Microsoft Research

Research Areas: Artificial Intelligence
Years: 2023
Number of Research Papers Evaluated: 210
Number of Research Papers After Evaluation: 65
Author: Dhruv Awasthi

CODEFUSION: A Pre-trained Diffusion Model for Code Generation (EMNLP, Nov 2023; Link)

Abstract:
Imagine a developer who can only change their last line of code — how often would they have to start writing a function from scratch before it is correct? Auto-regressive models for code generation from natural language have a similar limitation: they do not easily allow reconsid ering earlier tokens generated. We introduce CODEFUSION, a pre-trained diffusion code gen eration model that addresses this limitation by iteratively denoising a complete program con ditioned on the encoded natural language. We evaluate CODEFUSION on the task of natural language to code generation for Bash, Python, and Microsoft Excel conditional formatting (CF) rules. Experiments show that CODEFU SION (75M parameters) performs on par with state-of-the-art auto-regressive systems (350M 175B parameters) in top-1 accuracy and outper forms them in top-3 and top-5 accuracy, due to its better balance in diversity versus quality.

On comparing fair classifiers under data bias (NeurIPS, Nov 2023; Link)

Abstract:
In this paper, we consider a theoretical model for injecting data bias, namely, under-representation and label bias (Blum \& Stangl, 2019). We empirically study the effect of varying data biases on the accuracy and fairness of fair classifiers. Through extensive experiments on both synthetic and real-world datasets (e.g., Adult, German Credit, Bank Marketing, COMPAS), we empirically audit pre-, in-, and post-processing fair classifiers from standard fairness toolkits for their fairness and accuracy by injecting varying amounts of under-representation and label bias in their training data (but not the test data). Our main observations are: 1. The fairness and accuracy of many standard fair classifiers degrade severely as the bias injected in their training data increases; 2. A simple logistic regression model trained on the right data can often outperform, in both accuracy and fairness, most fair classifiers trained on biased training data, and 3. A few simple fairness techniques (e.g.,

reweighing, exponentiated gradients) seem to offer stable accuracy and fairness guarantees even when their training data is injected with under-representation and label bias. Our experiments also show how to integrate a measure of data bias risk in the existing fairness dashboards for real-world deployments.

Tree Prompting: Efficient Task Adaptation without Fine-Tuning (EMNLP, Nov 2023; Link)

Abstract:
Prompting language models (LMs) is the main interface for applying them to new tasks. However, for smaller LMs, prompting provides low accuracy compared to gradient-based finetuning. Tree Prompting is an approach to prompting which builds a decision tree of prompts, linking multiple LM calls together to solve a task. At inference time, each call to the LM is determined by efficiently routing the outcome of the previous call using the tree. Experiments on classification datasets show that Tree Prompting improves accuracy over competing methods and is competitive with fine-tuning. We also show that variants of Tree Prompting allow inspection of a model's decision-making process.

Enhancing Network Management Using Code Generated by Large Language Models (ACM, Nov 2023; Link)

Abstract:
Analyzing network topologies and communication graphs plays a crucial role in contemporary network management. However, the absence of a cohesive approach leads to a challenging learning curve, heightened errors, and inefficiencies. In this paper, we introduce a novel approach to facilitate a natural-language-based network management experience, utilizing large language models (LLMs) to generate task-specific code from natural language queries. This method tackles the challenges of explainability, scalability, and privacy by allowing network operators to inspect the generated code, eliminating the need to share network data with LLMs, and concentrating on application-specific requests combined with general program synthesis techniques. We design and evaluate a prototype system using benchmark applications, showcasing high accuracy, cost-effectiveness, and the potential for further enhancements using complementary program synthesis techniques.

Can Large Language Models Support Medical Facilitation Work? A Speculative Analysis (ACM, Nov 2023; Link)

Abstract:

Mobile messaging apps and SMS-based tools have been deployed to extend healthcare services beyond the clinic; peer support chat groups, consisting of patients and healthcare providers, can improve medication adherence. However, moderation can be burdensome for busy healthcare professionals who must respond to patients, provide accurate and timely information, and engage and build community among patients. In this paper, taking an ethnographic approach, we examine the moderation of chat groups for young people living with HIV in Kenya. We describe the roles and responsibilities of the moderator while striving to engage and build community among the participants and manage the group chat, highlighting the challenges they face. Grounded in the moderators' work, we explore how an LLM-enabled copilot could help or hinder group facilitation. In doing so, we contribute to discussions about the potential of Artificial Intelligence in supporting healthcare professionals.

Simple Data Sharing for Multi-Tasked Goal-Oriented Problems (NeurIPS, Oct 2023; Link)

Abstract:
Many important sequential decision problems — from robotics, games to logistics — are multi-tasked and goal-oriented. In this work, we frame them as Contextual Goal Oriented (CGO) problems, a goal-reaching special case of the contextual Markov decision process. CGO is a framework for designing multi-task agents that can follow instructions (represented by contexts) to solve goal-oriented tasks. We show that CGO problem can be systematically tackled using datasets that are commonly obtainable: an unsupervised interaction dataset of transitions and a supervised dataset of context-goal pairs. Leveraging the goal-oriented structure of CGO, we propose a simple data sharing technique that can provably solve CGO problems offline under natural assumptions on the datasets' quality. While an offline CGO problem is a special case of offline reinforcement learning (RL) with unlabelled data, running a generic offline RL algorithm here can be overly conservative since the goal-oriented structure of CGO is ignored. In contrast, our approach carefully constructs an augmented Markov Decision Process (MDP) to avoid introducing unnecessary pessimistic bias. In the experiments, we demonstrate our algorithm can learn near-optimal context-conditioned policies in simulated CGO problems, outperforming offline RL baselines.

Exploring the Boundaries of GPT-4 in Radiology (EMNLP, Oct 2023; Link)

Abstract:
The recent success of general-domain large language models (LLMs) has significantly changed the natural language processing paradigm towards a unified foundation model across domains and applications. In this paper, we focus on assessing the performance of GPT-4, the most capable LLM so far, on the text-based applications for radiology reports, comparing against state-of-the-art (SOTA) radiology-specific models. Exploring various

prompting strategies, we evaluated GPT-4 on a diverse range of common radiology tasks and we found GPT-4 either outperforms or is on par with current SOTA radiology models. With zero-shot prompting, GPT-4 already obtains substantial gains (≈ 10% absolute improvement) over radiology models in temporal sentence similarity classification (accuracy) and natural language inference (F1). For tasks that require learning dataset-specific style or schema (e.g. findings summarisation), GPT-4 improves with example-based prompting and matches supervised SOTA. Our extensive error analysis with a board-certified radiologist shows GPT-4 has a sufficient level of radiology knowledge with only occasional errors in complex context that require nuanced domain knowledge. For findings summarisation, GPT-4 outputs are found to be overall comparable with existing manually-written impressions.

Understanding Personalized Accessibility through Teachable AI: Designing and Evaluating Find My Things for People who are Blind or Low Vision (ASSETS, Oct 2023; Link)

Abstract:
The opportunity for artificial intelligence, or AI, to enable accessibility is rapidly growing, but widely impactful applications can be challenging to build given the diversity of user need within and across disability communities. Teachable AI systems give users with disabilities a way to leverage the power of AI to personalize applications for their own specific needs, as long as the effort of providing examples is balanced with the benefit of the personalization received. As an example, this paper presents the design and evaluation of Find My Things, an end-to-end application that can be taught by people who are blind or low vision to find their personal things. Through synthesis of the design process, this paper offers design considerations for the teaching loop that is so critical to realizing the power of teachable AI for accessibility.

Affective Conversational Agents: Understanding Expectations and Personal Influences (ArXiV, Oct 2023; Link)

Abstract:
The rise of AI conversational agents has broadened opportunities to enhance human capabilities across various domains. As these agents become more prevalent, it is crucial to investigate the impact of different affective abilities on their performance and user experience. In this study, we surveyed 745 respondents to understand the expectations and preferences regarding affective skills in various applications. Specifically, we assessed preferences concerning AI agents that can perceive, respond to, and simulate emotions across 32 distinct scenarios. Our results indicate a preference for scenarios that involve human interaction, emotional support, and creative tasks, with influences from factors such as emotional reappraisal and personality traits. Overall, the desired affective skills in AI

agents depend largely on the application's context and nature, emphasizing the need for adaptability and context-awareness in the design of affective AI conversational agents.

Will Code Remain a Relevant User Interface for End-User Programming with Generative AI Models? (ACM, Oct 2023; Link)

Abstract:
The research field of end-user programming has largely been concerned with helping non-experts learn to code sufficiently well in order to achieve their tasks. Generative AI stands to obviate this entirely by allowing users to generate code from naturalistic language prompts. In this essay, we explore the extent to which "traditional" programming languages remain relevant for non-expert end-user programmers in a world with generative AI. We posit the "generative shift hypothesis": that generative AI will create qualitative and quantitative expansions in the traditional scope of end-user programming. We outline some reasons that traditional programming languages may still be relevant and useful for end-user programmers. We speculate whether each of these reasons might be fundamental and enduring, or whether they may disappear with further improvements and innovations in generative AI. Finally, we articulate a set of implications for end-user programming research, including the possibility of needing to revisit many well-established core concepts, such as Ko's learning barriers and Blackwell's attention investment model.

Metadata-Based Detection of Child Sexual Abuse Material (IEEE, Oct 2023; Link)

Abstract:
Child Sexual Abuse Media (CSAM) is any visual record of a sexually explicit activity involving minors. Machine learning-based solutions can help law enforcement identify CSAM and block distribution. Yet, collecting CSAM imagery to train machine learning models has ethical and legal constraints. CSAM detection systems based on file metadata offer several opportunities. Metadata is not a record of a crime and, therefore, clear of legal restrictions. This paper proposes a CSAM detection framework consisting of machine learning models trained on file paths extracted from a real-world data set of over 1 million file paths obtained in criminal investigations. Our framework includes guidelines for model evaluation that account for data changes caused by adversarial data modification and variations in data distribution caused by limited access to training data, as well as an assessment of false positive rates against file paths from common crawl data. We achieve accuracies as high as 0.97 while presenting stable behavior under adversarial attacks previously used in natural language tasks. When evaluating the model on publicly available file paths from common crawl data, we observed a false positive rate of 0.002, showing that the model operating in distinct data distributions maintains low false positive rates.

LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression (Oct 2023; Link)

Abstract:
In long context scenarios, large language models (LLMs) face three main challenges: higher computational/financial cost, longer latency, and inferior performance. Some studies reveal that the performance of LLMs depends on both the density and the position of the key information (question relevant) in the input prompt. Inspired by these findings, we propose LongLLMLingua for prompt compression towards improving LLMs' perception of the key information to simultaneously address the three challenges. We conduct evaluation on a wide range of long context scenarios including single-/multi-document QA, few-shot learning, summarization, synthetic tasks, and code completion. The experimental results show that LongLLMLingua compressed prompt can derive higher performance with much less cost. The latency of the end-to-end system is also reduced. For example, on NaturalQuestions benchmark, LongLLMLingua gains a performance boost of up to 17.1% over the original prompt with ~4x fewer tokens as input to GPT-3.5-Turbo. It can derive cost savings of $28.5 and $27.4 per 1,000 samples from the LongBench and ZeroScrolls benchmark, respectively. Additionally, when compressing prompts of ~10k tokens at a compression rate of 2x-10x, LongLLMLingua can speed up the end-to-end latency by 1.4x-3.8x.

MEGA: Multilingual Evaluation of Generative AI (EMNLP, Oct 2023; Link)

Abstract:
Generative AI models have shown impressive performance on many Natural Language Processing tasks such as language understanding, reasoning and language generation. An important question being asked by the AI community today is about the capabilities and limits of these models, and it is clear that evaluating generative AI is very challenging. Most studies on generative LLMs have been restricted to English and it is unclear how capable these models are at understanding and generating text in other languages. We present the first comprehensive benchmarking of generative LLMs – MEGA, which evaluates models on standard NLP benchmarks, covering 16 NLP datasets across 70 typologically diverse languages. We compare the performance of generative LLMs including Chat-GPT and GPT-4 to State of the Art (SOTA) multilingual (both non-autoregressive and generative) models on these tasks to determine how well generative models perform compared to the previous generation of LLMs. We present a thorough analysis of the performance of models across languages and tasks and discuss challenges in improving the performance of generative LLMs on low-resource languages. We create a framework for evaluating generative LLMs in the multilingual setting and provide directions for future progress in the field.

On Surgical Fine-tuning for Language Encoder (EMNLP, Oct 2023; Link)

Abstract:
Fine-tuning all the layers of a pre-trained neural language encoder (either using all the parameters or using parameter-efficient methods) is often the de-facto way of adapting it to a new task. We show evidence that for different downstream language tasks, fine-tuning only a subset of layers is sufficient to obtain performance that is close to and often better than fine-tuning all the layers in the language encoder. We propose an efficient metric based on the diagonal of the Fisher information matrix (FIM score), to select the candidate layers for selective fine-tuning. We show, empirically on GLUE and SuperGLUE tasks and across distinct language encoders, that this metric can effectively select layers leading to a strong downstream performance. Our work highlights that task-specific information corresponding to a given downstream task is often localized within a few layers, and tuning only those is sufficient for strong performance. Additionally, we demonstrate the robustness of the FIM score to rank layers in a manner that remains constant during the optimization process.

Who's Harry Potter? Approximate Unlearning in LLMs (ArXiV, Oct 2023; Link)

Abstract:
Large language models (LLMs) are trained on massive internet corpora that often contain copyrighted content. This poses legal and ethical challenges for the developers and users of these models, as well as the original authors and publishers. In this paper, we propose a novel technique for unlearning a subset of the training data from a LLM, without having to retrain it from scratch.

We evaluate our technique on the task of unlearning the Harry Potter books from the Llama2-7b model (a generative language model recently open-sourced by Meta). While the model took over 184K GPU-hours to pretrain, we show that in about 1 GPU hour of finetuning, we effectively erase the model's ability to generate or recall Harry Potter-related content, while its performance on common benchmarks (such as Winogrande, Hellaswag, arc, boolq and piqa) remains almost unaffected. We make our fine-tuned model publicly available on HuggingFace for community evaluation. To the best of our knowledge, this is the first paper to present an effective technique for unlearning in generative language models.

Our technique consists of three main components: First, we use a reinforced model that is further trained on the target data to identify the tokens that are most related to the unlearning target, by comparing its logits with those of a baseline model. Second, we replace idiosyncratic expressions in the target data with generic counterparts, and leverage the model's own predictions to generate alternative labels for every token. These labels aim to approximate the next-token predictions of a model that has not been trained on the target

data. Third, we finetune the model on these alternative labels, which effectively erases the original text from the model's memory whenever it is prompted with its context.

ContiFormer: Continuous-Time Transformer for Irregular Time Series Modeling (NeurIPS, Oct 2023; Link)

Abstract:
Modeling continuous-time dynamics on irregular time series is critical to account for data evolution and correlations that occur continuously. The traditional methodologies including recurrent neural network or the Transformer model leverage inductive bias via powerful neural architectures to capture complex patterns. However, due to their discrete characteristic, they have limitations in generalizing to continuous-time data paradigm. Though Neural Ordinary Differential Equations (ODE) and their variants have shown promising results in dealing with irregular time series, they often fail to capture the intricate correlations within these sequences. It is challenging yet demanding to concurrently model the relationship between input data points and capture the dynamic changes of the continuous-time system. To tackle this problem, we propose ContiFormer that extends the relation modeling of vanilla Transformer to continuous domain, which explicitly incorporates the modeling abilities of continuous dynamics of Neural ODE with the attention mechanism of Transformers. We mathematically characterize the expressive power of ContiFormer and illustrated that, by curated designs of function hypothesis, many Transformer variants specialized in irregular time series modeling can be covered as a special case of ContiFormer. A wide range of experiments on both synthetic and real-world datasets have illustrated the superior modeling capacities and prediction performance of ContiFormer on irregular time series data.

Extensible Prompts for Language Models on Zero-shot Language Style Customization (NeurIPS, Oct 2023; Link)

Abstract:
We propose eXtensible Prompt (X-Prompt) for prompting a large language model (LLM) beyond natural language (NL). X-Prompt instructs an LLM with not only NL but also an extensible vocabulary of imaginary words. Imaginary words can help represent what NL words hardly describe, allowing a prompt to be more descriptive; also, they are designed to be out-of-distribution (OOD) robust so that they can be used like NL words in various prompts, distinguishing X-Prompt from soft prompt that is for fitting in-distribution data. To this end, we propose context-augmented learning (CAL) to learn imaginary words for general usability, enabling them to work properly in OOD (unseen) prompts. We conduct experiments that use X-Prompt for zero-shot language style customization as a case study. The promising results of X-Prompt demonstrate its potential of approaching advanced interaction between humans and LLMs to bridge their communication gap.

Deep Language Networks: Joint Prompt Training of Stacked LLMs using Variational Inference (NeurIPS, Oct 2023; Link)

Abstract:
We view large language models (LLMs) as stochastic \emph{language layers} in a network, where the learnable parameters are the natural language \emph{prompts} at each layer. We stack two such layers, feeding the output of one layer to the next. We call the stacked architecture a \emph{Deep Language Network} (DLN). We first show how to effectively perform prompt optimization for a 1-Layer language network (DLN-1). We then show how to train 2-layer DLNs (DLN-2), where two prompts must be learnt. We consider the output of the first layer as a latent variable to marginalize, and devise a variational inference algorithm for joint prompt training. A DLN-2 reaches higher performance than a single layer, sometimes comparable to few-shot GPT-4 even when each LLM in the network is smaller and less powerful.

The AI Revolution in Medicine GPT-4 and Beyond (Oct 2023; Link)

Abstract:
Just months ago, millions of people were stunned by ChatGPT's amazing abilities – and its bizarre hallucinations. But that was 2022. GPT-4 is now here: smarter, more accurate, with deeper technical knowledge. GPT-4 and its competitors and followers are on the verge of transforming medicine. But with lives on the line, you need to understand these technologies – stat.

What can they do? What can't they do – yet? What shouldn't they ever do? To decide, experience the cutting edge for yourself. Join three insiders who've had months of early access to GPT-4 as they reveal its momentous potential – to improve diagnoses, summarize patient visits, streamline processes, accelerate research, and much more. You'll see real GPT-4 dialogues – unrehearsed and unfiltered, brilliant and blundering alike– all annotated with invaluable context, candid commentary, real risk insights, and up-to-the-minute takeaways.

Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models (NeurIPS, Oct 2023; Link)

Abstract:
Large language models (LLMs) have achieved remarkable progress in solving various natural language processing tasks due to emergent reasoning abilities. However, LLMs have inherent limitations as they are incapable of accessing up-to-date information (stored on the

Web or in task-specific knowledge bases), using external tools, and performing precise mathematical and logical reasoning. In this paper, we present Chameleon, an AI system that mitigates these limitations by augmenting LLMs with plug-and-play modules for compositional reasoning. Chameleon synthesizes programs by composing various tools (e.g., LLMs, off-the-shelf vision models, web search engines, Python functions, and heuristic-based modules) for accomplishing complex reasoning tasks. At the heart of Chameleon is an LLM-based planner that assembles a sequence of tools to execute to generate the final response. We showcase the effectiveness of Chameleon on two multi-modal knowledge-intensive reasoning tasks: ScienceQA and TabMWP. Chameleon, powered by GPT-4, achieves an 86.54% overall accuracy on ScienceQA, improving the best published few-shot result by 11.37%. On TabMWP, GPT-4-powered Chameleon improves the accuracy by 17.0%, lifting the state of the art to 98.78%. Our analysis also shows that the GPT-4-powered planner exhibits more consistent and rational tool selection via inferring potential constraints from instructions, compared to a ChatGPT-powered planner.

ColDeco: An End User Spreadsheet Inspection Tool for AI-Generated Code (IEEE, Oct 2023; Link)

Abstract:
Code-generating large language models (LLMs) are transforming programming. Their capability to generate multi-step solutions provides even non-programmers a mechanism to harness the power of programming. Non-programmers typically use spreadsheets to manage tabular data, as it offers an intuitive understanding of data manipulation and formula outcomes. Considering that LLMs can generate complex, potentially incorrect code, our focus is on enabling user trust in the accuracy of LLM-generated code.

We present ColDeco, the first end user inspection tool for comprehending code produced by large language models for tabular data tasks. ColDeco integrates two new features for inspection with a grid-based interface. First, users can decompose a generated solution into intermediate helper columns to understand how the problem is solved. Second, users can interact with a filtered table of summary rows, which highlight interesting cases in the program. We evaluate our tool using a within-subjects user study (n=24) where participants are asked to verify the correctness of programs generated by a language model.
We found that while all features are independently useful and intuitive, participants preferred them in combination. Users especially noted the usefulness of helper columns, but wanted more transparency in how summary rows are generated to assist with understanding and trusting them. Users also highlighted the application of ColDeco in collaborative settings for explaining and understanding existing formulas.

Multimodal Agent — Localized Symbolic Knowledge Distillation for Visual Commonsense Models (NeurIPS, Oct 2023; Link)

Abstract:
Instruction following vision-language (VL) models like GPT-4 offer a flexible interface that supports a broad range of multimodal tasks in a zero-shot fashion. However, interfaces that operate on full images do not directly enable the user to “point to” and access specific regions within images. This capability is important not only to support reference-grounded VL benchmarks, but also, for practical applications that require precise within-image reasoning. We build Localized Visual Commonsense model which allows users to specify (multiple) regions-as-input. We train our model by sampling localized commonsense knowledge from a large language model (LLM): specifically, we prompt a LLM to collect common sense knowledge given a global literal image description and a local literal region description automatically generated by a set of VL models. This pipeline is scalable and fully automatic, as no aligned or human-authored image and text pairs are required. With a separately trained critic model that selects high quality examples, we find that training on the localized commonsense corpus expanded solely from images can successfully distill existing VL models to support a reference-as-input interface. Empirical results and human evaluations in zero-shot settings demonstrate that our distillation method results in more precise VL models of reasoning compared to a baseline of passing a generated referring expression.

Conservative State Value Estimation for Offline Reinforcement Learning (NeurIPS, Oct 2023; Link)

Abstract:
Offline reinforcement learning faces a significant challenge of value over-estimation due to the distributional drift between the dataset and the current learned policy, leading to learning failure in practice. The common approach is to incorporate a penalty term to reward or value estimation in the Bellman iterations. Meanwhile, to avoid extrapolation on out-of-distribution (OOD) states and actions, existing methods focus on conservative Q-function estimation. In this paper, we propose Conservative State Value Estimation (CSVE), a new approach that learns conservative V-function via directly imposing penalty on OOD states. Compared to prior work, CSVE allows more effective in-data policy optimization with conservative value guarantees. Further, we apply CSVE and develop a practical actor-critic algorithm in which the critic does the conservative value estimation by additionally sampling and penalizing the states \emph{around} the dataset, and the actor applies advantage weighted updates extended with state exploration to improve the policy. We evaluate in classic continual control tasks of D4RL, showing that our method performs better than the conservative Q-function learning methods and is strongly competitive among recent SOTA methods.

Language Models Augmented with Decoupled Memory (NeurIPS, Oct 2023; Link)

Abstract:

Existing large language models (LLMs) can only afford fix-sized inputs due to the input length limit, preventing them from utilizing rich long-context information from past inputs. To address this, we propose a framework, Decoupled-Memory-Augmented LLMs (DeMA), which enables LLMs to memorize long history. We design a novel decoupled network architecture with the original backbone LLM frozen as a memory encoder and an adaptive residual side-network as a memory retriever and reader. Such a decoupled memory design can easily cache and update long-term past contexts for memory retrieval without suffering from memory staleness. Enhanced with memory-augmented adaptation training, \our{} can thus memorize long past context and use long-term memory for language modeling. The proposed memory retrieval module can handle flexible context in its memory bank to benefit various downstream tasks, including memorizing long inputs for language modeling and caching many-shot demonstration examples for enhancing in-context learning. Experiments show that our method outperforms strong long-context models on ChapterBreak, a challenging long-context modeling benchmark, and achieves remarkable improvements on memory-augmented in-context learning over LLMs.The results demonstrate that the proposed method is effective in helping language models to memorize and utilize long-form contents.

Formalizing Natural Language Intent into Program Specifications via Large Language Models (Oct 2023; Link)

Abstract:
Informal natural language that describes code functionality, such as code comments or function documentation, may contain substantial information about a programs intent. However, there is typically no guarantee that a programs implementation and natural language documentation are aligned. In the case of a conflict, leveraging information in code-adjacent natural language has the potential to enhance fault localization, debugging, and code trustworthiness. In practice, however, this information is often underutilized due to the inherent ambiguity of natural language which makes natural language intent challenging to check programmatically. The “emergent abilities” of Large Language Models (LLMs) have the potential to facilitate the translation of natural language intent to programmatically checkable assertions. However, it is unclear if LLMs can correctly translate informal natural language specifications into formal specifications that match programmer intent. Additionally, it is unclear if such translation could be useful in practice. In this paper, we describe LLM4nl2post, the problem leveraging LLMs for transforming informal natural language to formal method postconditions, expressed as program assertions. We introduce and validate metrics to measure and compare different LLM4nl2post approaches, using the correctness and discriminative power of generated postconditions. We then perform qualitative and quantitative methods to assess the quality of LLM4nl2post postconditions, finding that they are generally correct and able to discriminate incorrect code. Finally, we find that LLM4nl2post via LLMs has the potential to be helpful in practice; specifications generated from natural language were able to catch 70 real-world historical bugs from Defects4J.

Co-audit: tools to help humans double-check AI-generated content (Oct 2023; Link)

Abstract:
Users are increasingly being warned to check AI-generated content for correctness. Still, as LLMs (and other generative models) generate more complex output, such as summaries, tables, or code, it becomes harder for the user to audit or evaluate the output for quality or correctness. Hence, we are seeing the emergence of tool-assisted experiences to help the user double-check a piece of AI-generated content. We refer to these as co-audit tools. Co-audit tools complement prompt engineering techniques: one helps the user construct the input prompt, while the other helps them check the output response. As a specific example, this paper describes recent research on co-audit tools for spreadsheet computations powered by generative models. We explain why co-audit experiences are essential for any application of generative AI where quality is important and errors are consequential (as is common in spreadsheet computations). We propose a preliminary list of principles for co-audit, and outline research challenges.

Feature Decoupling Alignment for Fine-tuning Pre-trained Models in Few-shot Learning (NeurIPS, Oct 2023; Link)

Abstract:
Due to the limited availability of data, existing few-shot learning methods trained from scratch fail to achieve satisfactory performance. In contrast, large-scale pre-trained models such as CLIP demonstrate remarkable few-shot and zero-shot capabilities. To enhance the performance of pre-trained models for downstream tasks, fine-tuning model on downstream data is frequently necessary. However, fine-tuning the pre-trained model jeopardizes its generalizability in the presence of distribution shift, while the limited number of samples in few-shot learning makes the model highly susceptible to overfitting. Consequently, existing methods for fine-tuning few-shot learning primarily focus on fine-tuning the model's classification head or introducing additional structure. This paper introduces a feature decoupled alignment (FD-Align) fine-tuning approach, aiming to maximize the preservation of category-related information during fine-tuning while retaining category-independent information to maintain the model's generalizability. Extensive experiments demonstrate the superior effectiveness of our approach in enhancing model performance compared to direct fine-tuning. Furthermore, we showcase the effectiveness of our approach on the OOD dataset by achieving excellent OOD performance for the fine-tuned model.

Meet in the Middle: A New Pre-training Paradigm (NeurIPS, Oct 2023; Link)

Abstract:

Most language models (LMs) are trained and applied in an autoregressive left-to-right fashion, assuming that the next token only depends on the preceding ones. However, this assumption ignores the potential benefits of using the full sequence information during training, and the possibility of having context from both sides during inference. In this paper, we propose a new pre-training paradigm with techniques that jointly improve the training data efficiency and the capabilities of the LMs in the infilling task. The first is a training objective that aligns the predictions of a left-to-right LM with those of a right-to-left LM, trained on the same data but in reverse order. The second is a bidirectional inference procedure that enables both LMs to meet in the middle. We show the effectiveness of our pre-training paradigm with extensive experiments on both programming and natural language models, outperforming strong baselines.

LayoutPrompter: Awaken the Design Ability of Large Language Models (NeurIPS, Oct 2023; Link)

Abstract:
Conditional graphic layout generation, which automatically maps user constraints to high-quality layouts, has attracted much attention in recent years. Despite good performance, recent work still suffers some pivotal challenges. First, the neural models customized for this task require a large amount of layout data for model training, which is time-consuming and expensive. Second, the previous approaches usually do not have strong cross-domain generalization ability (e.g., from UI to Document). In this work, we propose LayoutPrompter to address the aforementioned issues by simply prompting GPT-3 text-davinci-003 model with a few demonstration examples. By meticulously designing the prompting strategy, our approach can generate high-quality, cross-domain graphic layouts without any model training or fine-tuning. Although remarkably simple, the experiments show that LayoutPrompter is competitive with state-of-the-art approaches on five traditional conditional layout generation tasks, and even outperforms them on two metrics (Alignment and Overlap). Furthermore, we also extend our approach to solve two challenging problems that have more flexible constraints, namely Text-to-layout and Content-aware layout generation. The qualitative and quantitative results of our approach are superior to those of existing methods, demonstrating the effectiveness and generalization ability of LayoutPrompter on more difficult tasks, even without model training. Our code and prompts will be released.

Model-enhanced Vector Index (NeurIPS, Oct 2023; Link)

Abstract:
Embedding-based retrieval methods construct vector indices to search for document representations that are most similar to the query representations. They are widely used in document retrieval due to low latency and decent recall performance. Recent research indicates that deep retrieval solutions offer better model quality, but are hindered by

unacceptable serving latency and the inability to support document updates. In this paper, we aim to enhance the vector index with end-to-end deep generative models, leveraging the differentiable advantages of deep retrieval models while maintaining desirable serving efficiency. We propose Model-enhanced Vector Index (MEVI), a differentiable model-enhanced index empowered by a twin-tower representation model. MEVI leverages a Residual Quantization (RQ) codebook to bridge the sequence-to-sequence deep retrieval and embedding-based models. To substantially reduce the inference time, instead of decoding the unique document ids in long sequential steps, we first generate some semantic virtual cluster ids of candidate documents in a small number of steps, and then leverage the well-adapted embedding vectors to further perform a fine-grained search for the relevant documents in the candidate virtual clusters. We empirically show that our model achieves better performance on the commonly used academic benchmarks MSMARCO Passage and Natural Questions, with comparable serving latency to dense retrieval solutions.

U.S. Deaf Community Perspectives on Automatic Sign Language Translation (ACM, Oct 2023; Link)

Abstract:
Millions of Deaf and hard-of-hearing (DHH) people primarily use a sign language for communication, but there is a lack of adequate sign language interpreting to fill these communication needs. Development of automatic sign language translation (ASLT) systems could help translate between a sign language and spoken language in situations where human interpreters are unavailable, and recent advances in large multi-lingual language models may soon enable ASLT to become a reality. Despite the potential for ASLT, Deaf community perspectives on and requirements for such technologies are poorly understood. In this work, we conduct a survey of Deaf community perspectives in the U.S. on ASLT in order to inform the development of ASLT systems that meet user needs and minimize harms. Our results shed light on scenarios where DHH users in the U.S. might want to use ASLT, their performance expectations for ASLT in these scenarios, design preferences for ASLT interfaces,

EmFore: Online Learning of Email Folder Classification Rules (CIKM, Oct 2023; Link)

Abstract:
Modern email clients support predicate-based folder assignment rules that can automatically organize emails. Unfortunately, users still need to write these rules manually. Prior machine learning approaches have framed automatically assigning email to folders as a classification task and do not produce symbolic rules. Prior inductive logic programming (ILP) approaches, which generate symbolic rules, fail to learn efficiently in the online environment needed for email management. To close this gap, we present EmFORE, an online system that learns symbolic rules for email classification from observations. Our key insights to do this

successfully are: (1) learning rules over a folder abstraction that supports quickly determining candidate predicates to add or replace terms in a rule, (2) ensuring that rules remain consistent with historical assignments, (3) ranking rule updates based on existing predicate and folder name similarity, and (4) building a rule suppression model to avoid surfacing low-confidence folder predictions while keeping the rule for future use. We evaluate on two popular public email corpora and compare to 13 baselines, including state-of-the-art folder assignment systems, incremental machine learning, ILP and transformer-based approaches. We find that EmFORE performs significantly better, updates four orders of magnitude faster, and is more robust than existing methods and baselines.

Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies (MSR, Sep 2023; Link)

Abstract:
Log data can reveal valuable information about how users interact with web search services, what they want, and how satisfied they are. However, analyzing user intents in log data is not easy, especially for new forms of web search such as AI-driven chat. To understand user intents from log data, we need a way to label them with meaningful categories that capture their diversity and dynamics. Existing methods rely on manual or ML-based labeling, which are either expensive or inflexible for large and changing datasets. We propose a novel solution using large language models (LLMs), which can generate rich and relevant concepts, descriptions, and examples for user intents. However, using LLMs to generate a user intent taxonomy and apply it to do log analysis can be problematic for two main reasons: such a taxonomy is not externally validated, and there may be an undesirable feedback loop. To overcome these issues, we propose a new methodology with human experts and assessors to verify the quality of the LLM-generated taxonomy. We also present an end-to-end pipeline that uses an LLM with human-in-the-loop to produce, refine, and use labels for user intent analysis in log data. Our method offers a scalable and adaptable way to analyze user intents in web-scale log data with minimal human effort. We demonstrate its effectiveness by uncovering new insights into user intents from search and chat logs from Bing.

SCREWS: A Modular Framework for Reasoning with Revisions (ArXiV, Sep 2023; Link)

Abstract:
Large language models (LLMs) can improve their accuracy on various tasks through iteratively refining and revising their output based on feedback. We observe that these revisions can introduce errors, in which case it is better to roll back to a previous result. Further, revisions are typically homogeneous: they use the same reasoning method that produced the initial answer, which may not correct errors. To enable exploration in this space, we present SCREWS, a modular framework for reasoning with revisions. It is

comprised of three main modules: Sampling, Conditional Resampling, and Selection, each consisting of sub-modules that can be hand-selected per task. We show that SCREWS not only unifies several previous approaches under a common framework, but also reveals several novel strategies for identifying improved reasoning chains. We evaluate our framework with state-of-the-art LLMs (ChatGPT and GPT-4) on a diverse set of reasoning tasks and uncover useful new reasoning strategies for each: arithmetic word problems, multi-hop question answering, and code debugging. Heterogeneous revision strategies prove to be important, as does selection between original and revised candidates.

Textbooks Are All You Need II: phi-1.5 technical report (Sep 2023; Link)

Abstract:
We continue the investigation into the power of smaller Transformer-based language models as initiated by \textbf{TinyStories} — a 10 million parameter model that can produce coherent English — and the follow-up work on \textbf{phi-1}, a 1.3 billion parameter model with Python coding performance close to the state-of-the-art. The latter work proposed to use existing Large Language Models (LLMs) to generate “textbook quality” data as a way to enhance the learning process compared to traditional web data. We follow the “Textbooks Are All You Need” approach, focusing this time on common sense reasoning in natural language, and create a new 1.3 billion parameter model named \textbf{phi-1.5}, with performance on natural language tasks comparable to models 5x larger, and surpassing most non-frontier LLMs on more complex reasoning tasks such as grade-school mathematics and basic coding. More generally, \textbf{phi-1.5} exhibits many of the traits of much larger LLMs, both good — such as the ability to “think step by step” or perform some rudimentary in-context learning — and bad, including hallucinations and the potential for toxic and biased generations — encouragingly though, we are seeing improvement on that front thanks to the absence of web data. We open-source \textbf{phi-1.5} to promote further research on these urgent topics.

Generative AI for Programming Education: Benchmarking ChatGPT, GPT-4, and Human Tutors (ICER, Sep 2023; Link)

Abstract:
Generative AI and large language models hold great promise in enhancing computing education by powering next-generation educational technologies. State-of-the-art models like OpenAI’s ChatGPT [8] and GPT-4 [9] could enhance programming education in various roles, e.g., by acting as a personalized digital tutor for a student, a digital assistant for an educator, and a digital peer for collaborative learning [1, 2, 7]. In our work, we seek to comprehensively evaluate and benchmark state-of-the-art large language models for various scenarios in programming education.

Recent works have evaluated several large language models in the context of programming education [4, 6, 10, 11, 12]. However, these works are limited for several reasons: they have typically focused on evaluating a specific model for a specific education scenario (e.g., generating explanations), or have considered models that are already outdated (e.g., OpenAI's Codex [3] is no longer publicly available since March 2023). Consequently, there is a lack of systematic study that benchmarks state-of-the-art models for a comprehensive set of programming education scenarios.

In our work, we systematically evaluate two models, ChatGPT (based on GPT-3.5) and GPT-4, and compare their performance with human tutors for a variety of scenarios in programming education. These scenarios are designed to capture distinct roles these models could play, namely digital tutors, assistants, and peers, as discussed above. More concretely, we consider the following six scenarios: (1) program repair, i.e., fixing a student's buggy program; (2) hint generation, i.e., providing a natural language hint to the student to help resolve current issues; (3) grading feedback, i.e., grading a student's program w.r.t. a given rubric; (4) peer programming, i.e., completing a partially written program or generating a sketch for the solution program; (5) task creation, i.e., generating new tasks that exercise specific types of concepts or bugs; (6) contextualized explanation, i.e., explaining specific concepts or functions in the context of a given program.

Our study uses a mix of quantitative and qualitative evaluation to compare the performance of these models with the performance of human tutors. We conduct our evaluation based on 5 introductory Python programming problems with a diverse set of input/output specifications. For each of these problems, we consider 5 buggy programs based on publicly accessible submissions from geeksforgeeks.org [5] (see Figure 1); these buggy programs are picked to capture different types of bugs for each problem. We will provide a detailed analysis of the data and results in a longer version of this poster. Our preliminary results show that GPT-4 drastically outperforms ChatGPT (based on GPT-3.5) and comes close to human tutors' performance for several scenarios.

CodePlan: Repository-level Coding using LLMs and Planning

(NeurIPS, Sep 2023; Link)

Software engineering activities such as package migration, fixing errors reports from static analysis or testing, and adding type annotations or other specifications to a codebase, involve pervasively editing the entire repository of code. We formulate these activities as repository-level coding tasks.

Recent tools like GitHub Copilot, which are powered by Large Language Models (LLMs), have succeeded in offering high-quality solutions to localized coding problems. Repository-level coding tasks are more involved and cannot be solved directly using LLMs, since code within a repository is inter-dependent and the entire repository may be too large to fit into the prompt. We frame repository-level coding as a planning problem and present a task-agnostic framework, called CodePlan to solve it. CodePlan synthesizes a multi-step chain of edits (plan), where each step results in a call to an LLM on a code location with

context derived from the entire repository, previous code changes and task-specific instructions. CodePlan is based on a novel combination of an incremental dependency analysis, a change may-impact analysis and an adaptive planning algorithm.

We evaluate the effectiveness of CodePlan on two repository-level tasks: package migration (C#) and temporal code edits (Python). Each task is evaluated on multiple code repositories, each of which requires inter-dependent changes to many files (between 2-97 files). Coding tasks of this level of complexity have not been automated using LLMs before. Our results show that CodePlan has better match with the ground truth compared to baselines. CodePlan is able to get 5/6 repositories to pass the validity checks (e.g., to build without errors and make correct code edits) whereas the baselines (without planning but with the same type of contextual information as CodePlan) cannot get any of the repositories to pass them.

Grace: Language Models Meet Code Edits (ESEC, Sep 2023; Link)

Abstract:
Developers spend a significant amount of time in editing code for a variety of reasons such as bug fixing or adding new features. It has been an active yet challenging area of research due to the diversity of code edits and the difficulty of capturing the developer intent. In this work, we address these challenges by endowing pre-trained large language models (LLMs) with the knowledge of relevant prior associated edits, which we call the Grace (Generation conditioned on Associated Code Edits) method. The generative capability of the LLMs helps address the diversity in code changes and conditioning code generation on prior edits helps capture the latent developer intent. We evaluate two well-known LLMs, codex and CodeT5, in zero-shot and fine-tuning settings respectively. In our experiments with two datasets, Grace boosts the performance of the LLMs significantly, enabling them to generate 29% and 54% more correctly-edited code in top-1 suggestions relative to the current state-of-the-art symbolic and neural approaches, respectively.

We share the scripts, prompts, and instructions to access the finetuned models at https://aka.ms/GrACE-Code.

Frustrated with Code Quality Issues? LLMs can Help! (Sep 2023; Link)

Abstract:
As software projects progress, quality of code assumes paramount importance as it affects reliability, maintainability and security of software. For this reason, static analysis tools are used in developer workflows to flag code quality issues. However, developers need to spend extra efforts to revise their code to improve code quality based on the tool findings. In this work, we investigate the use of (instruction-following) large language models (LLMs) to

assist developers in revising code to resolve code quality issues. We present a tool, CORE (short for COde REvisions), architected using a pair of LLMs organized as a duo comprised of a proposer and a ranker. Providers of static analysis tools recommend ways to mitigate the tool warnings and developers follow them to revise their code. The proposer LLM of CORE takes the same set of recommendations and applies them to generate candidate code revisions. The candidates which pass the static quality checks are retained. However, the LLM may introduce subtle, unintended functionality changes which may go un-detected by the static analysis. The ranker LLM evaluates the changes made by the proposer using a rubric that closely follows the acceptance criteria that a developer would enforce. CORE uses the scores assigned by the ranker LLM to rank the candidate revisions before presenting them to the developer. We conduct a variety of experiments on two public benchmarks to show the ability of CORE: 1 to generate code revisions acceptable to both static analysis tools and human reviewers (the latter evaluated with user study on a subset of the Python benchmark), 2 to reduce human review efforts by detecting and eliminating revisions with unintended changes, 3 to readily work across multiple languages (Python and Java), static analysis tools (CodeQL and SonarQube) and quality checks (52 and 10 checks, respectively), and 4 to achieve fix rate comparable to a rule-based automated program repair tool but with much smaller engineering efforts (on the Java benchmark). CORE could revise 59.2% Python files (across 52 quality checks) so that they pass scrutiny by both a tool and a human reviewer. The ranker LLM reduced false positives by 25.8% in these cases. CORE produced revisions that passed the static analysis tool in 76.8% Java files (across 10 quality checks) comparable to 78.3% of a specialized program repair tool, with significantly much less engineering efforts.

Large Language Models Can Accurately Predict Searcher Preferences (Sep 2023; Link)

Abstract:
Relevance labels, which indicate whether a search result is valuable to a searcher, are key to evaluating and optimising search systems. The best way to capture the true preferences of users is to ask them for their careful feedback on which results would be useful, but this approach does not scale to produce a large number of labels. Getting relevance labels at scale is usually done with third-party labellers, who judge on behalf of the user, but there is a risk of low-quality data if the labeller doesn't understand user needs. To improve quality, one standard approach is to study real users through interviews, user studies and direct feedback, find areas where labels are systematically disagreeing with users, then educate labellers about user needs through judging guidelines, training and monitoring. This paper introduces an alternate approach for improving label quality. It takes careful feedback from real users, which by definition is the highest-quality first-party gold data that can be derived, and develops an large language model prompt that agrees with that data.

We present ideas and observations from deploying language models for large-scale relevance labelling at Bing, and illustrate with data from TREC. We have found large language models can be effective, with accuracy as good as human labellers and similar capability to pick the hardest queries, best runs, and best groups. Systematic changes to the

prompts make a difference in accuracy, but so too do simple paraphrases. To measure agreement with real searchers needs high-quality “gold” labels, but with these we find that models produce better labels than third-party workers, for a fraction of the cost, and these labels let us train notably better rankers.

Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation (Sep 2023; Link)

Abstract:
We study the problem of in-context learning (ICL) with large language models (LLMs) on private datasets. This scenario poses privacy risks, as LLMs may leak or regurgitate the private examples demonstrated in the prompt. We propose a novel algorithm that generates synthetic few-shot demonstrations from the private dataset with formal differential privacy (DP) guarantees, and show empirically that it can achieve effective ICL. We conduct extensive experiments on standard benchmarks and compare our algorithm with non-private ICL and zero-shot solutions. Our results demonstrate that our algorithm can achieve competitive performance with strong privacy levels. These results open up new possibilities for ICL with privacy protection for a broad range of applications.

Robust Situational Reinforcement Learning in Face of Context Disturbances (ICML, Sep 2023; Link)

Abstract:
In many real-world tasks, some parts of state features, called contexts, are independent of action signals, e.g., customer demand in inventory control, speed of lead car in autonomous driving, etc. One of the challenges of reinforcement learning in these applications is that the true context transitions can be easily exposed some unknown source of contamination, leading to a shift of context transitions between source domains and target domains, which could cause performance degradation for RL algorithms. However, existing methods on robust RL aim at learning robust policies against the deviations of the entire system dynamics. To tackle this problem, this paper proposes the framework of robust situational Markov decision process (RS-MDP) which captures the possible deviations of context transitions explicitly. To scale to large context space, we introduce the softmin smoothed robust Bellman operator to learn the robust Q-value approximately, and apply our RS-MDP framework to existing RL algorithm SAC to learn the desired robust policies. We conduct experiments on several robot control tasks with dynamic contexts and inventory control tasks to demonstrate that our algorithm can generalize better and more robust against deviations of context transitions, and outperform existing robust RL algorithms.

Multimodal Foundation Models: From Specialists to General-Purpose Assistants (CVPR, Sep 2023; Link)

Abstract:
This paper presents a comprehensive survey of the taxonomy and evolution of multimodal foundation models that demonstrate vision and vision-language capabilities, focusing on the transition from specialist models to general-purpose assistants. The research landscape encompasses five core topics, categorized into two classes. (i) We start with a survey of well-established research areas: multimodal foundation models pre-trained for specific purposes, including two topics — methods of learning vision backbones for visual understanding and text-to-image generation. (ii) Then, we present recent advances in exploratory, open research areas: multimodal foundation models that aim to play the role of general-purpose assistants, including three topics — unified vision models inspired by large language models (LLMs), end-to-end training of multimodal LLMs, and chaining multimodal tools with LLMs. The target audiences of the paper are researchers, graduate students, and professionals in computer vision and vision-language multimodal communities who are eager to learn the basics and recent advances in multimodal foundation models.

PACE-LM: Prompting and Augmentation for Calibrated Confidence Estimation with GPT-4 in Cloud Incident Root Cause Analysis (Sep 2023; Link)

Abstract:
Major cloud providers have employed advanced AI-based solutions like large language models to aid humans in identifying the root causes of cloud incidents. Despite the growing prevalence of AI-driven assistants in the root cause analysis process, their effectiveness in assisting on-call engineers is constrained by low accuracy due to the intrinsic difficulty of the task, a propensity for LLM-based approaches to hallucinate, and difficulties in distinguishing these well-disguised hallucinations. To address this challenge, we propose to perform confidence estimation for the predictions to help on-call engineers make decisions on whether to adopt the model prediction. Considering the black-box nature of many LLM-based root cause predictors, fine-tuning or temperature-scaling-based approaches are inapplicable. We therefore design an innovative confidence estimation framework based on prompting retrieval-augmented large language models (LLMs) that demand a minimal amount of information from the root cause predictor. This approach consists of two scoring phases: the LLM-based confidence estimator first evaluates its confidence in making judgments in the face of the current incident that reflects its “grounded-ness” level in reference data, then rates the root cause prediction based on historical references. An optimization step combines these two scores for a final confidence assignment. We show that our method is able to produce calibrated confidence estimates for predicted root causes, validate the usefulness of retrieved historical data and the prompting strategy as well as the generalizability across different root cause prediction models. Our study takes an important move towards reliably and effectively embedding LLMs into cloud incident management systems.

Community Economics for AI Powered Micro-Grid (Sep 2020; Link)

Abstract:
The focus of this research is 2-fold a) Investigate the potential of AI model optimization for increasing energy efficiency of micro-grids and thereby reducing economic burdens on low-income households and marginalized communities; b) Build better economic models focused on community resiliency and equity for the AI to effectively optimize towards those objectives.

To this effect, we are interested in investigating the potential of AI model optimization for increasing energy efficiency of micro-grids and thereby reducing economic burdens on low-income households and marginalized communities. We seek to build better community economic models focused on community resiliency and equity for the AI, and to effectively optimize towards those objectives.

PwR: Exploring the Role of Representations in Conversational Programming (Sep 2023; Link)

Abstract:
Large Language Models (LLMs) have revolutionized programming and software engineering. AI programming assistants such as GitHub Copilot X enable conversational programming, narrowing the gap between human intent and code generation. However, prior literature has identified a key challenge–there is a gap between user's mental model of the system's understanding after a sequence of natural language utterances, and the AI system's actual understanding. To address this, we introduce Programming with Representations (PwR), an approach that uses representations to convey the system's understanding back to the user in natural language. We conducted an in-lab task-centered study with 14 users of varying programming proficiency and found that representations significantly improve understandability, and instilled a sense of agency among our participants. Expert programmers use them for verification, while intermediate programmers benefit from confirmation. Natural language-based development with LLMs, coupled with representations, promises to transform software development, making it more accessible and efficient.

ChatGPT Empowered Long-Step Robot Control in Various Environments: A Case Application (IEEE, Aug 2023; Link)

Abstract:
This paper demonstrates how OpenAI's ChatGPT can be used in a few-shot setting to convert natural language instructions into a sequence of executable robot actions. The paper proposes easy-to-customize input prompts for ChatGPT that meet common requirements in practical applications, such as easy integration with robot execution systems and

applicability to various environments while minimizing the impact of ChatGPT's token limit. The prompts encourage ChatGPT to output a sequence of predefined robot actions, represent the operating environment in a formalized style, and infer the updated state of the operating environment. Experiments confirmed that the proposed prompts enable ChatGPT to act according to requirements in various environments, and users can adjust ChatGPT's output with natural language feedback for safe and robust operation. The proposed prompts and source code are open-source and publicly available at this https URL.

AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation (MSR, Aug 2023; Link)

Abstract:
We present AutoGen, an open-source framework that allows developers to build LLM applications via multiple agents that can converse with each other to accomplish tasks. AutoGen agents are customizable, conversable, and can operate in various modes that employ combinations of LLMs, human inputs, and tools.
Using AutoGen, developers can also flexibly define agent interaction behaviors. Both natural language and computer code can be used to program flexible conversation patterns for different applications. AutoGen serves as a generic infrastructure to build diverse applications of various complexities and LLM capacities. We provide many examples to build effective applications for domains ranging from mathematics, coding, question answering, operations research, online decision-making, entertainment, etc.

Constraint-aware and Ranking-distilled Token Pruning for Efficient Transformer Inference (KDD, Aug 2023; Link)

Abstract:
Deploying pre-trained transformer models like BERT on downstream tasks in resource-constrained scenarios is challenging due to their high inference cost, which grows rapidly with input sequence length. In this work, we propose a constraint-aware and ranking-distilled token pruning method ToP, which selectively removes unnecessary tokens as input sequence passes through layers, allowing the model to improve online inference speed while preserving accuracy. ToP overcomes the limitation of inaccurate token importance ranking in the conventional self-attention mechanism through a ranking-distilled token distillation technique, which distills effective token rankings from the final layer of unpruned models to early layers of pruned models. Then, ToP introduces a coarse-to-fine pruning approach that automatically selects the optimal subset of transformer layers and optimizes token pruning decisions within these layers through improved L0 regularization. Extensive experiments on GLUE benchmark and SQuAD tasks demonstrate that ToP outperforms state-of-the-art token pruning and model compression methods with improved accuracy and speedups. ToP reduces the average FLOPs of BERT by 8.1X while achieving

competitive accuracy on GLUE, and provides a real latency speedup of up to 7.4X on an Intel CPU. Code is available at https://github.com/microsoft/Moonlit/tree/main/ToP.

A Causality Inspired Framework for Model Interpretation (ACM, Aug 2023; Link)

Abstract:
A critical issue in eXplainable Artificial Intelligence (XAI) is determining whether explanations uncover the underlying causal factors for model behavior or merely show coincidental relationships. Failing to make this distinction can lead to incorrect understandings. To address this issue, we first understand the model interpretation through a causal lens. We find that the explanation scores of certain representative explanation methods align with the concept of average treatment effect in causal inference and evaluate their relative strengths and limitations from a unified causal perspective. Based on our observations, we outline the major challenges in applying causal inference to model interpretation, including identifying common causes that can be generalized across instances and ensuring that explanations provide a complete causal explanation of model predictions. We then present CIMI, a Causality-Inspired Model Interpreter, which addresses these challenges. CIMI has three modules: the causal sufficiency module and the causal intervention module ensure the explanations are both causally sufficient and generalizable, while the causal prior module facilitates easy learning. Our experiments show that CIMI provides superior and generalizable explanations and is useful for debugging and improving models.

Demonstration of CORNET: Learning Spreadsheet Formatting Rules by Example (VLDC, Aug 2023; Link)

Abstract:
Data management and analysis tasks are often carried out using spreadsheet software. A popular feature in most spreadsheet platforms is the ability to define data-dependent formatting rules. These rules can express actions such as “color red all entries in a column that are negative” or “bold all rows not containing error or failure”. Unfortunately, users who want to exercise this functionality need to manually write these conditional formatting (CF) rules. We introduce Cornet, a system that automatically learns such conditional formatting rules from user examples. Cornet takes inspiration from inductive program synthesis and combines symbolic rule enumeration, based on semi-supervised clustering and iterative decision tree learning, with a neural ranker to produce accurate conditional formatting rules. In this demonstration, we show Cornet in action as a simple add-in to Microsoft’s Excel. After the user provides one or two formatted cells as examples, Cornet generates formatting rule suggestions for the user to apply to the spreadsheet.

LightGlue: Local Feature Matching at Light Speed (ICCV, Aug 2023; Link)

Abstract:
We introduce LightGlue, a deep neural network that learns to match local features across images. We revisit multiple design decisions of SuperGlue, the state of the art in sparse matching, and derive simple but effective improvements. Cumulatively, they make LightGlue more efficient – in terms of both memory and computation, more accurate, and much easier to train. One key property is that LightGlue is adaptive to the difficulty of the problem: the inference is much faster on image pairs that are intuitively easy to match, for example because of a larger visual overlap or limited appearance change. This opens up exciting prospects for deploying deep matchers in latency-sensitive applications like 3D reconstruction. The code and trained models are publicly available at github.com/cvg/LightGlue.

Participatory prompting: a user-centric research method for eliciting AI assistance opportunities in knowledge workflows (PPIG, Aug 2023; Link)

Abstract:
Generative AI, such as image generation models and large language models, stands to provide tremendous value to end-user programmers in creative and knowledge workflows. Current research methods struggle to engage end-users in a realistic conversation that balances the actually existing capabilities of generative AI with the open-ended nature of user workflows and the many opportunities for the application of this technology. In this work-in-progress paper, we introduce participatory prompting, a method for eliciting opportunities for generative AI in end-user workflows. The participatory prompting method combines a contextual inquiry and a researcher-mediated interaction with a generative model, which helps study participants interact with a generative model without having to develop prompting strategies of their own. We discuss the ongoing development of a study whose aim will be to identify end-user programming opportunities for generative AI in data analysis workflows.

Retentive Network: A Successor to Transformer for Large Language Models (ArXiV, Jul 2023; Link)

Abstract:
In this work, we propose Retentive Network (RetNet) as a foundation architecture for large language models, simultaneously achieving training parallelism, low-cost inference, and good performance. We theoretically derive the connection between recurrence and attention. Then we propose the retention mechanism for sequence modeling, which supports three

computation paradigms, i.e., parallel, recurrent, and chunkwise recurrent. Specifically, the parallel representation allows for training parallelism. The recurrent representation enables low-cost O(1) inference, which improves decoding throughput, latency, and GPU memory without sacrificing performance. The chunkwise recurrent representation facilitates efficient long-sequence modeling with linear complexity, where each chunk is encoded parallelly while recurrently summarizing the chunks. Experimental results on language modeling show that RetNet achieves favorable scaling results, parallel training, low-cost deployment, and efficient inference. The intriguing properties make RetNet a strong successor to Transformer for large language models. Code will be available at this https URL.

Distilling Large Language Models for Biomedical Knowledge Extraction: A Case Study on Adverse Drug Events (Jul 2023; Link)

Abstract:
Large language models (LLMs), such as GPT-4, have demonstrated remarkable capabilities across a wide range of tasks, including health applications. In this paper, we study how LLMs can be used to scale biomedical knowledge curation. We find that while LLMs already possess decent competency in structuring biomedical text, by distillation into a task-specific student model through self-supervised learning, substantial gains can be attained over out-of-box LLMs, with additional advantages such as cost, efficiency, and white-box model access.

We conduct a case study on adverse drug event (ADE) extraction, which is an important area for improving care. On standard ADE extraction evaluation, a GPT-3.5 distilled PubMedBERT model attained comparable accuracy as supervised state-of-the-art models without using any labeled data. Despite being over 1,000 times smaller, the distilled model outperformed its teacher GPT-3.5 by over 6 absolute points in F1 and GPT-4 by over 5 absolute points.

Ablation studies on distillation model choice (e.g., PubMedBERT vs BioGPT) and ADE extraction architecture shed light on best practice for biomedical knowledge extraction. Similar gains were attained by distillation for other standard biomedical knowledge extraction tasks such as gene-disease associations and protected health information, further illustrating the promise of this approach.

Why Did the Chicken Cross the Road? Rephrasing and Analyzing Ambiguous Questions in VQA (ACL, Jul 2023; Link)

Abstract:
Natural language is ambiguous. Resolving ambiguous questions is key to successfully answering them. Focusing on questions about images, we create a dataset of ambiguous examples. We annotate these, grouping answers by the underlying question they address

and rephrasing the question for each group to reduce ambiguity. Our analysis reveals a linguistically-aligned ontology of reasons for ambiguity in visual questions. We then develop an English question-generation model which we demonstrate via automatic and human evaluation produces less ambiguous questions. We further show that the question generation objective we use allows the model to integrate answer group information without any direct supervision.

TACR: A Table Alignment-based Cell Selection Method for HybridQA (ACL, Jul 2023; Link)

Abstract:
Hybrid Question-Answering (HQA), which targets reasoning over tables and passages linked from table cells, has witnessed significant research in recent years. A common challenge in HQA and other passage-table QA datasets is that it is generally unrealistic to iterate over all table rows, columns, and linked passages to retrieve evidence. Such a challenge made it difficult for previous studies to show their reasoning ability in retrieving answers. To bridge this gap, we propose a novel Table-alignment-based Cell-selection and Reasoning model (TACR) for hybrid text and table QA, evaluated on the HybridQA and WikiTableQuestions datasets. In evidence retrieval, we design a table-question-alignment enhanced cell-selection method to retrieve fine-grained evidence. In answer reasoning, we incorporate a QA module that treats the row containing selected cells as context. Experimental results over the HybridQA and WikiTableQuestions (WTQ) datasets show that TACR achieves state-of-the-art results on cell selection and outperforms fine-grained evidence retrieval baselines on HybridQA, while achieving competitive performance on WTQ. We also conducted a detailed analysis to demonstrate that being able to align questions to tables in the cell-selection stage can result in important gains from experiments of over 90\% table row and column selection accuracy, meanwhile also improving output explainability.

Limitations of Language Models in Arithmetic and Symbolic Induction (ACL Jul 2023; Link)

Abstract:
Recent work has shown that large pretrained Language Models (LMs) can not only perform remarkably well on a range of Natural Language Processing (NLP) tasks but also start improving on reasoning tasks such as arithmetic induction, symbolic manipulation, and commonsense reasoning with increasing size of models. However, it is still unclear what the underlying capabilities of these LMs are. Surprisingly, we find that these models have limitations on certain basic symbolic manipulation tasks such as copy, reverse, and addition. When the total number of symbols or repeating symbols increases, the model performance drops quickly. We investigate the potential causes behind this phenomenon and examine a set of possible methods, including explicit positional markers, fine-grained computation steps, and LMs with callable programs. Experimental results show that none of these

techniques can solve the simplest addition induction problem completely. In the end, we introduce LMs with tutor, which demonstrates every single step of teaching. LMs with tutor is able to deliver 100% accuracy in situations of OOD and repeating symbols, shedding new insights on the boundary of large LMs in induction.

Pre-trained Language Models Can be Fully Zero-Shot Learners (ACL, Jul 2023; Link)

Abstract:
How can we extend a pre-trained model to many language understanding tasks, without labeled or additional unlabeled data? Pre-trained language models (PLMs) have been effective for a wide range of NLP tasks. However, existing approaches either require fine-tuning on downstream labeled datasets or manually constructing proper prompts. In this paper, we propose nonparametric prompting PLM (NPPrompt) for fully zero-shot language understanding. Unlike previous methods, NPPrompt uses only pre-trained language models and does not require any labeled data or additional raw corpus for further fine-tuning, nor does it rely on humans to construct a comprehensive set of prompt label words. We evaluate NPPrompt against previous major few-shot and zero-shot learning methods on diverse NLP tasks: including text classification, text entailment, similar text retrieval, and paraphrasing. Experimental results demonstrate that our NPPrompt outperforms the previous best fully zero-shot method by big margins, with absolute gains of 12.8% in accuracy on text classification and 18.9% on the GLUE benchmark.

Smart Word Suggestions for Writing Assistance (ACL, Jul 2023; Link)

Abstract:
Enhancing word usage is a desired feature for writing assistance. To further advance research in this area, this paper introduces “Smart Word Suggestions” (SWS) task and benchmark. Unlike other works, SWS emphasizes end-to-end evaluation and presents a more realistic writing assistance scenario. This task involves identifying words or phrases that require improvement and providing substitution suggestions. The benchmark includes human-labeled data for testing, a large distantly supervised dataset for training, and the framework for evaluation. The test data includes 1,000 sentences written by English learners, accompanied by over 16,000 substitution suggestions annotated by 10 native speakers. The training dataset comprises over 3.7 million sentences and 12.7 million suggestions generated through rules. Our experiments with seven baselines demonstrate that SWS is a challenging task. Based on experimental analysis, we suggest potential directions for future research on SWS. The dataset and related codes is available at this https URL.

Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions (ACL, Jul 2023; Link)

Abstract:
Large language models (LLMs) can be used to generate text data for training and evaluating other models. However, creating high-quality datasets with LLMs can be challenging. In this work, we explore human-AI partnerships to facilitate high diversity and accuracy in LLM-based text data generation. We first examine two approaches to diversify text generation: 1) logit suppression, which minimizes the generation of languages that have already been frequently generated, and 2) temperature sampling, which flattens the token sampling probability. We found that diversification approaches can increase data diversity but often at the cost of data accuracy (i.e., text and labels being appropriate for the target domain). To address this issue, we examined two human interventions, 1) label replacement (LR), correcting misaligned labels, and 2) out-of-scope filtering (OOSF), removing instances that are out of the user's domain of interest or to which no considered label applies. With oracle studies, we found that LR increases the absolute accuracy of models trained with diversified datasets by 14.4%. Moreover, we found that some models trained with data generated with LR interventions outperformed LLM-based few-shot classification. In contrast, OOSF was not effective in increasing model accuracy, implying the need for future work in human-in-the-loop text data generation.

TART: Improved Few-shot Text Classification Using Task-Adaptive Reference Transformation (ACl, Jul 2023; Link)

Abstract:
Meta-learning has emerged as a trending technique to tackle few-shot text classification and achieve state-of-the-art performance. However, the performance of existing approaches heavily depends on the inter-class variance of the support set. As a result, it can perform well on tasks when the semantics of sampled classes are distinct while failing to differentiate classes with similar semantics. In this paper, we propose a novel Task-Adaptive Reference Transformation (TART) network, aiming to enhance the generalization by transforming the class prototypes to per-class fixed reference points in task-adaptive metric spaces. To further maximize divergence between transformed prototypes in task-adaptive metric spaces, TART introduces a discriminative reference regularization among transformed prototypes. Extensive experiments are conducted on four benchmark datasets and our method demonstrates clear superiority over the state-of-the-art models in all the datasets. In particular, our model surpasses the state-of-the-art method by 7.4% and 5.4% in 1-shot and 5-shot classification on the 20 Newsgroups dataset, respectively.

MEMEX: Detecting Explanatory Evidence for Memes via Knowledge-Enriched Contextualization (ACL, Jul 2023; Link)

Abstract:
Memes are a powerful tool for communication over social media. Their affinity for evolving across politics, history, and sociocultural phenomena makes them an ideal communication vehicle. To comprehend the subtle message conveyed within a meme, one must understand the background that facilitates its holistic assimilation. Besides digital archiving of memes and their metadata by a few websites like knowyourmeme.com(opens in new tab), currently, there is no efficient way to deduce a meme's context dynamically. In this work, we propose a novel task, MEMEX – given a meme and a related document, the aim is to mine the context that succinctly explains the background of the meme. At first, we develop MCC (Meme Context Corpus), a novel dataset for MEMEX. Further, to benchmark MCC, we propose MIME (MultImodal Meme Explainer), a multimodal neural framework that uses common sense enriched meme representation and a layered approach to capture the cross-modal semantic dependencies between the meme and the context. MIME surpasses several unimodal and multimodal systems and yields an absolute improvement of ~ 4% F1-score over the best baseline. Lastly, we conduct detailed analyses of MIME's performance, highlighting the aspects that could lead to optimal modeling of cross-modal contextual associations.

On Improving Summarization Factual Consistency from Natural Language Feedback (ACL, Jul 2023; Link)

Abstract:
Despite the recent progress in language generation models, their outputs may not always meet user expectations. In this work, we study whether informational feedback in natural language can be leveraged to improve generation quality and user preference alignment. To this end, we consider factual consistency in summarization, the quality that the summary should only contain information supported by the input documents, for user preference alignment. We collect a high-quality dataset, DeFacto, containing human demonstrations and informational feedback in natural language consisting of corrective instructions, edited summaries, and explanations with respect to the factual consistency of the summary. Using our dataset, we study two natural language generation tasks: 1) editing a summary using the human feedback, and 2) generating human feedback from the original summary. Using the two tasks, we further evaluate if models can automatically correct factual inconsistencies in generated summaries. We show that the human-edited summaries we collected are more factually consistent, and pre-trained language models can leverage our dataset to improve the factual consistency of original system-generated summaries in our proposed generation tasks. We make the DeFacto dataset publicly available at GitHub.

Structure-Aware Language Model Pretraining Improves Dense Retrieval on Structured Data (ACL, Jul 2023; Link)

Abstract:
This paper presents Structure Aware Dense Retrieval (SANTA) model, which encodes user queries and structured data in one universal embedding space for retrieving structured data. SANTA proposes two pretraining methods to make language models structure-aware and learn effective representations for structured data: 1) Structured Data Alignment, which utilizes the natural alignment relations between structured data and unstructured data for structure-aware pretraining. It contrastively trains language models to represent multi-modal text data and teaches models to distinguish matched structured data for unstructured texts. 2) Masked Entity Prediction, which designs an entity-oriented mask strategy and asks language models to fill in the masked entities. Our experiments show that SANTA achieves state-of-the-art on code search and product search and conducts convincing results in the zero-shot setting. SANTA learns tailored representations for multi-modal text data by aligning structured and unstructured data pairs and capturing structural semantics by masking and predicting entities in the structured data. All codes are available at this https URL.