IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Chasing Language Perfection: A Harmonized Approach To Language Model Innovation

[1]Helly Yogeshkumar Raval, [2]Dr. Satyen M Parikh
[1]Research Scholar, [2]Executive Dean
[1]Faculty of Computer Applications,
[1]Ganpat University, Kherva, India

Abstract : AI-powered conversational agents are becoming increasingly important in numerous sectors, reshaping human-computer interaction. From Joseph Weizenbaum's ELIZA in the 1960s, which utilized early natural language processing (NLP). From basic conversation simulations to advanced models like GPT-2, Mistral, and LLaMA, these system’s capabilities have substantially evolved. Conversational agents can understand and answer complicated queries with human-like conversational depth. This research paper examines the evolution of QA systems from rule-based to cutting-edge deep learning techniques. Current systems still struggle with managing ambiguous queries, maintaining conversational context, and performing commonsense reasoning tasks**.** This research develops LLMHarmonized, a new self-learning agent that combines the capabilities of two prominent models, Mistral and LLaMA2, into a hybrid architecture known as the Self-Learning Intelligist Agent. Through interaction and feedback, this agent increases its accuracy, contextual comprehension, and ambiguity resolution. Extensive testing with benchmark datasets demonstrates the framework's usefulness. LLMHarmonized earns a high F1 score of 86.7 on the CoQA dataset, demonstrating its ability to keep context and provide correct replies throughout conversations. The system performs well on the NQ-OPEN dataset, which assesses ambiguity resolution in open-domain questions, with an F1 score of 34.9 and an Exact Match of 45.8, indicating its ability to handle complex queries. The agent excels in commonsense reasoning, scoring 61.7% accuracy on the CommonsenseQA dataset, indicating its ability to comprehend and apply everyday logic. LLMHarmonized's self-learning technique and sophisticated language models enhance contemporary QA methods. This study addresses issues in context understanding and ambiguity management, paving the way for future breakthroughs in conversational AI. It aims to create smarter, more adaptable, and human-like conversational agents.

***IndexTerms* -** Machine Learning, Natural Language Processing, Selflearning Question Answering, Large Language Models, Knowledge Bases, Hugging Face Transformers, Artificial Intelligence.

INTRODUCTION

The fast evolution of artificial intelligence has profoundly altered how people interact with machines, notably with the advent of sophisticated Question Answering (QA) systems. These technologies have proven indispensable in a variety of fields, allowing rapid access to information and improving the user experience. This research paper investigates a variety of QA systems, emphasizes the relevance of Large Language Models (LLMs), and digs into the reasons behind and importance of improving QA systems through novel ways. Question-answering systems, a subfield of Natural Language Processing (NLP), have evolved significantly since their inception in the 1960s. Starting with Joseph Weizenbaum's ELIZA in 1966, which simulated conversations using pattern matching, these systems have grown in sophistication. In 1995, ALICE (Artificial Linguistic Internet Computer Entity) marked a major leap by employing AIML (Artificial Intelligence Markup Language) for handling complex dialogues.

IJNRD International Research Journal Research Through Innovation

The early 2000s saw a shift towards statistical and data-driven methods. IBM's Deep Blue highlighted the power of computational algorithms, while the Text REtrieval Conference (TREC) introduced QA tracks in 2001 to enhance question-answering systems. This period also witnessed the rise of neural networks and pretrained language models (PLMs). Word embeddings like Word2Vec (2013) advanced semantic understanding, and by 2015, deep learning methods improved search engines' ability to understand user intent [3].

Google's BERT (2018) revolutionized contextual understanding in NLP tasks, setting the stage for even larger models like OpenAI's GPT-3 (2020). These advances enabled question-answering systems to perform more nuanced and coherent text generation. Today, they are used in diverse fields, including healthcare, banking, education, and personal assistance [4].

Modern QA systems use three main approaches [1]:

1. **Retrieval-based methods**: Rely on finding and presenting relevant information from large datasets.
2. **Generative methods**: Create answers by generating text based on learned context.
3. **Hybrid methods**: Combine retrieval and generative techniques for improved performance.

From rule-based systems to today's AI-driven solutions, the evolution of question-answering systems reflects the rapid advancements in NLP and machine learning. With continued innovation, these systems promise even more advanced and interactive human-computer communication in the future [4].

❖ Types of Question Answering Systems

QA systems are meant to analyze user inquiries and offer exact responses. They can be classified into numerous sorts depending on their underlying methodologies:

- **Information Retrieval-Based QA:** These systems rely on locating relevant documents or sections within a large corpus. They employ search algorithms to discover texts with potential replies, focusing on keyword matching and document rating [1].

- **Knowledge-based QA Systems:** These systems use structured data from knowledge bases or ontologies to deliver responses by querying factual information contained in databases. They excel in fields where data can be clearly classified into things and relationships.

- **Generative QA:** Using neural networks and deep learning, these systems can generate natural language replies. They grasp context and provide answers that are not directly mentioned in the source material, resulting in more conversational and nuanced responses.

- **Hybrid QA:** By combining features of retrieval-based and generative models, hybrid QA systems seek to capitalize on the benefits of several methodologies. They gather relevant data and then utilize generative models to create coherent, contextually appropriate responses.

- **Rule-based QA system:** These systems utilize predetermined language and logical rules to analyze inquiries and create replies. While successful in narrow areas, they lack flexibility and suffer with ambiguous or complicated queries [1].

❖ Applications for Question Answering Systems

Enhanced QA systems have extensive uses in different industries:

- **Customer Service Automation:** Automating replies to typical questions boosts productivity and customer satisfaction while lowering operating expenses.

- **Healthcare Information Retrieval:** Providing correct medical information allows healthcare providers and patients to make educated decisions.

- **Educational Question Answering:** Helping students and teachers by providing immediate access to educational materials and explanations.

- **Financial Advisory Services:** Providing customers with current financial insights and data helps them make informed investment decisions and analyze the market.

- **Legal Research and Compliance:** Streamlining the retrieval of legal papers and compliance information boosts the productivity of lawyers [2].

This research endeavors to push the boundaries of current QA systems by harnessing the power of LLMs and innovative integration techniques. Through this work, we aim to contribute valuable insights and practical solutions that address the evolving challenges in the field of question answering.

NEED OF THE STUDY

Despite developments in QA systems and LLMs, important difficulties have to be solved. Current QA systems frequently suffer with:

- **Handling Complex and Dynamic Queries:** Users demand accurate responses to increasingly complex inquiries, which may include numerous contexts or require current information.

- **Adaptability to New Information:** Because information is always changing, systems must be able to integrate new data rapidly without requiring substantial retraining.

- **Dependence on Large Labeled Datasets:** Many models require massive volumes of labeled data, which is both expensive and time-consuming togather.

- **Improving User Interaction:** There is a demand for QA systems that can better engage users, giving not just answers but also a satisfying interaction experience.

This research fulfills these objectives by presenting a novel strategy that combines LLMs with personalized techniques. By doing so, it aims to increase the accuracy, relevance, and adaptability of QA systems, resulting in higher user satisfaction and system efficiency [5][6][7][8][9].

RESEARCH METHODOLOGY

This study intends to improve Question Answering (QA) systems by combining Large Language Models (LLMs) with self-learning processes, hence minimizing the need for fine-tuning. The process consists of identifying acceptable pre-trained LLMs, executing self-learning procedures, and assessing the system using standard metrics to determine performance gains.

Research Approach

An experimental technique was used to create and assess a QA system that takes use of LLM capabilities via self-learning. The important steps are:

- **Model Selection:** Choosing acceptable pre-trained LLMs based on their design and intrinsic capacity to handle QA tasks without further fine-tuning.

- **Implementing Self-Learning Mechanisms:** Using techniques like reinforcement learning and continuous learning to allow the model to develop independently over time.

Tools & Techniques

- **Development Environment Programming Language:** Python was chosen for its rich support in machine learning and natural language processing.

- **Frameworks and Library:** Libraries such as Hugging Face Transformers were used to access pre-trained models and construct self-learning algorithms.

- **Hardware:** High-performance GPUs enabled efficient inference and real-time learning.

Natural Language Processing Techniques

- **Tokenization :** It is the division of text into manageable parts for processing.
- **Semantic Analysis:** Understanding the meaning and context of questions in order to create appropriate responses.
- **Attention Mechanisms:** Using transformer architectures' self-attention capabilities to capture connection patterns in data.

Pre-trained Language Models for QA

The following pre-trained LLMs were chosen for their excellent performance in language interpretation and generating tasks:

- **BERT**: It excels in contextual comprehension thanks to bidirectional encoding [18].
- **T5**: Uses a text-to-text framework that is appropriate for a variety of NLP activities, including QA [17].
- **GPT-2:** It is well-known for its ability to generate and produce cohesive text [16].
- **LLAMA 2**: Known for multimodal capabilities, handling both text and images [14].
- **Mistral**: Optimized for efficiency and performance in resource-constrained environments [15].

These models were utilized in their pre-trained form, with no extra fine-tuning. Their natural skills were harnessed and strengthened via self-learning techniques.

Self-Learning Mechanisms

- Reinforcement Learning

Objective: Allow the model to learn optimal replies by maximizing the incentives associated with accurate answers.
Method: Policies were implemented that allow the model to get feedback on its performance and alter its future actions to enhance results.
Application: The model learns from interactions and adjusts its responses based on the rewards or penalties received.

- The Continuous Learning

Approach: It allows the model to update its knowledge base progressively as new information is provided.
Techniques: Used online learning algorithms to adapt to changes without having to start from scratch.

Benefit: Ensures that the model remains current and correct over time, managing changing data and user requests efficiently.

User Interaction and Feedback Integration

Feedback Loops: Received real-time feedback from users on the relevancy and correctness of the offered responses.

Active Learning: The model recognizes areas of ambiguity and actively solicits further information or clarification from users.

Outcome: The model's performance improved by learning directly from user interactions, resulting in more customized and accurate replies.

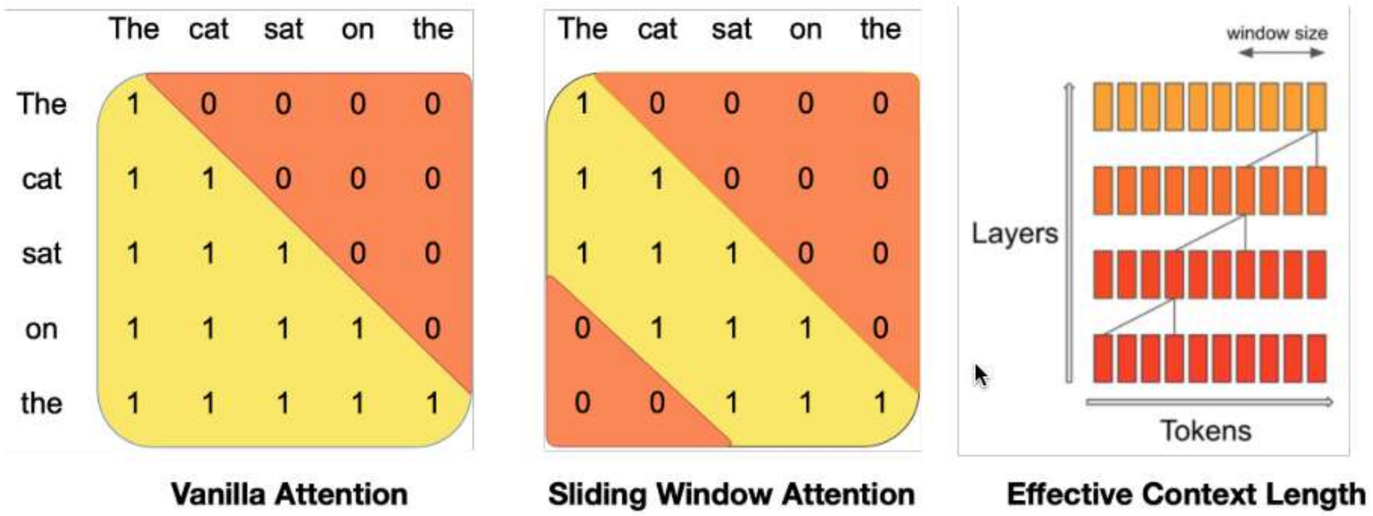
By combining pre-trained LLMs with self-learning processes and user input, this study created an improved QA system that can improve over time without the need for fine-tuning. The self-learning technique allowed the model to constantly adjust to new inputs and user preferences, increasing accuracy and pleasure. This research highlights the possibility for using self-learning in LLMs to improve QA systems and adds useful insights to the field of natural language processing.

Table 1. Database Information

Dataset	**Description**	**Total Number of Examples**	**Total QA Pairs**	**Total Answers**	**Notes**
CoQA [13]	Conversational Question Answering dataset consisting of dialogues.	127,000 dialogues	~1.4 million QA pairs	~1.4 million answers	Each dialogue contains multiple QA pairs.
Natural Questions [23]	Dataset with long passages and corresponding questions for extracting answers.	307,000 questions	307,000 QA pairs	307,000 answers	Each question is associated with one answer extracted from a passage.
Commonsense QA [8]	Dataset with multiple-choice questions that test commonsense reasoning.	12,000 questions	12,000 QA pairs	12,000 answers	Each question has 5 possible answer choices.

Mistral Model

The Mistral model is a cutting-edge Large Language Model (LLM) created by Mistral AI and released in September 2023. The flagship model, Mistral 7B, has 7 billion parameters and is intended to provide excellent performance while being efficient and accessible [15].



	The	cat	sat	on	the
The	1	0	0	0	0
cat	1	1	0	0	0
sat	1	1	1	0	0
on	1	1	1	1	0
the	1	1	1	1	1

Vanilla Attention

The	cat	sat	on	the
1	0	0	0	0
1	1	0	0	0
1	1	1	0	0
0	1	1	1	0
0	0	1	1	1

Sliding Window Attention

Effective Context Length

Figure 1. Mistral Model Architecture

Key features:

Despite being small as compared to larger models such as GPT-3 (175 billion parameters), Mistral 7B performs well on a variety of natural language processing (NLP) benchmarks. This efficiency makes it appropriate for applications that require strong performance but do not require a lot of processing resources.

Mistral uses an enhanced Transformer architecture with improvements such as efficient attention mechanisms and better training methodologies. These enhancements allow the model to efficiently perform linguistic tasks while decreasing computational cost [15].

Mistral 7B is open-source and released under a permissive license, encouraging widespread use and cooperation. Researchers and developers can use the model in their projects, alter it, and contribute to its development without facing substantial license constraints.

Adaptability and Flexibility: The model is intended to be easily adaptable to a variety of jobs with minimal fine-tuning. Its design enables smooth interaction with other NLP applications, including question-and-answer systems[15].

IJNRD Research Through Innovation

LLaMA 2

LLaMA 2 is Meta AI's sophisticated open-source large language model, which was launched in July 2023. Building on its predecessor, LLaMA 2 provides considerable improvements in scalability, speed, and accessibility, making it an excellent option for incorporation into question answering (QA) systems [14].

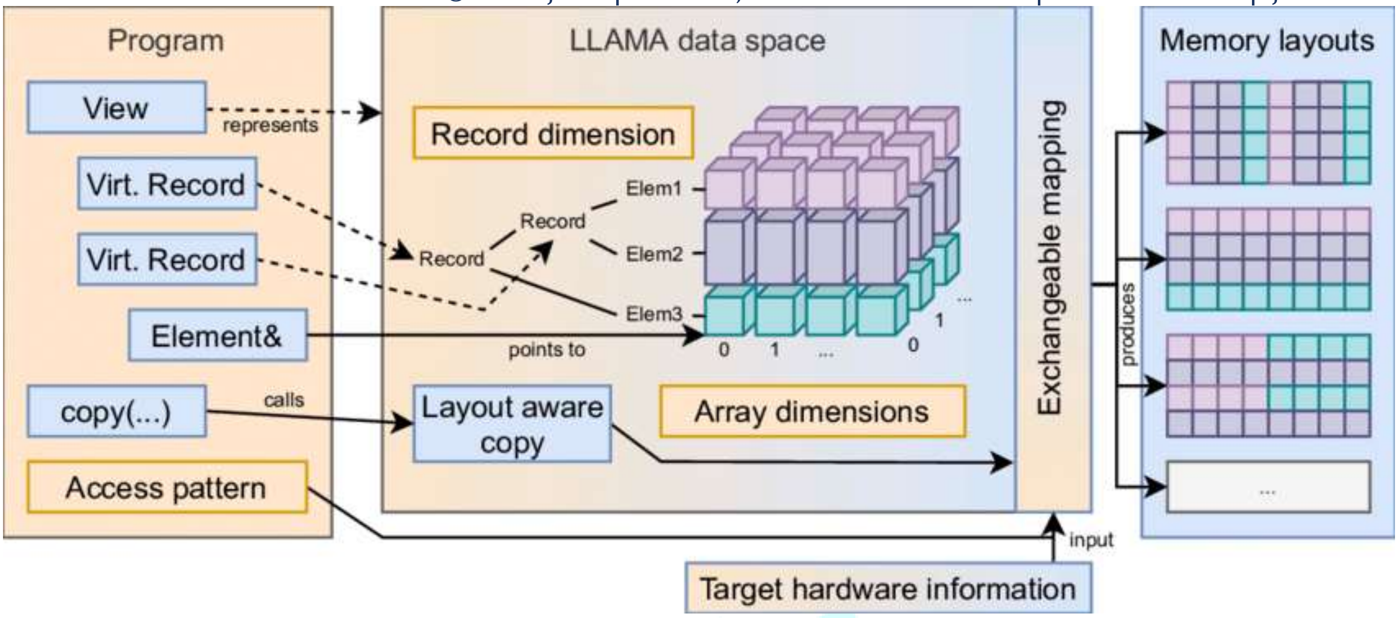


Figure 2. LLaMA 2 Model Architecture

Key features:

Parameters are available in three configurations: 7 billion (7B), 13 billion (13B), and 70 billion (70B).

Flexibility: Provides alternatives for balancing computational resource requirements and performance needs.

Efficiency Enhancements: Includes architectural improvements including Grouped Query Attention (GQA) and SwiGLU activation functions.

Enhanced Performance: These enhancements result in shorter inference times and greater usage of hardware resources while maintaining accuracy [14].

Extensive pretraining

Training Data: Trained on 2 trillion tokens of publicly available data, such as web pages, books, and code.
Broad Knowledge Base: The large dataset enables LLaMA 2 to perform effectively across a variety of linguistic tasks.

Fine-Tuned Variants for Dialogue:

LLaMA2-Chat: A conversational-specific version that improves the capacity to provide coherent and contextually relevant replies in discussion scenarios.
Safety Features: Contains mechanisms to minimize the creation of improper or hazardous material. Open Source License:
Accessibility: Released under a permissive community license, which permits research and commercial use with due credit.
Collaboration Encouraged: Promotes community participation and openness in model development [14].

IV. RESULTS AND DISCUSSION

Proposed Flow Architecture

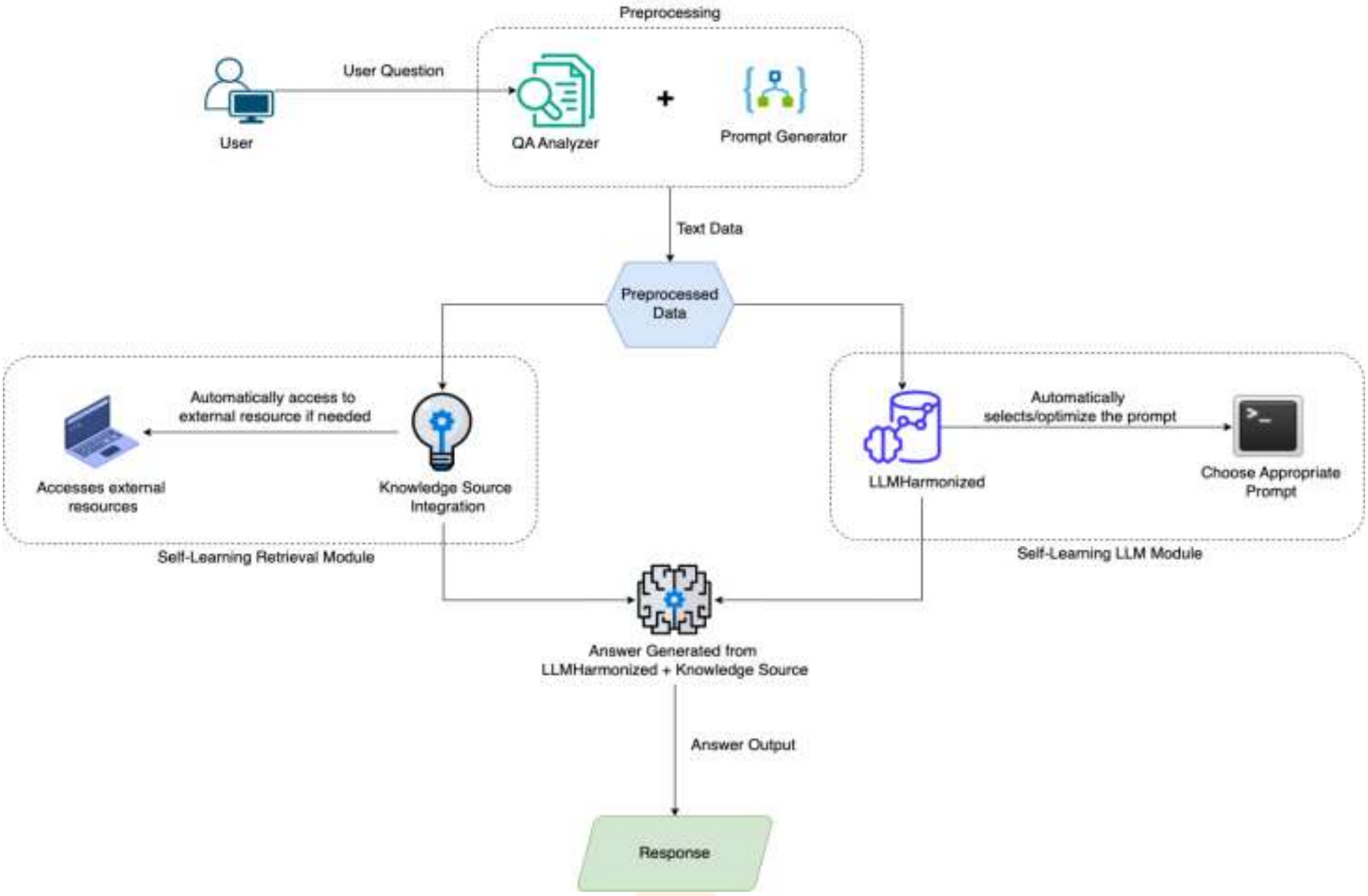

Preprocessing
User Question
User
QA Analyzer
+
Prompt Generator
Text Data
Preprocessed Data
Automatically access to external resource if needed
Accesses external resources
Knowledge Source Integration
Self-Learning Retrieval Module
Automatically selects/optimize the prompt
LLMHarmonized
Choose Appropriate Prompt
Self-Learning LLM Module
Answer Generated from LLMHarmonized + Knowledge Source
Answer Output
Response

Figure 3. Proposed Flow Architecture

User Input: The system receives unstructured natural language inquiries from the user and initiates the response process.

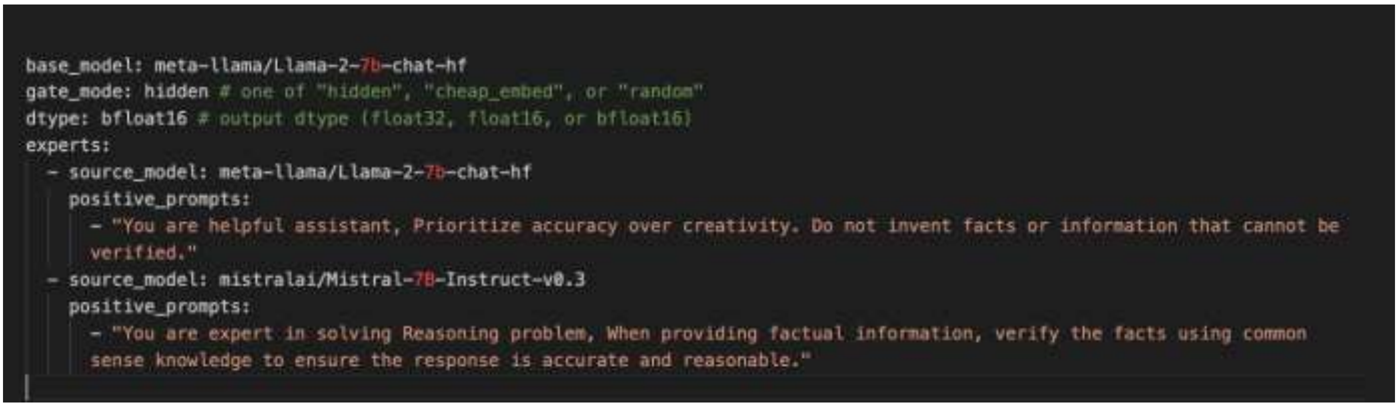
Preprocessing: Analyzes and converts the user's query into an appropriate format, detecting intent and directing it for optimal processing.

Few-Shot Examples: Provides the LLM with a short selection of relevant sample questions and replies to help it generate correct and context-aware responses.

The Retrieval Module searches and obtains pertinent information from knowledge bases to improve the LLM's capacity to provide accurate and educated responses.

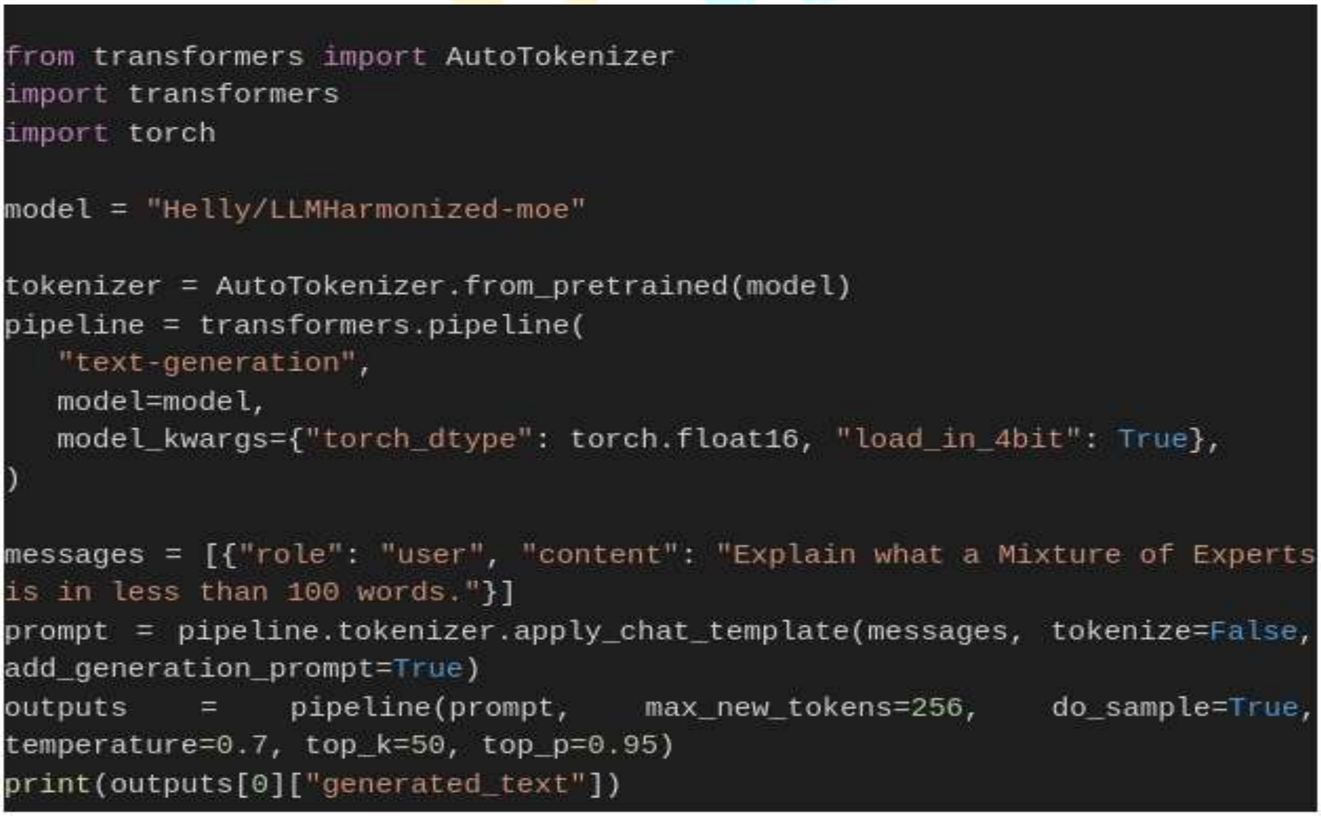
The LLM module generates the final, coherent answer by processing the preprocessed input and adding any recovered information as needed.

Implementation of the LLMHarmonized Model



```
base_model: meta-llama/Llama-2-7b-chat-hf
gate_mode: hidden # one of "hidden", "cheap_embed", or "random"
dtype: bfloat16 # output dtype (float32, float16, or bfloat16)
experts:
  - source_model: meta-llama/Llama-2-7b-chat-hf
    positive_prompts:
      - "You are helpful assistant, Prioritize accuracy over creativity. Do not invent facts or information that cannot be
        verified."
  - source_model: mistralai/Mistral-7B-Instruct-v0.3
    positive_prompts:
      - "You are expert in solving Reasoning problem, When providing factual information, verify the facts using common
        sense knowledge to ensure the response is accurate and reasonable."
```

Figure 4. Merge Implementation



```
from transformers import AutoTokenizer
import transformers
import torch

model = "Helly/LLMHarmonized-moe"

tokenizer = AutoTokenizer.from_pretrained(model)
pipeline = transformers.pipeline(
   "text-generation",
   model=model,
   model_kwargs={"torch_dtype": torch.float16, "load_in_4bit": True},
)

messages = [{"role": "user", "content": "Explain what a Mixture of Experts
is in less than 100 words."}]
prompt  = pipeline.tokenizer.apply_chat_template(messages, tokenize=False,
add_generation_prompt=True)
outputs    =    pipeline(prompt,    max_new_tokens=256,    do_sample=True,
temperature=0.7, top_k=50, top_p=0.95)
print(outputs[0]["generated_text"])
```

Figure 5. LLM Harmonized Implementation

Research Through Innovation

Input:

```
Explain what a Mixture of Experts is in less than 100 words.
```

Output:

```
A Mixture of Experts (MoE) is a machine learning technique that combines multiple independent experts' predictions to generate a more accurate forecast for a target variable. Each expert has their own model or algorithm that makes predictions based on different features or inputs, and the combined predictions are combined using a weighted average or other fusion method to produce the final output. The idea is to leverage the diversity of opinions and models from the experts to improve the overall forecasting accuracy.
```

Figure 6. Input and Output through proposed model

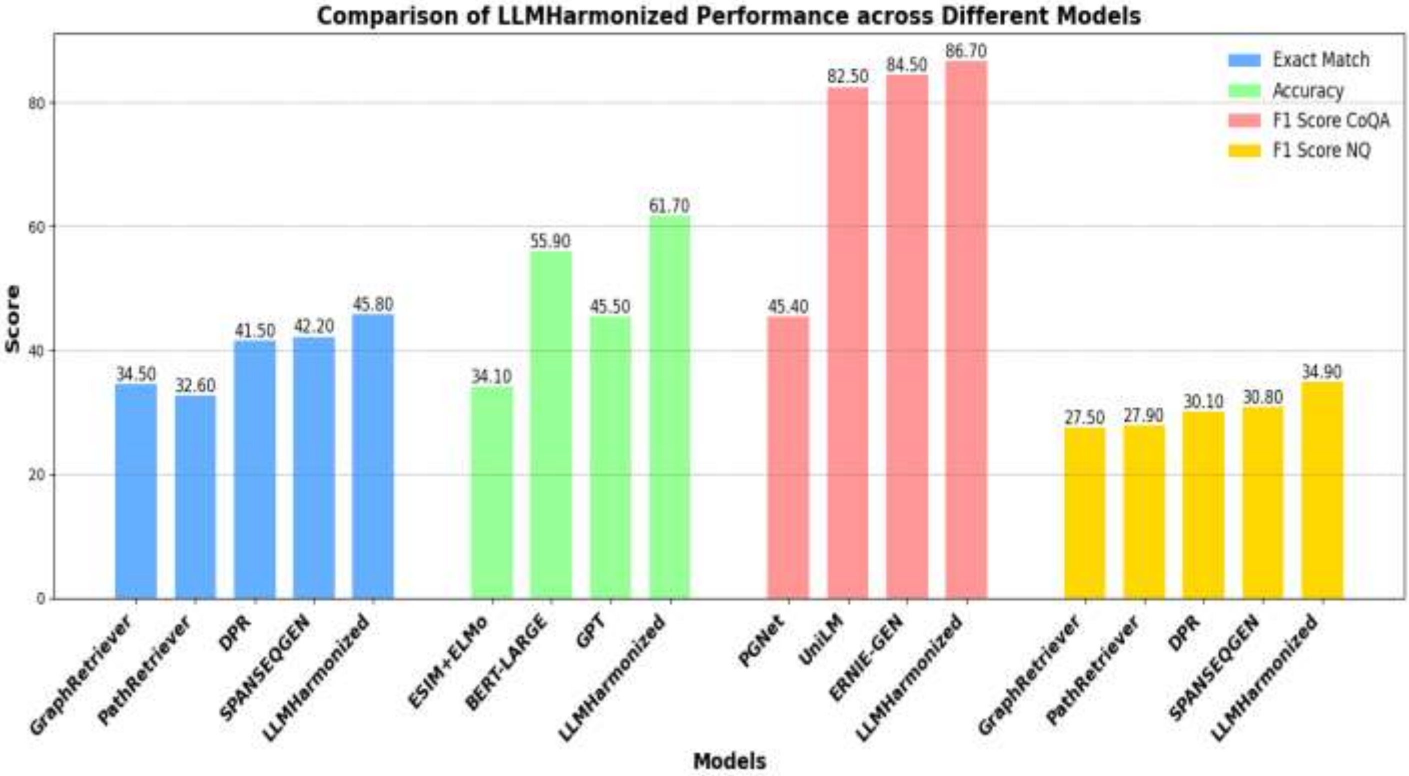


Figure 7. Result Comparison

RESULT SUMMARY

The study demonstrates notable improvements across various evaluation metrics and datasets for all three proposed versions using LLMHarmonized. Specifically, the model achieved an F1-score of 86.7, surpassing the single-model F1-score of 84.5. This indicates that the LLMHarmonized hybrid approach enhances both precision and response coverage on the CoQA dataset. The improvement reflects the model's ability to balance precision and recall, which is vital for maintaining relevance and coherence across different conversational turns.

In open-domain question-answering, the LLMHarmonized model performed well, achieving an F1-score of 34.9 and an Exact Match (EM) score of 45.8 on the NQ-OPEN dataset. This demonstrates the model's capability to resolve contextual ambiguities a typically challenging task and to provide answers even when responses are not perfectly precise. The increase in these metrics signifies a breakthrough in handling various linguistic phenomena, underscoring the model's potential to advance conversational AI technologies.

Furthermore, substantial performance differences were observed in commonsense understanding evaluations. On the COMMONSENSEQA dataset, the BERT-LARGE model achieved an accuracy of 55.9%, whereas the LLMHarmonized model significantly outperformed it with an accuracy of 61.7%. This considerable gap reveals that LLMHarmonized excels in reasoning over commonsense knowledge and context, enhancing its reliability in answering nuanced questions.

The study also highlights the importance of limiting hallucinations in generated responses. The LLMHarmonized model incorporates strict prompting and fact-checking mechanisms to prevent the creation of false or misleading information. Adopting this proactive strategy improves the model's reliability and provides more trustworthy outputs, which is especially critical in sensitive areas such as healthcare, education, and customer support.

Overall, the results across these benchmarks indicate significant benefits of the proposed LLMHarmonized model in reducing hallucinations, enhancing commonsense reasoning, disambiguation, and contextual understanding. These advancements make the model more reliable and effective for real-world question-answering systems and conversational agents.

REFERENCES

[1] Caballero, M. (2021). A brief survey of question answering systems. International Journal of Artificial Intelligence & Applications (IJAIA), 12(5).

[2] Stoilova, E. (2021). AI chatbots as a customer service and support tool. ROBONOMICS: The Journal of the Automated Economy, 2, 21.

[3] AbuShawar, B., & Atwell, E. (2015). ALICE chatbot: Trials and outputs. *Computación y Sistemas*, *19*(4), 625-632.

[4] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems.

[6] Min, S., Michael, J., Hajishirzi, H., & Zettlemoyer, L. (2020). AmbigQA: Answering ambiguous open-domain questions. arXiv preprint arXiv:2004.10645.

[7] Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., ... & Bowman, S. R. (2021). BBQ: A hand-built bias benchmark for question answering. arXiv preprint arXiv:2110.08193.

[8] Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2018). Commonsenseqa: A question answering challenge targeting commonsense knowledge. arXiv preprint arXiv:1811.00937.

[9] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[10] Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

[11] Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.

[12] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).

[13] Reddy, S., Chen, D., & Manning, C. D. (2019). *CoQA: A Conversational Question Answering Challenge*. Transactions of the Association for Computational Linguistics, 7, 249-266.

[14] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[15] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

[16] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[17] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), pp.5485-5551.

[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019b). ROBERTA: A robustly optimized BERT pretraining approach. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1907.11692.

[19] Raval, H. (2020). Limitations of existing chatbot with analytical survey to enhance the functionality using emerging technology. International Journal of Research and Analytical Reviews (IJRAR), 7(2).

[20] Raval, H. Y., Parikh, S. M., & Patel, H. R. (2022). Self-maintained health surveillance artificial intelligence assistant. In Handbook of Research on Lifestyle Sustainability and Management Solutions Using AI, Big Data Analytics, and Visualization (pp. 168-184).

[21] Goswami, S. A., Patel, K. D., Raval, H. Y., & Parikh, S. M. (2022, June). Taxonomy and Implications of Machine Learning for Internet of Things: Qualities, Uses and Algorithms. In International Conference on Signal & Data Processing (pp. 167-182). Singapore: Springer Nature Singapore.

[22] Raval, H., & Parikh, S. M. (2024). Benchmarking Coqa: A Study of Leading Question Answering Models.

[23] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... & Petrov, S. (2019). Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453-466.

International Research Journal IJNRD Research Through Innovation