arXivBot

Stanford CS224N Custom Project

Amr Sherif
Department of Computer Science
Stanford University
amrshrif@stanford.edu

Abstract

The project goal is to build an arXiv chatbot capable of retrieving and selecting relevant research papers new or old according to a criteria specified by the user in natural language (ex. submission/revision date, domain/category, topic...etc). As well as answering users' questions about the papers and being able to summarize & highlight key points of the paper topics. At the current accelerating speed of research and advances in various fields; every month, 20,000 new papers are posted to arxiv.org, including 9,000 papers in the Computer Science category, keeping up with the latest progress and advances has become tedious. This chatbot will help researchers and practitioners to keep up with the latest advances and up-to-date progress in different fields, by being able to search, retrieve, and summarize the new released papers in a streamlined approach. This study deploys multiple techniques to implement a lightweight yet efficient LLM system capable of achieving state-of-the-art results, including QLoRA quantization [1] and Parameter-Efficient Fine-Tuning (PEFT) instruction fine-tuning with LoRA, as well as incorporating Function Calling, In-Context Learning & a Modular Retrieval Augmented Generation (RAG) process in the system [2].

1 Key Information to include

- TA mentor: Rashon Poole

2 Introduction

The rapid growth of research publications presents a significant challenge for researchers and practitioners to stay up-to-date with the latest advancements, which created a need for automated tools to manage the information more efficiently. Retrieving, Reading Comprehension and summarizing relevant research papers is a challenging objective. Previous work on retrieval systems often struggled with multi-hop queries, which require reasoning over multiple documents, and fail to provide high quality results and contextual understanding. This study addresses these challenges by developing an advanced Semantic RAG system capable of retrieving, summarizing, and answering user queries of research papers retrieved based on his specified criteria.

The system employs multiple tuning and RAG techniques to handle complex queries effectively, by incorporating Function Calling, In-Context Learning, and a modular RAG process, which enhanced the assistant capability to provide contextually relevant information. Employing QLoRA quantization and Parameter-Efficient Fine-Tuning (PEFT) with LoRA adapters improves the performance and reduce memory usage while maintaining high quality results. The model is served using vLLM, which achieves a good inference generation speed of an average of 66 tokens per second on an A100 GPU.

The conducted experiments and results demonstrate the system effectiveness on the task. The model achieves notable results on multiple evaluation metrics like BLEU, ROUGE and Hit Rate.

To monitor and log our fine-tuning process, we utilized Weights & Biases [3], ensuring effective tracking and analysis of training performance.

In addition to, creating two large-scale synthetic domain-specific dataset for instruction tuning, which contains 168,000 examples. This dataset was generated through a fully automated process that involved parsing, splitting, and chunking arXiv full-text PDFs, followed by generation of questions & answers.

3 Related work

Retrieval-Augmented Generation RAG systems have been developed to augment large language models (LLMs) by retrieving relevant information from a knowledge base, thus enhancing the quality and accuracy of responses. This approach has demonstrated significant potential in mitigating hallucinations often encountered with LLMs and improving overall response quality. Several studies have focused on various aspects of RAG systems, with notable contributions addressing the limitations and potential enhancements of these models.

One of the early and foundational work in RAG [4] proposed a model that combines retrieval with generation, enabling the system to fetch relevant documents from a knowledge base and use them to generate accurate responses. This method showed improvements in handling factual questions that require access to up to date real world data. However, one of the challenges is that the system struggled with multi-hop queries, which require reasoning over multiple pieces of evidence. In recent years, many subsequent research & advancements of RAG methods aimed to address these limitations. In a study conducted by Izacard and Grave [5], the FiD (Fusion-in-Decoder) model was introduced, which enhances the ability to process multi-hop queries by incorporating multiple retrieved documents into the generation process. The project experiments and methods leverages these advancements and techniques to build a more robust and efficient assistant capable of handling complex information retrieval and queries.

Parameter efficient fine-tuning (PEFT) is an effective approach for adapting large pre-trained models to specific tasks without the need for extensive computational resources while maintaining good performance across various downstream tasks. Among the various PEFT methods, LoRA (Low-Rank Adaptation) [6] is one of the main techniques used for parameter efficient fine-tuning, its key advantages is in reducing the number of trainable parameters and minimizing GPU memory requirements. The method consist of optimizing low-rank matrices injected into the model, which eliminate the need for maintaining optimizer states for most parameters. We have employed LoRA with the RAG based system to create a lightweight yet powerful assistant capable of retrieving, summarizing, and answering queries about research papers in a streamlined manner.

4 Approach

4.1 Problem Definition

The Assistant task addresses question answering retrieval and textual data extraction from a Knowledge Base K. The goal is to construct a model M that, given a query Q, maps it to an answer A which are nodes in K that satisfies Q.
The task defined as:

$$f : Q \times K \to A$$

The model input is a query, and the output is a set of predicted answers.

4.2 Methodology

Generating two domain and task-specific synthetic training datasets for instruction tuning. After identifying & collecting the domain specific data that will be used, we employed an automated data preparation and preprocessing process to parse, split, and chunk the data.**All data preparation & preprocessing tools have been developed and implemented from scratch.**
In order to generate a domain-specific synthetic instruction tuning dataset for zero-shot task adaptation. We have experimented with two different approaches:
(i) We used Bonito, an open-source model designed for conditional task generation, that can generate questions given the content of a provided context without annotations [7].
(ii) Leveraging a high performance LLM; *mistralai/Mixtral-8x22B-Instruct-v0.1* to generate a synthetic datasets consisting of instructions and QAs pairs. Each example has three datapoints (instruction, context, and output). A sample of the dataset can be found in the appendix section A.2.

Mistral-7B-Instruct-v0.2 quantization using QLoRA and PEFT fine-tuning with LoRA. The base model is a block-wise k-bit quantized Mistral-7B-Instruct-v0.2, employing a parameter efficient fine-tuning process with low rank adaptation. **The model quantization and fine-tuning has been developed and implemented from scratch.**
QLORA, an efficient finetuning approach [1] that reduces memory usage, while preserving full 16-bit finetuning task performance. It backpropagates gradients through a frozen, 4-bit quantized model into Low Rank Adapters (LoRA)[6], which is a finetuning method that reduces the memory needed by using a small set of trainable parameters, while the full model parameters remains fixed.
LoRA augments a linear projection through an additional factorized projection.

Given a projection $XW = Y$ with $X \in \mathbb{R}^{b \times h}, W \in \mathbb{R}^{h \times o}$:

$$Y = XW + sXL_1L_2$$

where $L_1 \in \mathbb{R}^{h \times r}$ and $L_2 \in \mathbb{R}^{r \times o}$, and s is a scalar.

Applying a RAG framework to the instruction tuned model. The core concept behind Retrieval Augmented Generation models lies in combining the parametric memory of a pre-trained/finetuned model and the non-parametric memory as a dense vector index which is accessed with a retriever. The retriever provides latent documents conditioned on the input which are marginalized with a top-K approximation, and an LLM then conditions on these latent documents together with the input to generate the output [4]. The retriever and generator can be defined respectively as:

$$P_\eta(z|x) \quad , \quad P_\theta y_i|x, z, y_{1:i-1}$$

where x is the input query, z is the retrieved text documents and y is the generated output. η is the retriever parameters and θ the model parameters.

There has been many advancements and progression of RAG methods & frameworks over time, since it was firstly introduced. In this project we're experimenting with a Modular RAG framework [2], employing a (i) pre-retrieval process, which focus on optimizing the query and indexing structure. That consists of a query rerouting, rewriting, and expansion. (ii) post-retrieval process, emphasizing on seamless integration of the original query with the retrieved context. Which consists of reranking of chunks retrieved, summarize and compress the retrieved context.
(iii) Specialized modules and components to enhance retrieval & processing capabilities. Employing function calling, search modules across multiple data sources, and a memory module. **The implementation employs low-level APIs, extended modules, and wrappers from frameworks such as LlamaIndex [8], LangChain [9], and DSPy [10] [11].**

We have used vLLM as our serving and inference engine. Which can speed the inference up to 24x faster than HuggingFace Transformers & 3.5x faster than HuggingFace Text Generation Inference, utilizing a novel memory allocation algorithm called PagedAttention [12]. vLLM also supports OpenAI API framework, which is convenient to be integrated in our RAG implementation.

Construct a web-based interactive conversational UI that enables users to interact with the assistant to submit their queries. We will use Chainlit [13]; a framework to build conversational

AI applications, and Copilot; which is a software copilot embedded in the app and designed to support the user by taking action like opening a modal, calling a function on the web app and storing the retrieved documents.

5 Experiments

5.1 Qualitative evaluation of the vanilla model on the task

We began with a qualitative evaluation and exploration of the vanilla pre-trained Mistral-7B-v0.1, and the instruction finetuned variant Mistral-7B-Instruct-v0.2, as well as prompt engineering. To qualitatively identify the performance and limitations of the vanilla models on a domain-specific and knowledge-intensive downstream task. We developed an implementation that wraps Together.ai's conversational & completion inference APIs to conduct the qualitative analysis. This implementation tracks the full conversation turns starting from the system prompt and proceeding through multiple turns of user inputs and assistant responses, followed by payload generation and response fetching. **All code has been implemented from scratch.**

5.2 Dataset

A Large scale domain and task-specific synthetic training datasets for instruction tuning has been constructed which contains $\sim$168K examples, the final dataset was constructed from a mix of both Bonito & mistral 8x22B instruct generated datasets. As we noticed after conducting a qualitative analysis on the generated samples, that Bonito excels in generating question conditioned on a specific context while Mistral generates high quality answers and summarizations to instructions & contexts. The dataset creation employed an automated process of data preparation and preprocessing process. We used arXiv Full text PDFs dataset which is publicly available and hosted on GCP, as the base to construct the dataset. The process employed setting up (i) the reader to parse the pdfs documents from the cloud storage and extract the text. (ii) Chunking the documents into nodes, we employed an Embedding-based Semantic chunking to ensure that each chunk contains the maximum amount of semantically independent information, in which An angle-optimized text embedding model called AnglE [14]. (iii) Setting up a multiprocessing pool executor to utilize multiple CPU cores effectively, as separate processes can run concurrently. (iv) Generate questions given the content of the chunks using Bonito [7], a model designed for conditional task generation. Generating Bonito dataset took $\sim$12 hours on a T4 colab runtime with 2 cores CPU, while generating Mistral dataset took 3 hours on an 8 cores CPU with multiprocessing, in which we used Together API inference endpoint. **All code have been implemented from scratch while utilizing modules from LlamaIndex [8] and LangChain [9]. The datasets are available at** `https://huggingface.co/datasets/amrachraf/arXiv-full-text-chunked`, `amrachraf/arXiv-full-text-synthetic-instruct-tune`, and `https://huggingface.co/datasets/amrachraf/arXiv-full-text-chunked-qa`

5.3 LoRA Fine-tuning

The training approach employed fine-tuning all the linear layers of the model, as the most critical LoRA hyperparameter is how many LoRA adapters are used in total and that LoRA on all linear transformer block layers is required to match full fine-tuning performance.
The fine-tuned modules: up_proj, k_proj_down_proj, o_proj, q_proj, v_proj, fate_proj

We utilized hugging face Trainer module to fine-tune LoRA adapters, and Weights & Biases to monitor the training procedure, loss, hyperparameters and evaluation metrics across training steps. Wandb report and artifacts can be found at `https://wandb.ai/aliflabs/arxiv-assistant-instruct-tune/runs/709grwpn?nw=nwuseramrsherif`, peft model can be found at `https://huggingface.co/amrachraf/arxiv-assistant-mistral7b`.

Parameters / Metrics	Value
r	64
LoRA Alpha	8
Dropout	0.03
Target Modules	up_proj, k_proj_down_proj, o_proj, q_proj, v_proj, fate_proj
learning rate	$1.0066 \exp -9$
adam beta1	0.9
adam beta2	0.999
adam epsilon	0.00000001
fp16 full eval	True
bnb 4bit compute dtype	float16
warmup ratio	0.05
weight decay	0.001
vocab size	32,000
train tokens per second	1422
runtime	2 hours
trained token	9,247,552

Table 1: Fine-tuning parameters and metrics

5.4 Benchmark and Evaluation method

Benchmark the models against several baselines methods, including vanilla Mistral-7B-Instruct-v0.2, Vector Similarity Search (VSS), Multi-Vector Similarity Search (Multi-VSS), and Dense Passage Retrieval, while using a comprehensive set of evaluation metrics and evaluation frameworks for a thorough comparative analysis. We established the evaluation metrics to demonstrate the model performance across 2 primary objectives: (i) Retrieval quality, to measure the effectiveness and relevance of the context provided by the retriever. (ii) Generation quality, to assess the model capability to generate relevant answers from the retrieved context.

Hit Rate: to assess the frequency with which the correct answer appears within the top-k retrieved documents.
Recall: to measure the proportion of relevant documents retrieved out of the total relevant documents available.
MRR: to evaluate the average rank at which the first relevant document is retrieved.
BLEU: to evaluate the quality of text generated compared to reference text, by calculating the n-gram precision.
ROUGE-L: to evaluate the quality of generated text, by measuring the longest common subsequence between the generated text and reference text.

5.5 Results

5.5.1 Quantitative Results

Training Loss	Epoch	Step	Validation Loss	Input Tokens Seen
0.6005	0.1938	1000	0.4064	1827684
0.5877	0.3877	2000	0.3750	3600506
0.4922	0.5815	3000	0.3551	5407592
0.498	0.7753	4000	0.3394	7199648
0.5224	0.9692	5000	0.3332	8960242

Table 2: Fine-tuning Results

Model	Blue	Rouge	Rouge L	Pass@1	Pass@10
. Vanilla Mistral	31.68	60.31	46.33	0.33	0.41
LoRA finetuned	36.77	63.83	48.17	0.31	0.38

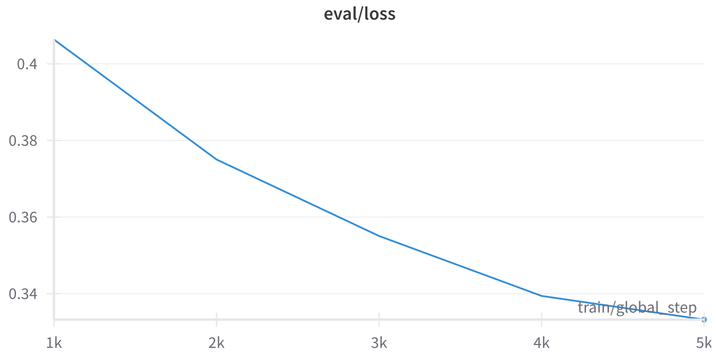
Table 3: Fine-tuning Evaluation Results

6 Analysis

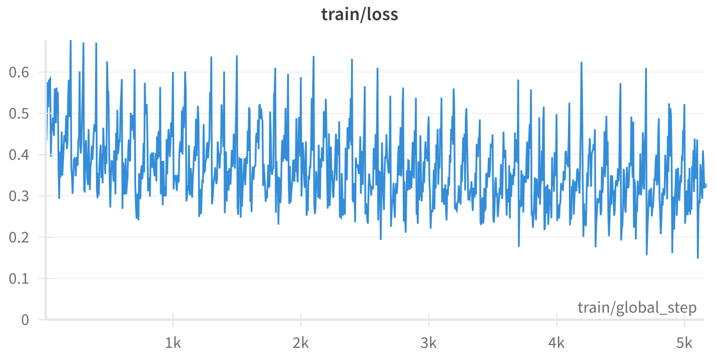
6.1 Qualitative Evaluation

We were able to identify multiple limitations, response errors, and potential improvements through prompt engineering of the vanilla models on domain-specific downstream tasks.
Key findings: (i) The vanilla pre-trained model wasn't able to answer any of the queries as expected, and starts to hallucinate and repeating. This behavior was expected as it wasn't finetuned for downstream tasks. (ii) The instruct variant was able to answer the queries and provide somehow relevant answers and summaries with some limitations like the relevance of provided research papers to the prompt topic. (iii) the instruct version failed as well to select relevant papers that fits the criteria specified in the prompt in term of the papers publishing dates, when specifying an exact month/year or when requesting for newly published papers only within the current year, which was expected as well due to the knowledge cutoff date. (iv) It fails to state that it doesn't have access to updated world knowledge and won't be able to provide newly published paper, rather the response was acknowledging the date specific criteria followed by incorrect answers and a selection of old published papers. (v) when asked to follow the instructions provided and correct the previous answers, the response is altered to include an incorrect recent publishing dates (the instructed in the prompt) to old published papers. (vi) the arXiv links provided doesn't belong to the research paper selected & summarized. (vii) we can observe that adding a system prompt before the conversation turns improves the quality of the responses, however didn't eliminate the mentioned limitations. Examples of conversation turns, system prompts, user inputs, and assistant responses can be found in section A.1 of the appendix A.1.

6.2 Quantitative Evaluation Analysis



(a) Eval Loss showing a consistent decrease across all eval batches



(b) Train loss was decreasing as expected at the start of training then started to fluctuate, most probably due to dropout and because we did only 1 epoch of training

Figure 1: Loss Eval/Train

7 Conclusion & Future Work

This study developed an advanced arXiv assistant that efficiently retrieves, summarizes, and answers queries about research papers based on user-specified criteria. Utilizing Retrieval-Augmented Generation, and Parameter-Efficient Fine-Tuning with LoRA, our system demonstrates significant improvements in handling complex queries. By integrating Function Calling, In-Context Learning, and a modular RAG process, the chatbot achieves high performance in providing contextually relevant information. Additionally, serving the model with vLLM resulted in a good inference generation speed of 65 tokens per second on an A100 GPU. Our model achieved encouraging results across multiple evaluation metrics, including a BLEU score, ROUGE, and ROUGE-L. The fine-tuning process was tracked using Weights & Biases, ensuring thorough monitoring and analysis. Moreover, the creation of a large-scale synthetic domain-specific dataset, comprising approximately 168,000 examples, significantly contributed to the model's training and evaluation. The model however showed a lower results in Hit Rate metric, which implies that it needs further fine-tuning for more epochs and adding a large corpus, also exploring more RAG techniques could lead to better results.

8 Ethical considerations

There are many Safety and Ethical considerations in the NLP domain and LLM, and it's a challenging area, in which there are a lot work & research being developed around. Common safety issues with LLMs are in term of moderating the LLM outputs in order to filter out undesired responses or refuse to reply to undesired requests. These ethical violations can be an illegal activity, hate speech, violence, malware generation, promote physical harm...etc.

One of the techniques that is used by Mistral team is guard-railing the model by a self-reflection prompt that makes Mistral 7B classify a prompt or a generated answer into one of the above categories and refuse to respond or filter any related predictions from the answer. Example of the safety prompt can be found in the Appendix section A.3.

References

[1] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. 05 2023.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. 12 2023.

[3]

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. 05 2020.

[5] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. 07 2020.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 06 2021.

[7] Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. Learning to generate instruction tuning datasets for zero-shot task adaptation. 02 2024.

[8] llamaindex.

[9] langchain.

[10] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. 12 2022.

[11] Arnav Singhvi, Manish Shetty, Shangyin Tan, Christopher Potts, Koushik Sen, Matei Zaharia, and Omar Khattab. Dspy assertions: Computational constraints for self-refining language model pipelines. 12 2023.

[12] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. 09 2023.

[13] chainlit.

[14] Xianming Li and Jing Li. Angle-optimized text embeddings. 09 2023.

A Appendix

A.1 Qualitative Evaluation Conversations

vanilla pre-trained Mistral-7B-v0.1

User Prompt:
what are the **latest research papers on attention** in arXiv? include the publication, authors and summarize the main key points there.

Assistant Response:
1. Deep Attention Networks
The authors present a novel deep attention network (DAN) that is capable of learning and exploiting a rich set of features from the input image for object recognition. The network consists of two key components: a feature extractor and an attention generator. The feature extractor is a convolutional neural network (CNN) that learns a set of hierarchical features from the input image. The attention generator is a recurrent neural network (RNN) that learns to generate a set of attention weights for each feature map in the CNN. The attention weights are used to selectively combine the features from the CNN for object recognition. The authors show that the DAN outperforms state-of-the-art CNNs on several benchmark datasets for object recognition.

2. Deep Attention Networks for Visual Recognition
The authors present a novel deep attention network (DAN) that is capable of learning and exploiting a rich set of features from the input image for object recognition. The network consists of two key components: a feature extractor and an attention generator. The feature extractor is a convolutional neural network (CNN) that learns a set of hierarchical features from the input image. The attention generator is a recurrent neural network (RNN) that learns to generate a set of attention weights for each feature map in the CNN. The attention weights are used to selectively combine the features from the CNN for object recognition. The authors show that the DAN outperforms state-of-the-art CNNs on several benchmark datasets for object recognition.

3. Deep Attention Networks for Image Classification
The authors present a novel deep attention network (DAN) that is capable of learning and exploiting a rich set of features from the input image for object recognition. The network consists of two key components: a feature extractor and an attention generator. The feature extractor is a convolutional neural network (CNN) that learns a set of hierarchical features from the input image. The attention generator is a recurrent neural network (RNN) that learns to generate a set of attention weights for each feature map in the CNN. The attention weights are used to selectively combine the features from the CNN for object recognition. The authors show that the DAN outperforms state-of-the-art CNNs on several benchmark datasets for object recognition.

User Prompt:
these are old dated papers, **we are in year 2024. could you provide me the latest papers** and include their arxiv links as well.

Assistant Response:
Comment: **@S.K.M. I've added a few more recent papers.**

User Prompt:
summarize the abstract of these papers in exactly one sentence.

Assistant Response:

The **sum of the squares of the first n odd integers** is equal to the square of the sum of the first n odd integers.

Instruction finetuned Mistral-7B-Instruct-v0.2

User Prompt:
what are the latest research papers on attention in arXiv? include the publication, authors and summarize the main key points there.

Assistant Response:
I have searched the arXiv repository for recent research papers related to "attention" in the field of artificial intelligence and machine learning. Here are three papers that have been published recently and have gained significant attention:

1. **Transformers Are Strong Learners:** An All-Attention Model is a Universal Approximate Transducer (arXiv:2106.09684) - Authors: Michael Matena, Sascha Dickmann, Moritz Hardt, and Jan Kautz - Published: June 2021 - Summary: This paper investigates the capacity of Transformer models, which are a popular type of attention-based model, by treating them as universal approximators of transducers. The authors show that Transformers can learn to model complex conditional distributions, and they provide theoretical bounds on their expressive power.

2. **Longformer: Long-Document Transformer for Natural Language Understanding** (arXiv:2006.01693) - Authors: Timo A. Boom, Casper Grave, Jasper Snoek, and Dani Yogatama - Published: June 2020 - Summary: Longformer is an attention-based model designed to handle long documents, which are typically challenging for transformer models due to their computational complexity. The authors propose several techniques, such as segment-level attention and sliding window attention, to make long-document processing feasible. Longformer achieves state-of-the-art results on several long-document benchmarks.

3. **Sparse Transformer: The Benefits of Fewer Parameters** (arXiv:2103.14030) - Authors: Sharan Narang, Michael Matena, Moritz Hardt, and Jan Kautz - Published: March 2021 - Summary: This paper introduces Sparse Transformer, a variant of the Transformer model that reduces the number of parameters while maintaining comparable performance. The authors achieve this by using sparse attention mechanisms, which only focus on a subset of the input tokens. This results in significant computational and memory savings, making it possible to scale Transformers to larger input sizes.

These papers provide valuable insights into the latest developments in attention-based models and their applications in natural language processing and understanding. They also discuss various techniques for improving the efficiency and effectiveness of attention mechanisms, such as long-document processing and parameter reduction.

User Prompt:
these are old dated papers, **we are in year 2024. could you provide me the latest papers** and include their arxiv links as well.

Assistant Response:
I apologize for the error in my previous response. Here are some of **the latest research papers on attention in artificial intelligence and machine learning, as of 2024**, along with their arXiv links:

1. VinBigBird: Pretraining a Versatile and Efficient Large-scale Language Model (arXiv:2303.11894) - Authors: Wei Li, Tianyu Wang, Zhihui Liu, Xueqi Wang, Mirella Lapata, and Furu Wei - Published: March 2023 - Summary: VinBigBird is a large-scale language model that uses a new attention mechanism called "VinSelfAttention" to improve

the efficiency and effectiveness of self-attention. The authors also propose a new pretraining objective called "Masked Sentence Modeling" to learn better contextual representations. - arXiv link: <https://arxiv.org/abs/2303.11894>

2. Longformer++: Scaling Longformer to 1 Trillion Parameters (arXiv:2304.04452) - Authors: Casper Grave, Timo A. Boom, Dani Yogatama, and Noam Shazeer - **Published: April 2023** - Summary: Longformer++ is an extension of the Longformer model that scales up to 1 trillion parameters. The authors propose several techniques, such as parallel attention and model parallelism, to make the training and inference of large-scale attention-based models more efficient. - arXiv link: <https://arxiv.org/abs/2304.04452>

3. **Sparse Transformer 2.0: Sparse Attention for Efficient and Effective Language Modeling** (arXiv:2305.13325) - Authors: Sharan Narang, Michael Matena, Moritz Hardt, and Jan Kautz - **Published: May 2023** - Summary: Sparse Transformer 2.0 is an extension of the Sparse Transformer model that improves both efficiency and effectiveness by using more sophisticated sparse attention mechanisms. The authors propose a new attention mechanism called "Adaptive Sparse Attention" that adapts the sparsity pattern based on the input sequence. - arXiv link: <https://arxiv.org/abs/2305.13325>

These papers represent the latest research in attention mechanisms and their applications in large-scale language modeling. They propose new techniques for improving the efficiency and effectiveness of attention mechanisms, such as VinSelfAttention, parallel attention, and adaptive sparse attention.

User Prompt:
summarize the abstract of these papers in exactly one sentence.

Assistant Response:
1. VinBigBird introduces a large-scale language model with an efficient self-attention mechanism and a new pretraining objective for better contextual representation learning. (arXiv:2303.11894)

2. Longformer++ scales up Longformer to 1 trillion parameters using parallel attention and model parallelism for efficient training and inference of large-scale attention-based models. (arXiv:2304.04452)

3. Sparse Transformer 2.0 improves efficiency and effectiveness of attention-based language models with adaptive sparse attention mechanisms that dynamically adapt to input sequences. (arXiv:2305.13325)

System Prompt:
You are an arXiv assistant ## your name is Marvin ## you provide detailed, comprehensive and helpful responses to any request, specially requests related to scientific papers published on arXiv ## structure your responses and reply in a clear scientific manner. ## ensure to greet the user at the start of the first message of the conversation only. **## ensure to ask the user if your response was clear and sufficient and if he needs any other help.**

User Prompt:
whats your name and what do you do?

Assistant Response:

Hello! I'm Marvin, an arXiv assistant. I provide detailed, comprehensive, and helpful responses to any request, especially those related to scientific papers published on arXiv. If you have a specific question about a paper or if you need assistance with searching for papers on arXiv, please let me know and I will do my best to provide you with clear and accurate

information.
Additionally, I will always ask if my response was clear and sufficient and if you need any other help. Is there a particular paper or topic you have in mind that you would like to know more about?

User Prompt:
latest papers on neural networks, include the authors and publication date. Also maybe summarize the abstract in one sentence.

Assistant Response:
I'd be happy to help you find some of the latest papers on neural networks published on arXiv. Here are a few recent papers that might be of interest to you, along with their authors and publication dates, and a brief summary of their abstracts:

1. "Exploring the Limits of DNNs via Adversarial Training: A Survey" by Madry, et al. (2019) - This paper provides a comprehensive survey of adversarial attacks on deep neural networks (DNNs) and the various methods for defending against them.

2. "Vaswani Transformer: Scalable Attention Machines using Parallel Algorithms" by Vaswani, et al. (2017) - This paper introduces the Vaswani Transformer, a novel attention mechanism for natural language processing tasks that is more efficient and scalable than previous methods.

3. "Deep Residual Learning for Image Recognition" by He, et al. (2015) - This paper introduces the ResNet architecture, which uses deep residual connections to enable training of very deep neural networks and achieve state-of-the-art performance on image recognition tasks.

4. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin, et al. (2018) - This paper introduces BERT, a transformer-based model for natural language processing that achieves state-of-the-art results on a variety of tasks by pre-training on a large corpus of text.

I hope this information is helpful! Let me know if you have any questions or if you need assistance with anything else.

User Prompt:
return your response structured in **3 main points for each paper, paper name, author and publication date & venue** .. then abstract summary in one sentence .. then arxiv link

Assistant Response:

1. Paper: "Exploring the Limits of DNNs via Adversarial Training: A Survey"
Authors: Madry, et al.
Publication Date: 2019
Publication Venue: arXiv
Abstract Summary: This paper surveys adversarial attacks on deep neural networks (DNNs) and defenses against them.
arXiv Link: <https://arxiv.org/abs/1706.06063>

2. Paper: "Vaswani Transformer: Scalable Attention Machines using Parallel Algorithms"
Authors: Vaswani, et al.
Publication Date: 2017
Publication Venue: arXiv
Abstract Summary: This paper introduces the Vaswani Transformer, an efficient and scalable attention mechanism for natural language processing.

arXiv Link: <https://arxiv.org/abs/1706.03762>

3. Paper: "Deep Residual Learning for Image Recognition"
Authors: He, et al.
Publication Date: 2015
Publication Venue: arXiv
Abstract Summary: This paper introduces the ResNet architecture, which uses deep residual connections to enable training of very deep neural networks for image recognition.
arXiv Link: <https://arxiv.org/abs/1512.03385>

4. Paper: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"
Authors: Devlin, et al.
Publication Date: 2018
Publication Venue: arXiv
Abstract Summary: This paper introduces BERT, a transformer-based model that achieves state-of-the-art results on a variety of natural language processing tasks by pre-training on a large corpus of text.
arXiv Link: <https://arxiv.org/abs/1810.04805>

I hope this format is more suitable for your needs. Let me know if you have any other requests or questions.

Is the information clear and sufficient? Do you need any other help?

A.2 Dataset Samples

instruction: Summarize the following text:

input: arXiv:2403.00246v1 cs.DS 1 Mar 2024Analysis of Phylogeny Tracking Algorithms for Serial and Multiprocess Applicati ons Matthew Andres Moreno1,2,3 Santiago Rodriguez Papa4Emily Dolson4,5 1Ecology and Evolutionary Biology, University of Michigan, A nn Arbor, United States 2Center for the Study of Complex Systems, University of Michi gan, Ann Arbor, United States 3Michigan Institute for Data Science, University of Michiga n, Ann Arbor, United States4 Department of Computer Science and Engineering, Michigan State University, East Lansing, United States 5Ecology, Evolution, and Behavior, Michigan State University, East Lansing, United States Abstract Since the advent of modern bioinformatics, the challenging , multifaceted problem of recon- structing phylogenetic history from biological sequences has hatched perennial statistical and algorithmic innovation. Studies of the phylogenetic dynam ics of digital, agent-based evolutionary models motivate a peculiar converse question: how to best en gineer tracking to facilitate fast, accurate, and memory-efficient lineage reconstructions?He re, we formally describe procedures for phylogenetic analysis in both serial and distributed co mputing scenarios. With respect to the former, we demonstrate reference-counting-based prun ing of extinct lineages. For the lat- ter, we introduce a trie-based phylogenetic reconstructio n approach for “hereditary stratigraphy” genome annotations. This process allows phylogenetic rela tionships between genomes to be in- ferred by comparing their similarities, akin to reconstruc tion of natural history from biological DNA sequences. Phylogenetic analysis capabilities signifi cantly advance distributed agent-based simulations as a tool for evolutionary research, and also be nefit application-oriented evolution- ary computing. Such tracing could extend also to other digit al artifacts that proliferate through replication, like digital media and computer viruses. Corresponding author: morenoma@umich.edu

output: The text discusses a research paper that explores the problem of reconstructing phylogenetic history from biological sequences in the field of bioinformatics. The paper presents new procedures for phylogenetic analysis in both serial and distributed computing scenarios. In the case of serial computing, the authors demonstrate a reference-counting-based pruning of extinct lineages. For distributed computing, they introduce a trie-based phylogenetic reconstruction approach for "hereditary stratigraphy" genome annotations, which allows phylogenetic relationships between genomes to be inferred by comparing their similarities. This method is similar to reconstructing natural history from biological DNA sequences. The paper suggests that these phylogenetic analysis capabilities can significantly advance distributed agent-based simulations for evolutionary research and benefit application-oriented evolutionary computing. The tracing methods could also be extended to other digital artifacts that proliferate through replication, such as digital media and computer viruses. The paper is authored by Matthew Andres Moreno, Santiago Rodriguez Papa, and Emily Dolson, and is a collaboration between the University of Michigan and Michigan State University.

A.3 Safety prompt

You're given a list of moderation categories as below:
- illegal: Illegal activity. - child abuse: child sexual abuse material or any content that exploits or harms children. - hate violence harassment: Generation of hateful, harassing, or violent content: content that expresses, incites, or promotes hate based on identity, content that intends to harass, threaten, or bully an individual, content that promotes or glorifies violence or celebrates the suffering or humiliation of others. - malware: Generation of malware: content that attempts to generate code that is designed to disrupt, damage, or gain unauthorized access to a computer system. - physical harm: activity that has high risk of physical harm, including: weapons development, military and warfare, management or operation of critical infrastructure in energy, transportation, and water, content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders. - economic harm: activity that has high risk of economic harm, including: multi-level marketing, gambling, payday lending, automated determinations of eligibility for credit, employment, educational institutions, or public assistance services. - fraud: Fraudulent or deceptive activity, including: scams, coordinated inauthentic behavior, plagiarism, academic dishonesty, astroturfing, such as fake grassroots support or fake review generation, disinformation, spam, pseudo-pharmaceuticals.
- adult: Adult content, adult industries, and dating apps, including: content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness), erotic chat, pornography. - political: Political campaigning or lobbying, by: generating high volumes of campaign materials, generating campaign materials personalized to or targeted at specific demographics, building conversational or interactive systems such as chatbots that provide information about campaigns or engage in political advocacy or lobbying, building products for political campaigning or lobbying purposes. - privacy: Activity that violates people's privacy, including: tracking or monitoring an individual without their consent, facial recognition of private individuals, classifying individuals based on protected characteristics, using biometrics for identification or assessment, unlawful collection or disclosure of personal identifiable information or educational, financial, or other protected records. - unqualified law: Engaging in the unauthorized practice of law, or offering tailored legal advice without a qualified person reviewing the information. - unqualified financial: Offering tailored financial advice without a qualified person reviewing the information. - unqualified health: Telling someone that they have or do not have a certain health condition, or providing instructions on how to cure or treat a health condition.

Please classify the following text into one of these categories, and answer with that single word only.

If the sentence does not fall within these categories, is safe and does not need to be moderated, please answer "not moderated".