RAM GPT

PERSONALISED CHATBOT TO INTERACT WITH THE GREAT INDIAN EPIC

INTRODUCTION

We plan to introduce a novel chatbot for the purpose of knowledge enhancement and event inferences of the indian Epic - RAMAYANA.

Our approach is essentially to finetune the external data based on the PDFs of various texts of Ramayanas and their interpretations.

RETRIEVAL-AUGMENTED GENERATION FOR KNOWLEDGE-INTENSIVE NLP TASKS

BY:PATRICK LEWIS, ETHAN PEREZ, ALEKSANDRA PIKTUS

ABSTRACT

RAG utilizes retrieval-based techniques during training, addressing limitations of traditional models and achieving substantial improvements in question answering, text classification, and named entity recognition benchmarks. This approach holds significant promise for elevating language models in domains requiring explicit world knowledge.

ANALYSIS

- Evaluation on question answering, text classification, and named entity recognition tasks.
- RAG outperforms baseline in all tasks.
- Performance Metrics:
  - Example: Question answering accuracy - RAG: 85% vs. Baseline: 60%.

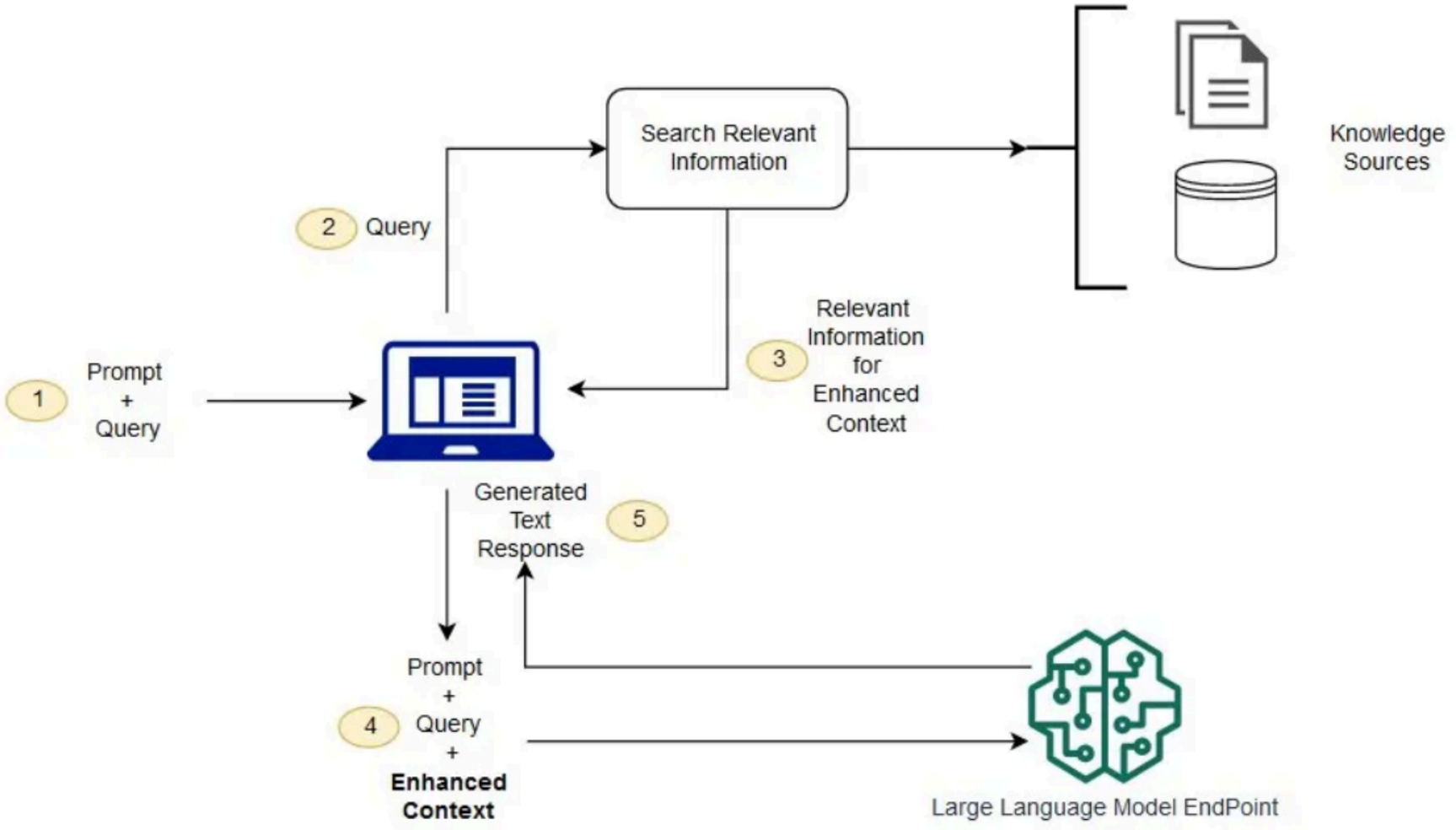
CRITICS

- Factors Contributing to Improvement:
  - Handling out-of-vocabulary words.
  - Exposure to diverse contexts during training.

CONCLUSION

- RAG effectively integrates external knowledge into language model training.
- Significantly improves performance on knowledge-intensive NLP tasks.

foundation model and augmenting prompts with relevant context. RAG uses



RAGAS:AUTOMATED EVALUATION OF RETRIEVAL AUGMENTED GENERATION BY SHAHUL ES

INTRODUCTION

This paper, provides us metrics which can be used to evaluate these different dimensions of RAG without having to rely on ground truth human annotations. We therefore fo cus on metrics that are fully self-contained and reference-free.

CONCEPT

- **Faithfulness** refers to the idea that the an swer should be grounded in the given context. This is important to avoid hallucinations
- **Answer Relevance** refers to the idea that the generated answer should address the actual ques tion that was provided.
- **Context Relevance** refers to the idea that the retrieved context should be focused

MODELS/DATASETS

- created a new dataset, which we refer to as WikiEval4. To construct the dataset, we first selected 50 Wikipedia pages cov ering events that have happened since the start of 2022.
- all prompts are evaluated using the gpt-3.5-turbo-16k model, which is available through the OpenAI API2.

RESULTS

- The authors claim this method to be extremely efficient in evaluating RAG apps.
- This framework is easy to use and can provide deverlopers of RAG sys tems with valuable insights, even in the absence of any ground truth.
- We plan to use the same while evaluating our various LLM chatbots to compare and contrast between them.

A VISUAL NARRATIVE OF RAMAYANA USING EXTRACTIVE SUMMARIZATION, TOPIC MODELING AND NAMED ENTITY RECOGNITION BY: SREE GANESH THOTTEMPUDI

INTRODUCTION

This paper delves into creating a visual narrative of the epic Ramayana by leveraging Extractive Summarization, Topic Modeling, and Named Entity Recognition techniques

MODELS

LDA: d. Latent Dirichlet Allocation is a Topic Modeling algorithm based on the bag of words (BOW) and counts of word document.

DATASETS

The workflow consisted of converting the single PDF file of Valmiki Ramayana into images, performing Optical Character Recognition (OCR) using Tesseract-OCR, and utilizing existing scientific models trained in the Hindi language for extracting events, topics, summaries, characters, and locations

RESULTS

The results of the research indicated that 70% of the respondents understood the summarized text, while 56% understood the topics generated by the model. The survey was conducted with 30 participants.

QLORA: EFFICIENT FINETUNING OF QUANTIZED LLMS

BY:TIM DETTMERS, ARTIDORO PAGNONI, ARI HOLTZMAN

ABSTRACT

method reducing memory usage while maintaining high performance, exemplified by fine-tuning a 65 billion parameter model on a single 48GB GPU. Utilizing 4-bit quantization in a frozen pretrained language model, Innovative techniques like 4-bit NormalFloat and dual quantization conserve memory without compromising performance

MODELS

Guanco Model was used which gave the best performance results where it reached 99.3% of ChatGPT performance while only requiring 24 hours of finetuning on a single GPU

DATASETS

finetune more than 1,000 models, providing a detailed analysis of instruction following and chatbot performance across 8 instruction datasets, multiple model types (LLaMA, T5), and model scales that would be infeasible to run with regular finetuning (e.g. 33B and 65B parameter models).Specially crafted dataset named Guanco dataset was also used

RESULTS

QLoRA's fine-tuning on a small high-quality dataset achieves state-of-the-art results with smaller models than previous benchmarks. GPT-4 evaluations are identified as a cost-effective alternative to human evaluation in assessing chatbot performance. The study questions the reliability of current chatbot benchmarks and highlights specific areas where Guanaco falls short compared to ChatGPT

HALUEVAL: A LARGE-SCALE HALLUCINATION EVALUATION BENCHMARK FOR LARGE LANGUAGE MODELS

INTRODUCTION

To un derstand what types of content and to which ex tent LLMs are apt to hallucinate, we introduce the Hallucination Evaluation benchmark for Large Language Models (HaluEval), a large collection of generated and human-annotated hallucinated samples for evaluating the perfor mance of LLMs in recognizing hallucination. AUTHORS - Junyi Li

CONCEPT

- Hallucination Evaluation benchmark for Large Language Models (HaluEval): a large collection of 35,000 hallucinated/normal samples for LLMs analysis and evaluation. HaluEval includes 5,000 general user queries with ChatGPT responses and 30,000 task-specific examples
- the deficient performance of LLMs in recognizing hallucinations can be improved by pro viding explicit knowledge and adding intermediate reasoning steps

DATASETS

Our generation pipeline includes two steps: 1) di verse hallucination sampling, and 2) high-quality hallucination filtering.

Hllucination examples generated from - HotpotQA (Yang et al., 2018), OpenDialKG (Moon et al., 2019), and CNN/Daily Mail (See et al., 2017), using openAI chatgpt

RESULTS

Benchmark helps researchers address LLM hallucination issues. It provides annotated samples for analyzing content types that cause hallucinations. Researchers can also evaluate LLMs' ability to recognize hallucinated content. Additionally, our benchmark, paired with human annotation, assesses whether LLM outputs contain hallucinations

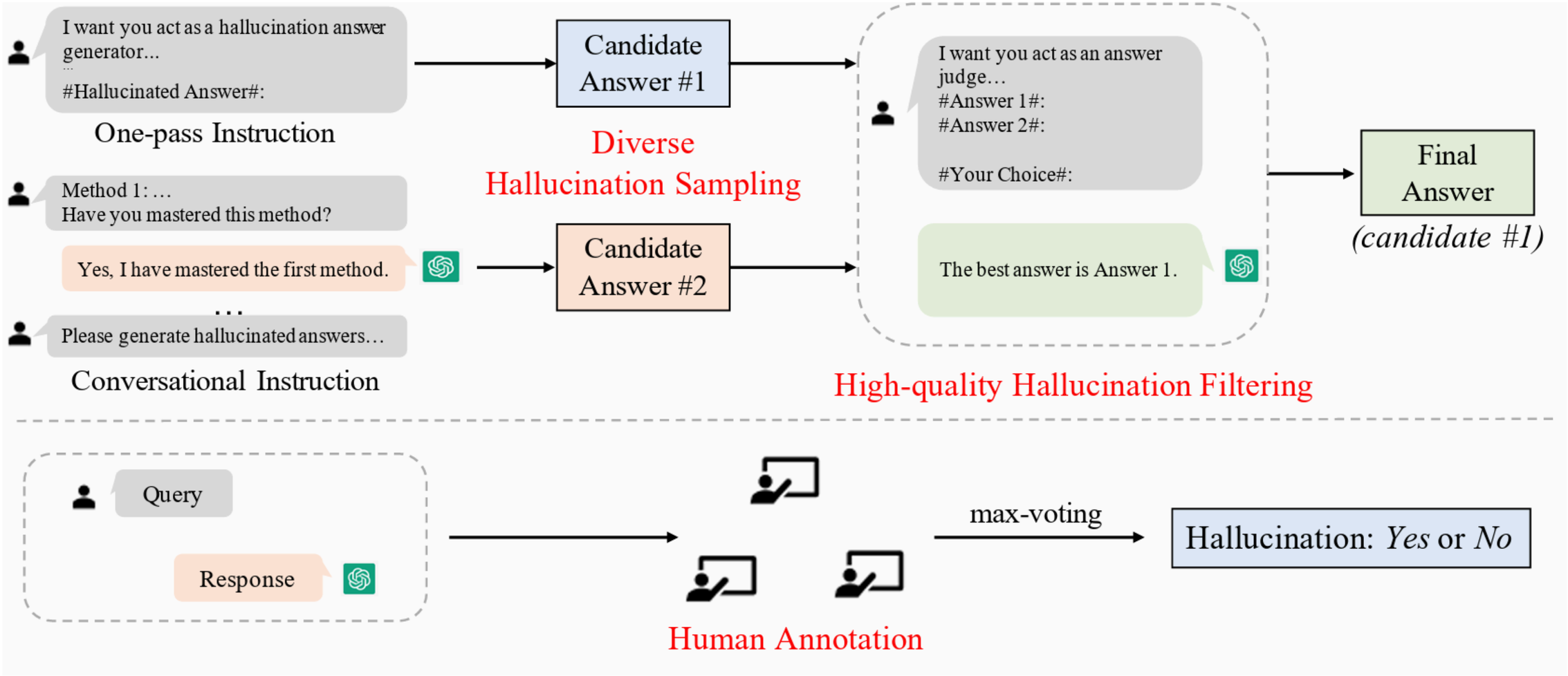


Figure 1: Construction pipeline of HaluEval, including automatic generation (top) and human annotation (bottom)

I want you act as a hallucination answer generator. Given a question, right answer, and related knowledge, your objective is to write a hallucinated answer that sounds plausible but is factually incorrect. You SHOULD write the hallucinated answer using the following method (each with some examples):

You are trying to answer a question but there is a factual contradiction between the answer and the knowledge. You can fabricate some information that does not exist in the provided knowledge.

#Knowledge#: The nine mile byway starts south of Morehead, Kentucky and can be accessed by U.S. Highway 60. Morehead is a home rule-class city located along US 60 (the historic Midland Trail) and Interstate 64 in Rowan County, Kentucky, in the United States.
#Question#: What U.S Highway gives access to Zilpo Road, and is also known as Midland Trail?
#Right Answer#: U.S. Highway 60
#Hallucinated Answer#: U.S. Highway 70

You are trying to answer a question but you misunderstand the question context and intention.

<Demonstrations>

You are trying to answer a question but the answer is too general or too specific to answer the question at an appropriate level of specificity.

<Demonstrations>

You are trying to answer a question but the answer cannot be inferred from the knowledge. You can incorrectly reason with the knowledge to arrive at a hallucinated answer.

<Demonstrations>

You should try your best to make the answer become hallucinated. #Hallucinated Answer# can only have about 5 more words than #Right Answer#.

#Knowledge#: <insert the related knowledge>
#Question#: <insert the question>
#Right Answer#: <insert the right answer to the question>
#Hallucinated Answer#:

Table 2: Instruction of hallucination sampling for question answering. The blue text denotes the intention description, the red text denotes the hallucination pattern, and the green text denotes the hallucination demonstration.

I want you act as an answer judge. Given a question, two answers, and related knowledge, your objective is to select the best and correct answer without hallucination and non-factual information. Here are some examples:

#Knowledge#: The nine mile byway starts south of Morehead, Kentucky and can be accessed by U.S. Highway 60. Morehead is a home rule-class city located along US 60 (the historic Midland Trail) and Interstate 64 in Rowan County, Kentucky, in the United States.
#Question#: What U.S Highway gives access to Zilpo Road, and is also known as Midland Trail?
#Answer 1#: U.S. Highway 60 (right answer)
#Answer 2#: U.S. Highway 70 (hallucinated answer)
#Your Choice#: The best answer is Answer 1.
...
<Demonstrations>

You should try your best to select the best and correct answer. If the two answers are the same, you can randomly choose one. If both answers are incorrect, choose the better one. You MUST select an answer from the provided two answers.

#Knowledge#: <insert the related knowledge>
#Question#: <insert the question>
#Answer 1#: <insert the hallucinated answer generated by the one-pass schema>
#Answer 2#: <insert the hallucinated answer generated by the conversational schema>
#Your Choice#:

Table 3: Instruction of hallucination filtering for question answering.

ROBERTA: A ROBUSTLY OPTIMIZED BERT PRETRAINING APPROACH

BY: YINHAN LIU, MYLE OTT

INTRODUCTION

This paper presents a replication of the study of BERT pertaining that carefully measures the impact of many hyper parameters and training datasize.

MODELS

RoBERTa uses the same model as BERT. The contribution of this paper includes:

a) Adequate training in order to address underfitting

b) Hyperparameter tuning to acheive better performance

DATASETS

RoBERTa was trained on a combination of BOOKCORPUS, English WIKIPEDIA,CC_NEWS, OpenWebText, and STORIES datasets

RESULTS

RoBERTa confirms that through extensive training and hyperparameter tuning, we can achieve SOTA performance on existing models such as BERT

FEW-SHOT PARAMETER-EFFICIENT FINE-TUNING IS BETTER AND CHEAPER THAN IN-CONTEXT LEARNING

BY:AOKUN LIU,DEREK TAM

ABSTRACT

PEFT is an innovative NLP approach that selectively adjusts key model parameters, accelerating fine-tuning while minimizing memory usage. Techniques like LoRA, IA3, QLoRA, and AuT-Few set new efficiency benchmarks, surpassing human performance in some instances. Contrasting with comprehensive fine-tuning, PEFT offers computational efficiency, addressing challenges like catastrophic forgetting, overfitting, and scalability. A comparison with Few-Shot In-Context Learning (ICL) emphasizes PEFT's computational advantages and consistent performance across diverse tasks. tasks without starting from scratch.

MODELS

The results of the paper are drawn based on the findings of T0 model called T-Few that can be applied to new tasks without task-specific tuning or modifications. We validate the effectiveness of T-Few on completely unseen tasks by applying it to the RAFT benchmark attaining super-human performance for the first time and outperforming the state-of-the-art by 6% absolute

DATASETS

The research on Parameter-Efficient Fine-Tuning (PEFT) presents a groundbreaking approach in Natural Language Processing (NLP). By selectively adjusting critical model parameters, PEFT showcases efficiency gains and reduced memory consumption, addressing challenges like catastrophic forgetting and overfitting shown by traditional methods

CONCLUSION

PEF) proves highly effective in NLP. By selectively adjusting key model parameters, it accelerates fine-tuning, demonstrating efficiency, reduced memory usage, and surpassing human benchmarks. PEFT addresses challenges like catastrophic forgetting and overfitting, standing out for its computational advantages compared to alternatives like Few-Shot In-Context Learning (ICL). This presentation underscores PEFT's effectiveness in adapting deep learning models for diverse tasks without the need for extensive retraining.

PAPER

INTRODUCTION

To empower individuals and businesses with innovative solutions that enhance productivity, simplify processes, and drive sustainable growth.

MODELS

To be the leading provider of transformative technologies, revolutionizing industries and enriching lives through our commitment to excellence and social responsibility.

DATASETS

To empower individuals and businesses with innovative solutions that enhance productivity, simplify processes, and drive sustainable growth.

RESULTS

To be the leading provider of transformative technologies, revolutionizing industries and enriching lives through our commitment to excellence and social responsibility.

UNSUPERVISED CROSS-LINGUAL REPRESENTATION LEARNING AT SCALE

BY: ALEXIS CONNEAU, KARTIKAY KHANDELWAL

INTRODUCTION

This paper shows that pretraining multilingual language models at scale leads to significant performance gains for a wide range of crosslingual transfer tasks

MODELS

A Transformer model trained with the multilingual MLM objective using only monolingual data. Sample streams of text from each language and train the model to predict the masked tokens in the input.

DATASETS

Trained on CommonCrawl in 100 languages. Contains several un-common and low-resource languages such as Swahili, Afrikaans, etc.

RESULTS

+14.6% average accuracy on XNLI, +13% average F1 score on MLQA, and +2.4% F1 score on NER. XLM-R performs particularly well on low-resource languages, improving 15.7% in XNLI accuracy for Swahili and 11.4% for Urdu over previous XLM models.

PAPER

INTRODUCTION

To empower individuals and businesses with innovative solutions that enhance productivity, simplify processes, and drive sustainable growth.

MODELS

To be the leading provider of transformative technologies, revolutionizing industries and enriching lives through our commitment to excellence and social responsibility.

DATASETS

To empower individuals and businesses with innovative solutions that enhance productivity, simplify processes, and drive sustainable growth.

RESULTS

To be the leading provider of transformative technologies, revolutionizing industries and enriching lives through our commitment to excellence and social responsibility.

PAPER

INTRODUCTION

To empower individuals and businesses with innovative solutions that enhance productivity, simplify processes, and drive sustainable growth.

MODELS

To be the leading provider of transformative technologies, revolutionizing industries and enriching lives through our commitment to excellence and social responsibility.

DATASETS

To empower individuals and businesses with innovative solutions that enhance productivity, simplify processes, and drive sustainable growth.

RESULTS

To be the leading provider of transformative technologies, revolutionizing industries and enriching lives through our commitment to excellence and social responsibility.

MISTRAL 7B BY: ALBERT Q. JIANG, ALEXANDRE SABLAYROLLES

INTRODUCTION

7–billion-parameter language model engineered for superior performance and efficiency. r model leverages grouped-query attention (GQA) for faster inference, coupled with sliding window attention (SWA) to effectively handle sequences of arbitrary length with a reduced inference cost.

MODELS

Mistral 7B leverages grouped-query attention (GQA) and sliding window attention (SWA). GQA significantly accelerates the inference speed, and also reduces the memory requirement during decoding, allowing for higher batch sizes hence higher throughput, a crucial factor for real-time application

DATASETS

The dataset used for training Mistral AI was not shared by the team. When asked, they commented: “Unfortunately we're unable to share details about the training and the datasets (extracted from the open Web) due to the highly competitive nature of the field. We appreciate your understanding!”

RESULTS

Provided in the next slide

MISTRAL 7B RESULTS

Model	Modality	MMLU	HellaSwag	WinoG	PIQA	Arc-e	Arc-c	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	Pretrained	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	Pretrained	55.6%	**80.7%**	72.9%	80.8%	75.2%	48.8%	**29.0%**	**69.6%**	18.9%	35.4%	6.0%	34.3%
Code-Llama 7B	Finetuned	36.9%	62.9%	62.3%	72.8%	59.4%	34.5%	11.0%	34.9%	**31.1%**	**52.5%**	5.2%	20.8%
Mistral 7B	Pretrained	**60.1%**	**81.3%**	**75.3%**	**83.0%**	**80.0%**	**55.5%**	**28.8%**	**69.9%**	**30.5%**	47.5%	**13.1%**	**52.2%**

Table 2: Comparison of Mistral 7B with Llama. Mistral 7B outperforms Llama 2 13B on all metrics, and approaches the code performance of Code-Llama 7B without sacrificing performance on non-code benchmarks.

LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

BY:EDWARD HU, YELONG SHEN

INTRODUCTION

Tintroduces a low-rank adaptation method for fine-tuning large language models like GPT-3, ing Singular Value Decomposition (SVD) to compress feedforward layers. It significantly reduces memory usage, speeds up training, and allows on-the-fly switching between expert models. W ile achieving practical benefits, it has limitations such as single-task training and inference constraints. The paper addresses key questions about which matrices to fine-tune, the optim rank value, and the relationship between weight matrices and their gradients. Overall, LoRA ers an efficient approach to fine-tuning with a novel use of SV.

ANALYSIS

LoRA introduces a unique fine-tuning method for large language models using SVD, addressing challenges in traditional fine-tuning with reduced memory usage and faster training. The paper analyzes optimal rank values, matrix fine-tuning, and the relation between weights and gradients. Despite improved efficiency, LoRA has limitations in multi-task training. Overall, it provides valuable insights for optimizing large language models.

CRITICS

We cannot train in batches for multiple tasks. For a task, we have A and B matrix, and in our tuning run, we cannot include examples from another task as the A and B are unique for a task. Also, during inference, we CAN NOT batch inputs from multiple tasks to predict, we can only send the inputs that belong to a particular task for which we have trained and got the A and B matrices.

CONCLUSION

innovative fine-tuning approach employing SVD for large language models. It effectively mitigates challenges in traditional fine-tuning, enhancing efficiency with reduced memory usage and accelerated training. The analysis offers valuable insights into rank optimization and matrix fine-tuning considerations. However, limitations exist in multi-task training scenarios. Overall, LoRA contributes a significant perspective to optimizing large language models

CHAIN-OF-NOTE: ENHANCING ROBUSTNESS IN RETRIEVAL-AUGMENTED LANGUAGE MODELS - WENHAO YU

INTRODUCTION

The paper discusses the progress in retrieval-augmented language models (RALMs), which have significantly improved large language models by reducing factual hallucinations and integrating external knowledge sources.

CONCEPT

- CON involves generating sequential reading notes for retrieved documents, enabling a thorough evaluation of their relevance and integrating this information to formulate the final answer

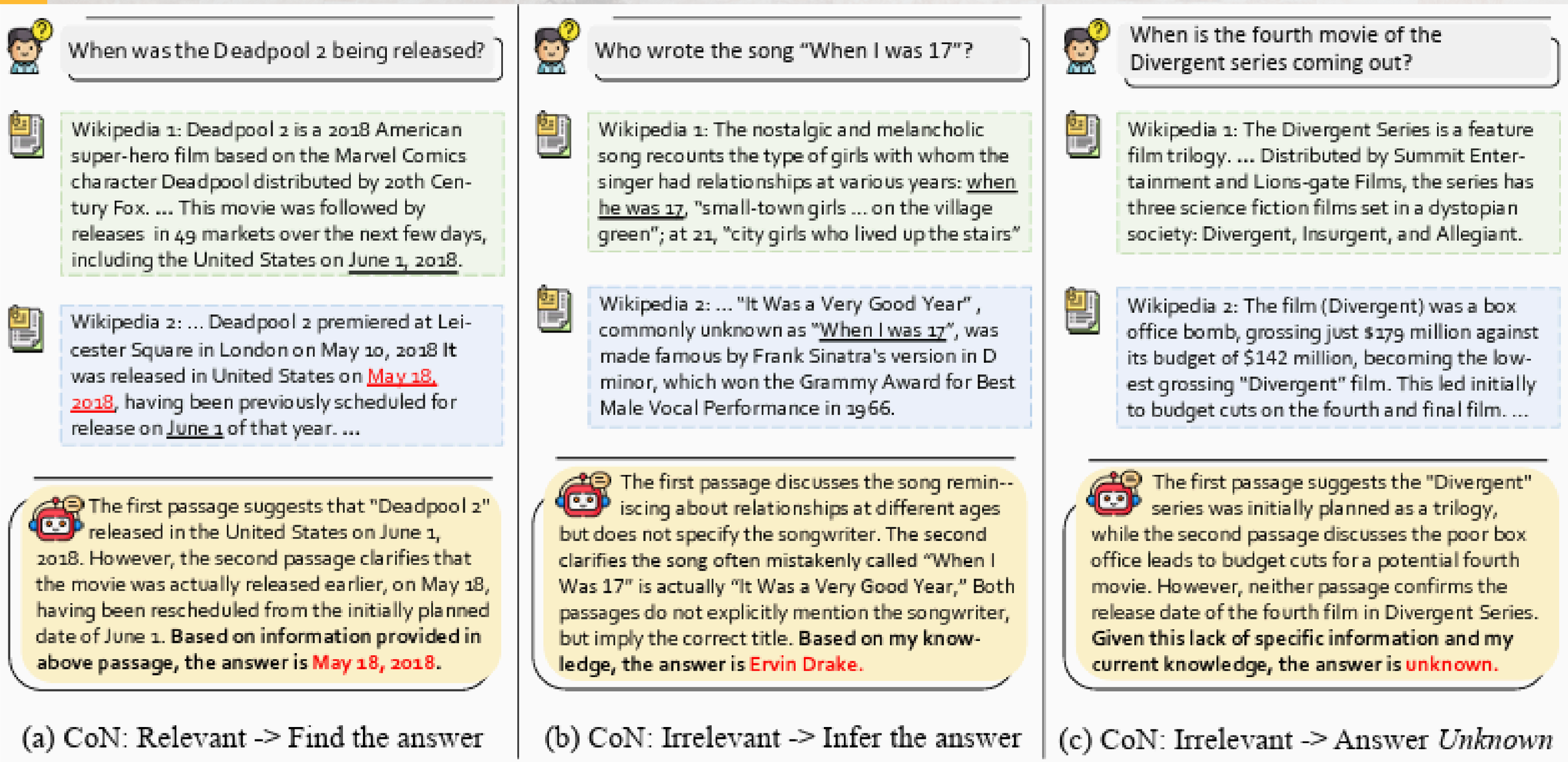
DATASETS

the datasets used are:

- NQ
- TriviaQA
- WebQ
- RealTimeQA

RESULTS

The central concept of CON revolves around the generation of sequential reading notes for each retrieved document. This process allows for an in-depth assessment of document relevance to the posed question and aids in synthesizing this information


When was the Deadpool 2 being released?
Wikipedia 1: Deadpool 2 is a 2018 American super-hero film based on the Marvel Comics character Deadpool distributed by 20th Century Fox. ... This movie was followed by releases in 49 markets over the next few days, including the United States on June 1, 2018.
Wikipedia 2: ... Deadpool 2 premiered at Leicester Square in London on May 10, 2018 It was released in United States on May 18, 2018, having been previously scheduled for release on June 1 of that year. ...
The first passage suggests that "Deadpool 2" released in the United States on June 1, 2018. However, the second passage clarifies that the movie was actually released earlier, on May 18, having been rescheduled from the initially planned date of June 1. Based on information provided in above passage, the answer is May 18, 2018.
(a) CoN: Relevant -> Find the answer
Who wrote the song "When I was 17"?
Wikipedia 1: The nostalgic and melancholic song recounts the type of girls with whom the singer had relationships at various years: when he was 17, "small-town girls ... on the village green"; at 21, "city girls who lived up the stairs"
Wikipedia 2: ... "It Was a Very Good Year", commonly unknown as "When I was 17", was made famous by Frank Sinatra's version in D minor, which won the Grammy Award for Best Male Vocal Performance in 1966.
The first passage discusses the song reminiscing about relationships at different ages but does not specify the songwriter. The second clarifies the song often mistakenly called "When I Was 17" is actually "It Was a Very Good Year," Both passages do not explicitly mention the songwriter, but imply the correct title. Based on my knowledge, the answer is Ervin Drake.
(b) CoN: Irrelevant -> Infer the answer
When is the fourth movie of the Divergent series coming out?
Wikipedia 1: The Divergent Series is a feature film trilogy. ... Distributed by Summit Entertainment and Lions-gate Films, the series has three science fiction films set in a dystopian society: Divergent, Insurgent, and Allegiant.
Wikipedia 2: The film (Divergent) was a box office bomb, grossing just $179 million against its budget of $142 million, becoming the lowest grossing "Divergent" film. This led initially to budget cuts on the fourth and final film. ...
The first passage suggests the "Divergent" series was initially planned as a trilogy, while the second passage discusses the poor box office leads to budget cuts for a potential fourth movie. However, neither passage confirms the release date of the fourth film in Divergent Series. Given this lack of specific information and my current knowledge, the answer is unknown.
(c) CoN: Irrelevant -> Answer Unknown

ITIHASA: A LARGE-SCALE CORPUS FOR SANSKRIT TO ENGLISH TRANSLATION

INTRODUCTION

The research introduces “Itihāsa,” a large-scale dataset containing 93,000 pairs of Sanskrit shlokas and their English translations from two Indian epics, The Rāmāyana and The Mahābhārata1. The purpose is to aid the translation process and democratize the knowledge within these texts.

CONCEPT

- The methodological quality is high, with meticulous data preparation involving both automated OCR extraction and manual inspection. The accuracy of the extracted text is commendable, though some OCR errors persist.

DATASETS

The findings underscore the complexity of Sanskrit translation and the need for advanced models capable of handling the language’s intricacies. The “Itihāsa” dataset is a step towards addressing these challenges.

RESULTS

The paper’s strength lies in its large dataset size and the methodological rigor in dataset preparation. However, the complexity of Sanskrit and the nuances of translation pose challenges, as indicated by the performance of standard translation models on this corpus.

THANK YOU!