FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Computer Science Research (IJCSR)
------- Article | Volume 3, Issue 3 | Pages 99-104 | ISSN XXXX-XXXX

AI-Driven Multilingual Document Analysis and Interaction System

Mohan B A[1*], Basavaraj G N[2], Karthik S A[3] and Rakesh N[4]
[1,2,4]Department of Information Science, BMS Institute of Technology and Management, Bengaluru, India
[3]Department of Computer Science & Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

***Correspondence:** *Mohan B A; ba.mohan@bmsit.in*

ABSTRACT- This research paper introduces a pioneering application that integrates the capabilities of Artificial Intelligence (AI) with document-based interactions. Users can effortlessly upload their preferred documents to the system, enabling AI analysis of the document's content. Following this analysis, users can pose questions related to the document's content. The application harnesses natural language understanding and AI-driven processing to provide comprehensive and insightful answers, effectively transforming static documents into dynamic repositories of interactive knowledge. By seamlessly integrating document processing and natural language comprehension, this paper aims to redefine and elevate the user experience of interacting with information. Whether utilized for research, learning endeavors, or exploration, this application stands out to enhance user engagement and, facilitating smarter and more interactive interactions with documents.

Keywords: LLM Models, NLP Integration, Document Analysis, AI-driven Interaction, Knowledge Extraction.

ARTICLE INFORMATION
Author(s): Mohan B A, Basavaraj G N, Karthik S A and Rakesh N;
Received: 10/07/2024; **Accepted**: 20/08/2024; **Published**: 15/09/2024;
e-ISSN: XXXX-XXXX;
Paper Id: IJCSR-030302;
Citation: 10.37391/IJCSR.030302

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Conventional document processing methods face significant challenges in processing tremendous amount of information, hidden in documents, extracting meaningful insights in the documents in an efficient manner. Shortcomings of these traditional methods, mainly due to an inability to manage the enormity and complexity of unstructured data, emphasizing the need for more advanced methodologies that can discover the information that is encapsulated within the documents. AI has caused a great revolution in the paradigms shift in document analysis and comprehension. The convergence of AI technologies—more particularly, Machine Learning (ML) and Natural Language Processing (NLP) with advanced document analysis techniques creates an opportunity for transformation the existing techniques. By harnessing capabilities of AI, the proposed work taps into ML algorithms and NLP models to help navigate the complex insights in unstructured textual data. The main goal here is to give users a more engaging experience than simply reading documents. The ability of ML-powered document analysis and NLP-enabled language understanding to convert the content of documents into easily grasped insights is sought in the proposed work. The transformation is focused on enabling smooth interactions, where the user can query document repositories in natural language and extract relevant information.

1.1 Comparative Analysis of Large Language Models for Structured and Unstructured Data in QA Systems

In smart Question and Answer (QA) systems, various large language models play a critical role in handling structured and unstructured data depending on their strengths and approaches. *Table-1* shows different models with the capabilities and applications. For structured data, models like SQLNet and TREQS, have shown proficiency. The SQLNet leverages deep learning approach to generate SQL queries directly from natural language, specifically catering to structured databases. It applies a sequence-to-sequence model to interpret user queries and translates them into executable SQL statements. This model is highly effective in environment where it is dealing with structured data. In contrast, for unstructured data, models such as BERT and its derivatives, like RoBERTa and GPT-3, have revolutionized text processing. BERT is trained on a huge amount of unstructured text, and so, by definition, it has a good understanding of the nuances of natural language; it does well in answering questions based on paragraphs, email, or articles. The strength of the models against each other in this case depends a lot on the data they process. For structured databases and generation of specific queries, SQLNet-like models with their specialized architectures are more suitable. On the other hand, BERT and GPT-3 give more flexibility and wider applicability for handling diversity in the formats of unstructured data at an additional cost in computational resources. In terms of performance metrics, models such as SQLNet have shown high accuracy in structured query language generation tasks, often surpassing 90% in benchmark datasets such as the Spider dataset. However, in unstructured environments, BERT-based models have achieved state-of-the-art results in NLP tasks, including a notable score of 84.5% in the SQuAD 2.0 Question Answering task. Different Large Language Models, each with their unique strengths, highlight

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
------- Article | Volume 3, Issue 3 | Pages 99-104 | ISSN XXXX-XXXX

the significance of selecting a suitable model depending on the data and precise needs of the QA system being created [1][2].

Table 1. Comparison of Large Language Models for QA

Model Name	Type of Data	Key Capabilities	Performance Metrics	Application Examples	Computational Resources Required
SQLNet	Structured	SQL Query Generation from Natural Language	Accuracy in SQL Query Generation (eg. >90 on Spider dataset)	Database querying, Business data analysis	Moderate
BERT	Unstructured	Contextual understanding of Text, Answer Extraction	F1 Score and Exact Match in SQuAD 2.0 (eg. 84.5% F1 Score)	Content summarization, Customer service automation	High
GPT-3	Unstructured	Text Generation Contextual Interface	Performance in Open-Domain Question Answering	Creative writing, Chatbots, Knowledge discovery	Very High
TREQS	Structured	Natural Language Interface to Databases	Precision and Recall in Structured Data Interpretation	Data analytics, Reporting tools	Moderate
RoBERTa	Unstructured	Improved text Understanding and Predictions	Enhanced Results in GLUE and SuperGLUE benchmarks	Sentiment analysis, Language translation	High

1.2 Comparison of Storage Techniques for Structured and Unstructured Data

In data storage domain, the selection of the appropriate technology depends on the nature of the data, weather it is structured or unstructured data. The *table 2* shows different storage techniques for both structured and unstructured data. The Relational databases such as MySQL and PostgreSQL have been the cornerstone of structured data storage, offering robust transactional support and efficient querying through SQL. Their table-based organization and ACID compliance make them ideal for applications that demand high data integrity, such as financial systems and customer record management. However, they often face scalability limitations when handling massive volumes of data or rapidly evolving schemas.

On the other hand, NoSQL databases, such as MongoDB and Cassandra, come in handy when dealing with unstructured or semi-structured data. Their schema-less nature makes them more flexible and capable of handling different data types without requiring a predefined structure for the same. These databases are scalable and hence appropriate for applications involving large-scale data storage, real-time analytics, and big data applications. Though providing flexibility and scalability, they often compromise on transactional consistency offered by relational databases.

Each of these storage methods has its different merits and best-fit scenarios. Choices among them depend on data types, scalability requirements, performance needs, and application requirements. As data grows in volume, variety, and velocity, the choice of appropriate storage solution becomes very critical in the architecture of solid and efficient data management systems.

Table 2. Comparison of Storage Techniques for Structured and Unstructured Data

Storage Technique	Data Type	Key Features	Scalability	Performance	Use Cases
Relational Databases (eg., MySql, PostgreSQL)	Structured	Table-based storage, ACID compliance, SQL querying	Moderate	High for structured queries	Financial systems, Customer records, Inventory management
NoSQL Databases (e.g., MongoDB, Cassandra)	Both (Primarily Unstructured)	Schema-less design, Horizontal scaling, Flexible data models	High	Varies; generally high for unstructured data	Big data applications, Real-time analytics, Content management
Data Warehouses (e.g., Amazon Redshift, snowflake)	Structured	Optimized for complex queries, High throughput for read operations	High	High for analytical queries	Business intelligence, Reporting, Data analysis
Object Storage (e.g., Amazon S3, Google Cloud Storage)	Unstructured	Highly scalable, Data stored as objects, API access	Very High	Moderate; optimized for durability and availability	Multimedia storage, Backup and archival, Big data storage
File Storage (e.g., Network Attached Storage)	Both	Hierarchical file system, Network based access	Moderate	High for specific file retrieval	Shared file systems, Document management, Media libraries
Graph Databases (e.g., Neo4, Amazon Neptune)	Structured (Relationship-focused)	Relationship mapping, Graph-based querying	Moderate	High for relationship-driven queries	Social Networks, Recommendation engines, Fraud detection
Document Databases (e.g., Elasticsearch, CouchDB)	Unstructured	Document-oriented, Full-text search capability	High	High for search-driven queries	Search engines, Logging systems, Content retrieval

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
------- Article | Volume 3, Issue 3 | Pages 99-104 | ISSN XXXX-XXXX

1.3 Components for Implementing a Multilingual System in QA

Developing a multilingual system in the QA framework is complex, and it involves different components as discussed in the Table-3. The starting point of this development process is language detection. The system would be able to quickly and precisely identify the language of the query with libraries such as FastText or Langid. These are followed by the translation service, critical to translating the query into a base language—like English—which the system can understand. These services call to retranslate the system's response back into the user's native language, thereby providing the user with a seamless experience. However, translation quality and latency remain the critical factor to keep effectiveness and efficiency in the system. NLP is the soul of a multilingual QA system. In most cases, multilingual NLP models like BERT Multilingual and XLM-R are pre-trained across a wide range of languages to capture the better understanding and text processing in multiple languages. These models are very language-agnostic in nature and provide quite a good range of language coverage, capturing contextual tones across different languages easily. Besides, cross-lingual transfer learning techniques can be used with these models to extend their capabilities further if limited amounts of training data are available for some languages. This approach borrows knowledge transferred from data-rich languages to help improve performance on less-represented languages [3][4][5].

This could also draw upon localized knowledge sources to then provide relevant and accurate answers, especially to region-specific questions. It would make use of regional databases and local news APIs offering up-to-date and relevant information tailored to specific language and cultures.

Finally, the QA systems should be designed with a user interface that supports multiple languages. This includes not only translation, but also support for Unicode to represent a variety of scripts, and considerations for right-to-left text in languages such as Arabic and Hebrew. More importantly, considerations are due for accessibility and usability of the interface to ensure that people across different language backgrounds could use the system easily.

This ensures that the multilingual QA system is well-integrated with different sub-components dealing with various aspects of language handling to provide a smooth and accurate experience to the user worldwide [6][7].

Table 3. Components for implementing a Multilingual System in QA

Component	Purpose	Key Technologies	Considerations
Language Detection	Identifying the language of the input query	Libraries like FastText, Langid	Accuracy, Speed
Translation Services	Translating queries and responses between languages	Google Translate API, Microsoft Translator	Quality of translation, Latency
Multilingual NLP Models	Processing and understanding queries in multiple languages	BERT Multilingual, XLM-R	Coverage of languages, Contextual understanding
Cross-Lingual Transfer Learning	Leveraging knowledge from one language to improve performance in another	Zero-shot learning techniques	Data availability, Model robustness
Localized Knowledge Sources	Providing accurate information for language-specific queries	Regional databases, Local news APIs	Relevance, Currency of information
User Interface	Offering multilingual support in user interactions	Unicode support, RTL text support	Accessibility, Ease of use

2. LITERATURE SURVEY

Authors Devlin et al. introduced Bidirectional Encoder Representations from Transformers, which was quite a revolutionary approach in NLP to take a step toward a paradigm shift in models for language representation. Unlike all the previous methods, BERT is based on the notion of pre-training deep bidirectional representations from unlabelled text, fundamentally changing how models understand the language context. The innovation here is that BERT conditions both the left and right contexts jointly in all its layers, a notable divergence from unidirectional or shallow bidirectional predecessors. Its strength does not lie merely in the conceptual simplicity; empirical evidence comes from good performance across a wide range of NLP tasks. Remarkably, BERT established new state-of-the-art results on many benchmarks, such as increasing the GLUE score to 80.5% (7.7-point improvement), and significant advances in MultiNLI accuracy as well as SQuAD question-answering metrics. Thus, the high flexibility of BERT for the maximum number of significantly differing tasks marks, indeed, very progressive development in this field. This breakthrough model in understanding and processing natural language sets a new standard for related question tasks and inference language, catalysing additional research and development in the field of AI-driven language processing[8].

In pursuit of improving natural language understanding (NLU), the most pioneering approach in that domain came in 2018 when handling traditional discriminatively trained models. In the course of their work in Improving Language Understanding by Generative Pre-Training, the authors managed to consider the inborn challenges of NLU tasks, such as textual entailment and question answering, as well as document classification under the constraint of scarce labelled data. The aim of their approach is generative pretraining of a language model on a large and diverse corpus of unlabelled text, followed by

discriminative fine-tuning on specific NLU tasks. In contrast, theirs is task-agnostic and actually beats the baseline discriminatively trained models for specific tasks in the order of nine out of twelve tasks chosen for the study. They report impressive deltas, of up to 8.9% for increasing commonsense reasoning and up to 5.7% for question-answering accuracy. Crucially to this approach, they use task-aware input transformation in the fine-tuning, which, in their words, enabled transfer learning to work with minimal architectural changes. it also set a new standard in the AI language processing domain by not only outpacing in the NLU benchmarks but also proving the efficacy of generative pretraining[9].

The vast domain of LLMs has been covered in its entirety in the paper titled "A comprehensive overview of Large Language Models", authored by Humza Naveed et al. This survey paper is a significant document for anyone interested in fast-evolving and multi-faceted LLMs in NLP. In this work, the authors discuss new advances related to the study of LLMs that include architectural innovations or training strategies, increasing context length, methodologies for fine-tuning, and multimodal LLMs. Equally important, the article discusses new applications for LLMs in the field of robotics—how LLMs combine with systems of robotic machines. The review also discusses some pointers on how to evaluate datasets, benchmark methods, and model efficiency. This work is invaluable in its comprehensive approach and gives an overview that is complete and self-contained. It will be useful not only as a structured survey but also as a reference for researchers and a resource for practitioners. The paper by Naveed et al. stitches together the rich and informative features of the current LLM research in a clear and incisive way on frontier issues. It helps in understanding complexities and guides future progress in the rapidly developing field[10].

A documentation of evolution and efficacy regarding deep learning in text classification has been presented in the paper "Deep Learning-Based Text Classification: A Comprehensive Review" by Shervin Minaee and colleagues. This comprehensive review has been a landmarker in this field, wherein the authors have gone into great depths to dissect more than 150 deep learning-based models which shaped and brought improvement to the task of text classification. A wide range of applications, from sentiment analysis, news categorization, question answering, and natural language inference, in most cases show that deep learning models have a better performance when compared against traditional machine learning approaches. One very valuable contribution of this work has been the comparative analysis of these models with regard to their technical contributions, similarity, and strength. This involves views on more than 40 popular datasets that are widely used in text classification research. It would be provided with a quantitative analysis for performance on popular benchmarks. That allows one to bring a critical viewpoint about the different deep learning strategies in text classification. Minaee et al.'s paper can thus be said to be a thorough guide, for both researchers and practitioners, through current achievements and an indication of future research directions to present comprehensive understanding and critical evaluation of the current state-of-the-art with respect to text classification of deep learning-based methods[11].

Of all the recent developments on sequence-to-sequence models, one worthy contribution is the work by Lewis et al. on "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". It contributes a new de-noising auto-encoder to pre-train sequence-to-sequence models, which has profound implications on natural language generation, translation, and comprehension. What is most innovative about BART is its training methodology—that of corrupting text with a host of noising functions and then reconstructing original text. This is a vanilla machine translation neural architecture, based on a transformer, combining the power of a BERT-like bidirectional encoder and a left-to-right decoder from GPT. This strategy enables the BART to generalize tasks along a spectrum of pre-training schemes, thus a powerful augmentation to the NLP toolkit at large. Some research has gone into the potential of Morse codes for communication, and it is majorly the case of disabled persons whose mode of expression involves eye blinking. This study focuses on using Morse code for communication, citing concerns over its usage by the target users. Use of Morse code depicts innovative strategies in assistive communication technologies that raise concerns over usability and user memorization[12].

AI and NLP in PDF-based Question Answering Systems: The smart question and answering system integrates BIM with AIOT through a BERT model. As previously stated, the said system includes three subsystems: NLP model training, the inference server, and the user interface. In the case of the former, the BERT model was trained with a text dataset containing messages on the BIM resource. On the other hand, the server communicated with the user interface for the processing of the query and the response. This system enables human-machine communication and enhances user knowledge of smart buildings, particularly for practitioners and researchers in the construction industry. By leveraging the NLP and BERT models, the system aims to facilitate accurate and efficient information retrieval related to BIM-AIOT integration[13].

Representative Vector Summarization for Large Unstructured Textual Data: The introduced RAG-assisted Representative Vector Summarization (RVS) addresses challenges related to performance decrease with increasing context lengths and the "Lost in the Middle" problem in large context window models. RVS selects a predefined number of representative text chunks from a non-parametric knowledgebase and applies a combined abstractive and extractive summarizing workflow to generate the final summary. The parameter value for selecting text chunks was determined based on the maximum affordable token limit. In addition, the model can identify keywords and their relative importance, creating visual representations of the document's content using word clouds and scatter plots. RVS is implemented in docGPT, a document intelligence program written in Python using the langchain framework[14].

AI-based Question Answering System using Particle Swarm Optimization and Fuzzy C-Means Clustering Algorithms: The

International Journal of Computer Science Research (IJCSR)
------- Article | Volume 3, Issue 3 | Pages 99-104 | ISSN XXXX-XXXX

smart QA system concerning power business scenarios classifies the question attributes through the particle swarm optimization algorithm and the fuzzy c-means clustering algorithm and matches the answers to the inquiries of the customer. The system collected questions raised by customers and uploaded them within the knowledge base. The pre-processed questions were classified using a particle swarm clustering algorithm. In the implementation, the fuzzy c-means clustering algorithm was used to ensure that the students got answers with maximum possibilities to their questions. It had low worst-case time complexity, which is up to 0.35 only. This approach proposed is AI-based and combines two algorithms to provide efficient and effective customer service. Its novelty lies in using priority information to address customer requests. The design of the system is focused on the improvement of customer service quality and operational efficiency in business scenarios related to power[15].

Fast and Memory-Efficient Attention Algorithm for Transformers: FlashAttention represents a novel attention algorithm aimed at enhancing the speed and memory efficiency of transformers. Its design involves employing hashing and low-rank approximation methods to estimate an attention matrix. This unique approach enables scalability in a linear manner with respect to the sequence length, which is a significant advantage for processing long sequences. Additionally, FlashAttention introduces "flash updates," enabling the rapid updating of the attention matrix in constant time, distinguishing it as faster than other approximate attention methods, such as Reformer and Performer. Compared with various transformer-based models, FlashAttention holds promise for enhancing the performance across a spectrum of tasks in natural language processing and computer vision[16].

Methods for Integrating Unstructured Data with Relational Databases: It extracts structured column values from unstructured sources and maps these to known database entities. This will then help in using statistical models for co-reference resolution and information extraction to identify entities stored in a database. The approach relies heavily on indexed access both to the database and to the unstructured world. Challenges range from the need for automated and domain-independent database-driven solutions to building bridges that will allow seamless querying. Indexing techniques have been extended for the similarity predicates needed during data integration. Integrating this, mapping is done from the unstructured webpages to rows and columns of existing databases. Of course, this kind of integration would easily broaden the potential for a far richer query interface than possible through traditional file-based processing of unstructured data[17].

3. PROPOSED SYSTEM

3.1 Structured Data Handling

Conversion to Markup Language: This involves the conversion of structured data, like databases and spreadsheets, into a markup language such as XML or JSON. This way, it is easier to parse and subsequently process by the system. Models of Structured Data Usage: Use the SQLNet model for generating queries. This is a model to be used for the conversion of natural language queries to SQL commands that are appropriate to extract information in markup format structured data. Integration of NLP techniques: NLP techniques are used to understanding of the context and relationships within the structured data, improving query results accuracy.

3.2 Unstructured Data Handling

NLP Models for Textual Data: Advanced NLP models such as BERT and GPT-3 are applied for unstructured textual data. These models comprehend the context and are able to respond coherently. It would imply implementation of the Optical Character Recognition technology to extract text from images, and then such text would be processed through an NLP model for relevant answer generation. Multilingual text handling: Develop multilingual versions of NLP models, for example, multilingual BERT or XML-R, to handle text in different languages. This will ensure that the system applies under different linguistic contexts.

3.3 Multilingual Support

Language Detection and Translation: Develop algorithms to identify the language that has been inputted. Translate the query to the base language and translate the response back to the native language of the user. This can be further improved by use of cross-lingual transfer learning, in languages with limited training data available, to develop better cross-lingual understanding.

3.4 User Interface

Interactive and multilingual interfaces: Ease of use by all users with different languages will be assured by designing a user-friendly interface with provision for use in multiple languages.

3.5 Feedback Mechanism

Feedback System: This allows users to rate answers and give suggestions for system improvement.

4. METHDOLOGY

The flow of data is depicted as shown in the *figure 1*. The structured/unstructured data is fed into the system, which starts with text splitter module converting data into chunks. This chunks are analysed to capture nuance, connections, and semantic meanings between the chunks and are stored in Vector DB. User queries are fed to sentence transformers which understands the query and context to answer it by referring the relevant docs in the Vector DB.

5. RESULTS

Vector databases have emerged as a critical component for managing high-dimensional data efficiently, particularly in applications such as QA with large language models.

The comparison of popular vector database systems along with shedding light on how they impact AI applications are discussed: The vector databases like Faiss, ChromaDB, Odrant (local mode), and PgVector are compared. Vector databases differ in how they manage their data on disks, leading to

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
------- Article | Volume 3, Issue 3 | Pages 99-104 | ISSN XXXX-XXXX

interesting variations in the client connection time, index search time, accuracy (recall), and storage footprint.

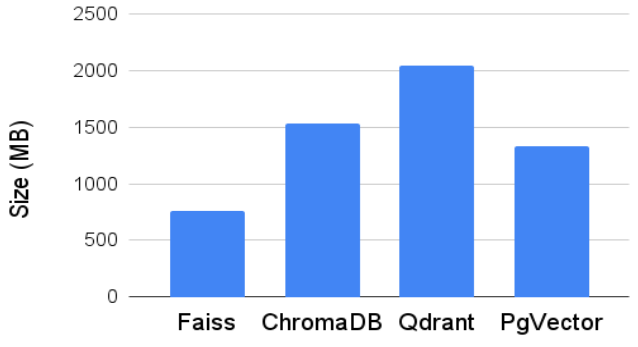

Figure 2. Storage Footprint

- Faiss did not have any data management capabilities. Therefore, EvaDB retains the Faiss index using its own format.
- ChromaDB manages vectors on the disk in a custom format, but maintains additional metadata in an SQLite database.
- Qdrant (local mode) stores both vectors and metadata in an SQLite database.
- PgVector stores both vectors and metadata in the PostGres database as an extension.

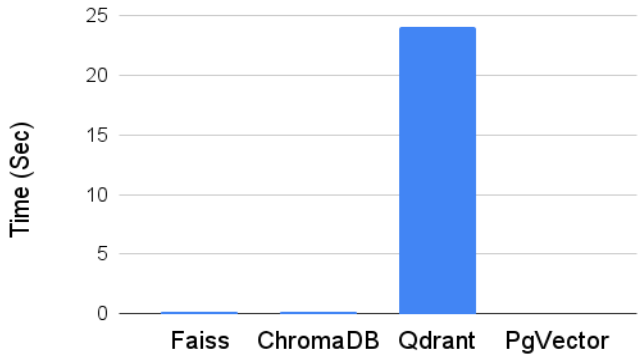

Figure 3. Client Connection Time / Index Initialization Time

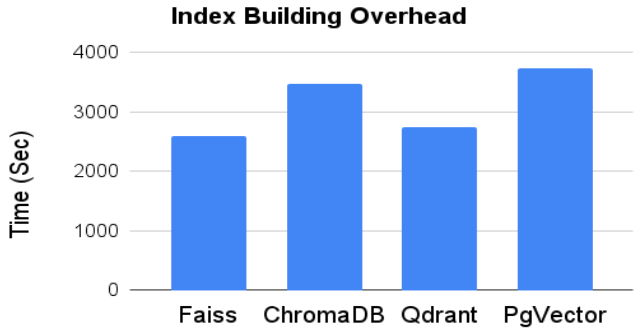

Figure 4. Index Building Time

The most striking observation here is that Qdrant (local mode) requires significantly more time for the client to connect to an existing index than other vector databases. This step occurs when a new client is started, and is a bottleneck when applications operate under an embedded architecture (i.e., not operating under a client-server architecture).

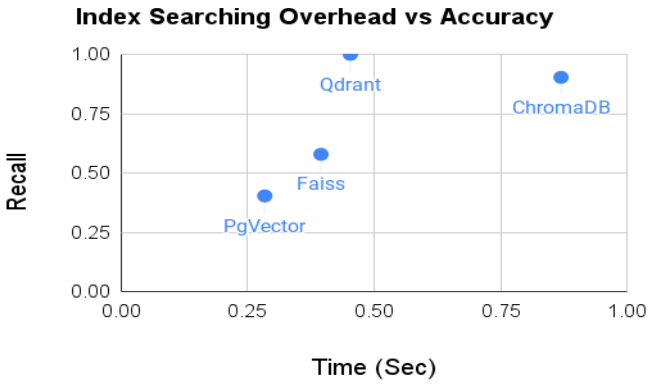

Figure 5. Index searching overhead Vs Accuracy

OpenAI's GPT models are trained to understand natural languages and code. These models respond with text, based on their inputs, also known as "prompts." A prompt is designed as a means of "programming" a GPT model; this is usually done through instructions or examples of successful task completion.

5.1 Model Evaluation

Evaluation of few pretrained models like GPT-4, GTP-3.5, LM-SOTA and SOTA are done based on certain parameters which is depicted in *table 4*.

- The pretrained base model of GPT-4 was evaluated on traditional language model benchmarks.
- Contamination checks were conducted over the test data to identify if there was any overlap with the training set.
- Few-shot prompting was used in all tests for the benchmarks.GPT-4 demonstrated superior performance compared with existing language models and previous state-of-the-art systems.
- GPT-4 did better than models that were often benchmark-specific crafted or extra training protocols. At test time.
- Performance was evaluated against GPT-4 and SOTA models, with training on that benchmark.
- GPT-4 has outperformed the existing models in all benchmarks except the DROP dataset and made the benchmark-specific training for all the datasets.
- GPT-4 has outperformed GPT-3.5 and other language models on a wide variety of languages, from low-resource ones like Latvian and Welsh to the semantic complexity of Swahili.

GPT-4 showed a radical enhancement of what the user wanted. To this end, based on a dataset of 5,214 prompts submitted to ChatGPT and the OpenAI API, the responses with GPT-4 were

favored compared to GPT-3.5 responses on 70.2% of the prompt.

Table 4. Evaluation of different models

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA - Best external model (incl. benchmark-specific tuning)
MMLU Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM	75.2% 5-shot Flan-PaLM
HellaSwag Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLaMA (validation set)	85.6% ALUM
AI2 Reasoning Challenge (ARC) Grade-school multiple choice science questions, Challenge-set	96.3% 25-shot	85.2% 25-shot	85.2% 8-shot PaLM	86.5% ST-MOE
WinoGrande Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	85.1% 5-shot PaLM	85.1% 5-shot PaLM
HumanEval Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM	65.8% CodeT + GPT-3.5
DROP Reading comprehension & arithmetic	80.9% 3-shot	64.1% 3-shot	70.8% 1-shotPaLM	88.4% QDGAT
GSM-8k Grade-school mathematics questions	92.0% 5-shot Chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva	87.3% Chinchilla + SFT + ORM-RL, ORM reranking

6. FUTURE ENCHANCEMENT

To advance QA system, we aim to optimize model performance for underrepresented languages through advanced NLP and machine learning techniques, ensuring more inclusive and comprehensive language coverage. Emphasis is placed on integrating the latest developments in deep learning for nuanced language comprehension and contextual awareness. We also plan to refine the user interface to make it more intuitive and adaptable to diverse languages and accessibility needs. Additionally, incorporating local knowledge sources will enable the system to offer more relevant region-specific responses.

Real-time processing and latency need to be improved, and this in turn requires use of efficient algorithms and hardware acceleration for fast handling of large volumes of data. Security of data and its privacy are major concerns and must be enhanced to be at par with global standards of data protection. Scalability and cloud integration will be necessary to handle the huge volume of queries and documents. Ultimately, it should evolve into a much more customizable and adaptive system that incorporates user feedback for continuous learning and system refinement. These are the keys to making our QA system more powerful, user-friendly, and versatile with respect to multilingual document analysis and interaction, therefore setting ever-expanding horizons of applicability for AI in the area.

7. CONCLUSION

The work presented in this paper represents a significant contribution towards fully AI based QA systems for processing both structured and unstructured data of diverse languages. The results from the experiments demonstrate the potential use of modern models (e.g., BERT and family) for multilingual data interpretation, but also highlight some practical hurdles in doing so. Although these models possess a good deal of language nuance, findings in this work highlight a critical gap in their performance consistency across diverse data formats and linguistic spectrums. By addressing these shortcomings, the development and inclusion of such models gives testament to how AI technologies for excellent language processing are continually improving. The importance of this research not just lies in its technical accomplishments but also in its broader implications. By enhancing the accessibility and comprehensibility of information across languages, the proposed system breaks all the barriers in global communication and information exchange. It has the potential to enhance access to knowledge, making it more feasible regardless of linguistic background.

The journey of enhancing and refining this system is continuous. As the field of AI and NLP rapidly evolves, staying ahead of the latest innovation and integrating them into our system will be crucial. The aim is to create a QA system that is not only technologically advanced but also socially inclusive, catering to a global audience with diverse needs and languages. In doing so, this research will contribute significantly to the advancement of AI applications in multilingual contexts, paving the way for more connected and accessible global communication networks.

REFERENCES

[1] S. CS224N, D. Project, G. K. Sullan, T. E. Truong, Mentor, and Y. Zhang, "Using Character Embedding and QANet to Improve Perfomance on Question-Answering Task (SQuAD)," 2022.

[2] K. Karpagam, K. Karpagam, A. Saradha, A. Saradha, and A. Saradha, "A framework for intelligent question answering system using semantic context-specific document clustering and Wordnet," Sadhana-academy Proceedings in Engineering Sciences, 2019, doi: 10.1007/s12046-018-1022-8.

International Journal of Computer Science Research (IJCSR)
------- Article | Volume 3, Issue 3 | Pages 99-104 | ISSN XXXX-XXXX

[3] A. Wang et al., "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," BlackboxNLP@EMNLP, 2019, doi: 10.18653/v1/w18-5446.

[4] G N, Basavaraj et.al. (2023). Hybrid Approach for Retail Store Auditing Using CRNN. 1-7. 10.1109/NMITCON58196.2023.10276301. Date of Conference: September 01-02, 2023.

[5] Reliability-driven time series data analysis in multiple-level deep Learning methods utilizing soft computing methods, Measurement: Sensors, Volume 24, Journal ISSN:2665-9174, https://doi.org/10.1016/j.measen.2022.100501, 29 September 2022. https://www.sciencedirect.com/science/article/pii/S2665917422001350

[6] G. Barlacchi et al., "FocusQA: Open-Domain Question Answering with a Context in Focus," Conference on Empirical Methods in Natural Language Processing, 2022, doi: 10.18653/v1/2022.findings-emnlp.381.

[7] S. Yoon, S. Yoon, E. Rhim, E. Rhim, D. Kim, and D. Kim, "Domain Question Answering System," 2015, doi: 10.5626/ktcp.2015.21.2.144.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," North American Chapter of the Association for Computational Linguistics, 2019, doi: 10.18653/v1/n19-1423.

[9] K. S. D. Ishwari et al., "Advances in Natural Language Question Answering: A Review."

[10] H. Naveed et al., "A Comprehensive Overview of Large Language Models," arXiv.org, 2023, doi: 10.48550/arxiv.2307.06435.

[11] S. Minaee, E. Cambria, and J. Gao, "Deep Learning–based Text Classification," ACM Computing Surveys (CSUR), vol. 54, pp. 1–40, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:235386502

[12] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.," arXiv: Computation and Language, 2019, doi: 10.18653/v1/2020.acl-main.703.

[13] G. Qiang, S. Tang, J. Hao, and L. Sarno, "A BIM and AIoT Integration Framework for Improving Energy Efficiency in Green Buildings," Construction Research Congress 2024, 2024, doi: 10.1061/9780784485262.059.

[14] S. Manathunga and Y. A. Illangasekara, "Retrieval Augmented Generation and Representative Vector Summarization for large unstructured textual data in Medical Education," arXiv.org, 2023, doi: 10.48550/arxiv.2308.00479.

[15] L. Wang, Y. Liu, X. Zhao, and Y. Xu, "Particle Swarm Optimization for Fuzzy c-Means Clustering," in 2006 6th World Congress on Intelligent Control and Automation, 2006, pp. 6055–6058. doi: 10.1109/WCICA.2006.1714243.

[16] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," 2022. [Online]. Available: https://arxiv.org/abs/2205.14135.

[17] S. Sarawagi and S. Sarawagi, "Models and indices for integrating unstructured data with a relational database," InternationalWorkshop on Knowledge Discovery in Inductive Databases, 2004, doi: 10.1007/978-3-540-3184.