model is fine-tuned on a large-scale audio-language dataset and features new instruction-tuning methods for complex reasoning in audio question-answering tasks, outperforming existing models in diverse auditory comprehension exercises.

https://arxiv.org/abs/2406.11768

367. EPO: Hierarchical LLM Agents with Environment Preference Optimization

This paper introduces a hierarchical framework for long-horizon decision-making using LLMs, decomposing complex tasks into subgoals and employing multimodal feedback for reward generation. The proposed Environment Preference Optimization method improves training signals for LLM-based agents, achieving top performance on the ALFRED dataset.

https://arxiv.org/abs/2408.16090

391. Beyond Embeddings: The Promise of Visual Table in Visual Reasoning

This paper introduces Visual Table, a new visual representation tailored for visual reasoning, which outperforms conventional visual embeddings by providing hierarchical descriptions of visual scenes along with instance-level world knowledge. The proposed method enhances the interpretability and controllable editing of visual representations while demonstrating significant improvements on various visual reasoning benchmarks and multimodal large language models.

http://arxiv.org/abs/2403.18252v2

418. Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale

This paper discusses the development of HuatuoGPT-Vision, a medical multimodal large language model (MLLM) that enhances medical visual knowledge by refining and denoising medical image-text pairs from PubMed. The resulting PubMedVision dataset, containing 1.3 million medical VQA samples, significantly improves the performance of medical MLLMs compared to existing datasets.

https://arxiv.org/abs/2406.19280

434. SURf: Teaching Large Vision-Language Models to Selectively Utilize Retrieved Information

The paper introduces SURf, a self-refinement framework that enhances Large Vision-Language Models (LVLMs) by teaching them to selectively utilize retrieved information when answering questions. Through experimentation across various tasks and datasets, the framework improves the models' ability to discern relevant from irrelevant information, thus enhancing their overall performance.

https://arxiv.org/abs/2409.14083

446. MIND: Multimodal Shopping Intention Distillation from Large Vision-language Models for E-commerce Purchase Understanding

The paper introduces MIND, a multimodal framework for deriving purchase intentions from large vision-language models in E-commerce environments. It emphasizes understanding human-centric intentions through a knowledge base compiled from extensive Amazon Review data and demonstrates improvements in intention comprehension tasks.

https://arxiv.org/abs/2406.10701

455. Empowering Backbone Models for Visual Text Generation with Input Granularity Control and Glyph-Aware Training

This paper addresses the limitations of diffusion-based text-to-image models in generating legible visual texts, particularly for English and Chinese. The authors propose enhancements through input granularity control and glyph-aware training to improve the models' performance in generating accurate and aesthetically pleasing visual text images.

http://arxiv.org/abs/2410.04439

460. mDPO: Conditional Preference Optimization for Multimodal Large Language Models

This paper presents mDPO, a conditional preference optimization approach for improving multimodal large language model alignment by addressing the unconditional preference problem. Through various experiments, mDPO is shown to effectively enhance performance by optimizing both language and image preferences in multimodal settings.

https://arxiv.org/abs/2406.11839

470. Pelican: Correcting Hallucination in Vision-LLMs via Claim Decomposition and Program of Thought Verification

Pelican introduces a novel framework to correct hallucinations in Large Visual Language Models by decomposing visual claims into sub-claims for detailed verification. The framework improves hallucination detection and mitigation through adaptive reasoning and is shown to decrease hallucination rates significantly across multiple benchmarks.

https://arxiv.org/abs/2407.02352

518. SignCLIP: Connecting Text and Sign Language by Contrastive Learning

SignCLIP is a novel approach that connects spoken language and sign language using contrastive learning techniques to project them into a unified space. It efficiently utilizes large-scale video-text pairs for sign language processing and demonstrates strong performance on various tasks, while providing insights into the linguistic structure of sign languages.

https://arxiv.org/abs/2407.01264

556. Efficient Temporal Extrapolation of Multimodal Large Language Models with Temporal Grounding Bridge

This paper presents the Temporal Grounding Bridge (TGB), a framework designed to enhance the temporal grounding capabilities of multimodal large language models (MLLMs) for interpreting long-form videos in response to queries. The authors demonstrate that TGB significantly improves performance across seven video benchmarks by introducing efficient multi-span temporal grounding algorithms and extending context window sizes without requiring additional annotations.

https://arxiv.org/abs/2402.16050

558. MMOE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts

This paper introduces the Multimodal Mixtures of Experts (MMoE) approach to enhance multimodal models by training separate expert models for different types of multimodal interactions. The method demonstrates improved performance on sarcasm and humor detection tasks, suggesting its applicability to a variety of models for better handling of real-world multimodal interactions.

https://arxiv.org/abs/2311.09580

735. OpenSep: Leveraging Large Language Models with Textual Inversion for Open World Audio Separation

OpenSep is a novel framework for automated audio separation that leverages large language models to overcome limitations of existing models in real-world audio scenarios. By using textual inversion and few-shot prompting, OpenSep effectively separates variable sources in unseen mixtures, demonstrating superior performance compared to state-of-the-art methods.

https://arxiv.org/abs/2409.19270

797. Vision-Language Model Fine-Tuning via Simple Parameter-Efficient Modification

This paper proposes ClipFit, a fine-tuning method for Vision-Language Models (VLMs) that optimally adjusts only specific parameters rather than the entire model to enhance performance. Experiments show that this method can improve zero-shot accuracy significantly without the need for extra parameters, while also providing insights into the changes in internal parameters during fine-tuning.

https://arxiv.org/abs/2409.16718

821. VHASR: A Multimodal Speech Recognition System With Vision Hotwords

This paper presents VHASR, a multimodal speech recognition system that enhances speech recognition performance by incorporating audio-related images. The proposed dual-stream architecture effectively utilizes visual hotwords and shows improved results over unimodal automatic speech recognition systems, achieving state-of-the-art performance in the multimodal ASR domain.

https://arxiv.org/abs/2410.00822

867. DocEditAgent: Document Structure Editing Via Multimodal LLM Grounding

DocEdit-v2 introduces a framework for end-to-end editing of document structures by grounding user requests onto textual and visual components of document images. Its three components allow for more accurate localization of edits and reformulation of commands, achieving significant performance improvements in editing tasks.

http://arxiv.org/abs/2410.16472v1

947. Reasoning Paths with Reference Objects Elicit Quantitative Spatial Reasoning in Large Vision-Language Models

This paper investigates the limitations of vision-language models (VLMs) in performing quantitative spatial reasoning with respect to object sizes and distances. It introduces a new benchmark and a prompting technique, SpatialPrompt, which significantly improves the reasoning capabilities of top-performing VLMs without additional data or model fine-tuning.

http://arxiv.org/abs/2409.09788v1

1030. PALM: Few-Shot Prompt Learning for Audio Language Models

This paper presents PALM, a novel method for few-shot prompt learning in Audio Language Models (ALMs), enhancing training efficiency by optimizing the feature space of a text encoder. The method demonstrates superior performance on multiple audio recognition datasets compared to existing baselines.

https://arxiv.org/abs/2409.19806

1072. Multimodal Self-Instruct: Synthetic Abstract Image and Visual Reasoning Instruction Using Language Model

This paper introduces a multi-modal self-instruct approach that leverages large language models to generate synthetic abstract images and visual reasoning instructions. It highlights the performance issues of current LMMs in understanding abstract images and spatial reasoning, presenting a new multimodal benchmark and showing improvements in tasks like chart understanding and map navigation after fine-tuning.

https://arxiv.org/abs/2407.07053

1092. IntCoOp: Interpretability-Aware Vision-Language Prompt Tuning

IntCoOp addresses the challenges of manual prompt engineering in image-text contrastive models by introducing an interpretable prompt-tuning method that incorporates compositional attributes. This approach improves performance across various tasks by aligning attribute-level biases with class embeddings, demonstrating superior results compared to existing frameworks.

https://arxiv.org/abs/2406.13683

1117. Whiteboard-of-Thought: Thinking Step-by-Step Across Modalities

This paper introduces a novel prompting technique called whiteboard-of-thought, enabling multimodal large language models to engage in visual reasoning by drawing and processing image-based reasoning steps. The method demonstrates significant improvements in accuracy for tasks involving visual and spatial reasoning compared to traditional chain-of-thought approaches.

https://arxiv.org/abs/2406.14562

1137. Eliciting In-Context Learning in Vision-Language Models for Videos Through Curated Data Distributional Properties

This paper presents a novel training paradigm called Emergent In-context Learning on Videos (eilev) that enhances vision-language models' (VLMs) in-context learning capabilities for analyzing and narrating videos. Through experiments, it demonstrates that eilev-trained models significantly outperform standard VLMs in few-shot video narration tasks, providing insights into the properties that enable effective in-context learning in this domain.

https://arxiv.org/abs/2311.17041

1153. IFCap: Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning

This paper presents IFCap, a novel approach to zero-shot captioning that bridges the modality gap by aligning text features with visually relevant features. It enhances caption quality through a Fusion Module and a Frequency-based Entity Filtering technique, outperforming existing state-of-the-art methods.

https://arxiv.org/abs/2409.18046

1158. Bayesian Example Selection Improves In-Context Learning for Speech, Text, and Visual Modalities

This paper introduces a novel Bayesian In-Context example Selection method (ByCS) aimed at improving in-context learning (ICL) for large language models across various modalities (speech, text, and images). The method leverages inverse inference based on Bayes' theorem to enhance the selection of in-context examples, demonstrating effectiveness in cross-tasking and cross-modality experiments.

https://arxiv.org/abs/2404.14716

1186. Retrieval-enriched zero-shot image classification in low-resource domains

This paper presents a novel method called CoRE, which enhances the performance of zero-shot low-resource image classification by using a retrieval-based strategy to incorporate relevant textual information. The proposed approach significantly improves classification outcomes in underrepresented domains without relying on synthetic data generation or model fine-tuning.

https://arxiv.org/abs/2411.00988

1191. Show and Guide: Instructional-Plan Grounded Vision and Language Model

This paper presents MM-PlanLLM, a multimodal language model designed to assist users with complex instructional tasks by integrating textual plans with visual data. The system employs a novel multitask-multistage training method to enhance its performance in multimodal interactions, enabling users to retrieve relevant video segments and generate step instructions based on their current progress.

https://arxiv.org/abs/2409.19074

1228. Holistic Evaluation for Interleaved Text-and-Image Generation

This paper introduces InterleavedBench, the first benchmark designed for evaluating interleaved text-and-image generation, addressing current gaps in evaluation methodologies. Additionally, it presents InterleavedEval, a reference-free metric that provides a comprehensive evaluation framework and shows strong correlations with human judgments.

https://arxiv.org/abs/2406.14643

133. By My Eyes: Grounding Multimodal Large Language Models with Sensor Data via Visual Prompting

This paper introduces a visual prompting approach for grounding multimodal large language models (MLLMs) with sensor data, significantly improving performance on sensory tasks. The method outperforms traditional text-based prompts by enhancing accuracy and reducing token costs through automated visualization generation for various sensory tasks.

https://arxiv.org/abs/2407.10385

143. Selective Vision is the Challenge for Visual Reasoning: A Benchmark for Visual Argument Understanding

This paper presents VisArgs, a dataset designed to benchmark and evaluate AI's ability to understand visual arguments, which are essential for persuasive communication in contexts like advertising. Experiments reveal significant performance gaps between human understanding of visual premises and that of current AI models, highlighting the challenges in visual reasoning tasks.

https://arxiv.org/abs/2406.18925

257. Visual Prompting in LLMs for Enhancing Emotion Recognition

The paper introduces a novel Set-of-Vision (SoV) approach for enhancing emotion recognition in Vision Large Language Models (VLLMs) by effectively utilizing spatial information from images. Experimental results show that SoV significantly improves the accuracy of face count and emotion categorization while maintaining the overall image context.

https://arxiv.org/abs/2410.02244

287. MaPPER: Multimodal Prior-guided Parameter Efficient Tuning for Referring Expression Comprehension

The paper presents MaPPER, a novel framework for Referring Expression Comprehension that uses Multimodal Prior-guided Parameter Efficient Tuning to enhance visual and language alignment without the computational burden of full model fine-tuning. It achieves superior accuracy with minimal tunable parameters by employing dynamic and local convolution adapters tailored for multimodal tasks.

https://arxiv.org/abs/2409.13609

351. Distilling Knowledge from Text-to-Image Generative Models Improves Visio-Linguistic Reasoning in CLIP

This work presents SDS-CLIP, a method that enhances the visio-linguistic reasoning capabilities of CLIP models by employing knowledge distillation techniques from text-to-image generative models. The proposed method demonstrates significant improvements in compositional reasoning tasks, indicating the effectiveness of distillation objectives in training image-text models.

https://arxiv.org/abs/2307.09233

385. From Local Concepts to Universals: Evaluating the Multicultural Understanding of Vision-Language Models

This paper addresses the suboptimal performance of vision-language models on non-western cultures due to inadequate representation in training datasets. It introduces the GlobalRG benchmark to evaluate multicultural understanding by conducting retrieval tasks for universal and culture-specific concepts across images from various countries.

https://arxiv.org/abs/2407.00263

439. DAMRO: Dive into the Attention Mechanism of LVLM to Reduce Object Hallucination

This paper presents DAMRO, a novel strategy aimed at reducing object hallucination in Large Vision-Language Models (LVLMs) by improving the attention mechanism in visual encoders and language model decoders. By filtering out high-attention background tokens, DAMRO significantly mitigates the issue of hallucination across different LVLM benchmarks.

http://arxiv.org/abs/2410.04514

559. OmAgent: A Multi-modal Agent Framework for Complex Video Understanding with Task Divide-and-Conquer

OmAgent is a framework designed for efficient multi-modal understanding of complex video content, addressing the challenges of extensive video processing that often leads to information loss. It utilizes a Divide-and-Conquer Loop and a robust tool-calling system to enhance query accuracy and autonomy while managing diverse video types and tasks.

https://arxiv.org/abs/2406.16620

793. Shaking Up VLMs: Comparing Transformers and Structured State Space Models for Vision & Language Modeling

This study compares Transformers and a structured state space model, Mamba, in Visual Language Models (VLMs) across various tasks such as captioning and question answering. Findings indicate Mamba performs well in tasks reliant on image summarization but struggles with explicit context retrieval, where Transformers excel.

http://arxiv.org/abs/2409.05395v2

864. FineCops-Ref: A new Dataset and Task for Fine-Grained Compositional Referring Expression Comprehension

This paper introduces FineCops-Ref, a new dataset designed to improve Fine-Grained Compositional Referring Expression Comprehension (REC) by incorporating varying difficulty levels and negative examples. The findings suggest existing models exhibit significant gaps in grounding performance, indicating a need for enhanced strategies in visual reasoning and cross-modal interactions.

https://arxiv.org/abs/2409.14750

944. On Efficient Language and Vision Assistants for Visually-Situated Natural Language Understanding: What Matters in Reading and Reasoning

This paper addresses the challenges of designing efficient language and vision models for visually-situated natural language understanding while maintaining performance and transparency. Through model optimization and strategic dataset formulation, significant improvements in inference throughput are achieved, with the intention to open-source the underlying code and datasets.

https://arxiv.org/abs/2406.11823

1059. TV-TREES: Multimodal Entailment Trees for Neuro-Symbolic Video Reasoning

TV-TREES introduces a novel multimodal entailment tree generator aimed at enhancing video understanding through interpretable joint-modality reasoning. The approach demonstrates state-of-the-art performance on the TVQA benchmark while promoting the evaluation of reasoning quality in multimodal contexts.

https://arxiv.org/abs/2402.19467

1125. Updating CLIP to Prefer Descriptions Over Captions

This paper proposes an update to the CLIP model in order to distinguish between image descriptions and captions, especially focusing on accessibility. The modified model uses parameter-efficient fine-tuning and a causal interpretability-inspired loss objective to improve the performance of describing images in a way that aligns better with the judgments of disabled users while maintaining general transfer capabilities.

https://arxiv.org/abs/2406.09458

1263. Mitigating Open-Vocabulary Caption Hallucinations

This paper addresses the issue of hallucinations in image captioning by proposing a framework that evaluates open-vocabulary object hallucinations using a new benchmark called OpenCHAIR. The proposed approach, MOCHa, effectively mitigates hallucinations without relying on closed vocabulary and improves caption generation by balancing fidelity and adequacy.

http://arxiv.org/abs/2312.03631v4

1062. Preserving Multi-Modal Capabilities of Pre-trained VLMs for Improving Vision-Linguistic Compositionality

This paper introduces a method, Fine-grained Selective Calibrated CLIP (FSC-CLIP), to enhance compositional understanding in pre-trained vision and language models while preserving their multi-modal capabilities. The approach effectively integrates local hard negative loss and selective calibrated regularization, achieving strong performance in both compositionality tasks and zero-shot multi-modal tasks.

http://arxiv.org/abs/2410.05210v1

44. GeoGPT4V: Towards Geometric Multi-modal Large Language Models with Geometric Image Generation

This paper presents GeoGPT4V, a novel pipeline that enhances the geometric capabilities of multi-modal large language models by generating a dataset of geometry problems with aligned text and images. The results show that this dataset significantly boosts the performance of models on geometric tasks, addressing the challenges of existing datasets that lack quality and alignment.

https://arxiv.org/abs/2406.11503

67. EFUF: Efficient Fine-Grained Unlearning Framework for Mitigating Hallucinations in Multimodal Large Language Models

This paper introduces the Efficient Fine-grained Unlearning Framework (EFUF) aimed at mitigating hallucinations in multimodal large language models (MLLMs) without the need for paired data. The method enhances alignment between images and text while preserving generation quality and reducing computational resources compared to traditional approaches.

https://arxiv.org/abs/2402.09801

105. HELPD: Mitigating Hallucination of LVLMs by Hierarchical Feedback Learning with Vision-enhanced Penalty Decoding

This paper introduces HELPD, a framework designed to mitigate hallucinations in Large Vision-Language Models (LVLMs) through a hierarchical feedback approach that considers both object and sentence semantics. The proposed method demonstrates significant improvements in reducing hallucinations and enhancing text generation quality across various benchmarks.

https://arxiv.org/abs/2409.20429

137. VIVA: A Benchmark for Vision-Grounded Decision-Making with Human Values

This paper introduces VIVA, a benchmark designed for vision-grounded decision-making that incorporates human values. It demonstrates the challenges large vision language models face in leveraging these values to make appropriate actions in diverse real-world scenarios.

https://arxiv.org/abs/2407.03000

144. Can visual language models resolve textual ambiguity with visual cues? Let visual puns tell you!

The paper presents UNPIE, a benchmark to assess the ability of machines to resolve lexical ambiguities using visual cues, specifically through the lens of puns. The study demonstrates that multimodal models outperform text-only models when visual context is provided, especially for more complex tasks.

https://arxiv.org/abs/2410.01023

154. African or European Swallow? Benchmarking Large Vision-Language Models for Fine-Grained Object Classification

This paper introduces exttt{FOCI}, a benchmark for fine-grained object classification in Large Vision-Language Models (LVLMs), highlighting a gap in the evaluation of these models for distinguishing between closely related categories. The authors demonstrate that CLIP models outperform LVLMs significantly, indicating a mismatch in the alignment for fine-grained classification and emphasizing a need for improved training data with detailed annotations.

https://arxiv.org/abs/2406.14496

194. Pixology: Probing the Linguistic and Visual Knowledge of Pixel-based Language Models

This paper investigates the capabilities of pixel-based language models, specifically the PIXEL model, in comparison to traditional subword-based models like BERT. The study reveals a significant performance gap between visual and linguistic understanding in the PIXEL model, offering insights for future improvements in pixel-based language model development.

https://arxiv.org/abs/2410.12011

369. UOUO: Uncontextualized Uncommon Objects for Measuring Knowledge Horizons of Vision Language Models

This paper introduces the UOUO benchmark for assessing the performance of Vision-Language Models (VLMs) on uncommon objects, emphasizing their limitations compared to larger models. The findings underscore the importance of addressing long-tail distributions in evaluating VLM capabilities, especially for rare object handling.

https://arxiv.org/abs/2407.18391

533. Read Anywhere Pointed: Layout-aware GUI Screen Reading with Tree-of-Lens Grounding

This paper proposes a new approach for screen reading of GUIs through a model called Tree-of-Lens (ToL), which utilizes a hierarchical layout tree based on user-indicated points. The ToL agent effectively understands content and spatial relationships in interfaces, showing promise against existing tools and tasks in mobile GUI navigation.

https://arxiv.org/abs/2406.19263

609. Kiss up, Kick down: Exploring Behavioral Changes in Multi-modal Large Language Models with Assigned Visual Personas

This study examines how multi-modal large language models (LLMs) can modify their behaviors to align with visual personas, focusing on negotiation behaviors influenced by perceived aggressiveness in avatar images. Results show that LLMs respond to visual cues in a manner akin to human interpersonal interactions, exhibiting varied aggression levels based on their own persona compared to that of their opponents.

https://arxiv.org/abs/2410.03181

619. Towards Online Continuous Sign Language Recognition and Translation

This paper presents a novel approach to online continuous sign language recognition (CSLR) aimed at reducing latency and memory usage associated with offline recognition methods. The proposed method involves developing a sign dictionary and using a sliding window technique for online recognition, achieving state-of-the-art performance across multiple benchmarks.

https://arxiv.org/abs/2401.05336

790. Tools Fail: Detecting Silent Errors in Faulty Tools

This paper proposes a framework to detect silent errors in tools used by large language models (LLMs) and presents an approach for failure recovery. The findings highlight the importance of a model's ability to identify and manage tool-related errors, which is critical as models increasingly act as tools in various tasks.

https://arxiv.org/abs/2406.19228

833. ActPlan-1K: Benchmarking the Procedural Planning Ability of Visual Language Models in Household Activities

This paper introduces ActPlan-1K, a benchmark designed to evaluate the procedural planning capabilities of visual language models (VLMs) in the context of household activities. It reveals that current VLMs struggle with generating human-level procedural plans, highlighting the need for enhanced reasoning in multi-modal and counterfactual planning scenarios.

http://arxiv.org/abs/2410.03907v1

898. VideoCLIP-XL: Advancing Long Description Understanding for Video CLIP Models

The paper introduces VideoCLIP-XL, a model designed to improve the understanding of long descriptions in video content, addressing the limitations of the existing CLIP models which focus on brief texts. It outlines the development of a large-scale dataset and new evaluation benchmarks to test the model's effectiveness in processing long-form video descriptions.

https://arxiv.org/abs/2410.00741

973. From LLMs to MLLMs: Exploring the Landscape of Multimodal Jailbreaking

This paper explores the vulnerabilities of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) to adversarial attacks, specifically focusing on jailbreaking techniques. It aims to provide an overview of current research in this area and identify future research directions to improve the robustness and security of MLLMs.

https://arxiv.org/abs/2406.14859

1016. Investigating and Mitigating Object Hallucinations in Pretrained Vision-Language (CLIP) Models

This paper investigates the issue of object hallucinations in CLIP models, highlighting that the problem persists even when the model functions in isolation. A novel method of counterfactual data augmentation is proposed to mitigate these hallucinations, improving the model's reliability as a visual encoder.

https://arxiv.org/abs/2410.03176

1075. Altogether: Image Captioning via Re-aligning Alt-text

This paper proposes a method called Altogether, which focuses on enhancing image captioning by re-aligning existing alt-text metadata to better reflect image content. The approach involves a systematic human annotation process to generate more informative captions, ultimately improving various visual recognition tasks.

https://arxiv.org/abs/2410.17251

1080. Investigating the Role of Instruction Variety and Task Difficulty in Robotic Manipulation Tasks

This paper introduces a framework to evaluate the generalization capabilities of multimodal models in robotic manipulation tasks, focusing on instruction variety and task difficulty. The findings indicate that current Transformer-based models exhibit weaknesses in robustness due to overfitting and the need for better integration of multimodal inputs.

https://arxiv.org/abs/2407.03967

1151. Visual Text Matters: Improving Text-KVQA with Visual Text Entity Knowledge-aware Large Multimodal Assistant

This paper presents VisTEL, a module that enhances visual text entity linking for knowledge-aware visual question answering (Text-KVQA) by leveraging advanced multimodal models and visual text recognition. It also introduces KaLMA, a large multimodal assistant that integrates knowledge with visual text entities to improve answer accuracy, achieving state-of-the-art results in the field.

https://arxiv.org/abs/2410.19144

1203. Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories

This study investigates how the visual representations of concrete and abstract concepts vary and how this variability can affect the performance of vision and multimodal models. Findings reveal that basic visual features outperform complex model features for classifying abstract versus concrete images, but that more complex models, like Vision Transformer, are better suited for nearest neighbor analysis.

http://arxiv.org/abs/2410.11657

1250. MIBench: Evaluating Multimodal Large Language Models over Multiple Images

This paper introduces MIBench, a benchmark designed to evaluate the performance of multimodal large language models (MLLMs) on multi-image scenarios. It categorizes multi-image tasks into three scenarios and reveals that current MLLMs perform well on single images but struggle with multi-image reasoning and complex learning tasks.

https://arxiv.org/abs/2407.15272

1257. Nearest Neighbor Normalization Improves Multimodal Retrieval

This paper introduces Nearest Neighbor Normalization (NNN), a technique that enhances the performance of contrastive models in multimodal image-text retrieval without requiring additional training. Through empirical tests across various models and datasets, NNN demonstrates significant improvements in retrieval metrics for both text and images.

https://arxiv.org/abs/2410.24114

106. TopViewRS: Vision-Language Models as Top-View Spatial Reasoners

This paper investigates the spatial reasoning capabilities of Vision-Language Models (VLMs) from a top-view perspective, crucial for human and agent navigation. It introduces the TopViewRS dataset to evaluate VLMs across various perception and reasoning tasks, revealing significant performance gaps compared to human capabilities and setting the stage for future improvements.

https://arxiv.org/abs/2406.02537

159. Does Object Grounding Really Reduce Hallucination of Large Vision-Language Models?

This paper investigates the impact of object grounding on the hallucination problem in large vision-language models (LVLMs), challenging previous claims that grounding reduces hallucinations. The authors conduct a comprehensive evaluation and conclude that grounding objectives have minimal influence on hallucination during open caption generation.

https://arxiv.org/abs/2406.14492

335. Quantifying the Gap Between Machine Translation and Native Language in Training for Multimodal, Multilingual Retrieval

The paper investigates the performance gaps in multilingual vision-language models, particularly focusing on how native perception affects the translation of image captions. It proposes and evaluates caption augmentation strategies to improve model performance, although some gaps still persist, highlighting ongoing challenges in this research area.

https://arxiv.org/abs/2410.02027

387. MMNeuron: Discovering Neuron-Level Domain-Specific Interpretation in Multimodal Large Language Model

This paper investigates domain-specific neurons in multimodal large language models (MLLMs), focusing on how these models process visual features alongside textual data. A proposed three-stage mechanism reveals that current MLLMs may not fully utilize domain-specific information, and improving this could enhance model accuracy significantly.

https://arxiv.org/abs/2406.11193

573. An image speaks a thousand words, but can everyone listen? On image transcreation for cultural relevance

This paper addresses the challenge of translating images for cultural relevance, focusing on building pipelines with generative models to culturally adapt images. The evaluation shows that current image editing models struggle significantly, with improvements possible by integrating large language models.

https://arxiv.org/abs/2404.01247

75. Pre-trained Language Models Do Not Help Auto-regressive Text-to-Image Generation

This paper investigates the effectiveness of pre-trained language models in auto-regressive text-to-image generation and finds that they offer limited assistance due to semantic differences between image and text tokens. Additionally, the simplicity of text tokens in image-text datasets degrades the capabilities of language models further.

https://arxiv.org/abs/2311.16201

213. VGBench: A Comprehensive Benchmark of Vector Graphics Understanding and Generation for Large Language Models

613. Large Language Models Know What is Key Visual Entity: An LLM-assisted Multimodal Retrieval for VQA

722. UNICORN: A Unified Causal Video-Oriented Language-Modeling Framework for Temporal Video-Language Tasks

996. In-Context Compositional Generalization for Large Vision-Language Models

1061. GRIZAL: Generative Prior-guided Zero-Shot Temporal Action Localization

1195. An Empirical Analysis on Spatial Reasoning Capabilities of Large Multimodal Models

112. TinyChart: Efficient Chart Understanding with Program-of-Thoughts Learning and Visual Token Merging

959. MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification

1086. Video-Text Prompting for Weakly Supervised Spatio-Temporal Video Grounding

1119. Self-Training Large Language and Vision Assistant for Medical

1133. Unveiling Multi-level and Multi-modal Semantic Representations in the Human Brain using Large Language Models

120. Tag-grounded Visual Instruction Tuning with Retrieval Augmentation

325. From Coarse to Fine: Impacts of Feature-Preserving and Feature-Compressing Connectors on Perception in Multimodal Models

801. ASL STEMpedia: Dataset and Benchmark for Interpreting STEM Articles

893. Generating Demonstrations for In-Context Compositional Generalization in Grounded Language Learning

922. Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering

972. Dual-oriented Disentangled Network with Counterfactual Intervention for Multimodal Intent Detection

998. Game on Tree: Visual Hallucination Mitigation via Coarse-to-Fine View Tree and Game Theory

1104. Unsupervised Discrete Representations of American Sign Language

1128. VIEWS: Entity-Aware News Video Captioning

1187. I-AM-G: Interest Augmented Multimodal Generator for Item Personalization

1243. M3D: MultiModal MultiDocument Fine-Grained Inconsistency Detection

1256. Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework

454. Efficient Vision-Language pre-training via domain-specific learning for human activities

488. MEANT: Multimodal Encoder for Antecedent Information

Computational Social Science and Cultural Analytics

337. Fine-Grained Detection of Solidarity for Women and Migrants in 155 Years of German Parliamentary Debates

This paper investigates the nuances of solidarity towards women and migrants in German parliamentary debates over 155 years, leveraging large language models for fine-grained analysis. It finds that while solidarity with migrants has historically outweighed anti-solidarity, the forms of solidarity have evolved, particularly towards more compassionate frameworks, demonstrating the potential and effectiveness of LLMs in social science research.

https://arxiv.org/abs/2210.04359

416. MAgIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration

This paper proposes a benchmarking framework to evaluate large language models in multi-agent environments, focusing on their cognitive abilities such as reasoning, collaboration, and adaptability. Through various games and scenarios, the study quantitatively assesses the performance of different models, revealing substantial capability gaps and noting improvements when enhancing models with probabilistic graphical methods.

https://arxiv.org/abs/2311.08562

532. Susu Box or Piggy Bank: Assessing Cultural Commonsense Knowledge between Ghana and the US

724. OATH-Frames: Characterizing Online Attitudes Towards Homelessness with LLM Assistants

This paper investigates public attitudes towards homelessness in the U.S. using large language models to analyze millions of Twitter posts, introducing a new typology called OATH-Frames to categorize critiques and responses. It showcases the efficiency of LLM-assisted annotation, revealing trends in societal perceptions of homelessness over time and across different populations, which can apply to other social issues as well.

https://arxiv.org/abs/2406.14883

945. Community-Cross-Instruct: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities

Community-Cross-Instruct introduces an unsupervised framework for generating instructions to align large language models with online communities, improving efficiency in representing these populations. The method allows for the automated and cost-effective surveying of diverse online communities by creating digital twins that reflect their language and attitudes.

https://arxiv.org/abs/2406.12074

1110. Virtual Personas for Language Models via an Anthology of Backstories

This paper introduces a method called 'Anthology' for conditioning large language models (LLMs) to specific virtual personas using open-ended life narratives. The approach improves the consistency and reliability of model responses in behavioral studies, showing significant enhancements in matching human response distributions.

https://arxiv.org/abs/2407.06576

59. HEART-felt Narratives: Tracing Empathy and Narrative Style in Personal Stories with LLMs

This paper explores the relationship between narrative style and empathy in personal stories through the use of LLMs and a newly introduced taxonomy called HEART. By analyzing narrative elements and collecting empathy judgments from a large crowd, it aims to provide insights into how narrative storytelling can evoke empathy.

https://arxiv.org/abs/2405.17633

329. Benchmarking Vision Language Models for Cultural Understanding

This paper introduces CulturalVQA, a visual question-answering benchmark that evaluates Vision Language Models (VLMs) on their cultural understanding through a curated collection of image-question pairs. The benchmarking reveals disparities in cultural comprehension across different regions and cultural facets, highlighting areas where VLMs need improvement.

https://arxiv.org/abs/2407.10920

384. Locating Information Gaps and Narrative Inconsistencies Across Languages: A Case Study of LGBT People Portrayals on Wikipedia

The paper presents the InfoGap method, which identifies information gaps and discrepancies in Wikipedia articles about LGBT individuals across different languages. It demonstrates large biases in factual coverage, revealing a tendency for negative portrayals to be emphasized in Russian Wikipedia, thus offering a new framework for comparative text analysis across languages.

https://arxiv.org/abs/2410.04282

910. WorryWords: Norms of Anxiety Association for 44,450 English Words

This paper presents WorryWords, a large-scale repository of word-anxiety associations derived from over 44,450 English words, showing high reliability in these associations. It highlights the application of WorryWords in studying anxiety relationships and tracking changes in text, while also being accessible for various research domains like psychology and NLP.

https://arxiv.org/html/2411.03966v1

7. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay

This paper investigates the social behaviors of LLM-based agents within the game Avalon, focusing on their collaboration and confrontation dynamics. A novel multi-agent framework is proposed, and its performance is evaluated based on game success and social interactions which suggest potential applications in dynamic social environments.

https://arxiv.org/abs/2310.14985

158. An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making

This paper presents GEDI, an electoral decision-making module for large language models that enhances diversity and reasoning capabilities through various voting mechanisms. The empirical analysis demonstrates that this approach improves multi-agent collaboration without complex system designs, showing robustness and effectiveness across benchmarks.

https://arxiv.org/abs/2410.15168

306. How Far Can We Extract Diverse Perspectives from Large Language Models?

This paper investigates the ability of large language models (LLMs) to generate a diverse range of human opinions on subjective topics, utilizing a criteria-based prompting technique. The findings reveal that LLMs can effectively produce diverse perspectives comparable to human output, particularly in contexts with varying levels of subjectivity.

https://arxiv.org/abs/2311.09799

327. The Computational Anatomy of Humility: Modeling Intellectual Humility in Online Public Discourse

This study explores the computational modeling of intellectual humility (IH) in online public discourse, specifically focusing on measuring this virtue at scale using LLM-based models. Results highlight the challenges and provide a foundation for future NLP research aimed at improving online interactions by fostering IH.

https://arxiv.org/abs/2410.15182

694. Enhancing Data Quality through Simple De-duplication: Navigating Responsible Computational Social Science Research

This paper analyzes 20 datasets used in natural language processing for Computational Social Science, identifying issues related to data duplication that lead to inconsistencies and compromises in model reliability. The authors propose new protocols and best practices to enhance dataset quality from social media sources.

https://arxiv.org/abs/2410.03545

846. Decoding Susceptibility: Modeling Misbelief to Misinformation Through a Computational Approach

This paper proposes a computational approach to model individuals' susceptibility to misinformation, addressing the limitations of self-reported beliefs. By incorporating various demographic and psychological factors, the study demonstrates significant alignment between the model's susceptibility scores and human judgments, revealing important relationships with misinformation sharing behavior on social media during COVID-19.

https://arxiv.org/abs/2311.09630

989. Unleashing the Power of Emojis in Texts via Self-supervised Graph Pre-Training

This paper presents a method to incorporate emojis in text analysis through a self-supervised graph pre-training framework, enhancing the semantic understanding of emojis and their interaction with words in social media texts. The proposed approach, validated on Xiaohongshu and Twitter datasets, outperforms previous methods in downstream tasks related to text and emoji co-modeling.

https://arxiv.org/abs/2409.14552

1063. FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture

FoodieQA is a new multimodal dataset designed to enhance the understanding of Chinese food culture by integrating both images and text. The study evaluates various models on tasks related to question answering using this dataset, revealing performance gaps particularly in vision-language models compared to large language models.

https://arxiv.org/abs/2406.11030

1069. Emotion Granularity from Text: An Aggregate-Level Indicator of Mental Health

This paper introduces computational measures of emotion granularity derived from social media text, providing an alternative to self-reports for understanding emotional differentiation among individuals. The findings suggest that lower levels of emotion granularity can be indicative of various mental health conditions, potentially aiding in the assessment of mental health based on textual analysis.

https://arxiv.org/abs/2403.02281

151. Surveying the Dead Minds: Historical-Psychological Text Analysis with Contextualized Construct Representation (CCR) for Classical Chinese

This paper introduces a pipeline for analyzing historical psychology in classical Chinese texts using Contextualized Construct Representations (CCR), which combines expert psychological knowledge with advanced NLP techniques. It presents a fine-tuning approach leveraging an innovative historical psychology corpus, demonstrating superior performance in measuring psychological constructs compared to existing methods.

https://arxiv.org/abs/2403.00509

354. On the Reliability of Psychological Scales on Large Language Models

This study investigates the reliability of psychological scales when applied to Large Language Models (LLMs), focusing on their capacity to exhibit consistent personality traits. The findings suggest that LLMs can reliably emulate different personalities using specific prompt instructions, which could have implications for social sciences research.

https://arxiv.org/abs/2305.19926

931. Still Not Quite There! Assessing Large Language Models for Comorbid Mental Health Diagnosis

This paper presents ANGST, a benchmark specifically designed for classifying depression-anxiety comorbidity from social media posts, allowing for multi-label classification. The study evaluates several language models, revealing that none reach high performance in this complex diagnostic task, with GPT-4 being the most capable yet still below 72% F1 score.

https://arxiv.org/abs/2410.03908

1184. Computational Meme Understanding: A Survey

196. Noise, Novels, Numbers. A Framework for Detecting and Categorizing Noise in Danish and Norwegian Literature

256. Deciphering Rumors: A Multi-Task Learning Approach with Intent-aware Hierarchical Contrastive Learning

661. Towards a Greek Proverb Atlas: Computational Spatial Exploration and Attribution of Greek Proverbs

827. A Closer Look at Multidimensional Online Political Incivility

869. Understanding Slang with LLMs: Modelling Cross-Cultural Nuances through Paraphrasing

872. Adaptive Axes: A Pipeline for In-domain Social Stereotype Analysis

1113. The Empirical Variability of Narrative Perceptions of Social Media Texts

1122. Detecting Online Community Practices with Large Language Models: A Case Study of Pro-Ukrainian Publics on Twitter

1226. Style-Shifting Behaviour of the Manosphere on Reddit

Speech processing and spoken language understanding

366. EH-MAM: Easy-to-Hard Masked Acoustic Modeling for Self-Supervised Speech Representation Learning

This paper introduces EH-MAM, a self-supervised learning method for speech representation that uses an adaptive masking strategy to progressively target harder regions during training. The method outperforms existing models on low-resource speech recognition tasks, demonstrating improved representation learning through selective masking.

https://arxiv.org/abs/2410.13179

430. Muting Whisper: A Universal Acoustic Adversarial Attack on Speech Foundation Models

This paper presents a novel acoustic adversarial attack on the Whisper speech recognition model, demonstrating how special tokens in its vocabulary can be exploited to manipulate its behavior. The authors propose a universal adversarial audio segment that effectively mutes the model's responses for over 97% of speech samples, clouding the implications for speech moderation and privacy protection.

https://arxiv.org/abs/2405.06134

466. EmoKnob: Enhance Voice Cloning with Fine-Grained Emotion Control

EmoKnob is a framework that enhances text-to-speech synthesis by allowing users to exert fine-grained control over the emotions and their intensity in generated speech. The framework demonstrates significant improvements in emotion expressiveness compared to existing commercial TTS services and introduces new evaluation metrics for assessing emotion control in speech synthesis.

https://arxiv.org/abs/2410.00316

562. ESC: Efficient Speech Coding with Cross-Scale Residual Vector Quantized Transformers

The Efficient Speech Codec (ESC) is introduced as a lightweight and parameter-efficient neural speech codec that outperforms traditional convolutional audio codecs through the use of cross-scale residual vector quantization and transformers. ESC achieves high-fidelity speech reconstruction with lower model complexity, thereby improving bitrate efficiency without compromising audio quality.

https://arxiv.org/abs/2404.19441

595. Speechworthy Instruction-tuned Language Models

This paper presents methods to align instruction-tuned language models with the speech domain using prompting strategies and a new speech-based preference learning dataset. The findings indicate that combining these methods enhances the suitability of responses generated by the models for speech applications.

https://arxiv.org/abs/2409.14672

1116. Continual Test-time Adaptation for End-to-end Speech Recognition on Noisy Speech

This paper introduces a Fast-slow Test-Time Adaptation (TTA) framework for end-to-end Automatic Speech Recognition (ASR), focusing on continual adaptation for improving performance on noisy speech data. The proposed Dynamic SUTA method employs an entropy-minimization approach and a dynamic reset strategy for better robustness against domain shifts, outperforming previous non-continual and continual TTA techniques on various datasets.

http://arxiv.org/abs/2406.11064v2

405. Advancing Test-Time Adaptation in Wild Acoustic Test Settings

This paper addresses the performance issues of Automatic Speech Recognition (ASR) systems in challenging wild acoustic environments, proposing a method called Confidence-Enhanced Adaptation for effective online Test-Time Adaptation (TTA). The proposed approach incorporates a confidence-aware weight scheme and consistency regularization to enhance ASR performance in various real-world acoustic conditions.

https://arxiv.org/html/2310.09505v2

9. Speaking in Wavelet Domain: A Simple and Efficient Approach to Speed up Speech Diffusion Model

This paper presents a novel method to enhance the training and inference speed of Denoising Diffusion Probabilistic Models (DDPMs) for speech synthesis by utilizing the wavelet domain. The approach not only accelerates the performance of speech synthesis but also shows versatility in speech enhancement tasks.

http://arxiv.org/abs/2402.10642v2

21. Scaling Properties of Speech Language Models

This paper investigates the scaling properties of Speech Language Models (SLMs), particularly their syntax and semantic capabilities as influenced by training compute. It finds that while SLMs improve with scale, they do so at a much slower rate compared to text-based Large Language Models (LLMs), and suggests the use of synthetic data to enhance semantic understanding.

https://arxiv.org/abs/2404.00685

30. EmphAssess : a Prosodic Benchmark on Assessing Emphasis Transfer in Speech-to-Speech Models

EmphAssess is a benchmark for assessing how well speech-to-speech models can encode and reproduce prosodic emphasis. It includes a new model called EmphaClass that classifies emphasis at the word or frame level across tasks such as speech resynthesis and translation.

https://arxiv.org/abs/2312.14069

224. AlignCap: Aligning Speech Emotion Captioning to Human Preferences

AlignCap introduces a new approach to Speech Emotion Captioning (SEC) that aligns its outputs with human preferences using large language models. The method addresses issues of hallucination and generalization in SEC, demonstrating improved performance in zero-shot tasks.

https://arxiv.org/abs/2410.19134

286. Cross-Domain Audio Deepfake Detection: Dataset and Analysis

This paper presents a new cross-domain dataset for audio deepfake detection, created to enhance the generalization of detection models against advanced text-to-speech systems. The authors demonstrate that their proposed methods achieve low error rates, indicating effective detection capabilities, although challenges remain with data compression affecting accuracy.

https://arxiv.org/abs/2404.04904

302. Improving Spoken Language Modeling with Phoneme Classification: A Simple Fine-tuning Approach

This paper presents a simple fine-tuning approach that improves spoken language modeling by utilizing phoneme classification. The authors demonstrate that this method achieves comparable lexical comprehension with significantly less data than traditional text-based models.

https://arxiv.org/abs/2410.00025

503. Task Arithmetic can Mitigate Synthetic-to-Real Gap in Automatic Speech Recognition

This paper presents a novel method called SYN2REAL task vector that uses task vector arithmetic to improve automatic speech recognition (ASR) performance by addressing the synthetic-to-real gap in ASR systems. The method shows a significant average improvement in word error rate when applied to the SLURP dataset, demonstrating its effectiveness in adapting ASR models to different text domains.

https://arxiv.org/abs/2406.02925

607. What is lost in Normalization? Exploring Pitfalls in Multilingual ASR Model Evaluations

This paper investigates the flaws in the text normalization routines used in multilingual automatic speech recognition (ASR) model evaluations, focusing on Indic language scripts. It highlights how these practices lead to skewed performance metrics and proposes improved normalization strategies that incorporate linguistic expertise.

https://arxiv.org/abs/2409.02449

690. Self-Powered LLM Modality Expansion for Large Speech-Text Models

This paper discusses the development of a self-powered large speech-text model (LSM) that integrates instruction-tuning with automatic speech recognition to effectively mitigate speech anchor bias. The research showcases improved performance in speech-based tasks by refining the training process, thus enhancing the fusion of speech and text modalities.

http://arxiv.org/abs/2410.03798

1149. Interventional Speech Noise Injection for ASR Generalizable Spoken Language Understanding

This paper proposes a novel augmentation method for enhancing spoken language understanding (SLU) models by injecting speech noise that is plausible for any ASR system, aiming to increase robustness against ASR errors. The experimental results suggest that this approach improves the generalizability of SLU models when exposed to diverse ASR environments.

https://arxiv.org/abs/2410.15609

771. Towards an Open-Source Speech Foundation Model for EU: 950,000 Hours of Open-Source Compliant Speech Data for EU Languages

858. AudioVSR: Enhancing Video Speech Recognition with Audio Data

Ethics, Bias, and Fairness

17. Studying and Mitigating Biases in Sign Language Understanding Models

This paper investigates biases in sign language understanding models using the ASL Citizen dataset, highlighting the importance of equitable technology access for the community. It applies bias mitigation techniques that successfully reduce performance disparities while maintaining accuracy, and shares demographic data to support future bias reduction efforts.

http://arxiv.org/abs/2410.05206v1

34. A Study of Nationality Bias in Names and Perplexity using Off-the-Shelf Affect-related Tweet Classifiers

This study analyzes biases in sentiment and emotion classification related to nationality as expressed in names, showing substantial influences on classifier predictions based on country names. The results suggest that these biases result from the training data of pre-trained language models, with significant implications for understanding language model behavior across different cultures.

http://arxiv.org/abs/2407.01834

41. FLIRT: Feedback Loop In-context Red Teaming

This paper presents an automatic red teaming framework called FLIRT that evaluates generative models for vulnerabilities, particularly in generating inappropriate content. By leveraging in-context learning and various attack strategies, the framework effectively exposes weaknesses in models like Stable Diffusion and improves the identification of toxic outputs in text generation.

https://arxiv.org/abs/2308.04265

72. Fairer Preferences Elicit Improved Human-Aligned Large Language Model Judgments

This paper investigates the preference biases in large language models (LLMs) when evaluating language generation quality, revealing their brittleness and skewed results. The authors propose a framework called ZEPO to optimize prompts for fairer preference decisions, which significantly improves the alignment of LLMs with human judgments without requiring labeled data.

https://arxiv.org/abs/2406.11370

195. GoldCoin: Grounding Large Language Models in Privacy Laws via Contextual Integrity Theory

GoldCoin is a novel framework that grounds Large Language Models in privacy laws using contextual integrity theory to enhance the understanding of privacy contexts. The framework creates synthetic scenarios based on relevant privacy statutes, improving LLMs' ability to identify privacy risks in judicial settings.

https://arxiv.org/abs/2406.11149

240. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration

The paper introduces Modular Pluralism, a framework that facilitates pluralistic alignment in large language models (LLMs) by integrating specialized community LMs to represent diverse preferences. Through extensive evaluation, the proposed approach enhances the responsiveness of LLMs to value-laden inputs and underrepresented communities, demonstrating improved pluralism objectives across various tasks and datasets.

https://arxiv.org/abs/2406.15951

243. STOP! Benchmarking Large Language Models with Sensitivity Testing on Offensive Progressions

The paper introduces the Sensitivity Testing on Offensive Progressions (STOP) dataset, aimed at evaluating biases in Large Language Models (LLMs) through a comprehensive and inclusive approach. Results show that existing models struggle with bias detection and that alignment with human judgments can significantly enhance model performance on sensitive tasks.

https://arxiv.org/abs/2409.13843

253. Order of Magnitude Speedups for LLM Membership Inference

This paper addresses the privacy vulnerabilities in Large Language Models (LLMs) caused by membership inference attacks (MIAs) and presents a method that allows for effective MIAs using low-cost ensemble models based on quantile regression. The proposed method achieves comparable accuracy to state-of-the-art techniques while drastically reducing computational costs, making privacy risk evaluations for LLMs more feasible.

https://arxiv.org/abs/2409.14513

425. Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning

This study examines the fairness implications of in-context learning (ICL) in large language models when processing tabular data. It proposes a method to enhance fairness by strategically selecting demonstrations, resulting in improved predictive performance without sacrificing accuracy.

https://arxiv.org/abs/2408.09757

427. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment

This paper investigates the vulnerability of Large Language Models (LLMs) used in zero-shot assessment scenarios, revealing that adversarial phrases can significantly inflate LLM scores. The study highlights the potential risks of deploying assessment LLMs in high-stakes environments due to their susceptibility to targeted manipulation.

https://arxiv.org/abs/2402.14016

471. Resampled Datasets Are Not Enough: Mitigating Societal Bias Beyond Single Attributes

This paper addresses societal bias in image-text datasets by removing spurious correlations and ensuring protected group independence from all attributes. The proposed method effectively reduces bias without affecting model performance in tasks such as multi-label image classification and captioning.

https://arxiv.org/abs/2407.03623

474. Humans or LLMs as the Judge? A Study on Judgement Bias

This paper explores the biases introduced when using human judges and large language models (LLMs) to evaluate LLM performance, highlighting various types of biases that affect both. It proposes a novel framework to study these biases and uses a curated dataset to demonstrate the vulnerabilities of human and LLM-based evaluations.

https://arxiv.org/abs/2402.10669

516. Words Matter: Reducing Stigma in Online Conversations about Substance Use with Large Language Models

This paper explores the stigma surrounding substance use disorders (SUD) on social media and how it affects treatment engagement. By analyzing Reddit posts and using large language models to transform stigmatizing language into empathetic expressions, the study offers a computational framework aimed at reducing stigma in online conversations about SUD.

https://arxiv.org/abs/2408.07873

546. Unlocking Memorization in Large Language Models with Dynamic Soft Prompting

This paper presents a novel method for estimating memorization in large language models using dynamic, prefix-dependent soft prompts that adapt to input changes. The proposed method significantly enhances the measurement of LLM memorization compared to previous techniques, achieving impressive relative improvements across various tasks.

https://arxiv.org/abs/2409.13853

612. Unlabeled Debiasing in Downstream Tasks via Class-wise Low Variance Regularization

This paper introduces a debiasing regularization technique that addresses the reintroduction of biases during the fine-tuning of pre-trained language models on downstream tasks. The proposed method does not require attribute labels and can target any attribute, significantly improving upon existing techniques that are limited to specific biases.

https://arxiv.org/abs/2409.19541

621. Split and Merge: Aligning Position Biases in LLM-based Evaluators

The paper introduces PORTIA, a system designed to align position biases in LLM-based evaluators, which typically show inconsistency when comparing answers. Through extensive testing, PORTIA significantly improves the consistency of various language models, enabling less advanced models to match high-performance standards while also addressing cost-efficiency issues.

https://arxiv.org/abs/2310.01432

656. LoRA-Guard: Parameter-Efficient Guardrail Adaptation for Content Moderation of Large Language Models

LoRA-Guard is a parameter-efficient adaptation method designed for content moderation of large language models on resource-constrained devices, such as mobile phones. It utilizes low-rank adapters to extract language features and improve moderation without degrading performance, achieving significantly lower parameter overhead than existing methods.

https://arxiv.org/abs/2407.02987

732. BEEAR: Embedding-based Adversarial Removal of Safety Backdoors in Instruction-tuned Language Models

The paper introduces BEEAR, a novel approach to mitigate safety backdoor attacks in instruction-tuned language models by adjusting model parameters based on embedding space perturbations. The methodology significantly decreases the success rate of these backdoor attacks while maintaining the utility of the model, marking a significant step forward in AI safety and security.

https://arxiv.org/abs/2406.17092

738. Large Language Models Are Involuntary Truth-Tellers: Exploiting Fallacy Failure for Jailbreak Attacks

This paper investigates how large language models struggle to generate fallacious reasoning, often resulting in the generation of truthful outputs instead. A novel jailbreak attack method is proposed that exploits this vulnerability to elicit harmful outputs while bypassing model safeguards.

https://arxiv.org/abs/2407.00869

750. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination

This study investigates linguistic bias in ChatGPT, focusing on ten English dialects and how the models respond to non-standard varieties. It reveals that both GPT-3.5 Turbo and GPT-4 reinforce stereotypes and discrimination against speakers of non-standard dialects, with GPT-4 showing improved but still biased responses compared to its predecessor.

https://arxiv.org/abs/2406.08818

812. Social Bias Probing: Fairness Benchmarking for Language Models

This paper introduces a new framework for assessing social biases in language models, going beyond binary association tests to evaluate disparate treatment of individuals based on their demographic characteristics. The study also presents SoFa, a large-scale benchmark that captures the complexities of biases and reveals more nuanced bias manifestations in language models than previously understood.

https://arxiv.org/abs/2311.09090

820. BiasAlert: A Plug-and-play Tool for Social Bias Detection in LLMs

BiasAlert is a novel tool designed to detect social bias in the outputs generated by Large Language Models (LLMs) in various scenarios, such as sentence completion and question answering. It improves upon previous bias detection methods by integrating external human knowledge and demonstrating superior performance in experiments.

https://arxiv.org/abs/2407.10241

878. Images Speak Louder than Words: Understanding and Mitigating Bias in Vision-Language Model from a Causal Mediation Perspective

This paper investigates how vision-language models (VLMs) learn and propagate biases, particularly emphasizing the significant role of image features in contributing to such biases. It introduces a causal mediation analysis framework to understand bias mechanisms and presents findings that suggest targeted interventions in the image encoder can effectively reduce bias without harming model performance.

https://arxiv.org/abs/2407.02814

952. How Susceptible are Large Language Models to Ideological Manipulation?

This paper explores how large language models (LLMs) can be influenced by ideological biases present in their training data, revealing their vulnerability to ideological manipulation. The study finds that exposure to ideologically driven samples can significantly alter the perceived ideology of LLMs, raising concerns about biased data and the need for safeguards against such influences.

https://arxiv.org/abs/2402.11725

982. Moral Foundations of Large Language Models

This paper analyzes the biases present in large language models (LLMs) through the lens of moral foundations theory (MFT), which categorizes human moral reasoning into dimensions such as care/harm and liberty/oppression. It highlights the potential risks and consequences of LLMs having a particular moral stance based on the context of prompts and their relation to human moral values.

https://arxiv.org/abs/2310.15337

986. From Descriptive Richness to Bias: Unveiling the Dark Side of Generative Image Caption Enrichment

This paper investigates the negative consequences of using Generative Caption Enrichment (GCE) in image captioning, highlighting increased gender bias and hallucination in enriched captions. Through comparison with standard captions, it reveals that models trained on GCE may exacerbate these issues significantly, urging caution in the pursuit of descriptive richness.

https://arxiv.org/abs/2406.13912

1014. User Inference Attacks on Large Language Models

This paper explores the privacy risks associated with fine-tuning large language models (LLMs) on user data, specifically through a new threat model called user inference. The authors conduct an extensive analysis of the vulnerability of users to these attacks, presenting both theoretical and empirical insights, while suggesting mitigation techniques such as differential privacy and data redundancy reduction.

https://arxiv.org/pdf/2310.09266

1090. Adaptable Moral Stances of Large Language Models on Sexist Content: Implications for Society and Gender Discourse

This paper explores how large language models (LLMs) exhibit moral reasoning when addressing sexist content, demonstrating their ability to argue from various ideological perspectives. It raises concerns about the nuanced capabilities of LLMs to both endorse and critique sexist language, emphasizing the need for careful monitoring and intervention designs to prevent misuse.

https://www.arxiv.org/abs/2410.00175

1091. DISCERN: Decoding Systematic Errors in Natural Language for Text Classifiers

DISCERN is a framework designed to interpret systematic biases in text classifiers by generating natural language descriptions of these biases. The framework improves classifier performance through augmented training sets and has been shown to enhance user interpretation of biases significantly.

https://arxiv.org/abs/2410.22239

1094. The Generation Gap: Exploring Age Bias Underlying in the Value Systems of Large Language Models

This paper investigates the presence of age bias in the values represented by Large Language Models (LLMs), revealing a tendency for these models to align more closely with younger age demographics. The research utilizes data from the World Value Survey across various value categories and finds that incorporating age identity in prompts does not adequately address these value discrepancies among different age groups.

https://arxiv.org/abs/2404.08760

1188. Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps

This study analyzes the performance gaps in multilingual automatic speech recognition (ASR) models, focusing specifically on gender disparities across multiple languages and conditions. The findings indicate that while state-of-the-art models exhibit gender performance gaps, these cannot be solely attributed to acoustic or lexical factors, highlighting the need for improved evaluation and accessibility in training data.

https://arxiv.org/abs/2402.17954

1197. Local Contrastive Editing of Gender Stereotypes

This paper introduces local contrastive editing to identify and edit subsets of weights in language models that encode gender stereotypes, enhancing understanding of bias in LMs. Through experiments, it shows that this method can successfully localize and modify less than 0.5% of weights related to gender bias, suggesting new strategies for controlling model properties.

https://arxiv.org/abs/2410.17739

1212. Safety Arithmetic: A Framework for Test-time Safety Alignment of Language Models by Steering Parameters and Activations

This paper introduces Safety Arithmetic, a training-free framework designed to enhance the safety alignment of large language models with human values across various scenarios. It aims to prevent harmful outputs and foster safe content generation while maintaining model utility, thereby outperforming current methods in ensuring safety.

https://arxiv.org/abs/2406.11801

1234. M3Hop-CoT: Misogynous Meme Identification with Multimodal Multi-hop Chain-of-Thought

The paper introduces the M3Hop-CoT framework for identifying misogynous memes in social media, addressing the challenges of detecting subtle cues targeting women. Using a combination of multimodal approaches and large language models, the framework demonstrates strong performance on various benchmark datasets, emphasizing cultural and emotional contextual understanding.

https://arxiv.org/abs/2410.09220

1235. GPT-4 Jailbreaks Itself with Near-Perfect Success Using Self-Explanation

The paper presents a novel jailbreaking method called Iterative Refinement Induced Self-Jailbreak (IRIS), which utilizes the reflective capabilities of large language models for effective jailbreaking. IRIS achieves high success rates in circumventing model safety restrictions by using self-explanation and fewer queries compared to previous methods.

https://arxiv.org/abs/2405.13077

115. CMD: a framework for Context-aware Model self-Detoxification

The CMD framework aims to improve detoxification in language models by incorporating context-awareness, enhancing both detoxification effectiveness and generation quality. It utilizes a two-phase approach of detoxifying the context followed by generating safe outputs using language models, validated by experiments on multiple LLMs showing superior performance.

https://arxiv.org/abs/2308.08295

164. Alignment-Enhanced Decoding: Defending via Token-Level Adaptive Refining of Probability Distributions

This paper presents Alignment-Enhanced Decoding (AED), a new defense mechanism against jailbreak attacks on large language models by addressing alignment failures. The method utilizes adaptive decoding and feedback from self-evaluation to produce safer and more helpful output distributions, validated through extensive experiments.

https://arxiv.org/abs/2408.07663

245. SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning

This paper introduces SOUL, a second-order optimization framework for unlearning in large language models (LLMs), addressing the need for ethical AI practices and compliance with data regulations. The proposed method outperforms traditional first-order methods, demonstrating its effectiveness in removing unwanted data influences while preserving model utility.

https://arxiv.org/abs/2404.18239

671. The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm

This paper explores the challenges of aligning AI systems' safety measures with diverse global and local preferences, particularly in non-homogeneous linguistic settings. It introduces a new framework for understanding AI alignment across multiple languages, providing insights into cross-lingual transfer and optimization strategies to minimize harms in global AI applications.

https://arxiv.org/abs/2406.18682

679. Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models

This paper introduces the Shortcut Suite, a test suite designed to evaluate how Large Language Models (LLMs) rely on shortcuts derived from dataset biases, impacting their robustness and performance. The experiments reveal that LLMs are prone to overconfidence and reliance on shortcuts, with broader implications for improving their generalization capabilities.

https://arxiv.org/abs/2410.13343

760. Holistic Automated Red Teaming for Large Language Models through Top-Down Test Case Generation and Multi-turn Interaction

The paper proposes HARM, a novel framework for holistic automated red teaming of large language models (LLMs), which emphasizes comprehensive test case coverage and multi-turn interaction dynamics. By using a top-down test case generation strategy along with reinforcement learning, this method reveals vulnerabilities in LLMs while providing insights for improving model alignment.

https://www.arxiv.org/abs/2409.16783

844. CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation

CopyBench introduces a benchmark for evaluating both literal and non-literal reproduction of copyrighted text in language model outputs. The study finds that while literal copying is rare, significant non-literal copying occurs, especially in larger models, and discusses the effectiveness of strategies to mitigate this issue.

https://arxiv.org/abs/2407.07087

960. FlipGuard: Defending Preference Alignment against Update Regression with Constrained Optimization

FlipGuard addresses the issue of regression in preference alignment for Large Language Models, proposing a constrained optimization method to prevent performance degradation after updates. By utilizing a customized reward system and enforcing congruence with pre-aligned models, it aims to maintain knowledge while aligning preferences effectively.

https://arxiv.org/abs/2410.00508

1035. Who is better at math, Jenny or Jingzhen? Uncovering Stereotypes in Large Language Models

This paper addresses the issue of stereotypes propagated by large language models (LLMs) and their effects on marginalized communities by introducing a dataset named GlobalBias. Through analysis, it is found that larger models exhibit higher levels of stereotypical outputs, regardless of instructions to avoid them.

http://www.arxiv.org/abs/2407.06917

1098. Evaluating Short-Term Temporal Fluctuations of Social Biases in Social Media Data and Masked Language Models

This paper investigates how social biases in language models, particularly Masked Language Models, fluctuate over time as they are trained on growing social media data. The findings indicate that while biases are prevalent, they generally remain stable with some exceptions, providing insights into demographic preferences within the training corpora.

https://arxiv.org/abs/2406.13556

1200. STAR: SocioTechnical Approach to Red Teaming Language Models

This paper presents STAR, a sociotechnical framework designed to enhance the safety of large language models through improved steerability and signal quality. It introduces parameterised instructions for human red teamers and emphasizes the importance of diverse viewpoints in improving reliability and sensitivity in annotation processes.

https://arxiv.org/abs/2406.11757

1265. ALVIN: Active Learning Via INterpolation

The paper presents ALVIN, a novel active learning method that mitigates the issue of spurious correlations in models by focusing on under-represented groups through intra-class interpolations. Experimental results indicate that ALVIN surpasses existing active learning techniques in both in-distribution and out-of-distribution scenarios across multiple NLP tasks.

https://arxiv.org/abs/2410.08972

785. Large Language Models Can Be Contextual Privacy Protection Learners

PrivacyMind introduces a paradigm for fine-tuning Large Language Models that protects sensitive personally identifiable information (PII) while incorporating domain-specific knowledge. The proposed methods, particularly instruction tuning, show effectiveness in safeguarding privacy during model inference through extensive experiments across various datasets.

https://arxiv.org/abs/2310.02469

13. Thinking Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models

This study explores the use of structured prompting techniques to debias language model outputs, making them more accessible to end-users by involving System 2 thinking. The research demonstrates that complex prompts significantly reduce bias in generated text while maintaining competitive performance on various tasks.

https://arxiv.org/abs/2405.10431

16. Systematic Biases in LLM Simulations of Debates

This study examines systematic biases in Large Language Models (LLMs) during simulations of political debates, highlighting their tendency to reflect social biases despite attempts to simulate human-like behavior. The findings emphasize the need for developing methods to mitigate these biases, improving the realism of LLM-based simulations.

https://arxiv.org/abs/2402.04049

27. Impeding LLM-assisted Cheating in Introductory Programming Assignments via Adversarial Perturbation

This paper evaluates the performance of large language models (LLMs) on introductory programming assignments and explores the use of adversarial perturbations to reduce their effectiveness in generating correct code. A user study shows that these perturbations significantly decrease the accuracy of code generated by LLMs, effectively hindering potential cheating in academic contexts.

https://arxiv.org/abs/2410.09318

29. On the Influence of Gender and Race in Romantic Relationship Prediction from Large Language Models

This paper investigates biases in large language models regarding gender and race in the context of romantic relationship predictions. It reveals that these models exhibit heteronormative biases and shows disparities in predictions based on character racial identities, particularly highlighting issues surrounding Asian names.

http://arxiv.org/abs/2410.03996

33. Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection

This paper investigates the vulnerability of large language models (LLMs) to prompt injection attacks, which can manipulate these models by embedding malicious instructions into their input. The authors establish a benchmark to evaluate the robustness of LLMs in following instructions while revealing significant weaknesses and calling for a shift towards enhancing models' comprehension of prompts instead of solely their instruction-following capabilities.

https://arxiv.org/abs/2308.10819

60. Eliminating Biased Length Reliance of Direct Preference Optimization via Down-Sampled KL Divergence

This paper addresses an issue with Direct Preference Optimization (DPO) in Large Language Models, specifically the problem of 'verbosity' linked to biased label reliance and algorithmic length dependence. The authors propose a new downsampling method, SamPO, to mitigate verbosity, showing improvements in performance across various benchmarks and datasets.

https://arxiv.org/abs/2406.10957

70. AgentReview: Exploring Peer Review Dynamics with LLM Agents

AgentReview is a novel framework that uses large language models to simulate peer review dynamics, addressing complexity in the review process and concerns about data privacy. The study highlights the influence of various biases on paper decisions, revealing significant variability due to reviewer biases.

https://arxiv.org/abs/2406.12708

85. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment

This paper introduces Controllable Preference Optimization (CPO), a method that addresses the alignment tax in AI by allowing models to optimize multiple objectives based on explicit human preferences. Experimental results demonstrate the effectiveness of CPO in improving alignment across various objectives, such as helpfulness, honesty, and harmlessness.

https://arxiv.org/abs/2402.19085

98. SHIELD: Evaluation and Defense Strategies for Copyright Compliance in LLM Text Generation

This paper addresses the legal challenges posed by large language models (LLMs) regarding copyright compliance, including their tendency to generate copyrighted text and the vulnerability to safeguard bypassing attacks. It introduces a comprehensive evaluation benchmark and a defense mechanism to reduce the generation of copyrighted material, while also providing a curated dataset for testing these solutions.

https://arxiv.org/abs/2406.12975

108. Evaluating Psychological Safety of Large Language Models

This paper evaluates the psychological safety of large language models (LLMs) using unbiased prompts and personality tests, revealing that many LLMs maintain darker personality traits despite safety fine-tuning. It recommends employing comprehensive psychological metrics to enhance the safety and well-being of LLMs through optimized fine-tuning methods.

https://arxiv.org/abs/2212.10529

157. ASETF: A Novel Method for Jailbreak Attack on LLMs through Translate Suffix Embeddings

The paper proposes an Adversarial Suffix Embedding Translation Framework (ASETF) designed to improve the efficacy and efficiency of jailbreak attacks on large language models (LLMs) by transforming continuous adversarial embeddings into coherent text. The method reduces computational overhead and generates multiple adversarial samples, achieving better attack success rates and text fluency across various LLMs, including black-box models.

https://arxiv.org/abs/2402.16006

187. ScalingFilter: Assessing Data Quality through Inverse Utilization of Scaling Laws

This paper introduces ScalingFilter, a method for assessing data quality in pre-training large language models without relying on a reference dataset, thus avoiding bias and improving semantic diversity. It finds that ScalingFilter enhances zero-shot performance on downstream tasks by evaluating text quality through perplexity differences between language models.

https://arxiv.org/abs/2408.08310

201. Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering

This paper investigates the quality of automatically generated counterspeech aimed at mitigating hate speech, particularly examining the trade-off between safety guardrails and argumentative richness. The study finds that while safety measures can negatively impact the quality of responses, targeting specific components of hate speech can enhance argumentative effectiveness.

http://arxiv.org/abs/2410.03466

209. Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators through a User-Centric Method

This paper examines the shortcomings of automated content moderation tools in meeting the needs of volunteer moderators, particularly in identifying toxic content. Through a model review and a user survey study, it highlights significant gaps in existing models' ability to flag violations of forum rules effectively, indicating a need for better moderation assistant tools.

https://arxiv.org/abs/2311.07879

228. Dissecting Fine-Tuning Unlearning in Large Language Models

This paper investigates the effectiveness of fine-tuning-based unlearning methods for large language models, revealing that they do not completely erase harmful knowledge embedded in the model. It demonstrates that these unlearning approaches impact the model's overall behavior and knowledge retrieval, suggesting limitations to current methods utilized for unlearning.

https://arxiv.org/abs/2410.06606

238. MiTTenS: A Dataset for Evaluating Gender Mistranslation

The paper introduces the MiTTenS dataset to evaluate gender mistranslation errors in translation systems, which can be harmful in various languages. Evaluation using this dataset reveals that both neural machine translation systems and foundation models are affected by gender mistranslations, even in high-resource languages.

https://arxiv.org/abs/2401.06935

260. Outcome-Constrained Large Language Models for Countering Hate Speech

This study develops outcome-constrained methods for generating counterspeech aimed at combating hate speech online by utilizing large language models. The effectiveness of these methods is evaluated based on their ability to produce conversations with low incivility and to discourage hateful behaviors.

https://arxiv.org/abs/2403.17146

272. A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners

This paper investigates the reasoning capabilities of large language models (LLMs), focusing on their reliance on token bias rather than genuine reasoning. The study employs a hypothesis-testing framework and synthetic datasets to reveal that LLMs struggle with logical reasoning tasks despite performing well on classic problems due to their dependence on superficial patterns.

https://arxiv.org/abs/2406.11050

275. Seeing the Forest through the Trees: Data Leakage from Partial Transformer Gradients

This paper investigates the vulnerability of distributed machine learning to data leakage through gradient inversion attacks, specifically focusing on partial gradients from Transformer models. The authors demonstrate that even a minimal number of parameters can expose training data, and they highlight the limited effectiveness of differential privacy in mitigating this risk.

https://arxiv.org/abs/2406.00999

279. How Does the Disclosure of AI Assistance Affect the Perceptions of Writing?

This paper investigates how the disclosure of AI assistance in writing impacts people's perceptions and evaluations of the quality of written content. The experimental study reveals that revealing AI's involvement often leads to decreased quality ratings and increased variability in individual assessments, influenced by the rater's confidence and familiarity with AI tools.

https://arxiv.org/abs/2410.04545

300. Pretraining Data Detection for Large Language Models: A Divergence-based Calibration Method

This paper introduces a divergence-based calibration method for detecting pretraining data utilized by large language models, addressing limitations of previous methods in classifying non-training texts. The proposed approach shows significant improvements over existing methods, validated through new benchmarks, particularly in the Chinese language.

https://arxiv.org/abs/2409.14781

324. Thinking Outside of the Differential Privacy Box: A Case Study in Text Privatization with Language Model Prompting

This paper critiques the integration of Differential Privacy (DP) in Natural Language Processing (NLP), highlighting its limitations and challenges, particularly in the context of text privatization using language model prompting. Through empirical experiments, the authors demonstrate the need for a more nuanced discussion about the utility and effectiveness of DP in NLP compared to non-DP approaches.

https://arxiv.org/abs/2410.00751

345. ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations

The paper evaluates the effectiveness of large language models in detecting offensive language in Chinese, particularly when confronted with cloaking perturbations such as homophonic substitutions and emoji transformations. It highlights significant performance drops in these models' ability to identify offensive content and emphasizes the need for improved detection techniques.

https://arxiv.org/abs/2406.12223

364. Personas as a Way to Model Truthfulness in Language Models

This paper investigates how large language models (LLMs) can represent and separate truth from falsehood in generated statements by leveraging personas formed in the pretraining data. The authors propose that these personas help the models infer truthfulness and provide evidence by demonstrating how truthful responses can be predicted and improved with fine-tuning on factual data.

https://arxiv.org/abs/2310.18168

375. An Audit on the Perspectives and Challenges of Hallucinations in NLP

This paper conducts an audit of how hallucinations in large language models are defined and discussed in the literature, highlighting a significant lack of consensus on the term `hallucination`. It also includes a survey of practitioners in the field to gather diverse perspectives, suggesting the need for clear definitions and frameworks to address the challenges posed by hallucinations in NLP.

https://arxiv.org/abs/2404.07461

393. Secured Weight Release for Large Language Models via Taylor Expansion

The paper presents TaylorMLP, a method that secures the ownership of large language models (LLMs) by transforming their weights into Taylor-series parameters, which helps prevent unauthorized use and maintains privacy concerns. Through empirical experiments, it demonstrates that this approach increases latency while effectively safeguarding the LLM weights from reconstruction by users.

https://arxiv.org/abs/2410.05331

401. Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis

This paper investigates the susceptibility of large language models (LLMs) to jailbreaking attacks, aiming to understand the reasons why certain strategies succeed in eliciting harmful outputs while others do not. It provides insights into the representation space of harmful and harmless prompts and proposes a new objective leveraging hidden representations to enhance the effectiveness of jailbreak attacks.

https://arxiv.org/abs/2406.10794

412. Aligning Large Language Models with Diverse Political Viewpoints

The paper discusses how to align large language models with diverse political viewpoints by leveraging a dataset of 100,000 comments from candidates running for the Swiss parliament. It presents methods for producing balanced overviews of these viewpoints, demonstrating improved representation compared to existing commercial models like ChatGPT.

https://arxiv.org/abs/2406.14155

438. KnowledgeSG: Privacy-Preserving Synthetic Text Generation With Knowledge Distillation From Server

KnowledgeSG is a client-server framework that improves synthetic text generation while addressing privacy concerns associated with large language models (LLMs). It utilizes differential privacy and model distillation to enhance performance and ensure data privacy during the generation process, validated through experiments in sensitive domains such as medicine and finance.

http://arxiv.org/abs/2410.05725v2

445. Bridging Modalities: Enhancing Cross-Modality Hate Speech Detection with Few-Shot In-Context Learning

This study investigates the transferability of hate speech detection between various modalities, specifically text-based tweets and vision-language memes, using few-shot in-context learning with large language models. The findings indicate that text-based examples can enhance the accuracy of detecting vision-language hate speech, underscoring the significance of cross-modality knowledge transfer.

https://arxiv.org/abs/2410.05600

461. Data Advisor: Data Curation with Foresight for Safety Alignment of Large Language Models

This paper introduces Data Advisor, a novel LLM-based method for dynamic data curation aimed at improving the safety alignment of large language models. By monitoring the quality of generated data and advising on needed improvements, Data Advisor enhances data quality and model safety without sacrificing utility.

http://arxiv.org/abs/2410.05269

493. ReCaLL: Membership Inference via Relative Conditional Log-Likelihoods

ReCaLL is a novel membership inference attack designed to detect the pretraining data of large language models by analyzing changes in conditional log-likelihoods. The study demonstrates that this method outperforms existing techniques and provides insights into LLMs' usage of membership information for inference.

https://arxiv.org/abs/2406.15968

506. PostMark: A Robust Blackbox Watermark for Large Language Models

PostMark introduces a new blackbox watermarking method for large language models that does not require access to model logits, allowing third-party implementation. It demonstrates robustness against paraphrasing attacks while evaluating the trade-off between watermark quality and text output quality.

https://arxiv.org/abs/2406.14517

508. On the Relationship between Truth and Political Bias in Language Models

This paper analyzes the relationship between truthfulness and political bias in language models, revealing that optimizing for truthfulness can lead to a left-leaning bias. The findings raise concerns about the datasets used for training reward models and the complexities of aligning models to be truthful and politically unbiased.

https://arxiv.org/abs/2409.05283

510. Statistical Uncertainty in Word Embeddings: GloVe-V

This paper introduces GloVe-V, a method to assess statistical uncertainty in word embeddings by providing scalable reconstruction error variance estimates for the GloVe model. The approach facilitates hypothesis testing and analysis of bias in word embeddings, enhancing the reliability of conclusions drawn in computational social science applications.

https://arxiv.org/abs/2406.12165

511. Annotation alignment: Comparing LLM and human annotations of conversational safety

This study examines the alignment between LLMs, specifically GPT-4, and human annotations regarding conversational safety across various demographic groups. The findings indicate a moderate correlation between model predictions and human ratings, suggesting the need for larger datasets to explore demographic disparities further.

https://arxiv.org/abs/2406.06369

513. The Factuality Tax of Diversity-Intervened Text-to-Image Generation: Benchmark and Fact-Augmented Intervention

This paper introduces DoFaiR, a benchmark for assessing the trade-off between diversity interventions and factual accuracy in Text-to-Image (T2I) generation, particularly regarding historical figures. It proposes Fact-Augmented Intervention (FAI) to enhance demographic factuality while maintaining diversity in generations by using historical data as contextual guidance.

https://arxiv.org/abs/2407.00377

575. Multimodal Clickbait Detection by De-confounding Biases Using Causal Representation Inference

This paper introduces a novel method for detecting clickbait by addressing biases using causal representation inference. By disentangling and utilizing various latent factors, the model aims to improve detection accuracy and generalization in the face of misleading content.

http://arxiv.org/abs/2410.07673

585. InferAligner: Inference-Time Alignment for Harmlessness through Cross-Model Guidance

InferAligner is a novel inference-time alignment method that uses cross-model guidance to ensure harmfulness alignment in large language models (LLMs). It demonstrates effectiveness in reducing Attack Success Rates for harmful instructions and jailbreak attacks, while maintaining performance in various downstream tasks.

https://arxiv.org/abs/2401.11206

592. The Lou Dataset - Exploring the Impact of Gender-Fair Language in German Text Classification

The Lou dataset introduces a collection of German text reformulations that promote gender-fair language and assesses its impact on various text classification tasks using language models. The findings indicate that gender-fair language influences model predictions, although existing evaluation rankings remain largely consistent.

https://arxiv.org/abs/2409.17929

598. Demystifying Verbatim Memorization in Large Language Models

This paper explores how Large Language Models (LLMs) memorize sequences verbatim, revealing that significant repetition is necessary for this memorization to occur. It also finds that better model checkpoints are more prone to verbatim memorization and that addressing this issue proves challenging without compromising model quality.

https://arxiv.org/abs/2407.17817

614. Towards Probing Speech-Specific Risks in Large Multimodal Models: A Taxonomy, Benchmark, and Insights

This paper proposes a taxonomy for speech-specific risks in large multimodal models (LMMs) and evaluates their effectiveness in detecting these risks, particularly those related to speech modality. The study finds current models struggle with identifying risks associated with paralinguistic cues, suggesting a need for improvement in detecting malicious or biased interactions in speech-based systems.

https://arxiv.org/abs/2406.17430

642. Universal Vulnerabilities in Large Language Models: Backdoor Attacks for In-context Learning

This paper investigates the vulnerabilities of large language models to backdoor attacks specifically in the context of in-context learning, without requiring fine-tuning. The authors introduce a new attack method called ICLAttack, which manipulates model behavior through poisoned context examples and prompts, achieving a high success rate across various model sizes.

https://arxiv.org/abs/2401.05949

696. Voices in a Crowd: Searching for clusters of unique perspectives

This paper proposes a framework that identifies and clusters minority perspectives from language models by analyzing annotator behavior without encoding annotator metadata. The framework effectively generates robust clusters that capture diverse opinions, validated through both qualitative and quantitative metrics.

https://arxiv.org/abs/2407.14259

713. $\texttt{ModSCAN}$: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities

The paper presents a framework named ModSCAN to measure and analyze stereotypical biases in large vision-language models (LVLMs) concerning gender and race across various scenarios. The findings indicate significant stereotype biases in popular LVLMs, suggesting these biases may arise from the training data, and highlight effective strategies to mitigate them.

http://arxiv.org/abs/2410.06967

716. Defending Against Social Engineering Attacks in the Age of LLMs

This study explores the dual role of Large Language Models (LLMs) in facilitating and defending against chat-based social engineering attacks. The authors propose a defense pipeline, ConvoSentinel, to enhance detection of malicious intent in conversational contexts, highlighting the need for advanced cybersecurity strategies leveraging LLMs.

https://arxiv.org/abs/2406.12263

794. Are LLMs Good Zero-Shot Fallacy Classifiers?

This paper investigates the capability of Large Language Models (LLMs) to classify fallacies in a zero-shot manner, aiming to enhance their generalization abilities without reliance on extensive labeled datasets. The authors propose innovative prompting schemes that significantly improve performance, particularly for smaller LLMs, and present evidence showing that LLMs can effectively function as fallacy classifiers even in out-of-distribution scenarios.

http://arxiv.org/abs/2410.15050v1

829. Applying Intrinsic Debiasing on Downstream Tasks: Challenges and Considerations for Machine Translation

This paper investigates the impact of intrinsic debiasing techniques on the performance of neural machine translation models, highlighting the challenges faced in aligning debiasing approaches with the goals of real-world applications. The authors present systematic evaluations that reveal significant effects of debiasing strategies on downstream tasks, particularly relating to the choice of embeddings and language variations.

https://arxiv.org/abs/2406.00787

884. Cultural Conditioning or Placebo? On the Effectiveness of Socio-Demographic Prompting

This paper examines the effectiveness of socio-demographic prompting in studying cultural biases within large language models (LLMs). It finds that most models demonstrate variability in their responses depending on the cultural cues in the prompt, questioning the reliability of such methods.

https://arxiv.org/abs/2406.11661

886. Hate Personified: Investigating the role of LLMs in content moderation pipeline for hate speech

This paper investigates how large language models (LLMs) can effectively moderate hate speech by analyzing their sensitivity to context such as geographical cues and persona attributes. The findings emphasize the variability in hate detection annotations when using these models, highlighting the importance of contextual factors for culturally sensitive applications.

https://arxiv.org/abs/2410.02657

900. Defining Knowledge: Bridging Epistemology and Large Language Models

This paper explores the concept of knowledge in the context of large language models (LLMs) like GPT-4 by reviewing epistemological definitions and conducting a survey among philosophers and computer scientists. It identifies gaps in current NLP research on knowledge and proposes evaluation protocols for assessing LLMs' knowledge claims based on philosophical frameworks.

https://arxiv.org/abs/2410.02499

904. The Instinctive Bias: Spurious Images lead to Hallucination in MLLMs

This paper investigates the visual illusion phenomenon in multi-modal large language models (MLLMs) caused by spurious images that are relevant yet inconsistent with given answers. It introduces CorrelationQA, a benchmark for evaluating MLLMs' robustness against misleading images, illustrating that major models are susceptible to this bias.

http://arxiv.org/abs/2402.03757v2

908. Distract Large Language Models for Automatic Jailbreak Attack

This paper presents a novel black-box jailbreak framework designed to exploit the vulnerabilities in large language models (LLMs) that are aligned with human values. Through extensive experiments, it demonstrates the effectiveness and scalability of the framework in bypassing existing defenses against jailbreaking attacks.

https://arxiv.org/abs/2403.08424

918. Intrinsic Self-correction for Enhanced Morality: An Analysis of Internal Mechanisms and the Superficial Hypothesis

This paper investigates the intrinsic moral self-correction mechanism in Large Language Models (LLMs) to explore scenarios where it is effective and its impact on the models' hidden states. The authors conclude that while self-correction improves output quality, it does not necessarily reduce immorality within the models' hidden states, suggesting that the moral self-correction may be more superficial than substantive.

https://arxiv.org/abs/2407.15286

939. RAFT: Realistic Attacks to Fool Text Detectors

The paper presents RAFT, a method for executing realistic black-box attacks on large language model detectors while maintaining the quality of the original text. The study demonstrates the effectiveness of RAFT against various detectors, revealing vulnerabilities and advocating for the development of more resilient detection mechanisms.

https://arxiv.org/abs/2410.03658

957. Granular Privacy Control for Geolocation with Vision Language Models

This paper discusses the privacy risks associated with Vision Language Models (VLMs) capable of geolocating images and identifying individuals in photos. It introduces a benchmark called GPTGeoChat to evaluate VLMs' ability to moderate geolocation information in dialogues, highlighting the necessary adjustments for improved performance in identifying sensitive information.

https://arxiv.org/abs/2407.04952

1002. What the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-centered Study

This study investigates the tangible impacts of gender bias in machine translation, highlighting how biased outputs can harm users and lead to significant gaps in service quality based on gender. Through a human-centered approach, data from 90 participants demonstrated that feminine post-editing requires more effort and incurs higher financial costs, urging reforms in evaluation methods to reflect these disparities.

https://arxiv.org/abs/2410.00545

1029. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation

This paper introduces the D3CODE dataset, a large-scale cross-cultural dataset for offensive language detection with extensive annotations from diverse social and cultural groups. It highlights significant regional variations in perceptions influenced by individual moral values, emphasizing the need for culturally sensitive NLP models.

https://arxiv.org/abs/2404.10857

1031. Annotator-Centric Active Learning for Subjective NLP Tasks

The paper presents Annotator-Centric Active Learning (ACAL), a method designed to improve the efficiency of human annotations in subjective NLP tasks by incorporating diverse annotator perspectives. The results demonstrate that ACAL enhances data efficiency and model evaluations while emphasizing the importance of having a diverse pool of annotators.

https://arxiv.org/abs/2404.15720

1034. Jailbreaking LLMs with Arabic Transliteration and Arabizi

This paper explores the vulnerabilities of Large Language Models (LLMs) to 'jailbreak' attacks using Arabic and its variations, specifically focusing on Arabic transliteration and Arabizi (chatspeak). The findings reveal that these forms can provoke unsafe outputs from models like OpenAI GPT-4 and Anthropic Claude 3 Sonnet, suggesting a need for enhanced safety protocols in multilingual contexts.

https://arxiv.org/abs/2406.18725

1066. DetoxLLM: A Framework for Detoxification with Explanations

DetoxLLM is a comprehensive detoxification framework designed to address the limitations of existing models by providing cross-platform capabilities and explaining the detoxification process. It introduces a pseudo-parallel corpus and demonstrates improved effectiveness and robustness against adversarial toxicity compared to previous models.

https://arxiv.org/abs/2402.15951

1221. Accurate and Data-Efficient Toxicity Prediction when Annotators Disagree

The paper introduces three innovative methods for predicting individual annotator toxicity ratings in NLP, focusing on integrating annotator-specific information. The study demonstrates that using demographic and survey-derived data significantly enhances prediction accuracy in subjective assessments, highlighting the complexities of annotator modeling.

https://arxiv.org/abs/2410.12217

231. Private Language Models via Truncated Laplacian Mechanism

This paper presents a novel method for private word embedding called the high dimensional truncated Laplacian mechanism, addressing privacy concerns in deep learning models for NLP. The proposed method theoretically offers lower variance than existing approaches and demonstrates competitive performance even when prioritizing privacy.

https://arxiv.org/abs/2410.08027

1045. Fuse to Forget: Bias Reduction and Selective Memorization through Model Fusion

This paper explores the use of model fusion not just for performance enhancement but also for reducing biases and unwanted knowledge in language models. The findings indicate that during fusion, shared knowledge among models is strengthened while unshared knowledge tends to be forgotten, showcasing its potential as a debiasing tool.

https://arxiv.org/abs/2311.07682

1147. Do LLMs Know to Respect Copyright Notice?

This paper investigates whether large language models (LLMs) respect copyright information present in user input, highlighting the risks of copyright infringement by these models. It also provides a benchmark dataset for evaluating LLMs' behaviors regarding copyright, underscoring the need for alignment with copyright regulations.

http://arxiv.org/abs/2411.01136v1

1174. Language is Scary when Over-Analyzed: Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts

This study explores misogyny detection through the lens of argumentative reasoning, focusing on how large language models (LLMs) understand and convey implicit misogynistic meanings. The research highlights the shortcomings of LLMs in reasoning about these comments and their reliance on internalized stereotypes rather than on inductive reasoning.

https://arxiv.org/abs/2409.02519

1230. The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced By Them Instead?

This paper investigates the efficiency and accuracy of Large Language Models (LLMs) in specialized tasks through a user study, revealing that while LLMs can improve task completion speed, they may also introduce biases that affect analytical depth. The findings highlight the tension between enhanced efficiency and the risk of biased analytical processes in human-LLM partnerships in fields such as policy studies.

https://arxiv.org/abs/2410.04699

10. Hateful Word in Context Classification

77. Mitigating Language Bias of LMMs in Social Intelligence Understanding with Virtual Counterfactual Calibration

126. An Inversion Attack Against Obfuscated Embedding Matrix in Language Model Inference

132. Oddballs and Misfits: Detecting Implicit Abuse in Which Identity Groups are Depicted as Deviating from the Norm

244. Hidden Persuaders: How LLM Political Bias Could Sway Our Elections

377. Reconstruct Your Previous Conversations! Comprehensively Investigating Privacy Leakage Risks in Conversations with GPT Models

565. Optimizing Language Models with Fair and Stable Reward Composition in Reinforcement Learning

620. Mitigate Extrinsic Social Bias in Pre-trained Language Models via Continuous Prompts Adjustment

848. XDetox: Text Detoxification with Token-Level Toxicity Explanations

1022. GuardBench: A Large-Scale Benchmark for Guardrail Models

1172. BiasWipe: Mitigating Unintended Bias in Text Classifiers through Model Interpretability

1198. De-Identification of Sensitive Personal Data in Datasets Derived from IIT-CDIP

1207. Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP

1225. FairFlow: Mitigating Dataset Biases through Undecided Learning for Natural Language Understanding

566. Fine-grained Pluggable Gradient Ascent for Knowledge Unlearning in Language Models

622. Integrating Argumentation and Hate-Speech-based Techniques for Countering Misinformation

670. PANDA: Persona Attributes Navigation for Detecting and Alleviating Overuse Problem in Large Language Models

1099. Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection

1144. RAt: Injecting Implicit Bias for Text-To-Image Prompt Refinement Models

1215. Sprout: Green Generative AI with Carbon-Efficient LLM Inference

97. Predicate Debiasing in Vision-Language Models Integration for Scene Graph Generation Enhancement

107. DA^3: A Distribution-Aware Adversarial Attack against Language Models

259. Leveraging Conflicts in Social Media Posts: Unintended Offense Dataset

262. Adaptive Immune-based Sound-Shape Code Substitution for Adversarial Chinese Text Attacks

681. Where Am I From? Identifying Origin of LLM-generated Content

754. An Analysis and Mitigation of the Reversal Curse

877. BaitAttack: Alleviating Intention Shift in Jailbreak Attacks via Adaptive Bait Crafting

1005. Revisiting the Robustness of Watermarking to Paraphrasing Attacks

1103. MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation

1166. PREDICT: Multi-Agent-based Debate Simulation for Generalized Hate Speech Detection

1260. Context-aware Watermark with Semantic Balanced Green-red Lists for Large Language Models

Machine Learning for NLP

20. Learning Planning-based Reasoning by Trajectories Collection and Process Reward Synthesizing

This paper presents a framework to enhance reasoning in large language models by optimizing planning-based reasoning through Direct Preference Optimization on collected trajectories. The findings show that the proposed method allows a model to outperform competitors in logical reasoning tasks despite traditional challenges in latency and scaling human annotations.

https://arxiv.org/abs/2402.00658

38. MSI-Agent: Incorporating Multi-Scale Insight into Embodied Agents for Superior Planning and Decision-Making

The paper introduces the Multi-Scale Insight Agent (MSI-Agent), which enhances planning and decision-making abilities of large language models (LLMs) by effectively summarizing and utilizing insights across different scales. Through a three-part pipeline consisting of an experience selector, insight generator, and insight selector, MSI demonstrates improved performance and robustness in decision-making tasks, especially in domain-shifting scenarios.

https://arxiv.org/abs/2409.16686

64. A Survey on In-context Learning

This paper surveys in-context learning (ICL) as a paradigm for enhancing the capabilities of large language models (LLMs) by making predictions based on context and examples. It discusses the definition of ICL, advanced techniques, application scenarios, and associated challenges while suggesting future research directions.

https://arxiv.org/abs/2301.00234

66. AMR-Evol: Adaptive Modular Response Evolution Elicits Better Knowledge Distillation for Large Language Models in Code Generation

This paper introduces the Adaptive Modular Response Evolution (AMR-Evol) framework, which enhances the knowledge distillation process for open-source language models in code generation. Through modular decomposition and adaptive response evolution, AMR-Evol achieves notable performance improvements in code benchmarks compared to existing methods.

https://arxiv.org/abs/2410.00558

102. MetaGPT: Merging Large Language Models Using Model Exclusive Task Arithmetic

This paper presents MetaGPT, a method for merging large language models through a multi-task learning framework that optimizes performance, computational efficiency, and data privacy. By employing model-exclusive task arithmetic, MetaGPT facilitates cost-effective implementation and achieves state-of-the-art results across various tasks without relying on extensive training data.

https://arxiv.org/abs/2406.11385

104. Retrieved Sequence Augmentation for Protein Representation Learning

This paper presents Retrieved Sequence Augmentation (RSA) for protein representation learning, demonstrating significant improvements in structure and property prediction tasks. RSA enhances protein language models by linking query sequences to similar sequences, improving efficiency and accuracy without requiring extensive alignment processes.

http://arxiv.org/abs/2302.12563v1

125. Advancing Process Verification for Large Language Models via Tree-Based Preference Learning

This paper presents Tree-based Preference Learning Verifier (Tree-PLV), a novel approach that improves the verification process in Large Language Models by focusing on step-level preferences rather than binary classifications. The method significantly enhances reasoning path evaluations, demonstrating substantial performance gains across various reasoning tasks compared to existing benchmarks.

https://arxiv.org/abs/2407.00390

130. FuseGen: PLM Fusion for Data-generation based Zero-shot Learning

FuseGen introduces a novel framework for data-generation based zero-shot learning that utilizes multiple pre-trained language models (PLMs) to improve the quality of synthetic datasets. By employing a criteria for subset selection and iterative data generation, FuseGen significantly enhances the performance of small task-specific models (STMs) across various tasks.

https://arxiv.org/abs/2406.12527

148. Instruction Pre-Training: Language Models are Supervised Multitask Learners

This paper proposes Instruction Pre-Training, a framework that enhances language models by using supervised multitask pre-training with instruction-response pairs. Experiments demonstrate that this method improves pre-trained models, even allowing smaller models to perform comparably to larger counterparts.

https://arxiv.org/abs/2406.14491

149. LEMoE: Advanced Mixture of Experts Adaptor for Lifelong Model Editing of Large Language Models

This paper presents LEMoE, an advanced Mixture of Experts adaptor designed for lifelong model editing in large language models (LLMs). It addresses the challenges of catastrophic forgetting and inconsistent routing, proposing techniques that improve the editing process while maintaining performance in batch editing tasks.

https://arxiv.org/abs/2406.20030

183. On Training Data Influence of GPT Models

This paper introduces GPTfluence, a novel approach for assessing the influence of training data on the performance of GPT models. It effectively demonstrates the impact of individual training instances across various tasks while providing robust generalization capabilities to unseen data, with its resources made publicly available.

http://arxiv.org/abs/2404.07840v3

221. PTD-SQL: Partitioning and Targeted Drilling with LLMs in Text-to-SQL

This paper presents PTD-SQL, a method that leverages Large Language Models (LLMs) by employing query group partitioning to improve reasoning capabilities in Text-to-SQL tasks. Experimental results show that models using PTD-SQL can achieve state-of-the-art performance, highlighting the potential parallels with human learning processes.

https://arxiv.org/pdf/2409.14082

226. PRompt Optimization in Multi-Step Tasks (PROMST): Integrating Human Feedback and Heuristic-based Sampling

This paper introduces PRompt Optimization in Multi-Step Tasks (PROMST), a framework that optimizes prompts for large language models (LLMs) in multi-step tasks by integrating human feedback and heuristic sampling. The approach provides significant improvements over existing methods by effectively leveraging human-designed feedback rules and a learned heuristic model for prompt performance prediction.

https://arxiv.org/abs/2402.08702

274. MuMath-Code: Combining Tool-Use Large Language Models with Multi-perspective Data Augmentation for Mathematical Reasoning

The paper introduces MuMath-Code, a model that enhances mathematical reasoning by combining tool-use Large Language Models (LLMs) with multi-perspective data augmentation. It demonstrates significant performance improvements through a two-stage training strategy, providing a new dataset and code for public use.

https://arxiv.org/abs/2405.07551

301. Scaling Synthetic Logical Reasoning Datasets with Context-Sensitive Declarative Grammars

This paper presents a declarative framework for generating logical reasoning problems that improves the training of language models to mimic theorem provers. By using context-sensitive rules and semantic constraints, the approach achieves state-of-the-art accuracy on logic datasets with a smaller model, enhancing reasoning capabilities while maintaining performance on natural language tasks.

https://arxiv.org/abs/2406.11035

304. Formality Favored: Unraveling the Learning Preferences of Large Language Models on Data with Conflicting Knowledge

This study analyzes how large language models (LLMs) learn from pretraining data that contains conflicting knowledge and how they exhibit preferences towards more formal texts. The findings suggest that LLMs develop learning preferences that resemble human biases, leading to improved performance when data features align with consistency and formality, resulting in insights that could influence future model training and assessment.

http://arxiv.org/abs/2410.04784

338. CItruS: Chunked Instruction-aware State Eviction for Long Sequence Modeling

The paper presents CItruS, a novel technique for long sequence modeling that integrates task-specific attention preferences into the state eviction process of hidden states in Transformer models. CItruS demonstrates improved performance in long sequence comprehension and retrieval while maintaining perplexity, and it is designed to enhance efficiency in memory usage.

https://arxiv.org/abs/2406.12018

340. C-LLM: Learn to Check Chinese Spelling Errors Character by Character

This paper introduces C-LLM, a character-level tokenization approach for Chinese Spell Checking (CSC) that addresses the shortcomings of existing models in handling character constraints. C-LLM demonstrates significant performance improvements over state-of-the-art methods for CSC, particularly in vertical domains.

https://arxiv.org/abs/2406.16536

341. PSC: Extending Context Window of Large Language Models via Phase Shift Calibration

This paper introduces Position Interpolation (PI), a method that significantly extends the context window of RoPE-based pretrained language models like LLaMA, allowing for efficient processing of longer documents without sacrificing quality on tasks within the original context window. It demonstrates that PI achieves this by accurately scaling input position indices instead of extrapolating beyond trained limits, thus providing a stable and effective solution for long-context applications.

https://arxiv.org/abs/2306.15595

397. Automatic Instruction Evolving for Large Language Models

This paper introduces Auto Evol-Instruct, an automated framework for evolving instruction datasets for large language models that minimizes human intervention. It achieves superior performance to human-designed methods across various benchmarks by iteratively improving its evolutionary strategies based on identified issues during the instruction evolution process.

https://arxiv.org/abs/2406.00770

414. Extending Context Window of Large Language Models from a Distributional Perspective

This paper presents a novel method for extending the context window of RoPE-based large language models by minimizing the disturbance of rotary angle distributions. The proposed approach significantly improves performance while maintaining consistency with the pre-training phase, achieving notable gains over existing methods.

https://arxiv.org/abs/2410.01490

435. UNO Arena for Evaluating Sequential Decision-Making Capability of Large Language Models

The paper introduces UNO Arena, a framework for evaluating the sequential decision-making capabilities of large language models (LLMs), using the card game UNO. The study employs various player types, including random players and LLMs, to assess performance, and proposes the TUTRI player to enhance decision-making through reflections on game history and strategies.

http://arxiv.org/abs/2406.16382v1

458. AdaSwitch: Adaptive Switching between Small and Large Agents for Effective Cloud-Local Collaborative Learning

The paper presents a framework called AdaSwitch that enables collaborative learning between local and cloud-based large language models (LLMs). By dynamically switching between smaller and larger models based on task complexity, the approach enhances performance and efficiency, achieving competitive results with lower computational costs.

https://arxiv.org/abs/2410.13181

459. CoBa: Convergence Balancer for Multitask Finetuning of Large Language Models

This paper introduces CoBa, a novel multi-task learning (MTL) approach for fine-tuning large language models that balances task convergence while minimizing computational overhead. Experimental results demonstrate improvements in task convergence and LLM performance, achieving up to 13% better results compared to existing baselines.

http://arxiv.org/abs/2410.06741

475. WPO: Enhancing RLHF with Weighted Preference Optimization

This paper introduces Weighted Preference Optimization (WPO), a method that improves reinforcement learning from human feedback (RLHF) by adjusting off-policy preference data to better simulate on-policy learning. WPO effectively addresses the distributional gap issue in preference data and demonstrates superior performance on instruction following benchmarks compared to existing methods.

https://arxiv.org/abs/2406.11827

489. A Thorough Examination of Decoding Methods in the Era of LLMs

This paper investigates different decoding methods and their significance in converting language models into functional task solvers. It highlights the task-dependent nature of decoding performance and the trade-offs between optimal results and practical implementation across diverse settings.

https://arxiv.org/abs/2402.06925

496. LIONs: An Empirically Optimized Approach to Align Language Models

This paper presents an analysis of various design choices in the training of language models, specifically focusing on how to enhance their instruction-following and conversational abilities. The findings suggest that certain techniques, such as sequence packing and loss masking, significantly improve the performance of trained language models beyond existing instruct models.

https://arxiv.org/abs/2407.06542

525. Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs

This paper presents strategies for optimizing prompts in multi-stage Language Model Programs, aiming to enhance their performance on NLP tasks without module-level labels. The proposed MIPRO algorithm improves accuracy across diverse programs, demonstrating significant effectiveness when applied to an open-source model.

https://arxiv.org/abs/2406.11695

528. Structure Guided Prompt: Instructing Large Language Model in Multi-Step Reasoning by Exploring Graph Structure of the Text

This paper presents the Structure Guided Prompt framework that enhances the multi-step reasoning abilities of Large Language Models (LLMs) by employing graph structures to organize and navigate information. Experiments demonstrate that this approach significantly improves LLM performance in reasoning tasks across diverse scenarios.

https://arxiv.org/abs/2402.13415

551. StablePrompt : Automatic Prompt Tuning using Reinforcement Learning for Large Language Model

This paper introduces StablePrompt, a method for automatic prompt tuning using Reinforcement Learning that enhances training stability while optimizing prompts for large language models. The proposed Adaptive Proximal Policy Optimization approach outperforms existing methods across tasks like text classification, question answering, and text generation.

https://arxiv.org/abs/2410.07652

583. Enhancing Language Model Factuality via Activation-Based Confidence Calibration and Guided Decoding

This paper proposes an activation-based confidence calibration method, ActCab, to improve the factuality of language models (LMs) by aligning their generation confidence with actual answer correctness. It introduces CoDec, a confidence-guided decoding strategy that enhances the output reliability of LMs, achieving better calibration performance and factuality in question-answering scenarios.

https://arxiv.org/abs/2406.13230

587. Fisher Information-based Efficient Curriculum Federated Learning with Large Language Models

This paper presents a Fisher Information-based Efficient Curriculum Federated Learning framework (FibecFed) designed to enhance the fine-tuning of Large Language Models (LLMs) in a decentralized environment. The proposed methods improve the effectiveness and efficiency of Federated Learning by selectively sampling data and updating parameters adaptively, yielding significant gains in accuracy and speed compared to existing techniques.

https://arxiv.org/abs/2410.00131

597. Fine-Tuning and Prompt Optimization: Two Good Steps that Work Better Together

This paper presents a novel approach for optimizing modular Language Model pipelines by simultaneously fine-tuning model weights and prompt templates. Experimental results demonstrate that this combined optimization strategy significantly improves performance in various downstream tasks compared to optimizing each component separately.

https://arxiv.org/abs/2407.10930

623. BPO: Supercharging Online Preference Learning by Adhering to the Proximity of Behavior LLM

This paper introduces BPO, an online Preference Optimization method that helps align large language models (LLMs) with human preferences by utilizing online training samples. It demonstrates significant improvements in performance when integrated with direct alignment from preferences (DAP) methods across various tasks, highlighting the importance of proper trust region construction for effective LLM alignment.

https://arxiv.org/html/2406.12168v1

626. ORPO: Monolithic Preference Optimization without Reference Model

This paper introduces a novel reference model-free preference optimization algorithm called ORPO, which enhances supervised fine-tuning (SFT) methods for language models to achieve better alignment with preferences. Empirical results demonstrate that fine-tuning various models with ORPO outperforms existing state-of-the-art models in performance metrics.

https://arxiv.org/abs/2403.07691

701. Lifelong Event Detection via Optimal Transport

This paper presents a novel approach to Lifelong Event Detection that addresses the challenge of catastrophic forgetting through optimal transport principles. The proposed method, LEDOT, integrates replay sets and prototype latent representations, demonstrating significant improvements over existing methods in event detection tasks.

https://arxiv.org/abs/2410.08905

703. FIRST: Teach A Reliable Large Language Model Through Efficient Trustworthy Distillation

This paper introduces FIRST, a new method to improve the trustworthiness of large language models by addressing mis-calibration issues that arise during fine-tuning. It demonstrates that by utilizing a small portion of a teacher model's knowledge effectively, the model can achieve better accuracy and reduced mis-calibration in a cost-efficient manner.

https://arxiv.org/abs/2408.12168

712. DKEC: Domain Knowledge Enhanced Multi-Label Classification for Diagnosis Prediction

DKEC introduces a novel approach for multi-label text classification in the medical domain by leveraging external medical knowledge through heterogeneous knowledge graphs. The proposed method significantly enhances diagnosis prediction, especially for few-shot classes, outperforming traditional state-of-the-art models by enabling smaller language models to achieve comparable results to larger ones.

https://arxiv.org/abs/2310.07059

731. Chain and Causal Attention for Efficient Entity Tracking

This paper explores the limitations of transformers in efficiently tracking entities in large language models, identifying a theoretical constraint related to the number of layers required for handling state changes. It proposes a novel attention mechanism that enables effective entity tracking with fewer layers and provides both theoretical and empirical validations of its approach.

http://arxiv.org/abs/2410.05565v1

751. Lifelong Knowledge Editing for LLMs with Retrieval-Augmented Continuous Prompt Learning

This paper presents RECIPE, a method designed for lifelong knowledge editing in large language models (LLMs), enhancing editing efficacy and inference efficiency. RECIPE employs a novel approach by converting knowledge into informative prompts and implementing a dynamic threshold mechanism with a Knowledge Sentinel for improved performance across various models.

https://arxiv.org/abs/2405.03279

758. Fewer is More: Boosting Math Reasoning with Reinforced Context Pruning

This paper introduces CoT-Influx, a method that enhances the reasoning abilities of Large Language Models (LLMs) in math through an efficient context pruning technique. By using a pruner to focus on concise examples, CoT-Influx significantly outperforms existing prompting methods without requiring fine-tuning.

https://arxiv.org/abs/2312.08901

823. Bridging Local Details and Global Context in Text-Attributed Graphs

This paper introduces GraphBridge, a framework that integrates local and global perspectives in text-attributed graphs by utilizing contextual textual information. It addresses scalability and efficiency challenges through a graph-aware token reduction module while achieving state-of-the-art performance across various models and datasets.

https://arxiv.org/abs/2406.12608

842. LMs learn governing principles of dynamical systems, revealing an in-context neural scaling law

This paper investigates the capacity of pretrained large language models (LLMs) to effectively understand and predict the behavior of dynamical systems through zero-shot task performance. The authors present findings that highlight both the model's accuracy in predicting time series without fine-tuning and a relationship between input context length and prediction accuracy, revealing a neural scaling law.

http://arxiv.org/abs/2402.00795v4

847. Layer by Layer: Uncovering Where Multi-Task Learning Happens in Instruction-Tuned Large Language Models

This paper explores the encoding of task-specific knowledge in large language models (LLMs) and the impact of instruction tuning across multiple NLP tasks. It uses matrix analysis tools to identify how LLMs transition from general representations to task-oriented representations, enhancing our understanding of multi-task learning and informing future research in this area.

https://arxiv.org/abs/2410.20008

861. Re-ReST: Reflection-Reinforced Self-Training for Language Agents

This paper introduces Reflection-Reinforced Self-Training (Re-ReST) for improving self-training in language agents by refining low-quality generated samples using a reflector mechanism. Extensive experiments show Re-ReST enhances performance across various tasks, such as multi-hop question answering and code generation, demonstrating its efficiency in generating high-quality training data without extensive human intervention.

https://arxiv.org/abs/2406.01495

871. Re-Reading Improves Reasoning in Large Language Models

This paper introduces Re2, a prompting method that enhances the reasoning capabilities of Large Language Models by having them re-read questions as input. The empirical study demonstrates its effectiveness across multiple reasoning benchmarks and shows the method's adaptability with various language models and prompting techniques.

https://arxiv.org/abs/2309.06275

880. SciAgent: Tool-augmented Language Models for Scientific Reasoning

This paper introduces SciAgent, a tool-augmented approach for enhancing the scientific reasoning capabilities of Large Language Models by providing them with scalable toolsets. The study involves the creation of a training corpus, MathFunc, and a benchmark, SciToolBench, to evaluate the effectiveness of LLMs with tool assistance, showing significant performance improvements over existing models.

https://arxiv.org/abs/2402.11451

887. Temporally Consistent Factuality Probing for Large Language Models

This paper introduces TeCFaP, a novel task for evaluating the factual consistency of Large Language Models (LLMs) in a temporal context. It proposes a new dataset and solution framework that combines instruction tuning and reinforcement learning to enhance this capability in LLMs.

https://arxiv.org/abs/2409.14065

890. LLaMA-MoE: Building Mixture-of-Experts from LLaMA with Continual Pre-Training

This paper presents LLaMA-MoE, a mixture-of-experts model built upon the LLaMA-2 architecture that aims to alleviate the challenges of training large language models by utilizing expert construction and continual pre-training. Empirical results show that the LLaMA-MoE models outperform similar dense models, demonstrating an effective scaling approach for large language models.

https://arxiv.org/abs/2406.16554

916. Scaling Laws for Linear Complexity Language Models

This study explores the scaling laws for linear complexity language models, testing three specific architectures against a conventional transformer baseline. The findings indicate that these models not only match the performance of transformer models but also demonstrate improved linguistic capabilities.

https://arxiv.org/abs/2406.16690

920. LM2: A Simple Society of Language Models Solves Complex Reasoning

The paper introduces LM2, a modular society of language models designed to enhance complex reasoning capabilities by coordinating decomposition, solution generation, and verification tasks among three specialized models. Experimental results demonstrate that LM2 outperforms existing methods in reasoning tasks, indicating a significant advancement in handling multi-step reasoning more effectively.

https://arxiv.org/abs/2404.02255

934. Model Editing Harms General Abilities of Large Language Models: Regularization to the Rescue

This paper addresses the unintended degradation of general abilities in large language models (LLMs) due to model editing techniques used to update their knowledge and reduce hallucinations. It proposes a new method named RECT that regularizes the editing process to mitigate these side effects while maintaining strong editing performance.

https://arxiv.org/abs/2401.04700

955. Fill In The Gaps: Model Calibration and Generalization with Synthetic Data

This paper presents a method for calibrating machine learning models using synthetic data to enhance their generalizability without sacrificing accuracy. By leveraging large language models for synthetic data generation, the authors report significant improvements in accuracy and calibration error across various NLP tasks.

https://arxiv.org/html/2410.10864v1

971. Zero-Shot Detection of LLM-Generated Text using Token Cohesiveness

This paper presents TOCSIN, a method for zero-shot detection of text generated by large language models (LLMs) using a feature called token cohesiveness. The proposed detection paradigm demonstrates effective performance across various datasets by enhancing existing zero-shot detectors with a practical evaluation approach.

https://arxiv.org/abs/2409.16914

974. Symbolic Working Memory Enhances Language Models for Complex Rule Application

This paper proposes augmenting large language models with an external working memory to enhance their performance in multi-step deductive reasoning involving complex rule application. The framework allows for iterative symbolic rule grounding and LLM-based rule implementation, demonstrating effectiveness across various scenarios and settings.

http://arxiv.org/abs/2408.13654v1

1001. Tree of Problems: Improving structured problem solving with compositionality

This paper presents Tree of Problems (ToP), a structured approach aimed at enhancing problem-solving capabilities in complex reasoning tasks using compositionality. The results demonstrate that ToP outperforms existing approaches like Tree of Thoughts and Chain-of-Thought prompting, particularly in tasks that can be broken down into identical subtasks.

https://arxiv.org/abs/2410.06634

1010. Dual-Space Knowledge Distillation for Large Language Models

This paper presents a dual-space knowledge distillation (DSKD) framework, which aims to improve the knowledge transfer from large language models (LLMs) to smaller models by unifying their output spaces. The proposed method enhances the KD process between models with different vocabularies and achieves better performance than existing frameworks on instruction-following tasks.

https://arxiv.org/abs/2406.17328

1036. Instruction Matters, a Simple yet Effective Task Selection Approach in Instruction Tuning for Specific Tasks

This paper presents a method for task selection in instruction tuning that optimizes performance for specific tasks by leveraging instruction information alone. The approach is shown to be more efficient than traditional task selection methods, leading to significant improvements in model performance on various benchmarks.

https://arxiv.org/abs/2404.16418

1074. DEM: Distribution Edited Model for Training with Mixed Data Distributions

The paper presents the Distribution Edited Model (DEM), which offers an efficient method for optimizing training with mixed data distributions for multi-task and instruction-following models. DEM achieves improved performance on various benchmarks while being significantly cheaper and more flexible than traditional data mixing methods.

https://arxiv.org/abs/2406.15570

1081. GPT vs RETRO: Exploring the Intersection of Retrieval and Parameter-Efficient Fine-Tuning

This paper explores the effectiveness of Parameter-Efficient Fine-Tuning (PEFT) methods and Retrieval-Augmented Generation (RAG) on both GPT and RETRO models, highlighting their performance across various configurations. Results indicate that while RETRO excels in zero-shot scenarios, GPT has superior performance potential when fine-tuned with PEFT techniques, with an optimal balance found in 8B parameter models.

https://arxiv.org/abs/2407.04528

1095. TempoFormer: A Transformer for Temporally-aware Representations in Change Detection

TempoFormer is a transformer-based model designed for dynamic representation learning that incorporates temporal awareness to better understand linguistic changes over time. It achieves state-of-the-art performance on three real-time change detection tasks by jointly modeling context and temporal dynamics.

http://www.arxiv.org/abs/2408.15689

1107. Jump Starting Bandits with LLM-Generated Prior Knowledge

This paper demonstrates how integrating Large Language Models (LLMs) with Contextual Multi-Armed Bandits can significantly improve personalized recommendation systems by simulating human behavior. The authors propose a novel initialization algorithm that uses LLM-generated datasets to jump-start the learning process and reduce online learning regret and data-gathering costs.

http://arxiv.org/abs/2406.19317v2

1129. Towards Aligning Language Models with Textual Feedback

The paper presents ALT, a method for aligning language models with textual feedback from users, aiming to leverage text's expressiveness for richer and more effective model alignment. The approach shows improved performance in tasks like toxicity reduction and summarization while requiring minimal tuning compared to traditional reinforcement learning methods.

https://arxiv.org/abs/2407.16970

1180. Adapters Mixup: Mixing Parameter-Efficient Adapters to Enhance the Adversarial Robustness of Fine-tuned Pre-trained Text Classifiers

The paper introduces AdpMixup, a novel approach that combines fine-tuning via adapters with adversarial augmentation through mixup to improve the robustness of fine-tuned pre-trained language models against adversarial attacks. The method dynamically utilizes known adversarial examples during prediction, achieving an optimal balance between training efficiency and resistance to attacks across multiple tasks and models.

https://arxiv.org/abs/2401.10111

1183. GottBERT: a pure German Language Model

GottBERT is a newly introduced German language model that outperforms existing German and multilingual models in various NLP tasks, specifically Named Entity Recognition and text classification. This model was pre-trained using the German portion of the OSCAR dataset and is intended to advance the field of German NLP by providing a strong baseline for further research.

https://arxiv.org/abs/2012.02110

1206. Preference-Guided Reflective Sampling for Aligning Language Models

This paper presents Preference-Guided Reflective Sampling (PRS), a novel sampling method designed to improve alignment between large language models (LLMs) and human user preferences. PRS employs a tree-based generation framework and adaptive self-refinement techniques to enhance sampling efficiency and response quality compared to traditional random sampling methods.

https://arxiv.org/abs/2408.12163

1237. Evaluating Concurrent Robustness of Language Models Across Diverse Challenge Sets

This study investigates the robustness of language models to input perturbations, aiming to improve their performance through effective fine-tuning strategies. It employs a comprehensive methodology to assess and enhance model reliability across different datasets and perturbations, especially focusing on large language models using chain of thought prompting.

https://arxiv.org/abs/2311.08662

1247. CELLO: Causal Evaluation of Large Vision-Language Models

This paper introduces CELLO, a novel dataset comprising 14,094 causal questions designed to improve the understanding of causality in large vision-language models (LVLMs). The study highlights that existing LVLMs face challenges with causal reasoning tasks despite advances, and proposes CELLO-CoT, a prompting strategy to enhance performance.

https://arxiv.org/abs/2406.19131

1253. Language Models as Compilers: Simulating Pseudocode Execution Improves Algorithmic Reasoning in Language Models

This paper introduces Think-and-Execute, a framework that improves algorithmic reasoning in language models by simulating pseudocode execution. It effectively discovers task-level logic and enhances reasoning compared to instance-specific methods, demonstrating that pseudocode can guide language models better than natural language instructions.

https://arxiv.org/abs/2404.02575

1261. Knowledge Graph Enhanced Large Language Model Editing

This paper introduces GLAME, a novel model editing method that enhances the capabilities of large language models (LLMs) by incorporating knowledge graphs. The proposed approach improves the generalization ability of LLMs by effectively tracking and integrating knowledge changes resulting from edits, as demonstrated through experiments on GPT-J and GPT-2 XL.

https://arxiv.org/abs/2402.13593

229. Dancing in Chains: Reconciling Instruction Following and Faithfulness in Language Models

This paper investigates the trade-off between instruction following and faithfulness in language models, revealing how tuning for one often degrades the other. The authors propose a method called ReSet, which improves outcomes by using less data effectively, addressing discrepancies in alignment training for language models.

https://arxiv.org/abs/2407.21417

252. ARES: Alternating Reinforcement Learning and Supervised Fine-Tuning for Enhanced Multi-Modal Chain-of-Thought Reasoning Through Diverse AI Feedback

The paper proposes a two-stage algorithm, ARES, which enhances multi-modal chain-of-thought reasoning by alternating between reinforcement learning and supervised fine-tuning. It demonstrates improved performance on datasets by utilizing detailed AI feedback for individual sentence contributions, resulting in better inference accuracy and rationale reasoning when compared to baseline models.

https://arxiv.org/abs/2407.00087

267. How Do Humans Write Code? Large Models Do It the Same Way Too

This paper introduces Human-Think Language (HTL), a method that integrates Program-of-Thought (PoT) and Chain-of-Thought (CoT) approaches for improved mathematical reasoning in large language models. HTL demonstrates significant enhancements in reasoning accuracy and transferability across various datasets, making it a promising framework for unified reasoning tasks.

https://arxiv.org/abs/2402.15729

344. Mitigating Frequency Bias and Anisotropy in Language Model Pre-Training with Syntactic Smoothing

This paper presents a method for mitigating frequency bias and anisotropy in language models during pre-training by inducing a syntactic prior over token representations. The proposed Syntactic Smoothing technique enhances performance on rare tokens and reduces the clustering of representations in high-dimensional space.

https://arxiv.org/abs/2410.11462

417. Position Engineering: Boosting Large Language Models through Positional Information Manipulation

This paper introduces position engineering, a novel technique aimed at enhancing the performance of large language models (LLMs) by manipulating positional information in prompts without altering the text itself. The authors demonstrate that position engineering significantly improves task performance in two scenarios: retrieval-augmented generation (RAG) and in-context learning (ICL).

https://arxiv.org/abs/2404.11216

450. Mixture-of-Subspaces in Low-Rank Adaptation

This paper introduces a new method called Mixture-of-Subspaces LoRA (MoSLoRA) for Low-Rank Adaptation which enhances performance across various tasks by mixing two subspaces. MoSLoRA has been found to outperform the traditional LoRA on different modalities, demonstrating its robustness and effectiveness.

https://arxiv.org/abs/2406.11909

521. LLM See, LLM Do: Leveraging Active Inheritance to Target Non-Differentiable Objectives

This paper investigates how synthetic data impacts large language models (LLMs) by analyzing the passive inheritance of model properties from the data source. It proposes the concept of active inheritance to intentionally guide the data generation process toward desired non-differentiable objectives, such as enhanced lexical diversity or reduced toxicity.

https://arxiv.org/abs/2407.01490

777. How Do Your Code LLMs perform? Empowering Code Instruction Tuning with Really Good Data

This paper investigates the construction of better code instruction tuning datasets and highlights the issue of data leakage affecting model performance on various benchmarks. It proposes a pruning strategy for selecting high-quality samples and presents XCoder, a family of models that achieves state-of-the-art performance with reduced training data.

https://www.arxiv.org/abs/2409.03810

787. Mixture-of-Skills: Learning to Optimize Data Usage for Fine-Tuning Large Language Models

This paper introduces Mixture-of-Skills (MoS), a reinforcement learning framework designed to optimize the use of data while fine-tuning large language models (LLMs) for diverse skills. The proposed method dynamically adjusts dataset focus based on the model's learning state, improving overall performance through effective dataset rebalancing.

https://arxiv.org/abs/2406.08811

1008. Which Programming Language and What Features at Pre-training Stage Affect Downstream Logical Inference Performance?

This research investigates the effect of programming languages and their features during the pre-training stage of language models on their performance in logical inference tasks. The findings show that models pre-trained with programming languages generally outperform those trained with natural languages, particularly in following instructions and logical reasoning abilities.

http://arxiv.org/abs/2410.06735

1150. Rethinking the Role of Proxy Rewards in Language Model Alignment

This paper investigates the role of proxy rewards in aligning Large Language Models (LLMs) with human values through a white-box approach. By implementing reverse reward engineering, the authors aim to create a reward function that effectively emulates human feedback, showing promising results against existing models without needing an explicit human feedback dataset.

http://arxiv.org/abs/2402.03469v3

19. RoTBench: A Multi-Level Benchmark for Evaluating the Robustness of Large Language Models in Tool Learning

RoTBench introduces a comprehensive benchmark for assessing the robustness of large language models in tool learning, evaluating their performance across various levels of environmental noise. The findings highlight a significant drop in model performance under noisy conditions and propose RoTTuning to improve robustness through diverse training environments.

https://arxiv.org/abs/2401.08326

32. On Sensitivity of Learning with Limited Labelled Data to the Effects of Randomness: Impact of Interactions and Systematic Choices

This paper investigates how controlled randomness factors affect performance when learning from limited labeled data in NLP tasks. It reveals that ignoring interactions between these randomness factors can lead to inconsistent results and provides insights into systematic choices that influence the effects of randomness.

https://arxiv.org/abs/2402.12817

68. Rethinking Pruning Large Language Models: Benefits and Pitfalls of Reconstruction Error Minimization

This paper investigates a new approach to pruning large language models (LLMs) by dividing them into submodels and sequentially pruning them while minimizing reconstruction error. However, it reveals that while reducing reconstruction error can improve performance, it might lead to overfitting, suggesting a need for better calibration data strategies to balance both reconstruction and generalization.

https://arxiv.org/abs/2406.15524

79. Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment

This paper investigates a method for zero-shot cross-lingual alignment of language models using a reward model trained on preference data in one language, which can then be applied to other languages. The results demonstrate that this approach enhances model alignment effectiveness in summarization and dialogue generation tasks, with improved human preference ratings for aligned models.

https://arxiv.org/abs/2404.12318

90. Tracking the perspectives of interacting language models

This paper introduces a method for representing the perspectives of individual large language models (LLMs) within a network of interacting LLMs to study information diffusion. It formalizes the concept of a communication network of LLMs and explores implications for the future of language model interactions as they become more integrated into society.

https://arxiv.org/abs/2406.11938

139. Self-Refine Instruction-Tuning for Aligning Reasoning in Language Models

This paper presents the Self-refine Instruction-tuning method to enhance the reasoning abilities of Smaller Language Models (SLMs) by leveraging demonstrations from robust Large Language Models (LLMs). The approach involves a two-stage process of instruction-tuning followed by self-refinement through preference optimization, resulting in significantly improved performance on reasoning tasks.

https://arxiv.org/abs/2405.00402

185. DocKD: Knowledge Distillation from LLMs for Open-World Document Understanding Models

The paper presents DocKD, a framework for improving visual document understanding (VDU) models by distilling knowledge from large language models (LLMs) using external document knowledge. Experimental results indicate that VDU models trained with DocKD-generated data perform comparably to those trained on human-annotated data in specific tasks, and excel in generalization on out-of-domain tasks.

https://arxiv.org/abs/2410.03061

192. How do Large Language Models Learn In-Context? Query and Key Matrices of In-Context Heads are Two Towers for Metric Learning

This paper investigates how in-context learning (ICL) in large language models works specifically for sentence classification tasks, revealing the critical role of 'in-context heads' in the learning process. The authors propose a hypothesis about the learning of similarity metrics within these heads and suggest methods to reduce biases observed in ICL performance.

http://arxiv.org/abs/2402.02872v3

230. Where is the signal in tokenization space?

This paper explores the concept of non-canonical tokenizations in Large Language Models (LLMs) and reveals that numerous potential tokenizations exist for the same input string. It demonstrates that by aggregating the probabilities of these non-canonical tokenizations, significant performance improvements can be achieved across various LLM evaluation benchmarks.

https://arxiv.org/abs/2408.08541

295. Unveiling the Lexical Sensitivity of LLMs: Combinatorial Optimization for Prompt Enhancement

This paper reveals that large language models (LLMs) are overly sensitive to minor lexical variations in task instructions, affecting their performance on downstream tasks. The authors propose a black-box combinatorial optimization framework for prompt lexical enhancement (COPLE) that mitigates this sensitivity and improves task performance through iterative lexical optimization based on feedback from proxy tasks.

https://arxiv.org/abs/2405.20701

303. Safely Learning with Private Data: A Federated Learning Framework for Large Language Model

This paper presents a Federated Learning framework, FL-GLM, for training large language models (LLMs) using private data while addressing privacy and efficiency concerns. By implementing techniques to prevent data leakage and enhance training efficiency, FL-GLM achieves performance comparable to centralized models while maintaining security in a federated context.

https://arxiv.org/abs/2406.14898

307. EXPLORA: Efficient Exemplar Subset Selection for Complex Reasoning

This paper introduces EXPLORA, an algorithm designed for efficient static exemplar subset selection to improve complex reasoning over text and hybrid sources. By significantly reducing the required number of calls to large language models, EXPLORA enhances performance while maintaining fast inference times.

http://arxiv.org/abs/2411.03877v1

346. Boosting Scientific Concepts Understanding: Can Analogies from Teacher Models Empower Student Models?

This paper explores how teacher language models can generate analogies that help student language models understand scientific concepts more effectively. The findings indicate that both teacher-generated and student-generated analogies can enhance performance in scientific question answering, revealing the potential for self-directed learning through analogical reasoning.

https://arxiv.org/abs/2406.11375

349. Investigating Mysteries of CoT-Augmented Distillation

This paper explores the mechanisms behind the performance improvements observed in model distillation through the use of 'chain of thought' (CoT) sequences. It finds that the positioning of CoT sequences influences results and that coherent reasoning is not necessary for improvements, suggesting that a few key tokens can suffice for effective training.

https://arxiv.org/abs/2406.14511

359. Focused Large Language Models are Stable Many-Shot Learners

This paper proposes FocusICL, a training-free method that improves many-shot learning in large language models by filtering unimportant content and using hierarchical attention. Experiments show that FocusICL enhances performance in in-context learning scenarios by ensuring models maintain focus on key information during task adaptation.

https://arxiv.org/abs/2408.13987

381. Can Large Language Models Learn Independent Causal Mechanisms?

This paper investigates the use of Independent Causal Mechanisms (ICMs) in Large Language Models (LLMs) to improve their robustness against distribution shifts and enhance generalization ability. The authors propose a new LLM architecture that incorporates multiple sparsely interacting modules, demonstrating that applying causal constraints can lead to better performance on abstract reasoning tasks.

https://arxiv.org/abs/2402.02636

408. Python is Not Always the Best Choice: Embracing Multilingual Program of Thoughts

This paper critiques the singular reliance on Python for the Program of Thoughts (PoT) approach, revealing that varying programming languages can lead to better solutions across different tasks and models. It introduces a new method called MultiPoT that leverages multiple programming languages, resulting in higher performance compared to using Python alone.

https://arxiv.org/abs/2402.10691

424. Gold Panning in Vocabulary: An Adaptive Method for Vocabulary Expansion of Domain-Specific LLMs

This paper introduces VEGAD, an adaptive method for identifying a valuable subset of vocabulary to enhance the performance of domain-specific large language models (LLMs). It demonstrates that optimizing vocabulary selection can significantly improve results on both specialized and general tasks, validated through experiments on Chinese datasets.

https://arxiv.org/abs/2410.01188

429. More Than Catastrophic Forgetting: Integrating General Capabilities For Domain-Specific LLMs

This paper addresses the challenge of catastrophic forgetting in Large Language Models (LLMs) when they are fine-tuned on domain-specific tasks, introducing a concept called General Capabilities Integration (GCI). The study proposes a method to harmonize general competencies and domain-specific knowledge using a modified attention mechanism and demonstrates its effectiveness through experiments in the legal domain.

https://arxiv.org/abs/2405.17830

444. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?

This paper investigates how fine-tuning large language models (LLMs) on new factual knowledge affects their performance and tendency to produce hallucinations. The findings suggest that while fine-tuned models can learn new knowledge, they become more prone to generating incorrect information over time, emphasizing the risks of introducing new knowledge through fine-tuning.

https://arxiv.org/abs/2405.05904

464. FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models

The proposed FedKIM approach integrates lightweight local models to inject healthcare knowledge into a centralized medical foundation model within a federated learning framework, preserving data privacy. This method enhances the model's capabilities to handle complex medical tasks across various modalities without direct access to sensitive data.

https://arxiv.org/abs/2408.10276

465. Retrieved In-Context Principles from Previous Mistakes

This paper introduces Retrieved In-Context Principles (RICP), a method designed to enhance Large Language Models (LLMs) by analyzing mistakes made in task performance, clustering them to derive insights for error prevention. RICP aims to improve the customization and error coverage of principles used during inference by retrieving relevant mistakes for generating question-specific guidance.

http://arxiv.org/abs/2407.05682

481. The Best Defense is Attack: Repairing Semantics in Textual Adversarial Examples

This paper introduces a novel approach called Reactive Perturbation Defocusing (Rapid) aimed at repairing semantics in adversarial examples generated against pre-trained language models. The authors demonstrate that Rapid effectively identifies and repairs the semantics of adversarial examples in various attack scenarios through extensive experimental results.

https://arxiv.org/abs/2305.04067

509. Can Active Label Correction Improve LLM-based Modular AI Systems?

This paper investigates the use of Active Label Correction (ALC) to improve LLM-based modular AI systems by reducing noise in training datasets using human feedback. The findings demonstrate that ALC can enhance performance significantly with fewer training examples across various NLP tasks.

https://arxiv.org/abs/2401.05467

550. Can Transformer Language Models Learn n-gram Language Models?

This paper investigates the capacity of transformers to learn different types of n-gram language models, linking theoretical capabilities with empirical results. The findings suggest that while classic techniques may outperform transformers in certain conditions, transformers show superior performance under specific parameter-sharing scenarios.

https://arxiv.org/abs/2410.03001

561. CommonIT: Commonality-Aware Instruction Tuning for Large Language Models via Data Partitions

This paper introduces CommonIT, a novel instruction tuning strategy that clusters instruction datasets into distinct groups to enhance the command adherence of large language models (LLMs). By ensuring training mini-batches consist solely of data from a single group, CommonIT improves the instruction-following capabilities of LLMs, resulting in an average performance increase across various domains and tasks.

http://arxiv.org/abs/2410.03077v1

571. I Learn Better If You Speak My Language: Understanding the Superior Performance of Fine-Tuning Large Language Models with LLM-Generated Responses

This paper investigates the performance improvements in fine-tuning large language models using responses generated by other LLMs as opposed to human-generated responses, particularly in reasoning tasks. It finds that the model's inherent familiarity with LLM-generated content enhances learning efficiency, which is supported by lower perplexity measures.

https://arxiv.org/abs/2402.11192

606. Stable Language Model Pre-training by Reducing Embedding Variability

This paper discusses the importance of stable pre-training for language models and proposes using Token Embedding Variability as a cost-efficient measure of stability. It introduces Multi-head Low-Rank Attention as a solution to mitigate gradient explosion, presenting empirical results from GPT-2 that show improvements in model stability and performance.

https://arxiv.org/abs/2409.07787

617. Paraphrase Types Elicit Prompt Engineering Capabilities

This paper investigates how variations in linguistic expression of prompts affect the performance of language models, specifically analyzing different paraphrase types. Results demonstrate that adapting prompts using certain linguistic features can improve model performance across various tasks, with notable gains linked to morphology and lexicon changes.

http://arxiv.org/abs/2406.19898v3

629. Code Prompting Elicits Conditional Reasoning Abilities in Text+Code LLMs

This paper presents a novel technique called code prompting that enhances conditional reasoning abilities in language models by transforming natural language problems into code. The results show significant performance improvements across various models and datasets, highlighting the importance of code formatting for effective reasoning in LLMs.

https://arxiv.org/abs/2401.10065

646. Puzzle Solving using Reasoning of Large Language Models: A Survey

This survey explores the capabilities of Large Language Models (LLMs) in solving puzzles, distinguishing between rule-based and rule-less puzzles. It critiques LLMs' performance against human-like reasoning, highlighting significant challenges and the need for novel strategies and datasets to improve their logical reasoning capacity.

https://arxiv.org/abs/2402.11291

654. What Are the Odds? Language Models Are Capable of Probabilistic Reasoning

This paper evaluates the probabilistic reasoning capabilities of language models (LMs) across various statistical tasks, including estimating percentiles and drawing samples. It highlights the potential of LMs to infer distributions, especially when provided with relevant context and benchmark datasets for systematic evaluation.

https://arxiv.org/abs/2406.12830

667. TheoremLlama: Transforming General-Purpose LLMs into Lean4 Experts

TheoremLlama is a framework designed to enhance the capabilities of general-purpose LLMs in formal theorem proving using Lean4. It introduces innovative methods for dataset generation, curriculum learning, and proof writing, demonstrating improved performance compared to existing LLM models like GPT-4.

http://arxiv.org/abs/2407.03203v2

680. ControlMath: Controllable Data Generation Promotes Math Generalist Models

This paper presents ControlMath, an iterative method for generating diverse math word problems to enhance the performance of generalist models in mathematical reasoning. By combining generated data with existing datasets, the approach demonstrates improved generalization capabilities across various problem domains.

https://arxiv.org/abs/2409.15376

693. Decoding with Limited Teacher Supervision Requires Understanding When to Trust the Teacher

This paper explores how small-scale large language models (LLMs) can effectively use limited teacher supervision to enhance their generative capabilities. It presents an algorithm that adaptively adjusts trust in LLM predictions based on the confidence level of the small-scale LLM, demonstrating improved decoding strategies across various models and datasets.

http://arxiv.org/abs/2406.18002v2

695. The Mystery of the Pathological Path-star Task for Language Models

The authors introduce the path-star task, which exposes limitations of language models in generating paths within a graph structure. They propose a regularization method and demonstrate that the task can be solved effectively with the right training approaches and model configurations.

https://arxiv.org/abs/2410.13779

744. How to Leverage Demonstration Data in Alignment for Large Language Model? A Self-Imitation Learning Perspective

This paper presents a new framework called Generalized Self-Imitation Learning (GSIL) for aligning large language models with offline demonstration data while avoiding complex adversarial training. The framework improves efficiency and performance in various benchmarks, demonstrating significant advancements in the field of imitation learning for large-scale NLP models.

https://arxiv.org/abs/2410.10093

748. DA-Code: Agent Data Science Code Generation Benchmark for Large Language Models

The paper introduces DA-Code, a benchmark designed to evaluate large language models (LLMs) on complex agent-based data science tasks that necessitate advanced coding skills and data processing abilities. Despite the development of a baseline model that outperforms existing frameworks, current leading LLMs only achieve a modest accuracy of 30.5%, highlighting the need for further improvements.

https://arxiv.org/abs/2410.07331

761. Householder Pseudo-Rotation: A Novel Approach to Activation Editing in LLMs with Direction-Magnitude Perspective

This paper presents Householder Pseudo-Rotation (HPR), a new method for editing the internal activations of large language models to enhance their performance while maintaining consistency in activation magnitudes. It addresses the limitations of existing activation editing techniques by focusing on the directions and magnitudes of activations, demonstrating improvements in safety benchmarks.

https://arxiv.org/abs/2409.10053

765. Consecutive Batch Model Editing with HooK Layers

The paper introduces CoachHooK, a novel model editing method that effectively supports both consecutive and batch editing of model behavior while being memory-friendly. CoachHooK requires minimal memory and has shown superior performance in experimental results compared to existing methods in various editing scenarios.

https://arxiv.org/abs/2403.05330

788. MolTRES: Improving Chemical Language Representation Learning for Molecular Property Prediction

MolTRES is a novel framework designed to enhance chemical language representation learning by addressing issues of overfitting and scalability in molecular property prediction. The framework utilizes generator-discriminator training and incorporates knowledge from scientific literature to improve the understanding of molecular structures.

https://arxiv.org/abs/2408.01426

817. Learn Beyond The Answer: Training Language Models with Reflection for Mathematical Reasoning

This paper presents a novel training technique called reflective augmentation, aimed at improving language models' understanding and performance in mathematical reasoning tasks through problem reflection. The proposed method enhances existing data augmentation strategies and has been shown to benefit performance in complex scenarios requiring deeper cognitive processing.

http://arxiv.org/abs/2406.12050v3

826. Commonsense Knowledge Editing Based on Free-Text in LLMs

This paper addresses the challenges of commonsense knowledge editing in large language models by introducing a new approach that focuses on free-text. The proposed Dynamics-aware Editing Method (DEM) effectively localizes and updates commonsense knowledge in LLMs, demonstrating superior performance in experimental settings.

https://arxiv.org/abs/2410.23844

834. Shortcuts Arising from Contrast: Towards Effective and Lightweight Clean-Label Attacks in Prompt-Based Learning

This paper discusses the vulnerabilities of prompt-based learning in pretrained language models, specifically regarding backdoor clean-label attacks. It introduces a method called Contrastive Shortcut Injection (CSI) which improves attack stealthiness and effectiveness at low poisoning rates while exploring the trade-offs involved.

https://arxiv.org/abs/2404.00461

857. Not Everything is All You Need: Toward Low-Redundant Optimization for Large Language Model Alignment

This paper presents a method called ALLO for optimizing large language model (LLM) alignment by reducing redundancy in the model's parameters. The proposed method improves convergence and performance by focusing on the most relevant neurons and using a structured learning process.

https://arxiv.org/abs/2406.12606

865. Exploring the Learning Capabilities of Language Models using LEVERWORLDS

This paper investigates the interplay between learning general structure rules and specific properties in various learning methods, especially focusing on sample efficiency. Through the new framework LeverWorlds, controlled experiments reveal that standard learning algorithms outperform Transformer models in sample efficiency for certain tasks, though Transformers succeed overall.

https://arxiv.org/abs/2410.00519

868. DogeRM: Equipping Reward Models with Domain Knowledge through Model Merging

DogeRM is a framework that enhances reward models in reinforcement learning from human feedback by integrating domain-specific knowledge through a model merging process. Experiments show that this approach improves performance across various benchmarks, indicating the potential for better alignment of large language models.

https://arxiv.org/abs/2407.01470

888. A Comparison of Language Modeling and Translation as Multilingual Pretraining Objectives

This paper addresses the effectiveness of multilingual pretraining objectives by comparing language modeling and translation in a controlled environment. It highlights that model architecture significantly influences the optimal pretraining objective and concludes that multilingual translation can be highly effective when applied appropriately.

https://arxiv.org/abs/2407.15489

892. Mitigating Training Imbalance in LLM Fine-Tuning via Selective Parameter Merging

This paper addresses training imbalances in supervised fine-tuning processes for Large Language Models (LLMs) by introducing a novel technique called 'parameter-selection merging.' The proposed method outperforms traditional approaches and is validated through comprehensive analysis on multiple datasets.

https://arxiv.org/abs/2410.03743

903. CMR Scaling Law: Predicting Critical Mixture Ratios for Continual Pre-training of Language Models

The paper introduces the Critical Mixture Ratio (CMR) for continually pre-training language models, establishing a power-law relation between training loss, mixture ratio, and token scale. It provides practical guidelines to optimize training resources, balancing general and domain-specific abilities for enhanced model performance.

https://arxiv.org/abs/2407.17467

905. Rationale-Aware Answer Verification by Pairwise Self-Evaluation

This paper addresses the limitations of current answer verification methods for solutions generated by large language models by focusing on the validity of rationales rather than just the correctness of final answers. It introduces REPS, a method for enhancing verifier training through pairwise self-evaluation, leading to improved performance on reasoning benchmarks.

http://arxiv.org/abs/2410.04838

915. Exploring the Compositional Deficiency of Large Language Models in Mathematical Reasoning Through Trap Problems

This paper investigates the compositional abilities of large language models (LLMs) in mathematical reasoning through a newly created dataset containing logical traps. The findings reveal that LLMs struggle to systematically combine mathematical knowledge with trap-related knowledge, indicating compositionality remains a significant challenge for these models.

https://arxiv.org/abs/2405.06680

977. Mentor-KD: Making Small Language Models Better Multi-step Reasoners

This paper introduces Mentor-KD, a method for enhancing smaller language models' reasoning capabilities by addressing issues in knowledge distillation from larger models. The approach involves using a mentor model to provide additional Chain-of-Thought annotations and soft labels, demonstrating effectiveness across various reasoning tasks.

https://arxiv.org/abs/2410.09037

992. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate

This paper introduces the Multi-Agent Debate (MAD) framework aimed at mitigating the Degeneration-of-Thought (DoT) problem in large language models (LLMs) by promoting divergent thinking during problem-solving. The MAD framework involves multiple agents engaging in a debate, ultimately enhancing the cognitive abilities of LLMs on complex reasoning tasks through iterative feedback.

https://arxiv.org/abs/2305.19118

1004. Is C4 Dataset Enough for Pruning? An Investigation of Calibration Data for LLM Pruning

This paper investigates the effectiveness of various calibration datasets for pruning large language models (LLMs) and finds that the widely used C4 dataset is not the optimal choice. The study reveals that the selection of calibration data significantly impacts pruning performance and highlights that arithmetic datasets can sometimes outperform traditional pre-training datasets.

http://arxiv.org/abs/2410.07461

1108. Adaptation Odyssey in LLMs: Why Does Additional Pretraining Sometimes Fail to Improve?

This paper explores why additional pretraining of large language models (LLMs) may not always yield performance improvements, highlighting a counterintuitive situation where it can degrade model effectiveness. The authors present empirical observations that relate the performance degradation to the similarity between the additional and original pretraining datasets.

http://arxiv.org/abs/2410.05581v2

1126. CmdCaliper: A Semantic-Aware Command-Line Embedding Model and Dataset for Security Research

This paper introduces CmdCaliper, a command-line embedding model, and a new dataset called CyPHER developed for cybersecurity research. The study highlights its effectiveness in handling command-line semantic similarity, outperforming larger existing models while also addressing data generation challenges in the domain.

https://arxiv.org/abs/2411.01176

1164. Mixture-of-Modules: Reinventing Transformers as Dynamic Assemblies of Modules

The paper introduces a novel architecture called mixture-of-modules (MoM) that rethinks the conventional depth order of Transformers, allowing any layer to process tokens based on its capabilities. MoM not only serves as a flexible framework for Transformers with demonstrated performance improvements over vanilla models, but also reduces redundancy in Transformer parameterization, achieving more efficient computations.

http://arxiv.org/abs/2407.06677v1

1196. Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models

This paper introduces a synthetic knowledge ingestion method named Ski, designed to enhance the capabilities of large language models (LLMs) in refining and injecting knowledge from various external sources. The method is empirically tested across different domains, showing significant improvements in knowledge representation and factual accuracy in question-answering tasks.

https://arxiv.org/abs/2410.09629

1204. Evaluating Large Language Models on Time Series Feature Understanding: A Comprehensive Taxonomy and Benchmark

This paper introduces a framework to evaluate Large Language Models (LLMs) on their understanding of time series data, establishing a taxonomy of time series features. By creating a diverse dataset and conducting experiments, the authors assess the capabilities and limitations of LLMs in analyzing time series across various domains.

https://arxiv.org/abs/2404.16563

1210. Rebuilding ROME : Resolving Model Collapse during Sequential Model Editing

This paper addresses the limitations of the Rank-One Model Editing (ROME) method, specifically the issue of disabling edits that cause model collapse during sequential editing. The authors present a new implementation called r-ROME that resolves these issues and improves the stability and performance of model editing tasks.

https://arxiv.org/abs/2403.07175

1240. On the Fragility of Active Learners for Text Classification

This paper discusses the effectiveness and reliability of active learning (AL) techniques in text classification, emphasizing the lack of guidance for practitioners in selecting appropriate AL methods. It presents a thorough evaluation of various AL techniques, leading to insights on their performance compared to random sampling and the role of different factors in their success.

https://arxiv.org/abs/2403.15744

1254. Coffee-Gym: An Environment for Evaluating and Improving Natural Language Feedback on Erroneous Code

This paper introduces Coffee-Gym, a reinforcement learning environment designed to improve models that give feedback on code editing. It highlights the importance of high-quality datasets for training and demonstrates that its feedback models significantly enhance the performance of open-source code language models.

https://arxiv.org/abs/2409.19715

1266. Filtered Direct Preference Optimization

This paper investigates the impact of text quality on models optimized with Direct Preference Optimization (DPO) in Reinforcement Learning from Human Feedback (RLHF). It introduces filtered direct preference optimization (fDPO), which improves model performance by filtering out lower-quality texts from the preference dataset during training.

https://arxiv.org/abs/2404.13846

1267. Instruction Fine-Tuning: Does Prompt Loss Matter?

This study analyzes the impact of prompt loss token weights on supervised instruction fine-tuning (SIFT), revealing that small non-zero weights enhance performance for short-completion tasks. The findings indicate a significant negative relationship between model performance and larger weights, recommending that SIFT providers maintain the option to use prompt loss token weights.

https://arxiv.org/abs/2401.13586

332. Context-Aware Assistant Selection for Improved Inference Acceleration with Large Language Models

This paper addresses the challenge of selecting draft models to assist large language models in order to improve their inference speed and efficiency. It presents a solution using contextual bandits to optimize model selection without needing prior knowledge of the draft models' construction.

https://arxiv.org/abs/2408.08470

822. A Fundamental Trade-off in Aligned Language Models and its Relation to Sampling Adaptors

This paper investigates the correlation between string quality as assessed by humans and its associated probability in language models, particularly those aligned with human preferences. It reveals a trade-off between the average reward and average log-likelihood when sampling from these aligned models, highlighting the influence of sampling adaptors on this balance.

https://arxiv.org/abs/2406.10203

990. Data Contamination Can Cross Language Barriers

This paper investigates a form of data contamination in large language models (LLMs) that can cross language barriers, which traditional detection methods fail to catch. The authors introduce a method to detect this contamination by analyzing LLM performance changes under altered benchmark conditions, demonstrating both the vulnerabilities in current systems and the potential for improved LLM multilingual capabilities.

https://arxiv.org/abs/2406.13236

190. SEEKR: Selective Attention-Guided Knowledge Retention for Continual Learning of Large Language Models

204. From Bottom to Top: Extending the Potential of Parameter Efficient Fine-Tuning

313. Adaption-of-Thought: Learning Question Difficulty Improves Large Language Models for Reasoning

328. Consistent Bidirectional Language Modelling: Expressive Power and Representational Conciseness

563. Breaking ReLU Barrier: Generalized MoEfication for Dense Pretrained Models

1161. Contrastive Classification via Linear Layer Extrapolation

1220. Dynamic Rewarding with Prompt Optimization Enables Tuning-free Self-Alignment of Language Models

515. Enhancing Reinforcement Learning with Intrinsic Rewards from Language Model Critique

541. Birdie: Advancing State Space Models with a Minimalist Architecture and Novel Pre-training Objectives

580. CARER - ClinicAl Reasoning-Enhanced Representation for Temporal Health Risk Prediction

161. MoDULA: Mixture of Domain-Specific and Universal LoRA for Multi-Task Learning

215. Performance-Guided LLM Knowledge Distillation for Efficient Text Classification at Scale

273. Bayesian Calibration of Win Rate Estimation with LLM Evaluators

298. Multi-Granularity History and Entity Similarity Learning for Temporal Knowledge Graph Reasoning

333. Teaching Small Language Models Reasoning through Counterfactual Distillation

666. Applying Contrastive Learning to Code Vulnerability Type Classification

678. Empowering Multi-step Reasoning across Languages via Program-Aided Language Models

775. Structured Optimal Brain Pruning for Large Language Models

832. MoCoKGC: Momentum Contrast Entity Encoding for Knowledge Graph Completion

852. Improving Discriminative Capability of Reward Models in RLHF Using Contrastive Learning

876. Revisiting Supervised Contrastive Learning for Microblog Classification

894. FAME: Factual Multi-task Model Editing Benchmark

907. IM-BERT: Enhancing Robustness of BERT through the Implicit Euler Method

967. Low-rank Subspace for Binding in Large Language Models

980. Can Large Language Models Enhance Predictions of Disease Progression? Investigating Through Disease Network Link Prediction

1000. Quantum Recurrent Architectures for Text Classification

1130. ATPO: Automatic Tree-Structured Prompt Optimization

1189. Enhancing Language Model Alignment: A Confidence-Based Approach to Label Smoothing

1205. Can LLMs Learn Uncertainty on Their Own? Expressing Uncertainty Effectively in A Self-Training Manner

1233. HCEG: Improving the Abstraction Ability of Language Models with Hierarchical Conceptual Entailment Graphs

160. Take Off the Training Wheels! Progressive In-Context Learning for Effective Alignment

163. PhiloGPT: A Philology-Oriented Large Language Model for Ancient Chinese Manuscripts with Dunhuang as Case Study

593. When Generative Adversarial Networks Meet Sequence Labeling Challenges

676. TL-CL: Task And Language Incremental Continual Learning

689. A Coordinate System for In-Context Learning

927. Explicit Memory Learning with Expectation Maximization

NLP Applications

63. A Reflective LLM-based Agent to Guide Zero-shot Cryptocurrency Trading

This paper presents CryptoTrade, an LLM-based agent designed to enhance cryptocurrency trading by integrating on-chain and off-chain data analysis. The system demonstrates improved trading performance over traditional methods and sets a new benchmark for the field.

https://arxiv.org/abs/2407.09546

203. LLM4Decompile: Decompiling Binary Code with Large Language Models

This paper introduces LLM4Decompile, a series of large language models specifically trained to convert binary code into high-level source code. The models significantly outperform existing tools, showcasing their effectiveness through enhanced readability and executability of decompiled code.

https://arxiv.org/abs/2403.05286

498. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery

This paper reviews and analyzes over 260 scientific large language models (LLMs), focusing on their architectures, pre-training techniques, and deployment in scientific discovery. It provides a comprehensive overview that reveals cross-field and cross-modal connections, enhancing understanding of LLMs in various scientific applications.

http://arxiv.org/abs/2406.10833v3

733. A Bayesian Approach to Harnessing the Power of LLMs in Authorship Attribution

This paper presents a Bayesian approach to authorship attribution using Large Language Models (LLMs), focusing on their ability to maintain long-range textual associations and provide probability outputs. The study achieves 85% accuracy in one-shot authorship classification while utilizing only pre-trained models, thereby establishing new benchmarks in the field of forensic linguistics.

https://arxiv.org/abs/2410.21716

784. KARL: Knowledge-Aware Retrieval and Representations aid Retention and Learning in Students

The paper presents KARL, a content-aware flashcard scheduling model that utilizes student data and card content to enhance learning efficiency. By employing deep knowledge tracing and training on a new dataset, KARL outperforms existing models and shows significant improvements in student recall predictions and learning outcomes.

http://arxiv.org/abs/2402.12291v3

1082. CoCoST: Automatic Complex Code Generation with Online Searching and Correctness Testing

The CoCoST framework improves the generation of complex code by integrating online searching for additional information and correctness testing. It enhances understanding and adaptability in real-world applications by serializing inputs/outputs and generating test cases, demonstrating significant improvements in code quality.

https://arxiv.org/abs/2403.13583

1135. Multi-expert Prompting Improves Reliability, Safety and Usefulness of Large Language Models

Multi-expert Prompting enhances the generation capabilities of large language models by simulating multiple experts' responses and selecting the best outputs. This method effectively improves the quality and safety of model outputs, demonstrating significant increases in truthfulness, factuality, and informativeness while reducing negative traits.

https://arxiv.org/abs/2411.00492

436. Middleware for LLMs: Tools Are Instrumental for Language Agents in Complex Environments

This paper explores the use of middleware tools to enhance the performance of large language models (LLMs) in complex environments, specifically focusing on knowledge bases and databases. By integrating these tools, the authors demonstrate a significant boost in the performance of GPT-4, indicating the viability of such approaches for real-world applications.

https://arxiv.org/abs/2402.14672

806. SecCoder: Towards Generalizable and Robust Secure Code Generation

This paper presents SecCoder, a method aimed at enhancing the security and robustness of code generated by large language models. It demonstrates significant improvements in generalizability and robustness of secure code generation compared to existing approaches, with measurable security gains on unseen test cases.

https://arxiv.org/abs/2410.01488

966. A Simple yet Effective Training-free Prompt-free Approach to Chinese Spelling Correction Based on Large Language Models

This paper presents a training-free, prompt-free method for Chinese spelling correction using large language models. It innovatively adapts the LLMs as traditional language models and incorporates a minimal distortion model to ensure fidelity between the input and output sentences.

https://arxiv.org/abs/2410.04027

146. Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course

This paper investigates the use of large language models as automatic evaluators for assignments in a university course with over 1,000 students. It finds that while students generally accept these evaluators, they also highlight issues with adherence to evaluation instructions and manipulation potential, leading to recommendations for their integration and improvement.

http://arxiv.org/abs/2407.05216

200. Unsupervised Human Preference Learning

This paper introduces a novel method for personalizing content generated by large language models without requiring fine-tuning, using small parameter models as preference agents. The approach significantly enhances the ability to adapt model outputs to individual user preferences, outdoing existing methods in efficiency and effectiveness on practical datasets.

http://arxiv.org/abs/2410.03731v3

233. Consistent Autoformalization for Constructing Mathematical Libraries

This paper addresses the challenge of autoformalization, which involves translating natural language mathematical content into formal language expressions. It proposes a coordinated approach utilizing most-similar retrieval augmented generation, denoising steps, and auto-correction mechanisms to enhance the consistency and reliability of autoformalization across various large language models.

https://arxiv.org/abs/2410.04194

292. LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing

This paper investigates the potential of large language models (LLMs) to assist researchers in the time-consuming task of paper (meta-)reviewing and evaluates LLM-generated reviews against human-written ones. The study also introduces the ReviewCritique dataset, providing insights into the effectiveness of LLMs in identifying issues in peer reviews.

http://arxiv.org/abs/2406.16253v3

472. RealVul: Can We Detect Vulnerabilities in Web Applications with LLM?

This paper introduces RealVul, the first framework utilizing large language models (LLMs) for detecting vulnerabilities specifically in PHP web applications. It addresses challenges in sample extraction and processing, significantly enhancing vulnerability detection performance through improved data synthesis methods and model training techniques.

https://arxiv.org/abs/2410.07573

505. AssistantBench: Can Web Agents Solve Realistic and Time-Consuming Tasks?

This paper evaluates the performance of language agents built on language models for executing realistic and time-consuming tasks on the web, introducing the benchmark AssistantBench. The study reveals significant limitations in current models, and it proposes a new web agent, SeePlanAct (SPA), that outperforms previous agents in this context.

https://arxiv.org/abs/2407.15711

589. Decoding Matters: Addressing Amplification Bias and Homogeneity Issue in Recommendations for Large Language Models

This paper introduces a new decoding method called Debiasing-Diversifying Decoding (D3) to mitigate amplification bias and homogeneity issues encountered in LLM-based recommendation systems. Experimental results show that D3 improves both the accuracy and diversity of the recommendations generated by LLMs.

https://arxiv.org/abs/2406.14900

632. CodeAgent: Autonomous Communicative Agents for Code Review

The paper introduces CodeAgent, a multi-agent Large Language Model system designed to automate the code review process, which is typically labor-intensive and collaborative. CodeAgent's effectiveness is demonstrated in tasks such as detecting inconsistencies, identifying vulnerabilities, validating code style, and suggesting revisions, setting a new state-of-the-art in code review automation.

http://arxiv.org/abs/2402.02172v5

759. Large Language Models in the Clinic: A Comprehensive Benchmark

This paper constructs a benchmark called ClinicBench to evaluate large language models (LLMs) in the clinical context, focusing on complex open-ended question-answering and various clinical reasoning tasks. The study includes both existing and novel datasets, extensive evaluations of multiple LLMs, and expert assessments of their clinical applicability.

https://arxiv.org/abs/2405.00716

799. Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation

This survey reviews the task of generating distractors for multiple-choice questions, detailing methods, datasets, and evaluation metrics in both text and multi-modal domains. It highlights the shift from traditional approaches to using advanced AI methods, such as neural networks and pre-trained language models, in creating effective distractors for educational assessments.

https://arxiv.org/abs/2402.01512

839. Are Large Language Models Good Classifiers? A Study on Edit Intent Classification in Scientific Document Revisions

This paper investigates the capability of large language models (LLMs) in classification tasks, specifically focusing on edit intent classification in scientific documents. Through extensive experimentation and the creation of a new dataset, this study reveals key insights on the application of LLMs for classification, highlighting their potential and the need for more systematic research in this area.

https://arxiv.org/abs/2410.02028

874. Human-LLM Hybrid Text Answer Aggregation for Crowd Annotations

This paper introduces a human-LLM hybrid text answer aggregation method to improve the quality of crowd annotations. It explores the utilization of Large Language Models as aggregators in text answer aggregation, demonstrating effective outcomes in collaborative settings.

https://arxiv.org/abs/2410.17099

917. Autoregressive Multi-trait Essay Scoring via Reinforcement Learning with Scoring-aware Multiple Rewards

This paper introduces Scoring-aware Multi-reward Reinforcement Learning (SaMRL) for automated essay scoring (AES) that simultaneously evaluates multiple traits and enhances feedback. SaMRL employs a novel approach integrating QWK-based rewards into model training, leading to improved scoring outcomes for various prompts.

https://arxiv.org/abs/2409.17472

928. Learning to Generate Writing Feedback via Language Model Simulated Student Revisions

The paper presents PROF, a system that generates writing feedback by learning from language model (LM) simulated student revisions. It demonstrates that this approach significantly improves students' revision performance on an economic essay assignment compared to baseline methods.

https://arxiv.org/abs/2410.08058

1037. Recurrent Alignment with Hard Attention for Hierarchical Text Rating

This paper presents a novel framework called Recurrent Alignment with Hard Attention (RAHA) for hierarchical text rating utilizing large language models (LLMs). By integrating a hard attention mechanism and a recurrent alignment strategy, the framework shows improved performance over existing methods in predicting hierarchical text ratings.

http://arxiv.org/abs/2402.08874v2

1088. Factuality of Large Language Models in the Year 2024

This survey critically analyzes the factual accuracy of large language models (LLMs) that have been instruction-tuned for conversational applications. It identifies the challenges in improving LLM factuality and suggests future research directions.

https://arxiv.org/html/2402.02420v2

1121. Defending Jailbreak Prompts via In-Context Adversarial Game

This paper presents the In-Context Adversarial Game (ICAG), a novel approach for defending against jailbreaking attacks on LLMs without requiring fine-tuning. The empirical results show that ICAG effectively reduces jailbreak success rates and is transferable to other LLMs, highlighting its potential as a versatile defense mechanism.

https://arxiv.org/abs/2402.13148

1154. SPREADSHEETLLM: Encoding Spreadsheets for Large Language Models

The paper presents SpreadsheetLLM, a novel encoding method aimed at enhancing large language models' understanding and reasoning capabilities with spreadsheets. It introduces a compression framework that significantly improves performance on spreadsheet tasks and validates its effectiveness in a new spreadsheet QA task.

https://arxiv.org/abs/2407.09025

1219. Assessing and Verifying Task Utility in LLM-Powered Applications

This paper introduces AgentEval, a framework aimed at assessing and verifying the utility of LLM-powered applications by aligning functionality with user needs. It demonstrates the effectiveness of AgentEval on various tasks and provides resources for reproducibility.

https://arxiv.org/abs/2405.02178

1244. MedAdapter: Efficient Test-Time Adaptation of Large Language Models Towards Medical Reasoning

MedAdapter is introduced as a unified post-hoc adapter for adapting large language models (LLMs) specifically for biomedical reasoning without the need for extensive computational resources or sharing sensitive data. The approach allows for effective test-time adaptation by fine-tuning a small adapter model, yielding performance improvements while addressing privacy concerns.

https://arxiv.org/abs/2405.03000

1245. EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records

EHRAgent is a large language model agent specifically designed to autonomously tackle complex reasoning tasks involving electronic health records (EHRs) by generating and executing code. It improves its performance through interactive coding and feedback mechanisms, achieving significant improvements in success rates on real-world datasets.

https://arxiv.org/abs/2401.07128

14. A Usage-centric Take on Intent Understanding in E-Commerce

This paper presents a novel approach to understanding user intents in E-Commerce, framing it as a natural language reasoning task. It identifies weaknesses in existing intent knowledge graphs and introduces a benchmark for evaluating product recommendation based on user intent.

https://arxiv.org/abs/2402.14901

258. IDEAW: Robust Neural Audio Watermarking with Invertible Dual-Embedding

This paper presents IDEAW, a robust neural audio watermarking method that embeds and accurately extracts messages from audio while improving resistance to various attacks. It addresses the challenges faced by existing methods in terms of capacity, imperceptibility, and efficient watermark locating using an invertible dual-embedding model.

https://arxiv.org/abs/2409.19627

264. PsyGUARD: An Automated System for Suicide Detection and Risk Assessment in Psychological Counseling

PsyGUARD is an automated system designed for detecting suicidal ideation and conducting risk assessments in online psychological counseling. The paper details the development of a fine-grained detection taxonomy and a large-scale dataset to improve automated crisis interventions in mental health services.

https://arxiv.org/abs/2409.20243

312. Investigating LLMs as Voting Assistants via Contextual Augmentation: A Case Study on the European Parliament Elections 2024

This paper investigates the use of large language models (LLMs) as voting advice applications for the European Parliament elections, evaluating their accuracy in predicting political stances. It also explores ways to enhance model performance through approaches like Retrieval-Augmented Generation and Self-Reflection, achieving an average accuracy of 82% with significant variations among political groups.

https://arxiv.org/abs/2407.08495

653. Enhancing High-order Interaction Awareness in LLM-based Recommender Model

This paper presents an enhanced LLM-based recommender model, ELMRec, which improves the interpretation of user-item interactions using whole-word embeddings, facilitating better recommendations without the need for graph pre-training. The model also addresses the issue of emphasizing earlier interactions over recent ones and introduces a reranking solution that outperforms existing methods in recommendation tasks.

http://arxiv.org/abs/2409.19979

856. MetaBench: Planning of Multiple APIs from Various APPs for Complex User Instruction

This paper presents exttt{AppBench}, a benchmark designed to test the ability of Large Language Models (LLMs) to plan and execute multiple APIs from various sources to fulfill complex user instructions. The study reveals the challenges of managing graph structures for API execution order and addressing permission constraints, showcasing that current LLMs struggle significantly with these tasks.

https://arxiv.org/html/2410.19743v1

859. ECCO: Can We Improve Model-Generated Code Efficiency Without Sacrificing Functional Correctness?

This paper introduces ECCO, a reproducible benchmark designed to evaluate the efficiency of program generation by large language models (LLMs) while maintaining functional correctness. The authors adapt and investigate various methods of program generation and editing, revealing that execution information can help preserve correctness while enhancing efficiency.

https://arxiv.org/abs/2407.14044

958. MedReadMe: A Systematic Study for Fine-grained Sentence Readability in Medical Domain

This paper presents a systematic study on fine-grained readability measurements specifically in the medical domain, introducing the MedReadMe dataset which includes readability ratings and complex span annotations for 4,520 sentences. The authors benchmark and enhance existing readability metrics using a variety of linguistic features, demonstrating that incorporating jargon identification significantly improves correlation with human assessments.

https://arxiv.org/abs/2405.02144

357. Evaluating LLMs for Targeted Concept Simplification for Domain-Specific Texts

This paper presents a novel task of "targeted concept simplification" to help adult readers comprehend domain-specific texts by focusing on difficult concepts in context. Through a new dataset, WikiDomains, and a benchmarking of various LLMs, the authors find that human evaluations and automated metrics show varying results, highlighting the complexity of supporting reading comprehension in specialized domains.

https://arxiv.org/html/2410.20763

994. CURE: Context- and Uncertainty-Aware Mental Disorder Detection

392. CareCorpus+: Expanding and Augmenting Caregiver Strategy Data to Support Pediatric Rehabilitation

1083. Sequential API Function Calling Using GraphQL Schema

1138. Framework for Robust and Scalable Text Watermarking

390. A Survey of AMR Applications

691. ABSEval: An Agent-based Framework for Script Evaluation

862. Effective Synthetic Data and Test-Time Adaptation for OCR Correction

991. Automated Essay Scoring: A Reflection on the State of the Art

Information Retrieval and Text Mining

47. LongEmbed: Extending Embedding Models for Long Context Retrieval

This paper presents LongEmbed, a method for extending the context window of embedding models up to 32k tokens without additional training, targeting applications requiring long inputs like legal contracts. The authors benchmark existing models on the newly constructed LongEmbed benchmark, showing significant improvements and proposing training-free strategies for enhancing performance in long context retrieval.

https://arxiv.org/abs/2404.12096

58. BlendFilter: Advancing Retrieval-Augmented Large Language Models via Query Generation Blending and Knowledge Filtering

BlendFilter enhances retrieval-augmented large language models (LLMs) by introducing query generation blending and knowledge filtering to tackle issues with complex inputs and noisy knowledge retrieval. Extensive experiments demonstrate the effectiveness of BlendFilter, achieving superior performance on multiple open-domain question answering benchmarks compared to state-of-the-art approaches.

https://arxiv.org/abs/2402.11129

80. Large Language Models as Foundations for Next-Gen Dense Retrieval: A Comprehensive Empirical Assessment

This paper conducts an empirical assessment of large language models (LLMs) as backbone encoders in dense retrieval tasks, showing their advantages in domain accuracy and generalization. The study evaluates various LLM configurations and finds that larger, extensively pretrained models improve performance on multiple retrieval tasks.

https://arxiv.org/abs/2408.12194

208. A Generic Method for Fine-grained Category Discovery in Natural Language Texts

This paper presents a novel method for fine-grained category discovery in natural language texts using coarse-grained supervision. The approach addresses limitations in existing methods by focusing on semantic similarities and introduces a centroid inference mechanism to enhance real-time applications.

https://arxiv.org/abs/2406.13103

223. Advancing Large Language Model Attribution through Self-Improving

The paper introduces START, a framework designed to enhance large language models' ability to generate text with proper citations, thereby reducing hallucinations and improving verifiability. By leveraging self-constructed synthetic training data and iterative supervision signals, START significantly improves attribution capability in information-seeking systems without the need for manual annotations.

http://arxiv.org/abs/2410.13298v1

373. Unifying Multimodal Retrieval via Document Screenshot Embedding

This paper introduces Document Screenshot Embedding (DSE), a retrieval method that uses screenshots of documents as input, eliminating the need for preprocessing and preserving all original information. The effectiveness of DSE is demonstrated against traditional retrieval methods, showcasing significant improvements in accuracy for both text-intensive and mixed-modality tasks.

https://arxiv.org/abs/2406.11251

402. Enhancing Legal Case Retrieval via Scaling High-quality Synthetic Query-Candidate Pairs

This paper presents an automated method for constructing synthetic query-candidate pairs to enhance legal case retrieval (LCR), addressing challenges such as data limitations and query formats. The resulting dataset, LEAD, significantly improves training for LCR models and achieves state-of-the-art performance on benchmarks.

http://arxiv.org/abs/2410.06581v1

406. Learning to Retrieve Iteratively for In-Context Learning

The paper introduces a novel framework called iterative retrieval, which optimizes the selection of retrieved items using policy optimization and reinforcement learning, specifically targeting in-context learning for semantic parsing tasks. The proposed method enhances a dense retriever's performance with minimal additional parameters, showing improved results in selecting ICL exemplars across various datasets and generalizing across different large language models.

https://arxiv.org/abs/2406.14739

407. Taxonomy-guided Semantic Indexing for Academic Paper Search

This paper presents a framework called Taxonomy-guided Semantic Indexing (TaxoIndex) designed to improve academic paper search by effectively matching queries and documents based on academic concepts. By organizing key concepts into a semantic index guided by academic taxonomies, TaxoIndex enhances the interpretability of results and can be integrated with existing dense retrieval systems.

https://arxiv.org/abs/2410.19218

491. FIRST: Faster Improved Listwise Reranking with Single Token Decoding

This paper introduces FIRST, a faster novel listwise LLM reranking method that improves efficiency and ranking accuracy by prioritizing highly relevant passages during training. Empirical results show a 50% acceleration in inference time while enhancing retrieval performance on the BEIR benchmark.

https://arxiv.org/abs/2406.15657

502. From RAG to Riches: Retrieval Interlaced with Sequence Generation

The paper introduces RICHES, a method that integrates retrieval and generation by directly decoding the contents of retrieved documents. This novel approach adapts to various tasks with minimal training, showcasing strong performance in open-domain question answering tasks.

https://arxiv.org/abs/2407.00361

688. Leveraging Estimated Transferability Over Human Intuition for Model Selection in Text Ranking

This paper introduces a novel approach named Adaptive Ranking Transferability (AiRTran) for model selection in text ranking, addressing the limitations of existing transferability estimation methods which are focused on classification tasks. The proposed method improves model selection efficiency without intensive fine-tuning, showing significant performance enhancements in challenging scenarios compared to traditional methods and human intuition.

https://arxiv.org/abs/2409.16198

755. Exploring the Practicality of Generative Retrieval on Dynamic Corpora

This paper investigates the effectiveness of Generative Retrieval (GR) methods in dynamic information retrieval scenarios where document collections are always changing. It highlights GR's adaptability to evolving knowledge and its computational efficiency compared to traditional Dual Encoders, making it promising for real-world applications in dynamic document environments.

https://arxiv.org/abs/2305.18952

778. MINT: A Benchmark for Evaluating Instructed Information Retrieval

This paper introduces MAIR, a massive benchmark for evaluating information retrieval models on a variety of tasks across multiple domains. It finds that instruction-tuned models generally outperform non-instruction-tuned models but still struggle with certain long-tail tasks.

http://arxiv.org/abs/2410.10127v1

840. LitSearch: A Retrieval Benchmark for Scientific Literature Search

The paper introduces LitSearch, a retrieval benchmark designed for scientific literature search queries, which consist of 597 realistic queries related to ML and NLP papers. The study evaluates various retrieval models, revealing that state-of-the-art dense retrievers significantly outperform traditional methods like BM25, with additional improvements from LLM-based reranking strategies.

https://arxiv.org/abs/2407.18940

845. Dense X Retrieval: What Retrieval Granularity Should We Use?

This paper investigates the impact of retrieval granularity on dense retrieval performance in open-domain NLP tasks. It introduces a novel retrieval unit called 'proposition,' which encapsulates individual factoids and demonstrates superior performance in both retrieval and downstream question-answering tasks when compared to traditional passage-level indexing.

https://arxiv.org/abs/2312.06648

997. Improving Zero-shot LLM Re-Ranker with Risk Minimization

This paper introduces a new framework, UR^3, that addresses biases in Large Language Models (LLMs) used for document re-ranking within Retrieval-Augmented Generation systems. By applying Bayesian decision theory, UR^3 aims to improve the accuracy of document ranking and enhance performance in question-answering tasks.

https://arxiv.org/abs/2406.13331

1013. PairDistill: Pairwise Relevance Distillation for Dense Retrieval

This paper introduces Pairwise Relevance Distillation (PairDistill), a novel approach that enhances the training of dense retrieval models by leveraging pairwise reranking for improved information retrieval performance. Experiments indicate that PairDistill achieves state-of-the-art results on various benchmarks, demonstrating its effectiveness in refining retrieval techniques.

https://arxiv.org/abs/2410.01383

1056. Link, Synthesize, Retrieve: Universal Document Linking for Zero-Shot Information Retrieval

This paper presents a novel Universal Document Linking (UDL) algorithm aimed at improving zero-shot information retrieval by linking similar documents to enhance synthetic query generation across diverse datasets. The empirical results demonstrate the effectiveness and universality of UDL, outperforming existing methods in zero-shot scenarios.

https://arxiv.org/abs/2410.18385

1241. BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers

BMRetriever is a series of dense retrieval models designed to improve biomedical text retrieval, utilizing unsupervised pre-training on large biomedical datasets and instruction fine-tuning. The models show significant parameter efficiency and superior performance in various biomedical tasks, with data and checkpoints made available for transparency and further research.

https://arxiv.org/abs/2404.18443

1242. Comparing Neighbors Together Makes it Easy: Jointly Comparing Multiple Candidates for Efficient and Effective Retrieval

This paper presents the Comparing Multiple Candidates (CMC) framework that improves the efficiency and effectiveness of candidate retrieval by using shallow self-attention layers to contextualize multiple candidate embeddings relative to a query. Experimental results demonstrate that the CMC framework significantly enhances recall and prediction accuracy while maintaining efficiency compared to traditional bi-encoder and cross-encoder methods.

https://arxiv.org/abs/2405.12801

25. Consolidating Ranking and Relevance Predictions of Large Language Models through Post-Processing

This paper presents a method to improve the relevance labeling of documents by consolidating outputs from large language models (LLMs) with their ranking performance. By utilizing pairwise ranking prompts, the proposed post-processing approach enhances both the accuracy of relevance labels and the overall ranking results.

https://arxiv.org/abs/2404.11791

45. Improved Learned Sparse Retrieval with Entity Vocabulary

This paper presents an approach that utilizes corpus-specific vocabularies to enhance the performance of learned sparse retrieval systems in terms of both effectiveness and efficiency. The work demonstrates significant improvements in retrieval quality and reduced latency by experimenting with BERT pre-training and vocabulary expansion processes.

https://arxiv.org/abs/2401.06703

449. Improving Retrieval-augmented Text-to-SQL with AST-based Ranking and Schema Pruning

This paper presents ASTReS, a method for improving Text-to-SQL semantic parsing by dynamically retrieving database information and using abstract syntax trees for few-shot examples. The proposed approach achieves enhanced efficiency and effectiveness in generating SQL queries over both monolingual and cross-lingual benchmarks, indicating potential avenues for future research.

https://arxiv.org/abs/2407.03227

15. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs

This study compares unsupervised fine-tuning and retrieval-augmented generation (RAG) methods for knowledge injection in large language models (LLMs). Results show that while fine-tuning offers improvements, RAG consistently outperforms it, especially in acquiring new factual information.

https://arxiv.org/abs/2312.05934

73. Learning Interpretable Legal Case Retrieval via Knowledge-Guided Case Reformulation

This paper presents KELLER, a knowledge-guided case reformulation approach designed to improve legal case retrieval by incorporating professional legal knowledge into large language models. The method has shown superior performance in retrieving relevant legal cases from complex queries compared to existing techniques.

https://arxiv.org/abs/2406.19760

113. Do We Need Language-Specific Fact-Checking Models? The Case of Chinese

This paper examines the need for language-specific fact-checking models, particularly for the Chinese language, due to limitations of existing translation-based and multilingual methods. It proposes a Chinese fact-checking system that incorporates contextual information and demonstrates improved performance and robustness against biases compared to other methods.

https://arxiv.org/abs/2401.15498

171. Evaluating D-MERIT of Partial-annotation on Information Retrieval

This paper evaluates the impact of partial-annotation in information retrieval and proposes D-MERIT, a curated passage retrieval evaluation set from Wikipedia. The study highlights the distortion that can arise in model rankings due to incomplete annotations and advocates for a balance between resource efficiency and reliable evaluation.

https://arxiv.org/abs/2406.16048

250. PromptReps: Prompting Large Language Models to Generate Dense and Sparse Representations for Zero-Shot Document Retrieval

This paper presents PromptReps, a method that utilizes large language models (LLMs) to perform zero-shot document retrieval by generating both dense and sparse representations from prompts without necessitating extensive training. Experimental results demonstrate that PromptReps can achieve effective retrieval performance comparable to state-of-the-art methods even when applied to large datasets.

https://arxiv.org/abs/2404.18424

353. Do You Know What You Are Talking About? Characterizing Query-Knowledge Relevance For Reliable Retrieval Augmented Generation

This paper presents a framework to evaluate query-knowledge relevance in retrieval augmented generation (RAG) systems, addressing the issues of hallucinations and misinformation in language models. By using goodness-of-fit tests, the authors develop methods for both online and offline assessment to enhance the reliability of RAG through improved query relevance detection.

https://arxiv.org/abs/2410.08320

447. ECON: On the Detection and Resolution of Evidence Conflicts

This paper explores the challenges posed by misinformation and evidence conflicts in decision-making systems driven by large language models (LLMs). It introduces a method for simulating these conflicts and evaluates various models for detecting and resolving them, revealing insights on their performance and behavior.

https://arxiv.org/html/2410.04068v1

453. Efficient Performance Tracking: Leveraging Large Language Models for Automated Construction of Scientific Leaderboards

This paper presents SciLead, a manually-curated Scientific Leaderboard dataset aimed at automating the construction of scientific leaderboards, addressing issues found in existing community-contributed datasets. The authors propose a framework utilizing large language models (LLMs) for leaderboard construction across various experimental settings that reflect real-world scenarios, highlighting the challenges in result value extraction from publications.

https://arxiv.org/abs/2409.12656

490. AGRaME: Any-Granularity Ranking with Multi-Vector Embeddings

This paper presents AGRaME, a flexible ranking mechanism that utilizes multi-vector embeddings to achieve any-granularity ranking from a single encoding level. It introduces a multi-granular contrastive loss for training and shows its application in improving citation addition within retrieval-augmented generation frameworks.

https://arxiv.org/abs/2405.15028

534. Ranking Manipulation for Conversational Search Engines

This paper examines how prompt injections can manipulate the ranking order of sources used by conversational search engines that employ Large Language Models (LLMs). It highlights vulnerabilities in LLMs that can be exploited by adversarial attacks, demonstrating significant variances in how different LLMs prioritize content.

https://arxiv.org/abs/2406.03589

579. $\texttt{MixGR}$: Enhancing Retriever Generalization for Scientific Domain through Complementary Granularity

This paper presents MixGR, a method designed to enhance the generalization of dense retrievers for retrieving scientific documents by addressing the challenges associated with varying query-document relationships. It demonstrates significant improvements in retrieval performance on scientific datasets and showcases its effectiveness in downstream scientific question-answering tasks.

https://arxiv.org/abs/2407.10691

610. ATM: Adversarial Tuning Multi-agent System Makes a Robust Retrieval-Augmented Generator

This paper introduces an Adversarial Tuning Multi-agent (ATM) system to enhance the performance of retrieval-augmented generation (RAG) in large language models by improving their ability to discern useful documents amidst noise. The method involves iterative tuning with an auxiliary attacker agent to robustly guide the generator's responses, leading to superior performance in question answering scenarios.

https://arxiv.org/abs/2405.18111

636. Improve Dense Passage Retrieval with Entailment Tuning

This paper proposes a method called entailment tuning to enhance dense passage retrieval by aligning relevance with the concept of entailment from NLI tasks. The method effectively integrates retrieval and NLI data to improve the embedding of dense retrievers, demonstrating its efficiency in various NLP applications.

https://arxiv.org/html/2410.15801

924. Is It Really Long Context if All You Need Is Retrieval? Towards Genuinely Difficult Long Context NLP

This paper argues that the broad classification of long-context tasks based solely on input length is inadequate and can obscure task difficulty. It proposes a new taxonomy based on two axes of difficulty: diffusion (the challenge of finding necessary information) and scope (the quantity of necessary information), and suggests that genuinely difficult long-context tasks are currently under-explored.

http://arxiv.org/abs/2407.00402v3

926. SEGMENT+: Long Text Processing with Short-Context Language Models

SEGMENT+ is a framework designed to enhance the capability of language models in processing long texts, aimed at improving performance on extensive document understanding and information extraction tasks. By utilizing structured notes and a filtering module, this approach allows efficient handling of extended inputs within limited context windows.

http://arxiv.org/abs/2410.06519

943. Deciphering the Interplay of Parametric and Non-Parametric Memory in RAG Models

This study investigates how the extsc{Atlas} retrieval-augmented generation (RAG) model uses both parametric and non-parametric knowledge during information processing. The findings reveal that the model prefers retrieved context over its pre-existing knowledge and identifies several mechanisms involved in how the model processes useful information from context.

http://arxiv.org/abs/2410.05162

969. ClimRetrieve: A Benchmarking Dataset for Information Retrieval from Corporate Climate Disclosures

This paper introduces ClimRetrieve, a dataset designed for evaluating information retrieval processes within corporate climate disclosures, focusing on the needs of sustainability analysts. It also explores the role of expert knowledge in enhancing information retrieval using embeddings, while discussing the limitations encountered in the application of these techniques.

http://arxiv.org/abs/2406.09818v3

981. Searching for Best Practices in Retrieval-Augmented Generation

This paper investigates retrieval-augmented generation (RAG) techniques to improve the integration of current information and response quality in specialized domains. The authors aim to identify optimal practices for implementing RAG, balancing performance and efficiency, particularly through multimodal retrieval methods.

https://arxiv.org/abs/2407.01219

993. Unveiling and Consulting Core Experts in Retrieval-Augmented MoE-based LLMs

This paper investigates the internal mechanisms of Mixture-of-Experts (MoE)-based Large Language Models (LLMs) in the context of Retrieval-Augmented Generation (RAG) tasks. The authors identify core expert activations that influence RAG behaviors, proposing strategies to enhance RAG effectiveness through controlled experiments across various datasets.

https://arxiv.org/abs/2410.15438

1101. Threshold-driven Pruning with Segmented Maximum Term Weights for Approximate Cluster-based Sparse Retrieval

31. On Fake News Detection with LLM Enhanced Semantics Mining

61. Bridging Cultures in the Kitchen: A Framework and Benchmark for Cross-Cultural Recipe Retrieval

81. A New Pipeline for Knowledge Graph Reasoning Enhanced by Large Language Models Without Fine-Tuning

336. MTA4DPR: Multi-Teaching-Assistants Based Iterative Knowledge Distillation for Dense Passage Retrieval

431. GENRA: Enhancing Zero-shot Retrieval with Rank Aggregation

736. Language Concept Erasure for Language-invariant Dense Retrieval

818. FinDVer: Explainable Claim Verification over Long and Hybrid-content Financial Documents

828. Leveraging BERT and TFIDF Features for Short Text Clustering via Alignment-Promoting Co-Training

836. RaTEScore: A Metric for Entity-Aware Radiology Text Similarity

23. An Experimental Analysis on Evaluating Patent Citations

123. Optimizing Code Retrieval: High-Quality and Scalable Dataset Annotation through Large Language Models

Discourse and Pragmatics

76. QUDSELECT: Selective Decoding for Questions Under Discussion Parsing

This paper presents QUDSELECT, a joint-training framework for parsing Questions Under Discussion (QUD) that employs selective decoding to enhance the structure of QUD dependency. The approach significantly improves upon previous QUD parsers by considering theoretical criteria for discourse relationships and achieves better performance in both human and automatic evaluations.

https://arxiv.org/abs/2408.01046

639. DECOR: Improving Coherence in L2 English Writing with a Novel Benchmark for Incoherence Detection, Reasoning, and Rewriting

DECOR introduces a novel benchmark for detecting and correcting incoherence in L2 English writing, which significantly aids language learners. The system utilizes expert annotations to improve automated coherence assessment and rewriting of incoherent sentences.

https://arxiv.org/abs/2406.19650

360. Reconsidering Sentence-Level Sign Language Translation

This paper critiques the common practice of treating sign language machine translation as a sentence-level task by examining the need for discourse-level context in understanding sign language. It presents a human baseline study that highlights the limitations of the current machine translation paradigm and the importance of thorough context in sign language processing.

https://arxiv.org/abs/2406.11049

684. GDTB: Genre Diverse Data for English Shallow Discourse Parsing across Modalities, Text Types, and Domains

This paper introduces GDTB, a new open-access multi-genre benchmark for PDTB-style shallow discourse parsing, addressing the limitations of existing datasets focused on a single domain. Experiments reveal that while the new dataset is compatible with PDTB, there is a notable drop in performance for out-of-domain data, which can be mitigated through joint training.

https://arxiv.org/abs/2411.00491

730. Improving Logical Fallacy Reasoning with Logical Structure Tree

This paper presents a method for detecting and classifying logical fallacies using a logical structure tree to represent the hierarchical logic flow in arguments. By incorporating this structured representation into large language models (LLMs), the approach enhances the precision and recall of fallacy reasoning tasks.

https://arxiv.org/abs/2410.12048

913. AutoPersuade: A Framework for Evaluating and Explaining Persuasive Arguments

AutoPersuade is a framework designed to evaluate and explain persuasive arguments by curating a dataset, developing a topic model for argument features, and predicting new arguments' effectiveness. It showcases its utility through an experimental study on veganism arguments, validating its predictions and explanation mechanisms with human participants.

https://arxiv.org/abs/2410.08917

1093. Scope-enhanced Compositional Semantic Parsing for DRT

This paper introduces the AMS parser, a compositional, neurosymbolic semantic parser designed for Discourse Representation Theory (DRT). It addresses the limitations of seq2seq models on complex sentences by implementing a novel mechanism for predicting quantifier scope, achieving improved accuracy in parsing.

http://arxiv.org/abs/2407.01899v2

1114. Which questions should I answer? Salience Prediction of Inquisitive Questions

This paper introduces QSALIENCE, a salience predictor for inquisitive questions, which enhances discourse comprehension by prioritizing which questions to answer based on their salience. It demonstrates that answering salient questions can improve summarization quality in news articles by aligning with theoretical frameworks in discourse processing.

https://arxiv.org/abs/2404.10917

1213. Communicating with Speakers and Listeners of Different Pragmatic Levels

This paper examines how different levels of pragmatic competence affect communication success, suggesting that matching reasoning levels between speakers and listeners improves outcomes. The study shows that explicit language benefits learners regardless of their pragmatic skills, and integrating pragmatic reasoning enhances communication performance.

https://arxiv.org/abs/2410.05851

1258. Rethinking Pragmatics in Large Language Models: Towards Open-Ended Evaluation and Preference Tuning

1047. Surprisal Curves of Discourse

530. The effects of distance on NPI illusive effects in BERT

1028. GOME: Grounding-based Metaphor Binding With Conceptual Elaboration For Figurative Language Illustration

1054. Argument Relation Classification through Discourse Markers and Adversarial Training

1067. Building a Multi-Platform, BERT Classifier for Detecting Connective Language

774. TRoTR: A Framework for Evaluating the Re-contextualization of Text Reuse

Machine Translation

55. Chain-of-Dictionary Prompting Elicits Translation in Large Language Models

This paper introduces CoD, a method that enhances large language models' translation capabilities for low-resource languages by utilizing chains of multilingual dictionaries. The experiments show significant improvements in translation performance, demonstrating that CoD surpasses traditional few-shot learning techniques.

https://arxiv.org/abs/2305.06575

69. LLMs Are Zero-Shot Context-Aware Simultaneous Translators

This paper demonstrates that large language models (LLMs) can effectively perform simultaneous machine translation (SiMT) tasks in a zero-shot manner, showing performance comparable to state-of-the-art systems. It highlights that providing minimal background information can significantly enhance performance, particularly in technical domains, indicating LLMs' potential for next-generation multilingual and context-aware translation systems.

https://arxiv.org/abs/2406.13476

111. PsFuture: A Pseudo-Future-based Zero-Shot Adaptive Policy for Simultaneous Machine Translation

This paper presents PsFuture, a novel zero-shot adaptive read/write policy for simultaneous machine translation that operates without additional training. The proposed method achieves performance comparable to existing baselines while improving the efficiency of translation quality and latency through a new training strategy called Prefix-to-Full (P2F).

https://arxiv.org/abs/2410.04075

860. Ladder: A Model-Agnostic Framework Boosting LLM-based Machine Translation to the Next Level

This paper introduces MT-Ladder, a model-agnostic tool that enhances the performance of general-purpose LLMs in machine translation using a novel hierarchical fine-tuning strategy. By refining translations with pseudo-refinement triplets derived from existing LLMs, MT-Ladder significantly improves translation quality while reducing reliance on extensive human annotations and computational resources.

https://arxiv.org/abs/2406.15741

1223. xCOMET-lite: Bridging the Gap Between Efficiency and Quality in Learned MT Evaluation Metrics

This paper presents xCOMET-lite, a compressed machine translation evaluation metric that retains quality while being computationally efficient. It explores methods like distillation, quantization, and pruning to create smaller models that show high correlation with human judgment, outperforming existing metrics.

http://arxiv.org/abs/2406.14553v1

289. Aligning Translation-Specific Understanding to General Understanding in Large Language Models

This paper addresses the misalignment between translation-specific and general understanding in large language models, proposing the DUAT method to enhance translation quality by incorporating general comprehension into the translation process. Experimental results demonstrate that DUAT improves translation quality significantly and reduces literal translations for complex concepts.

https://arxiv.org/abs/2401.05072

914. Towards Cross-Cultural Machine Translation with Retrieval-Augmented Generation from Multilingual Knowledge Graphs

This paper introduces XC-Translate, a large-scale benchmark for machine translation that focuses on culturally nuanced entity names, and proposes KG-MT, a novel method that uses a multilingual knowledge graph with a dense retrieval mechanism to improve translation quality. Experiments demonstrate that KG-MT significantly outperforms existing machine translation systems, showcasing its effectiveness in addressing the challenges of cross-cultural translation.

https://arxiv.org/abs/2410.14057

24. Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice?

This paper investigates the effectiveness of fine-tuning large language models (LLMs) for multilingual machine translation with minimal training data. It finds that while translation capabilities can remain strong with limited parallel sentences, the choice of language direction and data quality significantly impacts translation performance.

https://arxiv.org/abs/2404.14122

214. What do large language models need for machine translation evaluation?

This paper investigates the requirements for large language models (LLMs) to effectively evaluate machine translation quality, exploring the necessary translation information and prompting techniques. The analysis reveals crucial insights, especially the importance of reference translations, and evaluates various prompting strategies across multiple language pairs.

https://arxiv.org/abs/2410.03278

374. Neuron Specialization: Leveraging Intrinsic Task Modularity for Multilingual Machine Translation

This paper introduces Neuron Specialization as a method for improving multilingual machine translation by leveraging intrinsic task modularity to reduce interference between languages. It demonstrates that language-specific neurons can enhance knowledge transfer and reduce negative interactions within multilingual networks.

https://arxiv.org/abs/2404.11201

555. SCOI: Syntax-augmented Coverage-based In-context Example Selection for Machine Translation

This paper introduces a new strategy called Syntax-augmented COverage-based In-context example selection (SCOI) that enhances the performance of machine translation through improved in-context example selection by incorporating syntactic knowledge. The proposed method is shown to outperform other learning-free approaches in empirical evaluations using multilingual large language models.

https://arxiv.org/abs/2408.04872

630. Unveiling the Role of Pretraining in Direct Speech Translation

This study examines the impact of pretraining on direct speech-to-text translation systems facing data scarcity. It demonstrates that a model trained from scratch can match the performance of a pretrained model with an adjustment in the decoder cross-attention mechanism, resulting in improved efficiency.

https://arxiv.org/abs/2409.18044

641. PrExMe: Large Scale Prompt Exploration of Open Source LLMs for Machine Translation and Summarization Evaluation

The paper introduces PrExMe, a large-scale evaluation of open-source LLMs for machine translation and summarization using over 720 prompt templates across 6.6M evaluations. It analyzes the stability and variability of different prompting strategies, revealing significant impacts of minor changes in prompt formats on model rankings.

https://arxiv.org/abs/2406.18528

708. Optimizing Rare Word Accuracy in Direct Speech Translation with a Retrieval-and-Demonstration Approach

This paper addresses the challenge of translating rare words in direct speech translation (ST) models, proposing a retrieval-and-demonstration approach to improve accuracy. By adapting ST models to utilize retrieved examples, the authors achieve significant improvements in rare word translation accuracy over existing baselines.

http://arxiv.org/abs/2409.09009v2

803. Modeling User Preferences with Automatic Metrics: Creating a High-Quality Preference Dataset for Machine Translation

This paper introduces an approach to create a high-quality machine translation preference dataset (MT-Pref) by combining human quality assessments and automatic metrics. The proposed dataset significantly improves translation quality when utilized on TOWER models, demonstrating the importance of aligning model outputs with human preferences.

https://arxiv.org/abs/2410.07779

879. Mitigating the Language Mismatch and Repetition Issues in LLM-based Machine Translation via Model Editing

This paper addresses the challenges in machine translation using Large Language Models (LLMs), specifically focusing on language mismatch and repetition issues that degrade translation quality. It proposes a method involving model editing techniques to mitigate these problems, successfully demonstrating improved translation outcomes without sacrificing overall quality.

http://arxiv.org/abs/2410.07054

1017. Simultaneous Masking, Not Prompting Optimization: A Paradigm Shift in Fine-tuning LLMs for Simultaneous Translation

The paper presents SimulMask, a new method for fine-tuning large language models (LLMs) specifically for simultaneous translation tasks. Unlike existing methods that use prompting optimization, SimulMask employs a novel attention mask to model translation more efficiently, resulting in enhanced translation quality and reduced computational costs.

https://arxiv.org/abs/2405.10443

1033. Enhanced Hallucination Detection in Neural Machine Translation through Simple Detector Aggregation

This paper addresses the challenge of detecting hallucinations in neural machine translation systems by proposing a method that aggregates multiple detectors. The effectiveness of this aggregated detector marks an advancement towards making machine translation systems more reliable.

https://arxiv.org/abs/2402.13331

1131. DeMPT: Decoding-enhanced Multi-phase Prompt Tuning for Making LLMs Be Better Context-aware Translators

The paper introduces Decoding-enhanced Multi-phase Prompt Tuning (DeMPT) as a method to improve context-aware neural machine translation (NMT) by discriminately utilizing inter- and intra-sentence contexts. This approach outperforms traditional concatenation methods, enhancing the performance of large language models in modeling discourse effectively.

https://arxiv.org/abs/2402.15200

1177. Exploring Intrinsic Language-specific Subspaces in Fine-tuning Multilingual Neural Machine Translation

This paper presents a method for fine-tuning multilingual neural machine translation models by exploiting intrinsic language-specific subspaces with a fraction of the model's parameters. The proposed language-specific LoRA approach and architecture learning techniques demonstrate improved performance and efficiency in training across various language resources.

https://arxiv.org/abs/2409.05224

1248. Simultaneous Interpretation Corpus Construction by Large Language Models in Distant Language Pair

This paper introduces a method for constructing a simultaneous interpretation corpus using Large Language Models, aimed at improving simultaneous machine translation systems. The proposed LLM-SI-Corpus allows for training with lower latency while maintaining quality similar to existing datasets.

https://arxiv.org/abs/2404.12299

188. Word Alignment as Preference for Machine Translation

This paper addresses the issues of hallucination and omission in machine translation by focusing on word alignment in large language models. It proposes a method of preference optimization based on word alignment to improve the translation accuracy and validate this approach through empirical experiments.

https://arxiv.org/abs/2405.09223

600. Distributional Properties of Subword Regularization

This paper analyzes the biases in stochastic variants of subword tokenization schemes like BPE and MaxMatch, identifying that they tend to favor a narrow set of tokenizations for words. The authors propose a new algorithm for uniformly sampling tokenizations that enhances the effectiveness of subword regularization in machine translation tasks.

https://arxiv.org/abs/2408.11443

634. MMTE: Corpus and Metrics for Evaluating Machine Translation Quality of Metaphorical Language

This paper introduces a framework for evaluating machine translation quality specifically for metaphorical language, which has been largely overlooked by traditional metrics. It presents a multilingual metaphor corpus along with new evaluation metrics that assess aspects like metaphorical equivalence and emotional authenticity in translations.

https://arxiv.org/abs/2406.13698

802. Can Automatic Metrics Assess High-Quality Translations?

This paper examines the efficacy of automatic metrics in assessing translation quality, highlighting their limitations in distinguishing subtle differences among high-quality translations. The authors advocate for a shift towards a binary evaluation of correctness in translation to improve decision-making in practical scenarios.

https://arxiv.org/abs/2405.18348

294. Beyond Reference: Evaluating High Quality Translations Better than Human References

704. Domain adapted machine translation: What does catastrophic forgetting forget and why?

1152. How Good is my MT Metric? A Framework for the Interpretation of Metric Assessments

1238. Simul-MuST-C: Simultaneous Multilingual Speech Translation Corpus Using Large Language Model

278. Using Language Models to Disambiguate Lexical Choices in Translation

824. Building Resources for Emakhuwa: Machine Translation and News Classification Benchmarks

1102. Error Analysis of Multilingual Language Models in Machine Translation for Low-resource Languages: A Case Study of Amharic to English Bi-directional Machine Translation

1218. SpeechQE: Estimating the Quality of Direct Speech Translation

Dialogue and Interactive Systems

71. ChatRetriever: Adapting Large Language Models for Generalized and Robust Conversational Dense Retrieval

This paper introduces ChatRetriever, a model optimized for conversational dense retrieval by effectively adapting large language models to understand complex multi-turn interactions. Through a dual-learning approach that incorporates contrastive learning and masked instruction tuning, it demonstrates state-of-the-art performance across multiple conversational search benchmarks and robustness in diverse contexts.

https://arxiv.org/abs/2404.13556

119. Aligning Language Models to Explicitly Handle Ambiguity

The paper presents a novel approach called Alignment with Perceived Ambiguity (APA), which enables language models to effectively manage ambiguous user queries and improve their reliability. By addressing challenges related to ambiguity in natural language interactions, APA allows language models to detect ambiguous utterances and respond accurately while maintaining performance on clear queries.

https://arxiv.org/abs/2404.11972

135. CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search

The paper introduces CHIQ, a two-step method for improving query rewriting in conversational search using open-source large language models (LLMs). It demonstrates that CHIQ outperforms previous methods reliant on closed-source LLMs by effectively resolving ambiguities in conversation history.

https://arxiv.org/abs/2406.05013

271. CoEvol: Constructing Better Responses for Instruction Finetuning through Multi-Agent Cooperation

The paper presents CoEvol, a multi-agent cooperation framework designed to improve responses for instruction fine-tuning of large language models. Through an iterative debate-advice-edit-judge paradigm, it enhances data quality and demonstrates superior performance on evaluation benchmarks compared to existing methods.

https://arxiv.org/abs/2406.07054

281. Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs

This paper presents a method for creating personalized agents that leverage a user's smartphone memories through a technique called Retrieval-Augmented Generation on Editable Memory Graphs. The approach improves user interaction with AI assistants by enhancing their ability to recall and utilize personal data effectively, validated through real-world experiments.

http://arxiv.org/abs/2409.19401

310. Dialog2Flow: Pre-training Action-Driven Sentence Embeddings for Automatic Dialog Flow Extraction

This paper presents Dialog2Flow (D2F), a method for deriving structured workflows from unannotated dialogs using pre-trained soft-contrastive sentence embeddings that group utterances by their communicative functions. The D2F embeddings facilitate the automation of workflow extraction from dialogs and show improved performance compared to traditional methods and other sentence embeddings across various domains.

https://arxiv.org/abs/2410.18481

320. Synergizing In-context Learning with Hints for End-to-end Task-oriented Dialog Systems

This paper presents SyncTOD, a system that enhances end-to-end task-oriented dialog systems by integrating task-specific hints for improved performance even in low-data scenarios. The results show that SyncTOD outperforms existing LLM-based models while remaining competitive with full-data settings.

https://arxiv.org/abs/2405.15585

372. Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning

This paper presents One PEFT Per User (OPPU), a method for personalizing large language models (LLMs) through personalized parameter-efficient fine-tuning (PEFT), enabling users to customize LLM interactions based on individual preferences. OPPU addresses limitations of existing prompt-based methods and demonstrates superior performance across various tasks while adapting to changes in user behavior.

https://arxiv.org/abs/2402.04401

477. MetaReflection: Learning Instructions for Language Agents using Past Reflections

MetaReflection introduces an offline reinforcement learning technique to enhance the performance of Language Agents by utilizing experiential learnings from past trials. The method shows improvements in various complex tasks and requires fewer LLM calls compared to current state-of-the-art techniques.

https://arxiv.org/abs/2405.13009

478. Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors

The paper discusses the use of large language models (LLMs) as dialog tutoring systems to help scaffold students' problem-solving abilities, particularly in detecting and responding to student reasoning errors. Empirical evaluations demonstrate that their proposed solution improves the quality of feedback provided to students, leading to more accurate and effective tutoring interactions.

http://arxiv.org/abs/2407.09136

644. Beyond the Turn-Based Game: Enabling Real-Time Conversations with Duplex Models

This paper presents duplex models that enhance real-time interactions with large language models by allowing them to concurrently listen and respond to users, simulating more human-like conversations. The study demonstrates that adapting LLMs through a time-division-multiplexing strategy and fine-tuning on a specially crafted dataset significantly improves user satisfaction compared to traditional models.

https://arxiv.org/abs/2406.15718

697. Neeko: Leveraging Dynamic LoRA for Efficient Multi-Character Role-Playing Agent

Neeko is a framework designed to enhance multi-character role-playing by utilizing a dynamic low-rank adapter (LoRA) strategy for effective character adaptation. This method improves the performance and interaction quality of dialogue agents in multi-character scenarios.

https://arxiv.org/abs/2402.13717

710. TransferTOD: A Generalizable Chinese Multi-Domain Task-Oriented Dialogue System with Transfer Capabilities

This paper introduces TransferTOD, a multi-domain task-oriented dialogue system designed for effectively handling conversations across diverse scenarios in Chinese. The study highlights the capabilities of Large Language Models (LLMs) in improving dialogue systems through fine-tuning, showcasing the model's significant generalization performance and efficiency in data utilization.

https://arxiv.org/abs/2407.21693

929. Small LLMs Are Weak Tool Learners: A Multi-LLM Agent

This paper introduces a modular multi-LLM framework that improves tool learning by decomposing capabilities into distinct components (planner, caller, and summarizer), each handled by individual LLMs. It highlights the benefits of this approach over traditional single-LLM methods through performance evaluations on various tool-use benchmarks.

https://arxiv.org/abs/2401.07324

964. Inductive-Deductive Strategy Reuse for Multi-Turn Instructional Dialogues

This paper introduces a method for improving the diversity and depth of instructional dialogues by capturing complex rules and employing an inductive-deductive approach. Experimental results demonstrate that the proposed method generates more effective multi-turn instructional dialogues compared to existing models.

https://arxiv.org/abs/2404.11095

1124. MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models

The paper introduces MT-Eval, a benchmark specifically designed to evaluate the multi-turn conversational capabilities of large language models (LLMs). It highlights the shortcomings of existing benchmarks that focus solely on single-turn evaluations and explores the performance of various LLMs in multi-turn contexts.

https://arxiv.org/abs/2401.16745

1167. TokenVerse: Unifying Speech and NLP Tasks via Transducer-based ASR

TokenVerse is a unified Transducer-based model that integrates various speech and NLP tasks, improving automatic speech recognition (ASR) and individual task performance like speaker change detection and named entity recognition. This approach eliminates traditional cascaded pipelines by incorporating task-specific tokens for streamlined inference and better accuracy.

https://arxiv.org/abs/2407.04444

1192. Beyond Turn-Based Interfaces: Synchronous LLMs as Full-Duplex Dialogue Agents

This paper introduces Synchronous LLMs that enable full-duplex spoken dialogue by integrating time information into pre-trained models, facilitating dynamic turn-taking and overlapping speech. The proposed models significantly enhance dialogue meaningfulness while using a limited set of real-world dialogue data in combination with synthetic data.

https://arxiv.org/abs/2409.15594

1018. ToolPlanner: A Tool Augmented LLM for Multi Granularity Instructions with Path Planning and Feedback

The paper introduces ToolPlanner, a tool-augmented LLM framework designed to address the gap between overly detailed instructions and real-world scenarios by utilizing multi-granularity instructions and a two-stage reinforcement learning approach. Experimental results indicate significant improvements in task completion and instruction adherence compared to existing models, along with enhanced human evaluation metrics.

https://arxiv.org/abs/2409.14826

437. MORPHEUS: Modeling Role from Personalized Dialogue History by Exploring and Utilizing Latent Space

The paper proposes MORPHEUS, a framework that models roles in personalized dialogue generation by leveraging latent space derived from dialogue history, thereby eliminating the need for external role data. Experiments show that MORPHEUS effectively enhances role extraction and improves response coherence across multiple languages.

https://arxiv.org/abs/2407.02345

473. Unsupervised End-to-End Task-Oriented Dialogue with LLMs: The Power of the Noisy Channel

This paper proposes an unsupervised method for training task-oriented dialogue systems using only a defined API schema and unlabeled dialogues. Utilizing a noisy channel model and expectation-maximization, the approach significantly improves dialogue success rates without the need for costly annotations.

https://arxiv.org/abs/2404.15219

721. Learning from Feedback with Coupled Comprehension and Generation

This paper examines the integration of language comprehension and generation through a learning model that adapts based on user interaction feedback. The results demonstrate significant performance improvements and a more human-like language output owing to the coupling of these two capabilities.

http://arxiv.org/abs/2408.15992v1

1006. A Survey of Ontology Expansion for Conversational Understanding

This survey paper reviews current techniques in Ontology Expansion for improving conversational understanding in AI systems. It categorizes existing literature into new intent discovery, new slot-value discovery, and joint ontology expansion, highlighting emerging challenges and methodologies.

http://arxiv.org/abs/2410.15019v1

93. Watch Every Step! LLM Agent Learning via Iterative Step-level Process Refinement

This paper presents the Iterative step-level Process Refinement (IPR) framework aimed at enhancing the training of language model agents through detailed step-by-step guidance. The effectiveness of IPR is demonstrated through experiments on complex tasks, showing improvements in action efficiency and applicability across various models.

https://arxiv.org/abs/2406.11176

26. Strength Lies in Differences! Towards Effective Non-collaborative Dialogues via Tailored Strategy Planning

This paper presents TRIP, a novel approach aimed at enhancing non-collaborative dialogue agents through tailored strategic planning that takes into account the diverse characteristics of users. By evaluating TRIP on benchmark tasks, the study shows improvements in the capability of dialogue agents to engage effectively with a range of users.

https://arxiv.org/html/2403.06769v1

37. Personality-aware Student Simulation for Conversational Intelligent Tutoring Systems

This paper presents a framework for personality-aware student simulation in conversational Intelligent Tutoring Systems (ITSs), emphasizing the need for adaptive learning experiences based on individual student characteristics. It demonstrates that large language models can simulate diverse student responses tailored to varying language abilities and personality traits to enhance engagement and teaching strategies.

https://arxiv.org/abs/2404.06762

42. Successfully Guiding Humans with Imperfect Instructions by Highlighting Potential Errors and Suggesting Corrections

The paper presents HEAR, a system that improves human guidance in residential environments by highlighting potential instruction errors and offering corrective suggestions. Through user evaluation, HEAR demonstrated significant improvements in success rate and error reduction compared to standard instructions.

https://arxiv.org/abs/2402.16973

82. Towards Tool Use Alignment of Large Language Models

This paper presents the concept of Personality Alignment, which adapts large language models (LLMs) to align with the individual preferences and behaviors of users by utilizing a dataset based on psychometric principles. The authors propose an optimization method that improves the efficiency and effectiveness of LLMs to align with users' personality characteristics while requiring less computational resources.

https://arxiv.org/abs/2408.11779

138. Direct Multi-Turn Preference Optimization for Language Agents

This paper introduces Direct Multi-Turn Preference Optimization (DMPO), a novel loss function designed for training language agents in multi-turn tasks. It addresses the challenges of applying Direct Preference Optimization (DPO) in Reinforcement Learning settings by proposing strategies to manage the partition function and trajectory length discrepancies.

https://arxiv.org/abs/2406.14868

308. An LLM Feature-based Framework for Dialogue Constructiveness Assessment

This paper introduces an LLM feature-based framework for assessing dialogue constructiveness, combining interpretable feature-based models with neural approaches. The results show that the proposed framework outperforms or matches existing models while providing more robust prediction rules.

https://arxiv.org/abs/2406.14760

383. InterIntent: Investigating Social Intelligence of LLMs via Intention Understanding in an Interactive Game Context

This paper presents InterIntent, a framework for assessing the social intelligence of large language models (LLMs) through their intention understanding in a game environment. It reveals that while LLMs are proficient in selecting intentions, they struggle to infer the intentions of others, highlighting the significance of intention understanding in evaluating social intelligence.

https://arxiv.org/abs/2406.12203

456. Evaluating Character Understanding of Large Language Models via Character Profiling from Fictional Works

This paper presents a method for evaluating the character understanding capabilities of large language models (LLMs) through character profiling from fictional works. By constructing the CroSS dataset and assessing the generated character profiles, the authors demonstrate the potential of LLMs in role-playing applications and emphasize the importance of nuanced character understanding.

https://arxiv.org/abs/2404.12726

554. Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations

The paper presents Collaborative STORM (Co-STORM), an interactive system that allows users to learn by observing and participating in conversations among language model agents. Co-STORM facilitates the discovery of unknown unknowns through guided discourse and dynamic information organization, outperforming traditional search engines and chatbots in user preference and report quality.

https://arxiv.org/abs/2408.15232

591. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles

Roleplay-doh is a human-LLM collaboration pipeline designed to help mental health experts create AI patients for novice counselors, overcoming challenges related to data privacy and expert feedback collection. The approach demonstrates significant improvements in the realism and principle adherence of simulated interactions, making it a novel contribution to the field of roleplaying with AI in therapeutic contexts.

https://arxiv.org/abs/2407.00870

628. Do LLMs suffer from Multi-Party Hangover? A Diagnostic Approach to Addressee Recognition and Response Selection in Conversations

This paper evaluates systems' performance in understanding and responding to Multi-Party Conversations (MPCs) while addressing the challenges posed by their structural complexities. It introduces a diagnostic approach to response selection and addressee recognition, emphasizing the need for privacy-preserving data practices and revealing task-dependent variations in model sensitivity.

https://arxiv.org/abs/2409.18602

631. PCQPR: Proactive Conversational Question Planning with Reflection

The paper introduces Proactive Conversational Question Planning with Reflection (PCQPR), which enhances Conversational Question Generation (CQG) by focusing on guiding conversations toward specific conclusions rather than merely reacting to them. By utilizing a planning algorithm inspired by Monte Carlo Tree Search combined with large language models, PCQPR allows for the proactive generation of contextually relevant questions aimed at achieving desired conversational outcomes.

https://arxiv.org/abs/2410.01363

643. Repairs in a Block World: A New Benchmark for Handling User Corrections with Multi-Modal Language Models

The paper introduces the BlockWorld-Repairs dataset, which focuses on mutual misunderstandings in dialogues and the processing of Third Position Repairs by multi-modal language models. It evaluates current models on their ability to handle user corrections in instruction-following tasks and suggests improvements through specialized training regimes.

https://arxiv.org/abs/2409.14247

648. Red Teaming Language Models for Processing Contradictory Dialogues

This study introduces a novel task for processing contradictory dialogues by detecting and modifying self-contradictory statements. It presents a Red Teaming framework that utilizes a dataset of contradictory dialogues, along with explanatory labels to improve understanding and modification of such dialogues.

https://arxiv.org/abs/2405.10128

705. Enhancing AI Assisted Writing with One-Shot Implicit Negative Feedback

This paper discusses an AI-assisted writing system that uses one-shot implicit negative feedback, specifically when users do not select suggested replies. The proposed approach, named Nifty, improves text generation accuracy significantly by integrating user feedback, resulting in substantial performance gains on benchmark datasets.

https://arxiv.org/html/2410.11009v1

709. ACE: A LLM-based Negotiation Coaching System

ACE is a negotiation coaching system powered by large language models (LLMs) that provides users with feedback to enhance their bargaining skills. It utilizes a dataset of negotiation transcripts to identify mistakes and evaluate improvements in user performance compared to alternative feedback methods.

https://arxiv.org/abs/2410.01555

883. ESC-Eval: Evaluating Emotion Support Conversations in Large Language Models

The paper proposes ESC-Eval, a framework to evaluate Emotion Support Conversations (ESC) generated by Large Language Models (LLMs). It includes a role-playing agent for interaction, human evaluations of dialogues, and an automated scoring system called ESC-RANK, showing that ESC-oriented LLMs outperform general AI assistants but fall short of human performance.

https://arxiv.org/abs/2406.14952

906. On the Robustness of Editing Large Language Models

This paper investigates the robustness of model editing for large language models (LLMs), focusing on how well edited LLMs perform in practical communicative AI scenarios. The findings reveal that performance declines significantly with prompt rephrasing, highlighting challenges in effectively editing knowledge within LLMs.

https://arxiv.org/abs/2402.05827

936. MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations

MediTOD introduces a new English dataset for medical history taking conversations, addressing the lack of comprehensive and annotated dialogue datasets in this field. It provides a valuable resource for improving medical task-oriented dialogue systems, aiming to assist doctors while expanding access to care.

https://arxiv.org/abs/2410.14204

951. GDPO: Learning to Align Language Models with Diversity Using GFlowNets

This paper presents GDPO, a method that enhances diversity in language models while aligning their output with human preferences. It addresses issues of bias and overfitting in traditional preference alignment techniques through a new RL algorithm applied in offline settings, demonstrating improved performance in dialog generation and summarization tasks.

https://arxiv.org/abs/2410.15096

968. CoSafe: Evaluating Large Language Model Safety in Multi-Turn Dialogue Coreference

The paper introduces CoSafe, a novel evaluation framework focused on the safety of large language models (LLMs) in multi-turn dialogue coreference scenarios. It highlights the vulnerabilities of LLMs in this context by presenting an evaluation of five widely used models with varying success rates under safety attacks.

https://arxiv.org/abs/2406.17626

979. MP2D: An Automated Topic Shift Dialogue Generation Framework Leveraging Knowledge Graphs

The paper presents a framework called MP2D that generates dialogue datasets with natural topic transitions using knowledge graphs. It highlights the challenge of managing topic shifts in dialogue systems and provides a benchmark for evaluating models on this task.

https://arxiv.org/abs/2403.05814

1060. Unsupervised Extraction of Dialogue Policies from Conversations

This paper presents a novel method for unsupervised extraction of dialogue policies from conversations using Large Language Models and a graph-based flow network approach. By converting conversations into a unified intermediate representation, the authors enhance control for conversation designers while improving the efficiency of dialogue policy development.

https://arxiv.org/abs/2406.15214

1070. BLSP-Emo: Towards Empathetic Large Speech-Language Models

BLSP-Emo introduces a novel approach to create an empathetic speech-language model that understands semantics and emotions in speech. By leveraging existing ASR and SER datasets, the model demonstrates enhanced capabilities in generating empathetic responses during conversations.

https://arxiv.org/abs/2406.03872

1100. Grounding Language in Multi-Perspective Referential Communication

This paper presents a new task and dataset for generating and comprehending referring expressions between agents in multi-agent environments, emphasizing the importance of perspective-taking. The authors demonstrate that automated models still struggle compared to human agents in this regard, but improvements can be made by training models with communicative success in mind.

https://arxiv.org/abs/2410.03959

1143. Advancing Social Intelligence in AI Agents: Technical Challenges and Open Question

This paper discusses the technical challenges and open questions in the development of socially intelligent AI agents that can understand and interact based on human emotions, behaviors, and cognition. It emphasizes the importance of multidisciplinary approaches, particularly in natural language processing, to advance research in Social-AI.

http://arxiv.org/abs/2404.11023

1208. Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs

This paper explores the limitations of simulating social interactions using large language models (LLMs) with a focus on the impact of information asymmetry on their performance. It introduces an evaluation framework to analyze LLMs in both omniscient and non-omniscient settings, revealing that LLMs excel in unrealistic scenarios but face challenges in more accurate real-world conditions.

https://arxiv.org/abs/2403.05020

520. Ontologically Faithful Generation of Non-Player Character Dialogues

This paper presents KNUDGE, a task designed for the generation of dialogue trees for non-player characters in a video game, focusing on maintaining fidelity to game lore and quest specifications. The study evaluates neural generation models in producing complex dialogues and highlights the challenges of creating realistic interactions that reflect narrative detail.

https://arxiv.org/abs/2212.10618

674. Let Me Teach You: Pedagogical Foundations of Feedback for Language Models

The paper introduces FELT, a feedback framework for Large Language Models (LLMs), which draws on ideas from pedagogy to systematically categorize natural language feedback. By providing a structured taxonomy and mapping of feedback characteristics, this work aims to enhance the design of feedback mechanisms and spur new research directions in natural language feedback systems.

https://arxiv.org/abs/2307.00279

811. LLM Task Interference: An Initial Study on the Impact of Task-Switch in Conversational History

This paper investigates how task-switching during conversations with large language models (LLMs) affects their performance, revealing that while context can enhance task performance, it may also lead to deterioration when the task shifts. Through experiments involving 15 task switches across 5 datasets, the study highlights vulnerabilities and performance degradation associated with task interference in conversational LLMs.

https://arxiv.org/abs/2402.18216

309. Relevance Is a Guiding Light: Relevance-aware Adaptive Learning for End-to-end Task-oriented Dialogue System

485. Zero-shot Cross-domain Dialogue State Tracking via Context-aware Auto-prompting and Instruction-following Contrastive Decoding

616. What are the Generator Preferences for End-to-end Task-Oriented Dialog System?

1157. One-to-Many Communication and Compositionality in Emergent Communication

86. Mitigating Matthew Effect: Multi-Hypergraph Boosted Multi-Interest Self-Supervised Learning for Conversational Recommendation

263. Bootstrapped Policy Learning for Task-oriented Dialogue through Goal Shaping

605. More Insightful Feedback for Tutoring: Enhancing Generation Mechanisms and Automatic Evaluation

1175. Thoughts to Target: Enhance Planning for Target-driven Conversation

268. Retrospex: Language Agent Meets Offline Reinforcement Learning Critic

331. Analyzing Key Factors Influencing Emotion Prediction Performance of VLLMs in Conversational Contexts

410. Incomplete Utterance Rewriting with Editing Operation Guidance and Utterance Augmentation

545. Evaluating the Effectiveness of Large Language Models in Establishing Conversational Grounding

685. RA2FD: Distilling Faithfulness into Efficient Dialogue Systems

804. DC-Instruct: An Effective Framework for Generative Multi-intent Spoken Language Understanding

881. Global Reward to Local Rewards: Multimodal-Guided Decomposition for Improving Dialogue Agents

937. **YesBut

1193. QuBE: Question-based Belief Enhancement for Agentic LLM

1269. Follow:

189. Improving Multi-party Dialogue Generation via Topic and Rhetorical Coherence

1252. ABLE: Personalized Disability Support with Politeness and Empathy Integration

1214. RECANTFormer: Referring Expression Comprehension with Varying Numbers of Targets

Special Theme: Efficiency in Model Algorithms, Training, and Inference

35. Mitigating the Alignment Tax of RLHF

This paper investigates the alignment tax that occurs when aligning LLMs under Reinforcement Learning with Human Feedback, leading to forgotten abilities. The authors propose a method called Heterogeneous Model Averaging (HMA) to optimize the alignment-performance trade-off in NLP tasks, demonstrating its effectiveness across different RLHF algorithms.

https://arxiv.org/abs/2309.06256

43. Parameter-Efficient Sparsity Crafting from Dense to Mixture-of-Experts for Instruction Tuning on General Tasks

This paper presents parameter-efficient sparsity crafting (PESC), a method that adapts dense models into sparse ones using the mixture-of-experts architecture to improve instruction tuning. It demonstrates that PESC can enhance model capacity while significantly reducing computational costs and outperforming other model variants in general NLP tasks.

https://arxiv.org/abs/2401.02731

57. RoseLoRA: Row and Column-wise Sparse Low-rank Adaptation of Pre-trained Language Model for Knowledge Editing and Fine-tuning

The paper proposes RoseLoRA, a novel parameter-efficient fine-tuning method that implements row and column-wise sparse low-rank adaptation for pre-trained language models. This approach aims to selectively update only the most important parameters during knowledge editing and fine-tuning, ensuring efficiency while preserving the model's overall knowledge.

https://arxiv.org/abs/2406.10777

100. Rethinking Token Reduction for State Space Models

This paper presents a new post-training token reduction method specifically designed for State Space Models (SSMs), addressing the limitations of existing token reduction techniques that often lead to performance drops. The proposed method significantly improves accuracy and reduces computational and memory requirements for SSMs, demonstrating its effectiveness across multiple benchmarks.

https://arxiv.org/abs/2410.14725

134. Prefixing Attention Sinks can Mitigate Activation Outliers for Large Language Model Quantization

This paper presents a method called CushionCache to mitigate activation outliers in large language model (LLM) quantization, aiming to improve per-tensor activation quantization performance. The proposed strategy involves carefully selecting prompt token sequences to reduce maximum activation values for subsequent tokens, thus enhancing quantization efficiency.

https://arxiv.org/abs/2406.12016

197. QUIK: Towards End-to-end 4-Bit Inference on Generative Large Language Models

This paper introduces QUIK, a novel quantization approach that reduces the inference costs of Large Language Models (LLMs) by quantizing both weights and activations to 4 bits, enhancing computational efficiency. The implementation of QUIK leads to significant speedups in execution while maintaining accuracy, achieving up to 3.4 times improvement relative to FP16 execution on generative models like LLaMA and OPT.

https://arxiv.org/abs/2310.09259

316. Predicting Rewards Alongside Tokens: Non-disruptive Parameter Insertion for Efficient Inference Intervention in Large Language Model

This paper introduces Non-disruptive Parameter Insertion (Otter) to enhance inference in large language models by predicting calibration signals alongside outputs. Otter reduces space and time overhead significantly while integrating easily with current inference engines, maintaining access to original model responses.

https://arxiv.org/abs/2408.10764

319. Scaling Laws Across Model Architectures: A Comparative Analysis of Dense and MoE Models in Large Language Models

This paper investigates the scaling behavior of Dense and Mixture of Experts (MoE) models in large language models (LLMs), highlighting the transferability of scaling laws between different architectures. The findings indicate that MoE models showcase superior generalization with consistent scaling, offering insights into optimizing their training and deployment.

https://arxiv.org/abs/2410.05661

389. Efficient LLM Comparative Assessment: A Product of Experts Framework for Pairwise Comparisons

This paper presents a Product of Expert (PoE) framework for efficient comparative assessment using LLMs, which significantly reduces computational costs while maintaining high correlation with human judgments. By strategically utilizing a limited number of pairwise comparisons, the framework allows for effective ranking in natural language generation tasks, demonstrating substantial computational savings.

https://arxiv.org/abs/2405.05894

409. Advancing Adversarial Suffix Transfer Learning on Aligned Large Language Models

This paper presents DeGCG, a two-stage transfer learning framework that enhances the efficiency of adversarial suffix transferability in large language models. It addresses the shortcomings of the Greedy Coordinate Gradient algorithm and shows improvements in search processes and model performance through experiments on HarmBench.

https://www.arxiv.org/abs/2408.14866

422. EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees

EAGLE-2 introduces a context-aware dynamic draft tree technique that enhances the efficiency of Large Language Model inference, achieving significant speed improvements while maintaining output quality. This paper demonstrates that the acceptance rates for draft tokens can vary contextually, which this new method effectively leverages for faster processing.

https://arxiv.org/abs/2406.16858

467. VPTQ: Extreme Low-bit Vector Post-Training Quantization for Large Language Models

This paper presents Vector Post-Training Quantization (VPTQ), a method to achieve extremely low-bit quantization for large language models, enhancing their deployment efficiency. By employing Second-Order Optimization and reducing memory requirements, the proposed method demonstrated significant accuracy improvements and increased inference throughput over state-of-the-art techniques.

https://arxiv.org/abs/2409.17066

501. MOSEL: Inference Serving Using Dynamic Modality Selection

The paper introduces MOSEL, an automated inference serving system that dynamically selects input modalities for multi-modal machine learning models based on performance and accuracy requirements. MOSEL significantly improves system throughput and reduces job completion times while ensuring model quality.

https://arxiv.org/abs/2310.18481

517. Efficient Sequential Decision Making with Large Language Models

This paper introduces an efficient method for integrating large language models (LLMs) into sequential decision-making processes without incurring the costs associated with retraining or extensive prompting. The proposed approach utilizes online model selection algorithms, achieving significant performance improvements while minimizing the number of LLM calls made during the decision-making process.

https://arxiv.org/abs/2406.12125

535. Fast Forwarding Low-Rank Training

This paper introduces Fast Forward, a new optimization strategy that enhances low-rank adaptation methods for efficient finetuning of pretrained Language Models. It achieves significant reductions in training time and computational costs while maintaining model performance across various tasks.

https://arxiv.org/abs/2409.04206

576. Matryoshka-Adaptor: Unsupervised and Supervised Tuning for Smaller Embedding Dimensions

The paper introduces Matryoshka-Adaptor, a novel framework for tuning embeddings from Large Language Models that significantly reduces their dimensionality while maintaining performance levels, enhancing computational efficiency. The approach is applicable in both unsupervised and supervised settings and showcases effectiveness across various datasets, proving valuable for applications that typically face challenges due to high-dimensional embeddings.

http://arxiv.org/abs/2407.20243

602. Towards Fast Multilingual LLM Inference: Speculative Decoding and Specialized Drafters

This paper addresses the challenge of high inference time in multilingual settings for large language models (LLMs) by proposing a training method that utilizes speculative decoding and language-specific draft models. The results show significant speedup in inference time when validated across various languages compared to previous methods.

https://arxiv.org/abs/2406.16758

633. TroL: Traversal of Layers for Large Language and Vision Models

The paper presents TroL, a family of efficient large language and vision models that utilize a traversal of layers strategy to optimize their performance while reducing resources. By allowing layer reuse in a token-wise manner, TroL efficiently outperforms existing models despite its smaller size and competes with larger models in performance.

https://arxiv.org/abs/2406.12246

717. Heterogeneous LoRA for Federated Fine-tuning of On-Device Foundation Models

This paper introduces HetLoRA, a method for federated fine-tuning of on-device foundation models that addresses data and system heterogeneity. It enhances convergence speed and performance while ensuring computation efficiency suitable for deployment across heterogeneous devices.

https://arxiv.org/abs/2401.06432

718. Make Some Noise: Unlocking Language Model Parallel Inference Capability through Noisy Training

The paper introduces the Make Some Noise (MSN) training framework, which enhances parallel decoding capabilities of language models by adding noise during the training process. It demonstrates significant improvements in inference speed (2.3-2.7x) while maintaining model performance, making it a competitive alternative to existing methods.

https://arxiv.org/abs/2406.17404

739. Turn Waste into Worth: Rectifying Top-k Router of MoE

This paper introduces the Rectify-Router, an innovative approach to improve the efficiency of top-k routing in Sparse Mixture of Experts models by addressing issues of dropped tokens and excessive zero padding. The proposed solution includes Intra-GPU Rectification and Fill-in Rectification, which together enhance performance and accuracy over traditional routing methods.

https://arxiv.org/abs/2402.12399

752. A Learning Rate Path Switching Training Paradigm for Version Updates of Large Language Models

This paper presents a new training paradigm called learning rate path switching for version updates of Large Language Models (LLMs). The proposed method significantly reduces training costs while maintaining high pre-training performance by optimizing learning rate adjustments during updates.

https://arxiv.org/abs/2410.04103

808. Scalable Efficient Training of Large Language Models with Low-dimensional Projected Attention

This paper introduces Low-dimensional Projected Attention (LPA) as a method to enhance the efficiency and effectiveness of large language models (LLMs) by targeting reduced parameters in the attention layers. Experimental results demonstrate that LPA can reduce training time and improve performance metrics on downstream tasks compared to traditional Transformer models.

https://arxiv.org/abs/2411.02063

814. DynaThink: Fast or Slow? A Dynamic Decision-Making Framework for Large Language Models

This paper presents DynaThink, a framework enabling large language models to choose between fast and slow inference methods based on task complexity and confidence levels. Experiments show that this dynamic decision-making approach optimizes both efficiency and effectiveness in reasoning tasks.

https://arxiv.org/abs/2407.01009

831. SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers

This paper introduces SparseGrad, a selective method for efficiently fine-tuning Multi-Layer Perceptron (MLP) layers in Transformer models by maintaining only the most significant gradients. SparseGrad significantly reduces memory usage and outperforms existing PEFT methods when fine-tuning BERT, RoBERTa, and LLaMA-2 for natural language understanding and question-answering tasks.

https://arxiv.org/abs/2410.07383

835. GRASS: Compute Efficient Low-Memory LLM Training with Structured Sparse Gradients

Grass introduces a method for efficiently training large language models by using structured sparse gradients, which reduce memory usage and improve computational throughput. The approach significantly enhances the feasibility of pretraining large models on limited hardware without sacrificing performance.

https://arxiv.org/abs/2406.17660

897. InfiniPot: Infinite Context Processing on Memory-Constrained LLMs

InfiniPot introduces a KV cache control framework to facilitate long context processing in memory-constrained environments for Large Language Models (LLMs). It employs Continual Context Distillation to retain critical information effectively, significantly outperforming traditional models in various NLP tasks.

https://arxiv.org/abs/2410.01518

975. LLoCO: Learning Long Contexts Offline

LLoCO presents a novel method for efficiently processing long contexts in large language models by compressing context and fine-tuning with LoRA. It significantly enhances the effective context window and demonstrates major improvements in inference speed and token usage during long-document question answering.

https://arxiv.org/abs/2404.07979

987. Pruning via Merging: Compressing LLMs via Manifold Alignment Based Layer Merging

The paper introduces a novel model compression technique called Manifold-Based Knowledge Alignment and Layer Merging Compression (MKA), which merges similar layers in large language models (LLMs) to reduce their size while maintaining performance. MKA demonstrates substantial compression ratios on multiple benchmark datasets, outperforming traditional pruning methods, particularly when combined with quantization.

https://arxiv.org/abs/2406.16330

1015. HiFT: A Hierarchical Full Parameter Fine-Tuning Strategy

The HiFT method proposes a novel hierarchical fine-tuning strategy that only updates a subset of parameters during training, significantly reducing GPU memory usage while maintaining comparable performance to standard fine-tuning. This approach supports various optimizers and enables full-parameter fine-tuning of large models on limited hardware, showcasing its efficiency and effectiveness.

http://arxiv.org/abs/2401.15207v3

1027. A Simple and Effective L_2 Norm-Based Strategy for KV Cache Compression

This paper presents a method for compressing the Key-Value (KV) cache in large language models by leveraging the L_2 norm of key embeddings, leading to substantial reductions in memory usage without sacrificing accuracy. The proposed strategy demonstrates a 50% reduction in KV cache size for language modeling tasks and an impressive 90% for retrieval tasks, ensuring compatibility with existing efficient attention mechanisms.

http://arxiv.org/abs/2406.11430v4

1038. CHESS: Optimizing LLM Inference via Channel-Wise Thresholding and Selective Sparsification

This paper introduces CHESS, a novel method for optimizing inference in large language models through channel-wise thresholding and selective sparsification. Experimental results indicate that CHESS reduces performance degradation while speeding up inference by activating fewer parameters across multiple downstream tasks.

https://arxiv.org/abs/2409.01366

1043. Transformers are Multi-State RNNs

This paper demonstrates that decoder-only transformers can be perceived as unbounded multi-state RNNs and presents a novel compression method, TOVA, which improves the efficiency of these models by reducing the size of their key-value cache. The results indicate that TOVA achieves significant throughput improvements while maintaining performance close to full model sizes across various long-range tasks.

https://arxiv.org/abs/2401.06104

1068. ShadowLLM: Predictor-based Contextual Sparsity for Large Language Models

This paper presents ShadowLLM, a novel predictor for enforcing contextual sparsity in large language models (LLMs), which leads to significant improvements in accuracy and speed compared to existing methods. By focusing on input-dependent sparsity patterns rather than simple magnitude-based pruning, ShadowLLM enhances performance on models with billions of parameters.

https://arxiv.org/abs/2406.16635

1141. AlphaExpert: Assigning LoRA Experts Based on Layer Training Quality

This paper presents AlphaLoRA, a method for efficiently allocating LoRA experts based on the training quality of different layers in Large Language Models. By leveraging Heavy-Tailed Self-Regularization Theory, it offers a training-free strategy that mitigates redundancy and improves task performance on various benchmarks.

https://arxiv.org/pdf/2410.10054

1168. ApiQ: Finetuning of 2-Bit Quantized Large Language Model

ApiQ offers a novel framework for the memory-efficient finetuning of large language models by addressing the challenges posed by quantization methods that lead to loss of information and knowledge. The approach is shown to improve finetuning results across various tasks and bit-widths by maintaining activation precision and minimizing error propagation in LLMs.

https://arxiv.org/abs/2402.05147

1176. Scalable Data Ablation Approximations for Language Models through Modular Training and Merging

This paper presents a method for efficiently approximating data ablation studies for Large Language Models (LLMs) by training individual models on subsets of data and leveraging their evaluations for various combinations. The approach allows for significant training efficiency improvements while maintaining the ability to assess and mix data mixtures incrementally, thus optimizing model performance.

https://arxiv.org/abs/2410.15661

1178. Attention Score is not All You Need for Token Importance Indicator in KV Cache Reduction: Value Also Matters

The paper proposes a novel method called Value-Aware Token Pruning (VATP) for token importance evaluation in large language models by incorporating both attention scores and value vector norms. Extensive experiments show that VATP significantly outperforms existing methods based solely on attention scores across multiple tasks, highlighting the importance of value in token significance.

https://arxiv.org/abs/2406.12335

1232. RevMUX: Data Multiplexing with Reversible Adapters for Efficient LLM Batch Inference

This paper presents RevMUX, a parameter-efficient multiplexing framework designed to enhance the efficiency of large language model (LLM) inference without needing to retrain the entire model. The method allows for merging multiple inputs into a single composite input while still recovering individual samples for accurate classification, demonstrating improved performance across various datasets and LLM architectures.

http://arxiv.org/abs/2410.04519

1264. Initialization of Large Language Models via Reparameterization to Mitigate Loss Spikes

This paper addresses the problem of loss spikes in the pre-training of large language models, identifying a non-uniformity in parameter norms as a contributing factor. It proposes a novel weight scaling reparameterization technique (WeSaR) that stabilizes training by ensuring uniform parameter norms, leading to improved training speed and performance in Transformer models.

https://arxiv.org/abs/2410.05052

56. AdaZeta: Adaptive Zeroth-Order Tensor-Train Adaption for Memory-Efficient Large Language Models Fine-Tuning

The paper introduces the AdaZeta framework, which enhances the performance and convergence of memory-efficient Zeroth-order (MeZO) fine-tuning methods for large language models. Through a tensorized adapter and an adaptive query number schedule, AdaZeta improves estimation accuracy and mitigates divergence issues during training.

https://arxiv.org/abs/2406.18060

742. Ouroboros: Generating Longer Drafts Phrase by Phrase for Faster Speculative Decoding

Ouroboros is a novel method that enhances the speed of speculative decoding in large language models by generating longer drafts in a parallelized manner without the need for additional training. The proposed system can achieve significant speedups compared to both speculative and vanilla decoding while maintaining model performance, as demonstrated through various text generation tasks.

https://arxiv.org/abs/2402.13720

370. Optimized Speculative Sampling for GPU Hardware Accelerators

This paper presents an optimization of speculative sampling methods for GPU hardware accelerators that enhances sampling speed while maintaining accuracy. By concurrently computing portions of intermediate matrices and approximating probability distributions, the authors achieve significant improvements in profiling time for tasks in automatic speech recognition and summarization.

https://arxiv.org/abs/2406.11016

925. BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training

This paper introduces the Picky BPE algorithm, a refined method for optimizing tokenization during tokenizer training. The new method enhances vocabulary efficiency and addresses issues related to under-trained tokens without negatively impacting performance on downstream tasks.

http://arxiv.org/abs/2409.04599v1

941. FFN-SkipLLM: A Hidden Gem for Autoregressive Decoding with Adaptive Feed Forward Skipping

This paper presents FFN-SkipLLM, a novel input-adaptive feed-forward skipping strategy designed to optimize autoregressive decoding in large language models by enabling them to skip 25-30% of feed-forward blocks without significant performance loss. The proposed method addresses issues related to computational overload and inefficiencies in existing early-exit strategies during language generation tasks.

https://arxiv.org/abs/2404.03865

1112. Reasoning in Token Economies: Budget-Aware Evaluation of LLM Reasoning Strategies

This paper critiques traditional evaluations of reasoning strategies in large language models, highlighting the importance of considering computational budget alongside performance metrics. It introduces a new framework that reveals how more complex reasoning strategies do not necessarily outperform simpler baselines when accounting for compute resources.

https://arxiv.org/abs/2406.06461

468. Deterministic Weighted L* Algorithm

1169. Memorize Step by Step: Efficient Long-Context Prefilling with Incremental Memory and Decremental Chunk

1202. Memory-Efficient Fine-Tuning of Transformers via Token Selection

Information Extraction

141. AutoScraper: A Progressive Understanding Web Agent for Web Scraper Generation

This paper introduces AutoScraper, a hierarchical two-stage framework that generates web scrapers using large language models (LLMs) to optimize data extraction from diverse web pages. It addresses the limitations of existing wrappers and language agents by enhancing adaptability and scalability, and proposes a new metric for measuring scraper performance.

https://arxiv.org/abs/2404.12753

165. MiniConGTS: A Near Ultimate Minimalist Contrastive Grid Tagging Scheme for Aspect Sentiment Triplet Extraction

This paper introduces MiniConGTS, a minimalist tagging scheme for Aspect Sentiment Triplet Extraction that aims to improve pretrained representations and reduce computational overhead. It provides a novel token-level contrastive learning strategy, yielding superior performance compared to existing state-of-the-art techniques, particularly in the context of few-shot learning scenarios with GPT-4.

https://arxiv.org/abs/2406.11234

352. Learning from Natural Language Explanations for Generalizable Entity Matching

This paper addresses the challenge of entity matching by re-casting it as a conditional generation task, enabling the use of natural language explanations to enhance model performance. The proposed method shows significant improvements in generalization to new data, outperforming traditional supervised approaches and demonstrating model robustness through explanation utilization.

https://arxiv.org/abs/2406.09330

355. Contrastive Entity Coreference and Disambiguation for Historical Texts

This paper introduces a method for coreference resolution and disambiguation tailored for historical texts, addressing the challenge of identifying individuals not present in contemporary knowledge bases. By leveraging a large-scale training dataset and employing contrastive bi-encoder models, the proposed approach outperforms existing models on both historical and some modern entity disambiguation tasks.

https://arxiv.org/abs/2406.15576

419. ADELIE: Aligning Large Language Models on Information Extraction

This paper presents ADELIE, a model designed to enhance the alignment of large language models for information extraction (IE) tasks. Through the construction of a specialized alignment corpus and innovative training methods, ADELIE achieves state-of-the-art performance across various IE datasets while maintaining its general capabilities.

https://arxiv.org/abs/2405.05008

652. Major Entity Identification: A Generalizable Alternative to Coreference Resolution

This paper introduces Major Entity Identification (MEI) as an alternative to coreference resolution, focusing on identifying major entities specified in the input without relying on additional annotated data. The authors demonstrate that MEI models generalize well across domains, making it suitable for practical applications such as entity mention searching.

https://arxiv.org/abs/2406.14654

660. NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data

This paper introduces NuNER, a language representation model specifically designed for the Named Entity Recognition (NER) task using data annotated by Large Language Models. NuNER excels in few-shot scenarios by being fine-tuned effectively and demonstrates significant performance improvements over similar-sized models, highlighting the importance of pre-training data diversity.

https://arxiv.org/abs/2402.15343

726. SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents

The paper presents a new dataset called SciER, which focuses on entity and relation extraction from full-text scientific documents, particularly emphasizing datasets, methods, and tasks. It includes a detailed fine-grained tagging system and comprehensive experiments to evaluate state-of-the-art models, aiming to advance the field of scientific information extraction.

https://arxiv.org/abs/2410.21155

756. OneNet: A Fine-Tuning Free Framework for Few-Shot Entity Linking via Large Language Model Prompting

This paper introduces OneNet, a framework for few-shot entity linking that leverages the capabilities of Large Language Models without requiring fine-tuning. It includes components for entity simplification, contextual linking, and consistency checking, demonstrating superior performance compared to existing methods across multiple benchmarks.

http://arxiv.org/abs/2410.07549

816. Weak Reward Model Transforms Generative Models into Robust Causal Event Extraction Systems

This paper addresses the challenges in evaluating causal event extraction tasks by proposing a weak-to-strong supervision method that relies on less annotated data while maintaining high performance. The authors trained evaluation models to align generative models with human preferences through Reinforcement Learning, demonstrating effective transfer across datasets.

https://arxiv.org/abs/2406.18245

39. CoCoLoFa: A Dataset of News Comments with Common Logical Fallacies Written by LLM-Assisted Crowds

This paper presents CoCoLoFa, a highly detailed dataset consisting of 7,706 news comments annotated for logical fallacies, created through a combination of crowdsourcing and an LLM-assisted interface. The dataset enables the development and validation of fallacy detection models, demonstrating high detection and classification performance through BERT-based models.

http://arxiv.org/abs/2410.03457v1

51. In-context Contrastive Learning for Event Causality Identification

This paper introduces In-Context Contrastive Learning (ICCL), a model designed to improve Event Causality Identification (ECI) by incorporating contrastive learning for both positive and negative event demonstrations. ICCL demonstrates significant performance enhancements over existing state-of-the-art methods when evaluated on established datasets like EventStoryLine and Causal-TimeBank.

https://arxiv.org/abs/2405.10512

87. Advancing Event Causality Identification via Heuristic Semantic Dependency Inquiry Network

This paper introduces a new method, SemDI, for Event Causality Identification (ECI) that overcomes limitations of existing methods by capturing semantic dependencies within texts. It employs a unified encoder and a Cloze Analyzer to determine causal relations, achieving superior results compared to state-of-the-art techniques across three benchmarks.

https://arxiv.org/abs/2409.13621

178. SEER: Self-Aligned Evidence Extraction for Retrieval-Augmented Generation

The paper presents SEER, a model-based evidence extraction framework designed to optimize performance in Retrieval-Augmented Generation (RAG) by addressing existing shortcomings such as poor generalization and semantics deficiency. Extensive experiments demonstrate that SEER significantly enhances the performance of RAG by improving the faithfulness, helpfulness, and conciseness of extracted evidence, while also reducing evidence length.

http://arxiv.org/abs/2410.11315

548. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction

This paper presents a framework named Extract-Define-Canonicalize (EDC) for constructing knowledge graphs (KGs) from text using large language models. The EDC framework addresses challenges in applying LLMs for KGC, providing a flexible approach that allows both predefined and self-generated schemas, and achieves high-quality extraction without parameter tuning.

https://arxiv.org/abs/2404.03868

624. One2Set + Large Language Model: Best Partners for Keyphrase Generation

This paper introduces a generate-then-select framework for keyphrase generation (KPG) that leverages both a one2set model for candidate generation and a large language model (LLM) for selection, addressing challenges in recall and precision. Experimental results demonstrate that this approach outperforms state-of-the-art methods, particularly in the area of absent keyphrase prediction.

https://arxiv.org/abs/2410.03421

720. SPEED++: A Multilingual Event Extraction Framework for Epidemic Prediction and Preparedness

The paper presents SPEED++, a multilingual event extraction framework designed to extract epidemic-related information from social media posts in various languages to aid in epidemic prediction and preparedness. It introduces an extended ontology and a dataset comprising tweets in multiple languages while demonstrating the efficacy of zero-shot models for extracting epidemic information without prior training in local languages.

https://arxiv.org/abs/2410.18393

54. Large Language Models for Data Annotation: A Survey

This survey focuses on the use of Large Language Models (LLMs) for automating the data annotation process, which is traditionally labor-intensive and costly. It covers LLM-based annotation generation, assessment, and utilization, providing insights into challenges and strategies for improving data annotation efficiency.

https://arxiv.org/abs/2402.13446

217. C3PA: An Open Dataset of Expert-Annotated and Regulation-Aware Privacy Policies to Enable Scalable Regulatory Compliance Audits

This paper introduces C3PA, a novel open dataset containing expert-annotated privacy policies designed to aid in regulatory compliance audits, specifically targeting CCPA-related disclosure mandates. By utilizing this dataset, the paper aims to advance the development of tools for automated compliance auditing, which have been hampered by reliance on outdated and regulation-agnostic datasets.

http://arxiv.org/abs/2410.03925

388. Learning to Extract Structured Entities Using Language Models

This paper discusses advancements in information extraction using Language Models by proposing an entity-centric approach to evaluate model performance with new metrics. The authors introduce the Multistage Structured Entity Extraction (MuSEE) model, which enhances extraction tasks' effectiveness by decomposing them into multiple stages, confirmed through evaluations against baseline models.

https://arxiv.org/abs/2402.04437

463. Attribute or Abstain: Large Language Models as Long Document Assistants

This paper presents LAB, a benchmark evaluating large language models (LLMs) in the context of long documents, focusing on attribution and evidence-based responses. The findings reveal that citation methods perform well for larger models, while the quality of evidence influences response quality, underscoring different strategies for complex versus simple claims.

https://arxiv.org/abs/2407.07799

523. Text-Tuple-Table: Towards Information Integration in Text-to-Table Generation via Global Tuple Extraction

This paper introduces LiveSum, a benchmark dataset for generating summary tables from textual commentary, highlighting the challenges faced by LLMs in converting text to structured tables. The authors propose the T^3 pipeline to enhance table generation performance and demonstrate its effectiveness even without explicit training, addressing the limitations of previous methods.

https://arxiv.org/abs/2404.14215

540. Modeling Layout Reading Order as Ordering Relations for Visually-rich Document Understanding

This paper presents a new method for modeling layout reading order in visually-rich documents, arguing that traditional approaches fail to capture complete reading order information. By using ordering relations instead of permutations and creating a benchmark dataset, the authors demonstrate improved performance on various document understanding tasks.

https://arxiv.org/abs/2409.19672

655. MARE: Multi-Aspect Rationale Extractor on Unsupervised Rationale Extraction

The paper presents the Multi-Aspect Rationale Extractor (MARE), which enhances unsupervised rationale extraction by leveraging internal correlations between multiple aspects. MARE utilizes a Multi-Aspect Multi-Head Attention mechanism and multi-task training to achieve state-of-the-art performance on unsupervised rationale extraction benchmarks.

https://arxiv.org/abs/2410.03531

673. Explicit, Implicit, and Scattered: Revisiting Event Extraction to Capture Complex Arguments

This paper revisits event extraction by introducing implicit and scattered arguments, which existing frameworks fail to model adequately. It presents a novel dataset, DiscourseEE, to support the comprehensive extraction of these complex argument types in event modeling.

https://arxiv.org/abs/2410.03594

747. Grasping the Essentials: Tailoring Large Language Models for Zero-Shot Relation Extraction

This paper presents REPaL, a novel approach for zero-shot relation extraction that utilizes large language models to minimize the need for extensive annotated data. The method enhances performance by generating initial seed instances and refining relation learning through feedback mechanisms, demonstrating significant improvements on two datasets.

https://arxiv.org/abs/2402.11142

762. DynamicER: Resolving Emerging Mentions to Dynamic Entities for RAG

This paper introduces DynamicER, a solution for resolving emerging mentions to dynamic entities within knowledge bases, particularly in the context of retrieval-augmented generation (RAG). The proposed method involves a temporal segmented clustering approach that adapts to the evolving dynamics of entities and demonstrates improved performance on QA tasks compared to existing models.

https://arxiv.org/abs/2410.11494

763. Preserving Generalization of Language models in Few-shot Continual Relation Extraction

This paper presents a novel method aimed at enhancing Few-shot Continual Relation Extraction (FCRE) by using often-discarded language model heads to preserve prior knowledge while learning from limited data. The proposed approach employs a mutual information maximization strategy and leverages the capabilities of Large Language Models to improve overall model performance in this challenging domain.

https://arxiv.org/abs/2410.00334

819. Extracting Prompts by Inverting LLM Outputs

This paper proposes a method called output2prompt that extracts prompts from the outputs of language models without needing access to model internals or special queries. The method demonstrates zero-shot transferability across multiple large language models and introduces a memory-efficient sparse encoding technique.

https://arxiv.org/abs/2405.15012

919. ATAP: Automatic Template-Augmented Commonsense Knowledge Graph Completion via Pre-Trained Language Models

The paper proposes TAGREAL, a method that automatically generates prompts and retrieves information from text corpora to enhance open knowledge graph (KG) completion using pre-trained language models. The technique demonstrates state-of-the-art performance even with limited training data, outperforming existing methods across benchmark datasets.

https://arxiv.org/abs/2305.15597

988. Embedded Named Entity Recognition using Probing Classifiers

The paper introduces EMBER, an approach for streaming named entity recognition (NER) in decoder-only language models without the need for fine-tuning, allowing efficient token classification. Experiments demonstrate that EMBER achieves high token generation rates with minimal slowdown, improving computational costs for language model applications.

https://arxiv.org/abs/2403.11747

1003. Seg2Act: Global Context-aware Action Generation for Document Logical Structuring

Seg2Act is a generation-based method for extracting the hierarchical structure of documents by treating logical structure extraction as an action generation task. It improves document intelligence by using a global context-aware generative model that iteratively generates action sequences, showing superior performance in various experimental settings.

http://arxiv.org/abs/2410.06802

1049. Are Data Augmentation Methods in Named Entity Recognition Applicable for Uncertainty Estimation?

This paper explores how data augmentation techniques can enhance confidence calibration and uncertainty estimation in Named Entity Recognition (NER) tasks, particularly in high-stakes domains like healthcare and finance. The findings suggest that using data augmentation leads to better calibration and uncertainty measures, particularly in in-domain settings, and that lower perplexity in augmented data correlates with improved results.

https://arxiv.org/abs/2407.02062

1239. Is This a Bad Table? A Closer Look at the Evaluation of Table Generation from Text

This paper discusses the inadequacies of current metrics for evaluating the quality of generated tables from text and proposes a new evaluation strategy called TabEval. TabEval captures table semantics by comparing atomic statements derived from tables to ground truth statements, and is shown to correlate better with human judgments than existing methods.

https://arxiv.org/abs/2406.14829

1011. NoiseBench: Benchmarking the Impact of Real Label Noise on Named Entity Recognition

The paper introduces NoiseBench, a benchmark designed to study the impact of real label noise on named entity recognition (NER). By providing six types of real noise, the work highlights that this noise is more challenging than commonly used simulated noise, revealing significant shortcomings in current state-of-the-art models.

https://arxiv.org/abs/2405.07609

1087. A Fast and Sound Tagging Method for Discontinuous Named-Entity Recognition

This paper presents a new tagging method for identifying discontinuous named entities that leverages a weighted finite state automaton for inference. The method ensures soundness in predicted tag sequences while demonstrating competitive performance on biomedical datasets, emphasizing speed and simplicity over complexity.

https://arxiv.org/abs/2409.16243

1136. Will LLMs Replace the Encoder-Only Models in Temporal Relation Classification?

This paper investigates the performance of Large Language Models (LLMs) in the Temporal Relation Classification task, comparing them to encoder-only models like RoBERTa. It finds that LLMs with in-context learning significantly underperform smaller encoder models and explores reasons for this gap through explainable methods.

https://arxiv.org/abs/2410.10476

314. LogicST: A Logical Self-Training Framework for Document-Level Relation Extraction with Incomplete Annotations

588. Bio-RFX: Refining Biomedical Extraction via Advanced Relation Classification and Structural Constraints

1050. NeuroTrialNER: An Annotated Corpus for Neurological Diseases and Therapies in Clinical Trial Registries

49. Overcome Noise and Bias: Segmentation-Aided Multi-Granularity Denoising and Debiasing for Enhanced Quarduples Extraction in Dialogue

129. Integrating Structural Semantic Knowledge for Enhanced Information Extraction Pre-training

492. Exploring Nested Named Entity Recognition with Large Language Models: Methods, Challenges, and Insights

766. Topic-Oriented Open Relation Extraction with A Priori Seed Generation

95. Cross-domain NER with Generated Task-Oriented Knowledge: An Empirical Study from Information Density Perspective

399. Generative Models for Automatic Medical Decision Rule Extraction from Text

668. Multi-Level Cross-Modal Alignment for Speech Relation Extraction

772. Improving Knowledge Graph Completion with Structure-Aware Supervised Contrastive Learning

853. RoCEL: Advancing Table Entity Linking through Distinctive Row and Column Contexts

863. SRF: Enhancing Document-Level Relation Extraction with a Novel Secondary Reasoning Framework

962. MedCoT: Medical Chain of Thought via Hierarchical Expert

1025. LLMEdgeRefine: Enhancing Text Clustering with LLM-Based Boundary Point Refinement

103. Event Causality Identification with Synthetic Control

282. EVEDIT: Event-based Knowledge Editing for Deterministic Knowledge Propagation

855. Efficient Overshadowed Entity Disambiguation by Mitigating Shortcut Learning

901. TKGT: Redefinition and A New Way of Text-to-Table Tasks Based on Real World Demands and Knowledge Graphs Augmented LLMs

395. Knowledge-Centric Hallucination Detection

Generation

94. Standardize: Aligning Language Models with Expert-Defined Standards for Content Generation

The paper introduces Standardize, a framework for aligning language models with expert-defined standards to enhance content generation across various domains like education, healthcare, and engineering. Utilizing standards such as the CEFR and CCS, the research demonstrates significant improvements in the accuracy of content generated by language models when integrated with knowledge artifacts from these standards.

https://arxiv.org/abs/2402.12593

117. TCSinger: Zero-Shot Singing Voice Synthesis with Style Transfer and Multi-Level Style Control

This paper introduces TCSinger, a zero-shot singing voice synthesis model that utilizes style transfer and multi-level style control to generate singing voices with various unseen timbres and styles based on audio and text prompts. TCSinger achieves superior performance in synthesis quality, singer similarity, and style controllability through innovative modules designed for clustering style information, predicting style and duration, and enhancing generation details.

https://arxiv.org/abs/2409.15977

241. StyleRemix: Interpretable Authorship Obfuscation via Distillation and Perturbation of Style Elements

This paper presents StyleRemix, an interpretable authorship obfuscation technique that modifies specific stylistic elements of text while maintaining a low computational cost. StyleRemix is shown to outperform current state-of-the-art methods and is supported by the release of two valuable datasets for further research.

http://www.arxiv.org/abs/2408.15666

318. RSA-Control: A Pragmatics-Grounded Lightweight Controllable Text Generation Framework

RSA-Control introduces a pragmatics-grounded framework for controllable text generation that enhances attribute interpretation. It features a context-sensitive rationality adjustment and has shown promising results in maintaining fluency and consistency while achieving strong control of attributes.

https://www.arxiv.org/abs/2410.19109

382. MirrorStories: Reflecting Diversity through Personalized Narrative Generation with Large Language Models

This study introduces MirrorStories, a system utilizing Large Language Models to generate personalized narratives that reflect diverse identities, aiming to address the lack of representation in literature. The evaluation demonstrates that these personalized stories outperform generic narratives in engagement metrics and textual diversity, highlighting the models' capability to incorporate identity elements effectively.

https://arxiv.org/abs/2409.13935

386. Dynamic Multi-Reward Weighting for Multi-Style Controllable Generation

This paper explores methods for dynamically controlling multiple textual styles in text generation through multi-objective reinforcement learning. It presents a novel approach to weighting rewards that enhances style control while preserving linguistic quality, outperforming static approaches.

https://arxiv.org/abs/2402.14146

469. Towards Verifiable Text Generation with Evolving Memory and Self-Reflection

This paper introduces VTG, a framework for Verifiable Text Generation, which addresses issues of factual inaccuracies in language models by incorporating evolving memory and self-reflection mechanisms. Through extensive experiments, VTG demonstrates significant improvements over existing baseline methods in producing verifiable and accurate textual content.

http://arxiv.org/abs/2312.09075v3

484. Explaining and Improving Contrastive Decoding by Extrapolating the Probabilities of a Huge and Hypothetical LM

This paper investigates contrastive decoding (CD), a method for enhancing open-ended text generation using a smaller language model. It proposes a new unsupervised decoding technique called Asymptotic Probability Decoding (APD), which improves upon CD by better extrapolating probabilities from various sized models, achieving state-of-the-art results in factuality and commonsense question answering tasks.

https://arxiv.org/abs/2411.01610

514. CleanGen: Mitigating Backdoor Attacks for Generation Tasks in Large Language Models

This paper presents CleanGen, a novel defense mechanism against backdoor attacks in generation tasks performed by large language models (LLMs). CleanGen functions by detecting and replacing suspicious tokens during inference, resulting in significantly lower attack success rates while maintaining response helpfulness for benign queries.

https://arxiv.org/abs/2406.12257

527. Synchronous Faithfulness Monitoring for Trustworthy Retrieval-Augmented Generation

This paper presents SynCheck, a novel monitoring system designed to improve the trustworthiness of retrieval-augmented language models by detecting unfaithful outputs during generation. Additionally, it introduces FOD, a decoding algorithm that enhances the faithfulness of generated text, achieving significant improvements in performance across various tasks.

https://arxiv.org/abs/2406.13692

578. Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation

This paper introduces Contextualized Sequence Likelihood (CSL), a new confidence scoring method that enhances the reliability of sequence probabilities in natural language generation tasks by utilizing attention values from large language models. CSL improves upon existing confidence measures, demonstrating significant accuracy benefits in various question-answering datasets and across different LLMs.

https://arxiv.org/abs/2406.01806

745. Style-Specific Neurons for Steering LLMs in Text Style Transfer

The paper introduces sNeuron-TST, a method for steering large language models (LLMs) in text style transfer (TST) by utilizing style-specific neurons. The approach aims to improve the stylistic diversity of generated text while addressing the fluency issues that arise from deactivating certain source-style neurons, demonstrated through empirical results on multiple benchmarks.

http://arxiv.org/abs/2410.00593v1

798. ECIS-VQG: Generation of Entity-centric Information-seeking Questions from Videos

This paper presents a model for generating entity-centric information-seeking questions from videos, addressing the lack of focus on entities in existing question generation systems. A new dataset, VideoQuestions, consisting of YouTube videos and annotated questions, is introduced, along with a model architecture that incorporates Transformers and multimodal signals for improved question generation performance.

http://arxiv.org/abs/2410.09776v1

999. Label Confidence Weighted Learning for Target-level Sentence Simplification

This paper presents a novel approach called Label Confidence Weighted Learning (LCWL) for multi-level sentence simplification, focusing on generating simplified sentences with varying language proficiency levels. Experimental results show that LCWL significantly outperforms existing unsupervised methods, demonstrating its effectiveness in encoder-decoder models for text simplification tasks.

https://arxiv.org/abs/2410.05748

1046. Collective Critics for Creative Story Generation

This paper introduces the CritiCS framework, which enhances creative story generation by integrating a collective revision mechanism that involves both LLM critics and human writers. The framework significantly improves story creativity and reader engagement while maintaining coherence through iterative plan refinement and story generation stages.

https://arxiv.org/abs/2410.02428

1109. Not All Contexts Are Equal: Teaching LLMs Credibility-aware Generation

This paper introduces a framework called Credibility-aware Generation (CAG) that addresses the issue of flawed information in Retrieval-Augmented Generation (RAG) by enabling models to discern and process information based on its credibility. The proposed model demonstrates improved performance over traditional methods by effectively utilizing credibility for generation in real-world scenarios, reducing the impact of noisy input documents.

https://arxiv.org/html/2404.06809v1

1255. Improving Minimum Bayes Risk Decoding with Multi-Prompt

This paper proposes multi-prompt decoding to enhance the performance of instruction fine-tuned LLMs by utilizing a bank of diverse prompts to capture various generation methods. By implementing Minimum Bayes Risk (MBR) decoding, the study demonstrates improvements in generation quality across multiple tasks, models, and metrics due to a more extensive candidate space.

https://arxiv.org/abs/2407.15343

365. Satyrn: A Platform for Analytics Augmented Generation

This paper presents Satyrn, an analytics augmented generation (AAG) method that combines structured data analysis with large language models (LLMs) to produce accurate and fluent reports. Experiments showed that Satyrn generates reports with over 86% claim accuracy, outperforming a leading model in terms of factual accuracy.

http://arxiv.org/abs/2406.12069v2

599. AmbigNLG: Addressing Task Ambiguity in Instruction for NLG

The paper presents AmbigNLG, a task aimed at resolving task ambiguity in instructions for Natural Language Generation (NLG), which negatively affects the performance of Large Language Models (LLMs). It introduces an ambiguity taxonomy and a dataset, showing significant improvements in generated text alignment with user expectations through comprehensive experiments.

https://arxiv.org/abs/2402.17717

640. Text2Chart31: Instruction Tuning for Chart Generation with Automatic Feedback

This paper introduces Text2Chart31, a new dataset for chart generation that includes 31 unique plot types and 11.1K tuples of descriptions, code, data tables, and plots. The authors propose a reinforcement learning-based instruction tuning technique that enhances model performance in data visualization tasks without requiring human feedback.

https://arxiv.org/abs/2410.04064

706. Atomic Self-Consistency for Better Long Form Generations

This paper introduces Atomic Self-Consistency (ASC), a technique designed to enhance long-form responses generated by language models by merging relevant subparts from multiple samples. It demonstrates that ASC performs better than previous methods by improving both the recall and precision of information in responses, with experimental validation across various datasets.

https://arxiv.org/abs/2405.13131

1039. Semformer: Transformer Language Models with Semantic Planning

The paper presents Semformer, a new method for training Transformer language models that incorporates semantic planning into the token prediction process, avoiding the pitfalls of shortcut learning prevalent in traditional methods. Through a focus on a planning task and pretraining from scratch, Semformer demonstrates improved performance metrics and capabilities in summarization tasks.

http://arxiv.org/abs/2409.11143v1

1053. Generation with Dynamic Vocabulary

This paper presents a novel dynamic vocabulary for language models that allows for the generation of arbitrary text spans, thereby improving both the quality and efficiency of text generation. The dynamic vocabulary can be easily integrated into various applications and demonstrates significant enhancements in citation generation for question answering tasks.

https://arxiv.org/abs/2410.08481

1058. Prove Your Point!: Bringing Proof-Enhancement Principles to Argumentative Essay Generation

This paper presents a two-stage framework for argumentative essay generation (AEG) that focuses on enhancing logical consistency and proof validity in the generated texts. By using a large language model to construct pseudo-labels for claims and employing a tree planning approach, the framework aims to produce essays that are more logically coherent and persuasive than previous models.

https://arxiv.org/abs/2410.22642

99. MatchTime: Towards Automatic Soccer Game Commentary Generation

This paper presents MatchTime, a model for automatically generating soccer game commentary, which aims to enhance viewer experience. The authors created a high-quality dataset by correcting misalignments in existing datasets and demonstrated that their model, MatchVoice, achieves state-of-the-art performance in commentary generation through extensive experimentation.

https://arxiv.org/abs/2406.18530

270. Retrieve-Plan-Generation: An Iterative Planning and Answering Framework for Knowledge-Intensive LLM Generation

The paper proposes the Retrieve-Plan-Generation (RPG) framework to enhance the relevance of responses from large language models (LLMs) using an iterative process of planning and answering. By utilizing a multi-task prompt-tuning method, RPG helps LLMs manage both generating plans and refining answers based on relevant information, significantly improving performance on knowledge-intensive tasks.

https://arxiv.org/abs/2406.14979

512. DiVERT: Distractor Generation with Variational Errors Represented as Text for Math Multiple-choice Questions

The paper presents DiVERT, a novel method for generating high-quality distractors for math multiple-choice questions by representing the underlying errors as text. DiVERT outperforms state-of-the-art methods and produces error labels that are comparable to those created by human educators, demonstrating the effectiveness of its approach to automated distractor generation.

https://arxiv.org/abs/2406.19356

601. DataTales: A Benchmark for Real-World Intelligent Data Narration

DataTales presents a benchmark for evaluating language models in the task of transforming complex tabular data into coherent narratives, highlighting the analytical challenges involved in data narration. The benchmark includes 4.9k financial reports linked to market data, revealing the need for models to enhance their understanding of specialized terminology and analytical depth.

https://arxiv.org/abs/2410.17859

618. VLEU: a Method for Automatic Evaluation for Generalizability of Text-to-Image Models

This paper introduces the Visual Language Evaluation Understudy (VLEU), a new metric designed to evaluate the generalizability of Text-to-Image (T2I) models. VLEU quantifies a model's ability to handle diverse textual prompts and provides a way to compare different T2I models during their evaluation and finetuning.

https://arxiv.org/abs/2409.14704

727. Analysis of Plan-based Retrieval for Grounded Text Generation

This paper investigates the application of plan-based retrieval mechanisms to enhance grounded text generation and mitigate hallucinations in language models. The proposed method aims to improve the relevance and attribution of information in generated content by leveraging planning capabilities alongside retrieval systems.

https://arxiv.org/abs/2408.10490

767. Related Work and Citation Text Generation: A Survey

This paper provides a survey of related work and citation text generation, highlighting the significance and challenges of the task in the context of literature review writing. It also discusses the evolution of the RWG task and its connection to state-of-the-art NLP models.

https://arxiv.org/abs/2404.11588

800. Evaluating n-Gram Novelty of Language Models Using Rusty-DAWG

This paper investigates the novelty of texts generated by language models compared to their training data by evaluating the probability of training n-grams and the proportion of n-grams that are novel. The authors develop a novel search tool called Rusty-DAWG to facilitate this process and find that LM-generated text tends to be less novel than human-written text for larger n-grams, while exploring various factors affecting this novelty.

https://arxiv.org/abs/2406.13069

870. Unlocking Anticipatory Text Generation: A Constrained Approach for Large Language Models Decoding

This paper introduces a constrained approach to improve text generation by Large Language Models, targeting issues like toxicity and hallucination while ensuring adherence to given prompts. Through extensive experimentation across multiple tasks, the proposed future-constrained generation framework shows significant effectiveness in guiding the generation process.

https://arxiv.org/abs/2312.06149

953. Measuring Psychological Depth in Language Models

This paper introduces the Psychological Depth Scale (PDS) to evaluate the emotional and narrative complexity of creative stories generated by language models. It demonstrates that LLMs can produce stories that evoke similar levels of empathy and engagement as those written by humans, highlighting the need to shift focus from purely objective metrics to subjective impact.

https://arxiv.org/abs/2406.12680

978. Are Large Language Models Capable of Generating Human-Level Narratives?

This paper evaluates the storytelling capabilities of large language models (LLMs), highlighting their deficiencies compared to human narrative writing in suspense, diversity, and emotional depth. A novel computational framework is presented to analyze narratives through various aspects, with findings that suggest integrating discourse features can significantly improve LLM-generated stories.

https://arxiv.org/abs/2407.13248

1040. DocCGen: Document-based Controlled Code Generation

DocCGen is a framework for improving code generation from natural language to domain-specific languages by leveraging detailed documentation. It enhances the process through a two-step mechanism that detects relevant libraries and applies schema rules from the documentation, consistently yielding better results in reducing errors during code generation for structured languages.

https://arxiv.org/abs/2406.11925

1073. DataNarrative: Automated Data-Driven Storytelling with Visualizations and Texts

This paper presents DataNarrative, a framework for generating automated data-driven storytelling that combines both visualizations and text. It introduces a novel task for data story generation and demonstrates a multiagent system using Large Language Models for improving the coherence and quality of the generated narratives.

https://arxiv.org/abs/2408.05346

1190. Contrastive Policy Gradient: Aligning LLMs on sequence-level scores in a supervised-friendly fashion

This paper introduces Contrastive Policy Gradient (CoPG), a new reinforcement learning algorithm that enables large language models to optimize policies using off-policy data without the reliance on importance sampling. The proposed method is demonstrated in experiments for fine-tuning large language models on summarization tasks, focusing on aligning model outputs with human judgment using flexible reward structures.

https://arxiv.org/abs/2406.19185

1096. Pron vs Prompt: Can Large Language Models already Challenge a World-Class Fiction Author at Creative Text Writing?

This paper explores whether Large Language Models (LLMs) like GPT-4 can compete with a world-class novelist, Patricio Pron, in creative text writing. Through a contest involving manual assessments, the results show that LLMs currently fall short of matching the creativity level of top human writers.

https://arxiv.org/abs/2407.01119

1097. Evaluating Diversity in Automatic Poetry Generation

This paper evaluates the diversity in poetry generated by automatic systems by comparing it to human-generated poetry across various dimensions such as structural, lexical, and semantic aspects. The findings indicate that current models often lack diversity and suggest that certain modeling techniques can enhance the creative output significantly.

https://arxiv.org/abs/2406.15267

255. F^{2}RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation

568. On the In-context Generation of Language Models

768. Curriculum Consistency Learning for Conditional Sentence Generation

899. CorrSynth - A Correlated Sampling Method for Diverse dataset Generation from LLMs

1148. SpecHub: Provable Acceleration to Multi-Draft Speculative Decoding

48. Making Large Language Models Better Reasoners with Orchestrated Streaming Experiences

234. Contextual and Parametric Knowledge: More Context, More Focus

297. Induct-Learn: Short Phrase Prompting with Instruction Induction

553. Multi-pass Decoding for Grammatical Error Correction

686. Subjective Topic meets LLMs: Unleashing Comprehensive, Reflective and Creative Thinking through the Negation of Negation

749. Leveraging Context-aware Prompting for Commit Message Generation

902. Free your mouse! Command Large Language Models to Generate Code to Format Word Documents

963. Varying Sentence Representations via Condition-Specified Routers

1216. Do LLMs Plan Like Human Writers? Comparing Journalist Coverage of Press Releases with LLMs

1222. Adversarial Text Generation using Large Language Models for Dementia Detection

88. Exploring Union and Intersection of Visual Regions for Generating Questions, Answers, and Distractors

127. MantisScore: A Reliable Fine-grained Metric for Video Generation

433. Divide and Conquer Radiology Report Generation via Observation Level Fine-grained Pretraining and Prompt Tuning

577. KNN-Instruct: Automatic Instruction Construction with K Nearest Neighbor Deduction

675. Unknown Claims: Generation of Fact-Checking Training Examples from Unstructured and Structured Data

723. Story Morals: Surfacing value-driven narrative schemas using large language models

995. PepRec: Progressive Enhancement of Prompting for Recommendation

Linguistic theories, Cognitive Modeling and Psycholinguistics

53. Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs

This paper investigates how language models learn about rare grammatical phenomena, specifically the Article+Adjective+Numeral+Noun (AANN) construction. The findings suggest that language models can generalize from more common constructions to learn these rare phenomena, particularly when there is increased variability in the input data.

https://arxiv.org/abs/2403.19827

122. Decoding the Echoes of Vision from fMRI: Memory Disentangling for Past Semantic Information

This study explores how the human brain encodes and retrieves visual memories during continuous visual processing, introducing a new task called Memory Disentangling to decode past information from fMRI signals. A novel disentangled contrastive learning method is proposed to mitigate interference in fMRI signal decoding, showing effective separation of current and past visual information.

http://arxiv.org/abs/2409.20428

167. With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models

This study explores how multimodal large language models (LLMs) and vision language models (VLMs) understand sound symbolism, using classical psycholinguistic tasks and comparing their performance to human judgements. Findings indicate that these models can recognize certain sound-meaning associations but require more information than humans, with model size influencing their understanding of linguistic iconicity.

https://arxiv.org/abs/2409.14917

179. On The Role of Context in Reading Time Prediction

This paper explores the role of context in reading time prediction through a new method that orthogonalizes contextual predictors from language models. It reveals that the impact of context on reading times may be less significant than previously thought, suggesting a need for reevaluation in the understanding of context in language processing.

https://arxiv.org/abs/2409.08160

198. Fine-Grained Prediction of Reading Comprehension from Eye Movements

This paper explores the potential of predicting reading comprehension from eye movement data during reading tasks. Utilizing multimodal language models, the authors demonstrate that eye movements can provide valuable signals for this prediction, highlighting the complexity of the task and the effectiveness of their approach.

https://arxiv.org/abs/2410.04484

522. RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs

This paper presents RuBLiMP, a benchmark for evaluating the grammatical knowledge of language models in Russian through the use of linguistic minimal pairs. The dataset contains 45k sentence pairs focusing on various grammatical phenomena, highlighting the limitations of language models in understanding structural relations and advanced grammatical aspects compared to humans.

https://arxiv.org/abs/2406.19232

539. Development of Cognitive Intelligence in Pre-trained Language Models

This paper explores the emergent cognitive abilities of large pre-trained language models (PLMs) and their alignment with human cognitive development under a developmental framework. The study evaluates several PLMs across various cognitive tasks and finds a notable trend where model performance aligns significantly with human cognition during specific training phases.

https://arxiv.org/abs/2407.01047

698. SLANG: New Concept Comprehension of Large Language Models

This research presents SLANG, a benchmark aimed at improving large language models' understanding of slang and memes on the Internet by autonomously integrating novel data. The approach employs causal inference to enhance comprehension of evolving linguistic shifts without continual retraining, showing significant improvements in precision and relevance.

https://arxiv.org/abs/2401.12585

725. AnaloBench: Benchmarking the Identification of Abstract and Long-context Analogies

This paper introduces AnaloBench, a benchmark designed to assess the analogical reasoning capabilities of language models (LMs) using human-like analogical thinking processes. The findings indicate that while scaling up LMs improves performance to some extent, it does not significantly assist in tasks that require recalling relevant analogies from extensive information or in complex scenarios.

https://arxiv.org/abs/2402.12370

769. A Systematic Analysis of Large Language Models as Soft Reasoners: The Case of Syllogistic Inferences

This paper systematically investigates the reasoning abilities of Large Language Models (LLMs) in the context of syllogistic reasoning. It explores the effects of chain-of-thought reasoning, in-context learning, and supervised fine-tuning on LLM performance, revealing that certain techniques enhance inference validity while mitigating biases.

https://arxiv.org/abs/2406.11341

932. The Odyssey of Commonsense Causality: From Foundational Benchmarks to Cutting-Edge Reasoning

This paper presents a comprehensive survey on commonsense causality, highlighting its importance for human intelligence and decision-making, particularly in legal contexts. It bridges the gap in research through a systematic overview of taxonomies, benchmarks, and proposed future research directions in this field.

https://arxiv.org/abs/2406.19307

1009. Why do objects have many names? A study on word informativeness in language use and lexical systems.

This paper investigates the diverse lexical choices available for referring to the same object, focusing on the interplay between informativeness and communication context. It proposes a novel measure of word informativeness and analyzes color naming in English and Mandarin, concluding that optimal lexical systems allow multiple names to enhance communication effectiveness.

https://arxiv.org/abs/2410.07827

1020. How to Compute the Probability of a Word

This paper discusses the accurate computation of word probabilities from language models, which is essential for understanding perplexity and surprisal. It highlights common errors in recent linguistic studies due to incorrect methodologies in probability computations, demonstrating their significant impact on research outcomes.

https://arxiv.org/abs/2406.14561

1024. Language models and brains align due to more than next-word prediction and word-level information

This paper investigates the alignment between pretrained language models and human brain activity during language comprehension. It suggests that the alignment is influenced by more than just next-word prediction and explores alternative shared mechanisms involved in this process.

https://arxiv.org/abs/2212.00596

1105. Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models

This paper investigates the theory of mind (ToM) capabilities of large language models (LLMs) by introducing two datasets to evaluate perception inference and perception-to-belief inference. The findings indicate that while LLMs are proficient at perception inference, they struggle with perception-to-belief inference, necessitating the development of a method called PercepToM to enhance their performance in ToM tasks.

https://arxiv.org/abs/2407.06004

222. Conditional and Modal Reasoning in Large Language Models

This paper investigates the logical reasoning capabilities of twenty-nine large language models, particularly focusing on their ability to handle conditionals and epistemic modals. The findings reveal that while zero-shot chain-of-thought prompting improves their reasoning, significant gaps remain compared to human judgments, indicating weaknesses in LLMs' logical reasoning abilities.

https://arxiv.org/abs/2401.17169

526. Reverse-Engineering the Reader

This paper explores the reverse engineering of language models to align them with human cognitive processes by optimizing them based on psychometric data. Through a novel technique, the authors demonstrate that aligning models to reading times improves psychometric predictive power, but negatively impacts model performance on standard NLP tasks.

https://arxiv.org/abs/2410.13086

206. MTLS: Making Texts into Linguistic Symbols

415. Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions

938. Scaling Cognitive Limits: Identifying Working Memory Limits in LLMs

950. Do LLMs learn a true syntactic universal?

1042. The Emergence of Compositional Languages in Multi-entity Referential Games: from Image to Graph Representations

1231. Is Child-Directed Speech Effective Training Data for Language Models?

283. Predicting Nonnative Sentence Processing with L2LMs

773. Contribution of Linguistic Typology to Universal Dependency Parsing: An Empirical Investigation

1146. Learnability of Indirect Evidence in Language Models

Question Answering

714. Large Language Models Can Self-Correct with Key Condition Verification

This paper presents the ProCo framework, which enhances large language models' (LLMs) ability to self-correct by verifying responses through a simple key condition masking method. Experimental results demonstrate significant improvements in accuracy across various reasoning tasks compared to prior self-correction methods.

https://arxiv.org/abs/2405.14092

153. QUITE: Quantifying Uncertainty in Natural Language Text in Bayesian Reasoning Scenarios

The paper introduces QUITE, a dataset designed for Bayesian reasoning scenarios that includes natural language nuances in stating probabilities and uncertainties. The findings highlight that neuro-symbolic models are superior in reasoning tasks compared to traditional large language models, and the dataset is released for further research.

http://arxiv.org/abs/2410.10449

205. CoTKR: Chain-of-Thought Enhanced Knowledge Rewriting for Complex Knowledge Graph Question Answering

This paper presents CoTKR, a Chain-of-Thought Enhanced Knowledge Rewriting method aimed at improving Knowledge Graph Question Answering (KGQA) by generating reasoning traces and enhancing knowledge representation. It also introduces a training strategy called PAQAF to optimize the feedback from the QA model, resulting in significant performance improvements over existing methods.

https://arxiv.org/abs/2409.19753

227. Empowering Large Language Model for Continual Video Question Answering with Collaborative Prompting

This paper addresses the challenges of continual learning in Video Question Answering (VideoQA) by introducing a novel approach called Collaborative Prompting (ColPro), which helps mitigate catastrophic forgetting in large language models. It demonstrates that ColPro outperforms existing methods on the NExT-QA and DramaQA datasets, showcasing its effectiveness in integrating various types of prompting to enhance model performance in dynamic video content environments.

https://arxiv.org/abs/2410.00771

321. REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering

The paper introduces REAR, a Relevance-Aware Retrieval-augmented framework for enhancing open-domain question answering by improving the relevance assessment of retrieved documents. By developing a new architecture with a ranking mechanism and refined training methods, REAR leverages external knowledge more effectively, demonstrating superior performance compared to existing methods in several QA tasks.

https://arxiv.org/abs/2402.17497

378. Right for Right Reasons: Large Language Models for Verifiable Commonsense Knowledge Graph Question Answering

The paper proposes R3, a commonsense Knowledge Graph Question Answering (KGQA) methodology that enhances the verifiability of reasoning processes in Large Language Models (LLMs). It addresses issues of hallucination in answering commonsense questions while providing experimental validation across various tasks, demonstrating improved performance over existing methods.

https://arxiv.org/abs/2403.01390

400. Encoding and Controlling Global Semantics for Long-form Video Question Answering

This paper presents a new approach to long-form video question answering (videoQA) by introducing a state space layer (SSL) to better integrate global semantics. The framework improves performance by overcoming the limitations of adaptive frame selection and introduces benchmarks for rigorous evaluation of videoQA systems.

https://arxiv.org/abs/2405.19723

544. TraveLER: A Modular Multi-LMM Agent Framework for Video Question-Answering

The paper introduces TraveLER, a modular framework designed to enhance video question-answering by using multiple agents to locate important information and evaluate their findings adaptively. This method allows for dynamic replanning when insufficient information is initially gathered, resulting in improved performance on various VideoQA benchmarks.

https://arxiv.org/abs/2404.01476

560. Enhancing Pre-Trained Generative Language Models with Question Attended Span Extraction on Machine Reading Comprehension

This paper introduces the Question-Attended Span Extraction (QASE) module, which enhances pre-trained generative language models in Machine Reading Comprehension (MRC) tasks. QASE allows these models to outperform extractive strategies while maintaining computational efficiency and achieves state-of-the-art results across various datasets.

http://arxiv.org/abs/2404.17991v3

770. Pre-training Cross-lingual Open Domain Question Answering with Large-scale Synthetic Supervision

This paper presents a unified approach to cross-lingual open domain question answering with a single encoder-decoder model, addressing the challenges of retrieval and answer generation in different languages. The authors introduce a self-supervised training method using linked Wikipedia pages, showing that their model, CLASS, excels in both supervised and zero-shot settings, outperforming existing methods that rely on machine translation.

https://arxiv.org/abs/2402.16508

805. KnowTuning: Knowledge-aware Fine-tuning for Large Language Models

The paper proposes a knowledge-aware fine-tuning method called KnowTuning, aimed at enhancing large language models' ability to utilize knowledge in answering questions. Through both fine-grained knowledge augmentation and coarse-grained knowledge comparison, KnowTuning improves the factual correctness and logical coherence of responses, validated by experiments on various datasets.

https://arxiv.org/abs/2402.11176

813. Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models

This paper introduces Chain-of-Noting (CoN), a new method to improve the robustness of retrieval-augmented language models (RALMs) by enabling better handling of irrelevant data and unknown queries. CoN generates sequential reading notes for retrieved documents, which enhances the model's ability to evaluate relevance and produce accurate answers, leading to significant performance improvements in open-domain QA tasks.

https://arxiv.org/abs/2311.09210

1023. Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering

This paper proposes a method called Generate-on-Graph (GoG) to enhance question answering with Incomplete Knowledge Graphs (IKG), addressing limitations of existing approaches that assume complete knowledge graphs. The GoG framework allows Large Language Models (LLMs) to generate new factual triples needed for answering questions by leveraging both internal and external knowledge sources.

https://arxiv.org/abs/2404.14741

1236. RE-RAG: Improving Open-Domain QA Performance and Interpretability with Relevance Estimator in Retrieval-Augmented Generation

The RE-RAG framework enhances open-domain question answering by introducing a relevance estimator that assesses the usefulness of retrieved contexts for answering questions. It utilizes weak supervision to train the estimator and proposes novel decoding strategies to improve performance and interpretability of generated responses.

https://arxiv.org/abs/2406.05794

1259. LongRAG: A Dual-perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering

This paper introduces LongRAG, a novel retrieval-augmented generation model designed to improve accuracy in long-context question answering. LongRAG addresses the challenges of existing models by enhancing understanding of global information and factual details, significantly outperforming prior approaches in experimental validations.

https://arxiv.org/abs/2410.18050

347. Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation

This paper introduces MIRAGE, a method for enhancing answer attribution in retrieval-augmented generation systems for question answering. MIRAGE utilizes model internals to connect context-sensitive answer tokens with relevant documents, achieving high agreement with human attribution and robust citation quality in multilingual settings.

https://arxiv.org/abs/2406.13663

625. Unlocking Markets: A Multilingual Benchmark to Cross-Market Question Answering

This paper presents a novel task called Multilingual Cross-market Product-based Question Answering (MCPQA), which leverages information from multiple e-commerce marketplaces to answer product-related questions in a multilingual context. It introduces a large dataset from 17 marketplaces and demonstrates that incorporating cross-market information improves performance in product-related question answering tasks.

https://arxiv.org/abs/2409.16025

1209. A Simple LLM Framework for Long-Range Video Question-Answering

This paper presents LLoVi, a framework for long-range video question-answering that effectively combines visual captioners with large language models to improve performance on video understanding tasks. By breaking down the problem into short and long-range processing stages, the authors demonstrate significant accuracy improvements on several benchmarks, outperforming prior state-of-the-art methods.

https://arxiv.org/abs/2312.17235

1251. ZEBRA: Zero-Shot Example-Based Retrieval Augmentation for Commonsense Question Answering

ZEBRA is a zero-shot example-based question answering framework that improves commonsense reasoning in language models by integrating knowledge retrieval and reasoning without additional training or templates. The evaluation on multiple benchmarks showcases its ability to enhance performance and interpretability, outperforming existing approaches.

http://arxiv.org/abs/2410.05077

101. Triad: A Framework Leveraging a Multi-Role LLM-based Agent to Solve Knowledge Base Question Answering

The paper presents Triad, a framework that employs a multi-role LLM-based agent to address knowledge base question answering (KBQA) tasks. By assigning distinct roles to the agent, the framework achieves superior performance on benchmark datasets compared to existing systems.

https://arxiv.org/abs/2402.14320

110. Self-Bootstrapped Visual-Language Model for Knowledge Selection and Question Answering

This paper presents a novel framework for enhancing visual question answering (VQA) by integrating a visual-language model with a retrieval-augmented generation approach. The proposed self-bootstrapping method significantly improves the retrieval and selection of relevant knowledge for answering complex questions, achieving state-of-the-art performance on the OK-VQA benchmark.

https://arxiv.org/abs/2404.13947

147. Seemingly Plausible Distractors in Multi-Hop Reasoning: Are Large Language Models Attentive Readers?

This paper investigates the multi-hop reasoning capabilities of Large Language Models (LLMs) and their tendency to exploit simplifying cues in benchmarks. It presents a new challenging benchmark that generates misleading reasoning paths, finding a significant decrease in performance when these are introduced.

https://arxiv.org/abs/2409.05197

199. Efficient Retriever for Multi-Hop Retrieval Question Answerin

The paper introduces EfficientRAG, a new approach to multi-hop question answering that improves on existing retrieval-augmented generation methods by reducing the reliance on large language model calls. Experimental results show that EfficientRAG outperforms previous methods across several datasets, demonstrating its effectiveness in handling complex queries.

https://arxiv.org/abs/2408.04259

212. Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism

The paper introduces a method called Learn to Refuse (L2R) to improve the reliability of large language models (LLMs) by incorporating a refusal mechanism, allowing LLMs to decline answering questions that exceed their knowledge scope. By utilizing a structured knowledge base that fetches validated knowledge, the authors demonstrate enhanced controllability and reliability through qualitative and quantitative analyses.

https://arxiv.org/abs/2311.01041

249. RAG-QA Arena: Evaluating Domain Robustness for Long-form Retrieval Augmented Question Answering

This paper introduces RAG-QA Arena, an evaluation framework for long-form retrieval augmented question answering that addresses the limitations of current datasets. The authors create a new dataset, Long-form RobustQA, consisting of human-written long-form answers, and demonstrate its effectiveness in comparing model-generated answers with human judgments.

https://arxiv.org/abs/2407.13998

504. Learning to Correct for QA Reasoning with Black-box LLMs

This paper presents a novel approach called CoBB that aims to enhance the reasoning capabilities of black-box large language models (LLMs) for question answering (QA) tasks. By utilizing a trained adaptation model and optimizing the selection of training pairs, CoBB improves reasoning accuracy without needing access to detailed output probabilities.

https://arxiv.org/abs/2406.18695

537. Attribute Diversity Determines the Systematicity Gap in VQA

This paper investigates the systematicity gap in visual question answering (VQA), where there is a performance difference when reasoning about familiar versus new combinations of object attributes. The results highlight that increasing the diversity of training data, rather than just the quantity, effectively reduces this systematicity gap, indicating that distinct attribute combinations seen during training enhance model systematicity.

https://arxiv.org/abs/2311.08695

594. Evidence-Focused Fact Summarization for Knowledge-Augmented Zero-Shot Question Answering

The paper introduces EFSum, a framework designed to enhance the performance of Large Language Models (LLMs) in zero-shot Question Answering (QA) by summarizing facts from Knowledge Graphs (KGs). EFSum addresses challenges in existing methods, such as evidence density and clarity, by optimizing an LLM for summarization through distillation and preference alignment, resulting in improved QA performance.

https://arxiv.org/abs/2403.02966

741. CommVQA: Situating Visual Question Answering in Communicative Contexts

CommVQA is a Visual Question Answering (VQA) dataset that incorporates communicative contexts, showing how the asking of questions is influenced by contextual information and prior knowledge. It highlights the limitations of existing VQA models that work with isolated image-question pairs and reveals high rates of hallucination and context misalignment in generated answers.

https://arxiv.org/abs/2402.15002

807. Nash CoT: Multi-Path Inference with Preference Equilibrium

This paper introduces Nash CoT, a multi-path inference framework that enhances reasoning accuracy in Large Language Models by using question-related role templates. It aims to balance role-specific and general generation in LLMs to maintain diverse reasoning without increasing inference costs, demonstrated through various inference tasks.

https://arxiv.org/abs/2407.07099

873. ERVQA: A Dataset to Benchmark the Readiness of Large Vision Language Models in Hospital Environments

This paper introduces the ERVQA dataset for benchmarking the capabilities of Large Vision Language Models in hospital environments, specifically through the Visual Question Answering (VQA) task. The analysis reveals the complexity of healthcare scenarios, emphasizing the need for specialized solutions and presents detailed error trends based on model properties.

http://arxiv.org/abs/2410.06420

889. Can LLMs replace Neil deGrasse Tyson? Evaluating the Reliability of LLMs as Science Communicators

This paper evaluates the reliability of large language models (LLMs) as science communicators through a novel dataset and benchmarking suite, focusing on their ability to answer nuanced scientific questions. The findings reveal that while GPT-4 Turbo outperforms many open-access models, it still exhibits significant reliability issues, highlighting the challenges in using LLMs for effective science communication.

https://arxiv.org/abs/2409.14037

956. Adaptive Question Answering: Enhancing Language Model Proficiency for Addressing Knowledge Conflicts with Source Citations

This paper addresses the challenge of resolving knowledge conflicts in Question Answering tasks by proposing a novel approach that includes source citations even in ambiguous settings with multiple valid answers. The authors introduce new datasets and metrics to facilitate research and evaluation in this area, aiming to enhance the trustworthiness and interpretability of QA systems.

https://arxiv.org/abs/2410.04241

985. Enhancing Post-Hoc Attributions in Long Document Comprehension via Coarse Grained Answer Decomposition

This paper explores a novel method for enhancing post-hoc attribution in long document comprehension by employing a coarse-grained approach to answer decomposition. The proposed method aims to improve the mapping of generated answers back to their source documents, providing a thorough examination of various attribution techniques.

http://arxiv.org/abs/2409.17073v2

1026. CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures

The paper presents the Multilingual CasiMedicos-Arg dataset, which is the first medical question answering dataset enriched with argumentative structures and explanations provided by doctors. It contains 558 clinical cases in four languages and aims to aid in improving explanation skills in AI education by demonstrating how competitive baselines perform on this dataset for the argument mining task.

https://arxiv.org/abs/2410.05235

1052. Towards Faithful Knowledge Graph Explanation Through Deep Alignment in Commonsense Question Answering

This paper addresses the challenge of generating faithful explanations in commonsense question answering by identifying key factors like LM-KG misalignment. It introduces the LM-KG Fidelity metric and the LKDA algorithm to enhance explanation fidelity and demonstrates its effectiveness through experiments on relevant datasets.

https://arxiv.org/abs/2310.04910

1194. COMPACT: Compressing Retrieved Documents Actively for Question Answering

This paper introduces CompAct, a framework designed to actively condense extensive documents while retaining vital information for question answering tasks. The proposed method shows significant improvements in performance and compression rates on multi-hop question-answering benchmarks, indicating a flexible integration with existing retrieval systems.

https://arxiv.org/abs/2407.09014

1201. Do great minds think alike? Investigating Human-AI Complementarity for Question Answering

This paper explores the comparative problem-solving abilities of humans and AI in question-answering tasks through a novel framework named CAIMIRA. The findings reveal strengths and weaknesses of both humans and AI systems in different reasoning skills, suggesting areas for future AI development to better complement human cognitive abilities.

https://arxiv.org/abs/2410.06524

62. RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models

This paper introduces RULE, a method to enhance the factual accuracy of Medical Large Vision Language Models (Med-LVLMs) using Retrieval-Augmented Generation (RAG). The approach addresses issues related to context retrieval and over-reliance on retrieved information, achieving an average improvement of 47.4% in factual accuracy across multiple medical datasets.

https://arxiv.org/abs/2407.05131

677. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?

This paper evaluates the effectiveness of domain-adaptive pretraining (DAPT) for medical large language models (LLMs) and vision-language models (VLMs), revealing that most medical models do not outperform their general-purpose counterparts in medical question-answering tasks. The results suggest that general-domain models may already possess significant medical knowledge and provide guidance for future research methodologies in this field.

https://arxiv.org/abs/2411.04118

961. StorySpark: Expert-Annotated QA Pairs with Real-World Knowledge for Children Storytelling

The paper introduces StorySparkQA, a dataset of 5,868 expert-annotated question-answer pairs designed to enhance children's story-based learning by infusing real-world knowledge. It addresses the inadequacies of existing QA datasets in capturing expert thinking during interactive story reading, providing a framework for generating effective QA pairs that extend beyond the narrative.

http://arxiv.org/abs/2311.09756v3

912. LONGAGENT: Achieving Question Answering for 128k-Token-Long Documents through Multi-Agent Collaboration

266. DVD: Dynamic Contrastive Decoding for Knowledge Amplification in Multi-Document Question Answering

394. TimeR^4 : Time-aware Retrieval-Augmented Large Language Models for Temporal Knowledge Graph Question Answering

1163. SciDQA: A Deep Reading Comprehension Dataset over Scientific Papers

91. MAR: Matching-Augmented Reasoning for Enhancing Visual-based Entity Question Answering

1140. You Make me Feel like a Natural Question: Training QA Systems on Transformed Trivia Questions

1249. Training-free Deep Concept Injection Enables Language Models for Video Question Answering

442. An Empirical Study of Multilingual Reasoning Distillation for Question Answering

815. Revisiting Automated Evaluation for Long-form Table Question Answering in the Era of Large Language Models

Semantics: Lexical, Sentence-level Semantics, Textual Inference and Other areas

791. Pcc-tuning: Breaking the Contrastive Learning Ceiling in Semantic Textual Similarity

The paper addresses the limitations in achieving higher performance in Semantic Textual Similarity (STS) using contrastive learning, identifying a ceiling of 87.5 for Spearman's correlation scores. It introduces a novel method called Pcc-tuning, which utilizes Pearson's correlation coefficient as a loss function to significantly enhance model performance with fewer annotated samples.

https://arxiv.org/abs/2406.09790

156. To Word Senses and Beyond: Inducing Concepts with Contextualized Language Models

This paper introduces Concept Induction, an unsupervised task for learning sets of concepts from words by generalizing Word Sense Induction. It employs a bi-level approach that uses both local and global perspectives, demonstrating their benefits in improving performance on lexical tasks.

https://arxiv.org/abs/2406.20054

663. Advancing Semantic Textual Similarity Modeling: A Regression Framework with Translated ReLU and Smooth K2 Loss

This paper introduces a regression framework to improve Semantic Textual Similarity (STS) modeling through two novel loss functions: Translated ReLU and Smooth K2 Loss. It demonstrates enhanced performance across several STS benchmarks, addressing limitations of existing contrastive learning approaches.

https://arxiv.org/abs/2406.05326

1229. FOLIO: Natural Language Reasoning with First-Order Logic

FOLIO is a newly introduced dataset designed to evaluate natural language reasoning capabilities using first-order logic (FOL), consisting of 1,430 examples annotated for logical correctness. The paper discusses the benchmark results of various state-of-the-art language models on this dataset, highlighting the challenges even advanced models like GPT-4 face in reasoning tasks.

https://arxiv.org/abs/2209.00840

140. In Search of the Long-Tail: Systematic Generation of Long-Tail Inferential Knowledge via Logical Rule Guided Search

This paper introduces LINK, a framework for generating long-tail inferential knowledge aimed at addressing the weaknesses of large language models in handling rare examples. It provides a dataset called LINT, evaluating LLMs on this dataset and revealing significant performance drops on long-tail data, emphasizing the need for robust evaluation methods.

https://arxiv.org/abs/2311.07237

404. Liar, Liar, Logical Mire: A Benchmark for Suppositional Reasoning in Large Language Models

This paper presents TruthQuest, a benchmark for evaluating suppositional reasoning in large language models using knights and knaves logical puzzles. The study reveals significant challenges faced by models like Llama 3 and Mixtral-8x7B in reasoning about truthfulness and the implications of statements.

https://arxiv.org/abs/2406.12546

531. Enhancing Systematic Decompositional Natural Language Inference Using Informal Logic

This paper introduces a consistent and theoretically grounded approach to annotating decompositional entailment to enhance natural language inference. The authors demonstrate that their new dataset and methods significantly improve the accuracy and proof quality of reasoning engines in textual inference tasks.

https://arxiv.org/abs/2402.14798

687. Experimental Contexts Can Facilitate Robust Semantic Property Inference in Language Models, but Inconsistently

This paper explores how experimental contexts can enhance the property inheritance capabilities of language models, emphasizing that these enhancements can yield inconsistent results. It highlights that, although language models show improved behavior with specific in-context examples, they can also fall back on superficial heuristics when tasks are modified slightly.

https://arxiv.org/abs/2401.06640

783. Where am I? Large Language Models Wandering between Semantics and Structures in Long Contexts

921. Towards a Semantically-aware Surprisal Theory

290. FOOL ME IF YOU CAN! An Adversarial Dataset to Investigate the Robustness of LMs in Word Sense Disambiguation

776. Automatically Generated Definitions and their utility for Modeling Word Meaning

1076. VerifyMatch: A Semi-Supervised Learning Paradigm for Natural Language Inference with Confidence-Aware MixUp

Syntax: Tagging, Chunking and Parsing

479. On Eliciting Syntax from Language Models via Hashing

The paper presents a method for unsupervised parsing through the use of binary representations to infer syntactic structures from raw text. It proposes an upgraded CKY algorithm that integrates syntax and lexicon in a unified manner and showcases competitive performance in deriving parsing trees from pre-trained language models.

https://arxiv.org/abs/2410.04074

635. Revisiting Supertagging for faster HPSG parsing

This paper introduces new supertaggers trained on English grammar-based treebanks to improve the speed and accuracy of HPSG parsing. It demonstrates that SVM and neural methods significantly enhance parsing performance while presenting diverse datasets which reveal varying results in parser evaluation.

https://arxiv.org/abs/2309.07590

665. Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation

The paper introduces a new model, Segment any Text (SaT), for robust and efficient sentence segmentation that is adaptable to various domains. It achieves high performance across multiple datasets while being less reliant on punctuation and significantly faster than existing methods.

https://arxiv.org/abs/2406.16678

645. Strengthening Structural Inductive Biases by Pre-training to Perform Syntactic Transformations

This paper proposes enhancing the structural inductive bias of Transformers through intermediate pre-training, which focuses on syntactic transformations such as going from active to passive voice. The results show that this approach aids in few-shot learning for syntactic tasks and improves generalization in semantic parsing.

https://arxiv.org/abs/2407.04543

659. Dependency Graph Parsing as Sequence Labeling

This paper proposes a novel approach to parsing syntactic dependencies by extending sequence labeling to accommodate graph-based representations, enabling it to handle complexities like reentrancy and cycles. Experimental results show that this method achieves competitive accuracy with a high level of efficiency compared to state-of-the-art parsers.

https://arxiv.org/abs/2410.17972

235. Semantic Training Signals Promote Hierarchical Syntactic Generalization in Transformers

Multilinguality and Language Diversity

296. SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages

SEACrowd is a comprehensive initiative designed to address the scarcity of high-quality multimodal datasets for Southeast Asian languages, focusing on nearly 1,000 indigenous languages. By providing standardized corpora and benchmarks across various tasks, SEACrowd aims to enhance the representation and performance of AI models in the region.

https://arxiv.org/abs/2406.10118

380. Understanding and Mitigating Language Confusion in LLMs

This paper examines the issue of language confusion in large language models (LLMs), specifically their struggle to consistently generate text in a user's preferred language. The authors introduce the Language Confusion Benchmark (LCB) to evaluate this phenomenon across multiple languages and propose methods to mitigate these failures.

https://arxiv.org/abs/2406.20052

396. Revealing the Parallel Multilingual Learning within Large Language Models

This study investigates the in-context learning capabilities of multilingual large language models (LLMs) through a method called Parallel Input in Multiple Languages (PiM), which improves comprehension significantly. It reveals a surprising effect where using multiple languages inhibits certain neurons, leading to more precise activation, supporting a theory akin to synaptic pruning in neuroscience.

https://arxiv.org/abs/2403.09073

441. Breaking Language Barriers: Cross-Lingual Continual Pre-Training at Scale

This paper investigates cross-lingual continual pretraining of Large Language Models (LLMs), focusing on its efficiency and effectiveness compared to traditional training methods. The findings highlight that continual pretraining allows for faster convergence and the potential for significant resource savings while facilitating the transferability of LLMs across languages.

http://arxiv.org/abs/2407.02118

451. A Large-Scale Investigation of Human-LLM Evaluator Agreement on Multilingual and Multi-Cultural Data

This paper investigates the evaluation of multilingual Large Language Models (LLMs) by assessing both human and LLM evaluator agreement across various Indic languages. It finds that while human and LLM evaluations agree well in pairwise comparisons, discrepancies arise in direct assessments, revealing biases in evaluations, particularly within GPT-based models.

https://arxiv.org/abs/2406.15053

457. Getting More from Less: Large Language Models are Good Spontaneous Multilingual Learners

This paper investigates the spontaneous multilingual alignment improvement of Large Language Models (LLMs) when instruction-tuned with question translation data. It demonstrates that LLMs can enhance their multilingual capabilities and generalization across languages even without annotated answers.

https://arxiv.org/abs/2405.13816

570. Towards Robust Speech Representation Learning for Thousands of Languages

This paper presents XEUS, a Cross-lingual Encoder for Universal Speech, which significantly enhances speech representation learning for over 4000 languages using self-supervised learning techniques. It demonstrates improved models for multilingual speech data by incorporating a novel dereverberation objective, resulting in state-of-the-art performance on various benchmarks despite a lesser pre-training data requirement.

https://arxiv.org/abs/2407.00837

572. PreAlign: Boosting Cross-Lingual Transfer by Early Establishment of Multilingual Alignment

The paper introduces PreAlign, a framework designed to enhance the multilingual alignment of large language models during their initial pretraining phase. By establishing multilingual alignment early on, PreAlign significantly improves cross-lingual transfer capabilities and effectiveness across model sizes in various contexts.

https://arxiv.org/abs/2407.16222

604. Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models

This paper introduces Cross-lingual Expert Language Models (X-ELM) as a solution to the underperformance of multilingual models compared to monolingual ones. X-ELM separately trains models on language subsets, which improves adaptation to new languages, performance, and reduces training hardware requirements.

https://arxiv.org/abs/2401.10440

662. Unraveling Babel: Exploring Multilingual Activation Patterns of LLMs and Their Applications

This paper investigates the multilingual activation patterns of large language models (LLMs) by analyzing internal neuron activities when processing various languages. It introduces a method for fine-grained analysis and demonstrates that insights from these activation patterns can enhance sparse activation and model pruning techniques.

https://arxiv.org/abs/2402.16367

682. ReadMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment

This paper introduces ReadMe++, a multilingual multi-domain dataset aimed at improving readability assessment across languages and domains. With evaluations of various language models, it highlights the dataset's ability to enhance few-shot prompting and reveals insights into the performance of unsupervised methods.

https://arxiv.org/abs/2305.14463

729. RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs

This paper presents a scalable method for generating high-quality multilingual feedback data aimed at improving alignment techniques for multilingual large language models (LLMs). The study successfully enhances multilingual preference optimization, achieving superior performance in 23 languages and contributing valuable insights into cross-lingual transfer benefits.

https://arxiv.org/abs/2407.02552

743. 1+1>2: Can Large Language Models Serve as Cross-Lingual Knowledge Aggregators?

This paper explores the use of Large Language Models (LLMs) as cross-lingual knowledge aggregators and presents methods to improve their multilingual performance. The proposed approach focuses on reducing performance disparities across languages through knowledge aggregation, highlighting the potential of LLMs in enhancing multilingual capabilities.

https://arxiv.org/abs/2406.14721

1065. Exploring Intra and Inter-language Consistency in Embeddings with ICA

This paper investigates the intra and inter-language consistency of semantic axes in word embeddings by using Independent Component Analysis (ICA). It establishes a statistical framework to ensure the reliability and universality of these axes across languages and within a single language.

https://arxiv.org/abs/2406.12474

1123. Multilingual Topic Classification in X: Dataset and Analysis

This paper presents X-Topic, a multilingual dataset designed for tweet topic classification across four languages, addressing the challenges of categorizing social media content in a multilingual context. The authors conduct an analysis to compare the effectiveness of existing language models in handling this diverse dataset.

https://arxiv.org/abs/2410.03075

1165. No Culture Left Behind: ArtELingo-28, a Benchmark of WikiArt with Captions in 28 Languages

ArtELingo-28 is a vision-language benchmark that includes captions in 28 languages, focusing on emotional diversity across cultures. The study emphasizes the importance of multilinguality and presents baseline results for various machine learning conditions dealing with cross-lingual transfer.

https://arxiv.org/abs/2411.03769

1268. Entity Insertion in Multilingual Linked Corpora: The Case of Wikipedia

This paper addresses the challenge of inserting links into multilingual content, specifically in Wikipedia, by developing a framework named LocEI and its multilingual variant XLocEI. The proposed frameworks effectively identify suitable positions for link insertion across 105 languages and outperform existing methods, demonstrating the importance of entity insertion models for content editors.

https://arxiv.org/abs/2410.04254

236. When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages

This paper explores the effects of multilinguality on language modeling performance across 250 languages, revealing that adding multilingual data improves performance for low-resource languages but can hinder high-resource languages. As dataset sizes grow, the study highlights a potential 'curse of multilinguality' where more multilingual data leads to decreased effectiveness due to limited model capacity.

https://arxiv.org/abs/2311.09205

247. An Analysis of Multilingual FActScore

This paper analyzes the performance of FActScore, a metric for assessing the factuality of texts generated by Large Language Models, across multiple languages, noting significant limitations in multilingual contexts. It introduces a novel dataset and finds that LLMs behave differently in fact extraction and scoring tasks depending on the language and the quality of the knowledge sources used.

https://arxiv.org/abs/2406.19415

315. Concept Space Alignment in Multilingual LLMs

This paper investigates the implicit vector space alignment in multilingual large language models (LLMs) and evaluates their generalization capabilities across different languages. It highlights that while larger models show quality alignments, they still struggle with generalization for languages of dissimilar typology and abstract concepts.

https://arxiv.org/abs/2410.01079

317. NLEBench+NorGLM: A Comprehensive Empirical Analysis and Benchmark Dataset for Generative Language Models in Norwegian

This paper presents a detailed evaluation and benchmarking dataset for generative language models in Norwegian, an underrepresented language in NLP research. It addresses the limitations of existing models on Norwegian tasks and offers insights into the effectiveness of pre-trained models and multi-task learning approaches for language understanding and generation.

https://arxiv.org/abs/2312.01314

330. Methods of Automatic Matrix Language Determination for Code-Switched Speech

This paper explores the concept of Matrix Language in code-switched speech using the Matrix Language Frame theory for automatic Matrix Language Identity determination. It finds that predictors from audio outperform traditional language identification methods, revealing preferences for non-English languages in mixed-language utterances.

https://arxiv.org/abs/2410.02521

542. Is It Good Data for Multilingual Instruction Tuning or Just Bad Multilingual Evaluation for Large Language Models?

This paper investigates the alignment of multilingual large language models with their intended use, specifically questioning the impact of using translated versus native instruction data on model performance. The authors highlight the shortcomings of current evaluation practices and suggest that regularization can help improve model outcomes in certain contexts.

https://arxiv.org/abs/2406.12822

1159. Investigating Multilingual Instruction-Tuning: Do Polyglot Models Demand for Multilingual Instructions?

This paper investigates the effects of instruction-tuning multilingual LLMs on parallel instruction-tuning benchmarks across several Indo-European languages. The findings reveal that instruction-tuning on parallel datasets enhances the model's cross-lingual capabilities and challenges the validity of the Superficial Alignment Hypothesis in general cases.

https://arxiv.org/abs/2402.13703

933. Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups

This paper investigates the effectiveness of large language models (LLMs) in Complex Word Identification (CWI) tasks across multilingual and multidomain setups, emphasizing their performance under various training conditions. The findings indicate that while LLMs show potential, they often fail to outperform smaller, existing methods used for CWI, lexical complexity prediction, and multi-word expressions.

https://arxiv.org/abs/2411.01706

1057. Efficient Unseen Language Adaptation for Multilingual Pre-Trained Language Models

683. GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text

323. On Mitigating Performance Disparities in Multilingual Speech Recognition

Phonology, Morphology and Word Segmentation

421. Lexically Grounded Subword Segmentation

This paper presents innovations in tokenization and subword segmentation, leveraging unsupervised morphological analysis and subword embeddings to achieve efficient and lexically aware segmentation methods. The proposed approaches demonstrate improved performance in morphological tasks such as part-of-speech tagging, and exhibit significant enhancements in segmentation precision across multiple languages.

https://arxiv.org/abs/2406.13560

1064. A Two-Step Approach for Data-Efficient French Pronunciation Learning

This paper proposes a novel two-step method for learning French pronunciation that effectively operates under data constraints. The approach combines grapheme-to-phoneme transcription and post-lexical processing to address the challenges of limited pronunciation data while maintaining effective phonological analysis.

https://arxiv.org/abs/2410.05698

1170. A Morphology-Based Investigation of Positional Encodings

This paper investigates the relationship between morphological complexity and the effectiveness of positional encodings in transformer-based language models across various languages and tasks. The study finds that positional encoding becomes less significant as the morphological complexity of a language increases, indicating a need to expand our understanding of positional encoding in linguistic contexts.

https://arxiv.org/abs/2404.04530

672. Subword Segmentation in LLMs: Looking at Inflection and Consistency

849. Optimizing Chinese Lexical Simplification Across Word Types: A Hybrid Approach

1055. Getting The Most Out of Your Training Data: Exploring Unsupervised Tasks for Morphological Inflection

1156. Automatic sentence segmentation of clinical record narratives in real-world data