IMQuAD: A Tamil QA Dataset for the Question-Answering Systems Evaluated on FidelityFit: An Enhanced Contextual Evaluation Metric for QA Systems

Niveditha S*, Paavai Anand G

***Abstract*—Recent breakthroughs in language modeling have significantly enhanced Natural Language Processing, particularly in Reading Comprehension. However, training non-English Question Answering (QA) systems remains challenging due to limited datasets. To address this, we introduce IMQuAD (Iyarkkai maruttuvam Question Answering Dataset), a manually crafted, linguistically scrutinized dataset focused on Naturopathy for Tamil QA. Leveraging IMQuAD, we utilized MuRIL, a BERT model, for Tamil QA tasks. Our results show that the model performed better on IMQuAD (78%) compared to SQuAD (32%) and CHAII (62%). Additionally, we pioneered FidelityFit, a novel evaluation metric assessing QA dataset accuracy with unparalleled precision. IMQuAD and FidelityFit contribute to advancing non-English QA systems, demonstrating the potential for improved language understanding and applications in various domains, including healthcare and education. Our work paves the way for further research in Tamil NLP and beyond.**

***Index Terms*—Comprehensive Question Answering, Dataset, Tamil, Evaluation metrics**

I. Introduction

RECENT research focuses on creating question-and-answer (QA) systems for Tamil and other Indic languages. [1] developed an extractive QA system for Tamil utilizing XLM-RoBERTa, whereas Rubika Murugathas and Uthayasanker Thayasivam [2] presented a domain-specific QA production system for Tamil historical texts. Ram Vignesh Namasivayam and Manjusha Rajan [3] investigated many response prediction methods in Tamil and Hindi paragraphs, such as zero-shot transfer and fine-tuning of multilingual models. They also explored the development of Tamil QA datasets through translation. Dhruv Kolhatkar and Devika Verma [4] surveyed Indian language quality assurance, analyzing datasets, methodologies, and cutting-edge models for resource-constrained languages such as Tamil, Urdu, Marathi, and Hindi. These works demonstrate the increased interest in building QA systems for Indic languages, which face obstacles such as limited datasets and the necessity for language-specific techniques. [5] addressed leveraging multilingual BERT models and cross-lingual learning strategies to solve the scarcity of large Tamil datasets. Pandian and Geetha [6] suggested a Conditional Random Field model to classify Tamil questions that uses morpheme properties to identify expected answer kinds. [7] addressed the difficulty of code-mixed QA by creating a system that can handle queries combining English and Indian languages such as Tamil. They used deep learning algorithms like RNNs and HANs to classify questions. [8] examined several techniques for semantic-level QA and information retrieval, noting that, while many systems exist for English, few work for native languages such as Tamil. These studies illustrate continuing attempts to improve QA capabilities for Tamil and other low-resource languages by combining NLP, machine learning, and data analytics.

One important job for evaluating a machine's comprehension of natural language is answering questions. Question answering and text comprehension remain challenging tasks for machines, especially in low-resource languages like Tamil. English Question Answering models have made significant progress thanks to datasets like SQuAD1.1, SQuAD2.0, and CoQA. The unavailability of annotated datasets in various languages, including Tamil, has hindered the generation of Question Answering models tailored to specific languages. Language modeling progress has improved Reading Comprehension results, but datasets are scarce, costly, and mostly English-native, highlighting the need for more diverse datasets. Standard datasets are essential for algorithmic research, but there's a lack of Tamil datasets for natural language processing, making it difficult for researchers to compare performance between models. To address this, we introduce IMQuAD, a Tamil Reading Comprehension dataset similar to SQuAD1.1. IMQuAD is a new Tamil question-answering dataset that meets the standard of the SQuAD dataset, providing a large amount of learning data for machine reading comprehension tasks and enabling researchers to evaluate their models' performance objectively. This dataset aims to bridge the gap for Tamil language Question Answering models.

Manuscript received November 18, 2024; revised August 21 2025.

Niveditha S is a PhD candidate of Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, Chennai - 26, India (corresponding author to provide phone: +91 9944492526, e-mail: nivedits@srmist.edu.in).

Paavai Anand G is an Assistant Professor of Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, Chennai - 26, India (e-mail: paavaiag@srmist.edu.in.)

II. EXISTING DATASETS

This section conducts a comparative analysis of diverse datasets employed across various languages worldwide, including English-specific, non-English, and multilingual datasets.

A. English datasets

SQuAD[9], a well-known English question-answering dataset with more than 100,000 questions created by crowd workers using Wikipedia articles as source material. Annotators were tasked with creating questions related to the articles and identifying the corresponding answers. The dataset's second version includes more than 50,000 cleverly designed unanswerable questions that mimic answerable ones, adding a new layer of complexity. The TREC-8 [10] The Question Answering track marked the first large-scale evaluation of systems designed to generate answers to questions, and the evaluation methodology was examined to understand its limits and potential improvements.[11] The paper presents a novel syntax-driven approach to question answering, using a probabilistic quasi-synchronous grammar to learn soft alignments and significantly outperform state-of-the-art baselines on the TREC dataset. The WikiQA [12] and MS Marco [13] datasets were created by leveraging questions from Bing search engine users. Annotators were presented with the top ten search results for each question and tasked with finding the answer within the documents or indicating that the answer was not present. WikiQA consists of approximately 3,000 questions with answer sentences from Wikipedia pages; however, MS Marco has 100,000 questions and free-form responses. The Natural Questions (NQ) dataset [14], comprising over 300,000 examples, was generated by sampling Google search engine questions. For every question, annotators looked through the top five Google search results, noting the answer on the relevant pages or marking it as null if it was not discovered. Two conversational QA datasets, QuAC [15] and CoQA[16], contain dialogues between questioners and answerers. CoQA has 127,000+ question-answer pairs, created by having crowd workers discuss a passage and ask questions. NewsQA[17], based on CNN articles, has more than 100,000 question-answer (QA) pairs, which is developed in three stages: questioners ask questions based on headlines, answerers find answers in the article, and validators select an optimal answer from the set else decline all. Datasets like HotpotQA[18], which require multi-step question answering, necessitate reasoning across multiple documents to provide accurate responses. This dataset was meticulously crafted by leveraging knowledge bases, web documents, and the collective efforts of crowdsourcing. Comprising an impressive 113,000 questions pertinent to Wikipedia articles, HotpotQA requires the capacity to locate and synthesize information from multiple documents to find the perfect answer. Moreover, in the context of multi-hop datasets, QA systems must also possess the ability to pinpoint the specific paragraphs that serve as the foundation for the answer, thereby ensuring a comprehensive understanding of the subject matter. Table I provides a succinct compilation of various datasets carefully crafted for English Question Answering gives a comprehensive snapshot of these important resources.

B. Non-English Dataset

There are two ways to build non-English QA datasets: converting English datasets or creating native datasets from scratch. Translating English datasets, like SQuAD, into the required language using machine interpretation is faster and easier, but may produce problematic artifacts, such as keeping the original word sequence or using overly formal language. This can result in translated text that differs significantly from the native text. Researchers have used translation to build QA datasets for Spanish, Arabic, Korean, and Italian, but some argue against this approach due to its limitations. The diversity of languages and cultures leads to distinct question patterns and topics, which may not be captured in translated datasets. Unique perspectives and queries underscore the need for native QA datasets, carefully crafted by annotators familiar with the language and culture. The construction of native QA datasets, such as SberQuAD (Russian) [25], DRCD (Chinese), KorQuAD (Korean)[26], and FQuAD (French)[27] generally follows the SQuAD format. Table II compiles another collection of scrupulously crafted datasets for other language Question-Answering tasks, offering an insightful glimpse into these highly regarded resources.

C. Multilingual QA Datasets

Due to the time and resources needed, creating large-scale Question Answering datasets developed for non-English languages is a unique and difficult task. Cross-lingual QA datasets have been created to allow training in one language and transfer to another to address this. Notable examples include MLQA (seven languages)[36], MMQA (Hindi and English)[37], TyDi (English and 10 other languages)[38], XQA (English and eight other languages)[39]. These datasets facilitate zero-shot learning and have shown promising results.

A further compilation of expertly curated datasets for the Multi-lingual Question Answering dataset is presented in Table III, providing a valuable window into the richness and diversity of these esteemed resources.

III. DATASET CURATION

We are pleased to announce the release of a comprehensive dataset called IMQuAD (Iyarkkai maruttuvam Question Answering Dataset) [43] for Tamil language question answering to facilitate research and development in Natural Language Processing (NLP).

Our dataset is structured similarly to renowned benchmark datasets such as SQuAD and CHAII, which ensures seamless integration and comparability. We are pleased to announce the launch of a comprehensive dataset called IMQuAD (Iyarkkai maruttuvam Question Answering Dataset) [43] for the Question Answering System in the Tamil language, which aims to facilitate research and development in the field of Natural Language Processing (NLP). Our dataset is structured similarly to renowned benchmark datasets such as SQuAD and CHAII, which ensures seamless integration and comparability.

The questions and answers in our dataset were carefully curated from the published book. [44] to ensure accuracy, relevance, and diversity. Figure 1 shows the technique used to create the dataset. This dataset is an important step towards bridging the gap in NLP resources for Tamil and enabling

TABLE I
COMPREHENSION DATASET FOR THE ENGLISH LANGUAGE

Dataset Name	Type of Questionnaire	No. of Question Answers	Evaluation Metrics	Reference
MCTest	Multiple Choice Questions	2000	Manual curation, Grammar Test	[19]
WIKIQA	Open Domain QA	3047	Questions - LCLR, PV Answers – Precision, Recall, F1 Score	[12]
SQuAD1.0	Comprehension	100000+	Diversity in answers, Reasoning required to answer questions, Stratification by syntactic divergence.	[9]
NEWSQA	Machine Comprehension dataset	100,000	F1, Exact Match (EM) score, BLEU, and CIDEr	[17]
TriviaQA	Machine Comprehension dataset	95K	F1, Exact Match (EM) score	[20]
SearchQA	Machine Comprehension dataset	140K	TF-IDF Max	[21]
RACE	Machine Comprehension dataset	100,000	Accuracy	[22]
MS MARCO	Machine Comprehension dataset	1,010,916	BLEU, pa-BLEU	[13]
QuAC	Machine Comprehension dataset	14K	F1 Score	[15]
SQuAD 2.0	Machine Comprehension dataset	50,000	F1 Score	[23]
HotpotQA	Multi-hop Questions	113k	F1, Exact Match (EM) score	[18]
Natural Questions	Question Answer Pair	307,373	F1 Score	[14]
DROP	Reading comprehension	96K	F1 Score	[24]
CoQA	Conversational QA	127K	F1 Score	[16]

TABLE II
COMPREHENSION DATASET FOR NON-ENGLISH LANGUAGES DATASET

Dataset Name	Language	Type of Questionnaire	No. of Question Answers	Evaluation Metrics	Reference
DuReader	Chinese	Open-domain Chinese machine reading comprehension	200K	BLEU-4 and Rouge-L	[28]
DRCD	Chinese	Chinese machine reading comprehension	30,000+	F1 Score	[29]
KorQuAD1.0	Korean	Machine Comprehension	70,000+	F1, Exact Match (EM) score	[26]
SQuAD-es v.1.1.	Spanish	Automatically translated QA from SQuAD	100,000	F1, Exact Match (EM) score	[30]
Arabic Reading Comprehension Dataset (ARCD)	Arabic	Open domain Factual Arabic question answering	1,395	F1 Score	[31]
SberQuAD	Russian	Reading Comprehension	50K	F1 Score	[25]
FQuAD	French	Reading Comprehension	60,000+	F1, Exact Match (EM) score	[27]
PersianQuAD	Persian	Reading Comprehension	20,000	F1, Exact Match (EM) score	[32]
UQuAD1.0	Urdu	Machine Reading Comprehension	49K	F1 Score	[33]
UQA	Urdu	Text Comprehension	136211	F1, Exact Match (EM) score	[34]
ParSQuAD	Persian	Machine Comprehension	95192	F1, Exact Match (EM) score	[35]

researchers and developers to develop more accurate and effective question-answer models for this language.

A. Data Description

Our dataset contains an extensive collection of 509 comprehensive Question-Answer pairs that have been carefully compiled to enable innovative research. As shown in Table IV, each set of triples is carefully structured to give a robust foundation for NLP modelling. Each Question-Answer pair is structured as follows:

1. Context (C): A passage of text providing the necessary background information.

2. Question (Q): A well-defined query, posed to elicit a specific response.

3. Answer (A): A concise and accurate response, directly addressing the question.

To ensure linguistic consistency, our dataset is written entirely in Tamil. For ease of use, the dataset is stored in a plain text format (.txt), allowing seamless integration into various NLP projects.

Please note that the dataset is encoded in UTF-8 format due to the use of the Tamil script. To ensure correct display

TABLE III
COMPREHENSION DATASET FOR MULTILINGUAL QA DATASET

Dataset Name	Language	Type of Questionnaire	No. of Question Answers	Evaluation Metrics	Reference
MMQA	English, Hindi	multi-domain, multi-lingual question answering	5495	MRR, BLEU	[37]
XQA	English, German, French, Portuguese, Russian, Chinese, Tamil, Polish, Ukrainian	Open Domain QA	90,610	F1, Exact Match (EM) score	[39]
MLQA	English, German, Arabic, Vietnamese, Spanish, Simplified Chinese, and Hindi	Multidisciplinary aligned extractive quality assessment	English - 12K instances Other Languages - 5K each	F1, Exact Match (EM) score	[36]
TYDI QA	English, Finnish, Arabic, Bengali, Indonesian, Russian, Japanese, Kiswahili, Korean, Thai, Telugu	Question Answer Pair	204K	F1 Score	[38]
XOR QA	Japanese, Bengali, Korean, Finnish, Arabic, Russian, Telugu	Answering Open-Retrieval Questions in Cross-Language	40K	F1 Score, BLEU	[40]
MKQA	26 Languages	Evaluation of open-domain questions answering	Ten thousand QA pairs matched in twenty-six typologically distinct languages.	F1 Score	[41]
No Specific Name	Hindi and Marathi	Reading Comprehension dataset	28,000	EM, Rouge-2, Rouge-L, BLEU	[42]
MMQA	English, Hindi	multi-domain, multi-lingual question answering	5495	MRR, BLEU	[37]
XQA	English, German, French, Portuguese, Russian, Chinese, Tamil, Polish, Ukrainian	Open Domain QA	90,610	F1, Exact Match (EM) score	[39]

TABLE IV
SAMPLE CONTEXT, QUESTION, AND ANSWER IN IMQUAD DATASET

context	இருசக்கர, நான்கு சக்கர வாகனங்களின் மிகுதியான போக்குவரவு, பல்வேறு ஆலைகள் பெருக்கம், மரங்கள் மிகுதியாக வெட்டப்படுதல் ஆகியவற்றால் காற்று மாசுபடுகிறது. இத்தகைய மாசுள்ள தூய்மையிலாத காற்றினை நாம் சுவாசிப்பதால், நம் உடலில் குருதி கேடடுற்று நோய் ஏற்படுகிறது. மேலும் சமைத்த தானிய உணவை பெரும்பாலும் உண்பதால், உடலில் சளி மிகுதியாகி, காற்றுப் பையாகிய நுரையீரலில் காற்று மிகுதியாக இருப்பதற்குப் பதில், சளி மிகுதியாக உள்ளது. இதனால், நுரையீரலில் காற்றின் பரிமாண அளவு குறைந்து, நோய் ஏற்படுகிறது.
question	எது காற்றுப் பை என்று கூறப்படுகிறது?
answer_text	நுரையீரல்

TABLE V.
STATISTICS OF IMQUAD WITH OTHER BENCHMARK DATASETS

Dataset	IMQuAD	SQuAD 1.1	CHAII
	Train	Train/dev	Train
No. of questions	509	87,599 / 10,570	368
No. of unique paragraphs	125	18,896 / 2,067	368
Tokens' count			
Mean paragraph length	115.81	116.6 / 122.8	1303.69
Mean question length	5.73	10.1 / 10.2	4.77
Mean answer length	2.02	3.16 / 2.9	1.92
Characters' count			
Mean paragraph length	1020.38	735.8 / 774.3	12335.19
Mean question length	46.30	59.6 / 60.0	39.59
Mean answer length	15.31	20.2 / 18.7	13.27

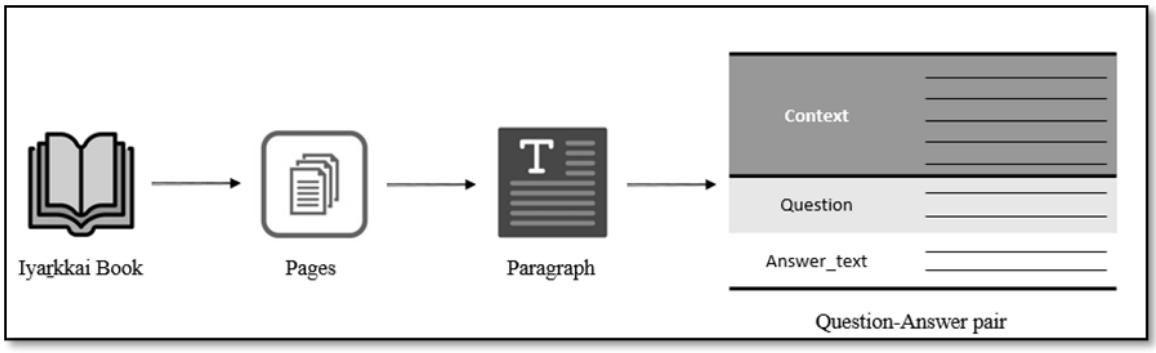

Fig 1. IMQUAD Dataset Creation

and accessibility, users must install a Tamil font on their system before using our dataset.

B. Dataset Validation

The book [44] delves into the principles of naturopathy, offering a comprehensive guide on how to achieve optimal health through a balanced diet, regular exercise, and natural remedies. Additionally, it provides in-depth discussions on over 100 common diseases, along with their corresponding solutions, empowering readers to take control of their well-being. Table V shows that the number of questions and the length of questions/answers of IMQuAD are compared with other benchmark datasets.

In particular, IMQuAD has more questions than CHAII and fewer than SQuAD. The length of questions and answers is also longer than that of CHAII and shorter than that of SQuAD, highlighting the unique characteristics of IMQuAD within the dataset landscape.

To validate the integrity and accuracy of our dataset, we conducted a rigorous evaluation process. We enlisted the expertise of a linguistic specialist, who brought their domain knowledge to assess the dataset's linguistic quality.

Specifically, the expert evaluated the following aspects:

Language correctness: The specialist reviewed the grammar, syntax, and semantics of the context paragraphs, questions, and answers to ensure they conform to standard linguistic norms.

Question formation: The expert assessed whether the questions were well-formed, clear, and unambiguous, and whether they accurately reflected the content of the context paragraphs.

Answer accuracy: The specialist verified whether the retrieved answers accurately responded to the questions and whether they were supported by the context paragraphs.

By incorporating the expert's feedback, we refined our dataset to ensure it meets high standards of linguistic quality, accuracy, and reliability.

IV. ALGORITHM

The algorithm shown in Fig. 2 is the BERT-based QA system that works by preprocessing input text, generating contextualized embedding, predicting answer start and end tokens, extracting the answer, and refining the output.

The Pre-train function structures the input by inserting [CLS] at the beginning of the question and [SEP] between the context, question, and answer. The same is depicted in equation (1).

$$[CLS]Context\ [SEP]Question\ [SEP]Answer\ [SEP] \rightarrow Pre-train\ (Context, Question, Answer) \quad (1)$$

As shown in equation (2), the BERT tokenizer is used to tokenize the formatted sequence, yielding c_i for the i^{th} context token, q_j for the j^{th} question token, and a_k for the k^{th} answer token.

The contextualized embedding is generated for token representation. Using the Embedding function, the embedding vectors of tokens are generated as shown in equation (3).

$$[I[CLS], c1, c2, .. cn, I[SEP], q1, q2, .. qn, I[SEP], a1, a2, .., am, I[SEP]] \rightarrow tokenize([CLS]\ context\ [SEP]\ question\ [SEP]\ answer\ [SEP]) \quad (2)$$

$$[EMBI[CLS], EMBc1, EMBc2, .. EMBcn, EMBI[SEP], EMBq1, EMBq2, .. EMBqn, EMBI[SEP], EMBa1, EMBa2, .., EMBam, EMBI[SEP]] \rightarrow Embedding\ ([I[CLS], c1, c2, .. cn, I[SEP], q1, q2, .. qn, I[SEP], a1, a2, .., am, I[SEP]]) \quad (3)$$

The embedded vector is then passed to encoders as shown in equation (4).

The vital component of the BERT system is the transformer encoder, responsible for processing input text and generating contextualized representations. Transformer encoders use self-attention layers instead of RNN (recurrent neural networks) to process input tokens in parallel. Each self-attention layer calculates attention weights, allowing the model to contextualize long-range dependencies.

$$inputs \rightarrow [EMBI[CLS], EMBc1, EMBc2, .. EMBcn, EMBI[SEP], EMBq1, EMBq2, .. EMBqn, EMBI[SEP], EMBa1, EMBa2, .., EMBam, EMBI[SEP]] \quad (4)$$

Transformer models can analyze variable-length input sequences effectively and produce state-of-the-art results in a variety of NLP applications according to this architecture. ith self-attention layer generates three vectors for each embedding vector, namely Query vector (Qi), which represents context for token consideration, Key vector (Ki), which represents tokens to attend, and Value vector (Vi), which represents information to extract from attended tokens.

The self-attention mechanism is capable of concentrating on essential tokens attributable to these vectors and contextualizing token representations. The computation of these vectors is shown in equations (5), (6), and (7)

$$Q_i \leftarrow emb_j * W_{Qi} \quad (5)$$
$$K_i \leftarrow emb_j * W_{Ki} \quad (6)$$
$$V_i \leftarrow emb_j * W_{Vi} \quad (7)$$

The self-attention mechanism calculates the output vector Zi by applying softmax to the attention scores. The dot product of the query vector Qi and the key vector Ki, scaled by the square root of Ki's dimensionality, is used to calculate attention scores.

These scores are normalized to a probability distribution using the softmax function, indicating the importance of each token. Finally, the softmax output is multiplied by the Value vector Vi to extract relevant information and contextualize token representations. The same is depicted in equation (8).

$$Z_i \leftarrow \text{softmax}\,(Q_i * K_i\text{^T}/\text{sqrt}(\dim(K_i))) * V_i \quad (8)$$

As shown in equation (9), Multi-Head Attention takes in Query, Key, and Value matrices and splits them into multiple attention "heads". Each head computes self-attention scores. [45] and outputs a vector, which is then concatenated. The concatenated output is linearly transformed using a learnable weight matrix Wo.

$$\text{Multi} - \text{Head}\,(Q, K, V) = \text{Concat}\,(h_1, h_2, .. h_h)W_o \quad (9)$$

Where

$h_i = Attention\,(QW_i^Q, KW_i^k, VW_i^V)$ [45]

To create a new vector emb_j^{new}, the Z vector is routed across a fully connected network. The network uses

Pre-train : **context, question, answer**
Input : **question**
Output : **answer**

[CLS] Context [SEP] Question [SEP] Answer [SEP] → Pre-train (Context, Question, Answer)
$[I_{[CLS]}, c_1, c_2, ..c_n, I_{[SEP]}, q_1, q_2, ..q_n, I_{[SEP]}, a_1, a_2, .., a_m, I_{[SEP]}]$ → tokenize([CLS] context [SEP] question [SEP] answer [SEP])
$[EMB_{I[CLS]}, EMB_{c1}, EMB_{c2}, ..EMB_{cn}, EMB_{I[SEP]}, EMB_{q1}, EMB_{q2}, ..EMB_{qn}, EMB_{I[SEP]}, EMB_{a1}, EMB_{a2}, .., EMB_{am}, EMB_{I[SEP]}]$ → Embedding ($[I_{[CLS]}, c_1, c_2, ..c_n, I_{[SEP]}, q_1, q_2, ..q_n, I_{[SEP]}, a_1, a_2, .., a_m, I_{[SEP]}]$)
inputs → $[EMB_{I[CLS]}, EMB_{c1}, EMB_{c2}, ..EMB_{cn}, EMB_{I[SEP]}, EMB_{q1}, EMB_{q2}, ..EMB_{qn}, EMB_{I[SEP]}, EMB_{a1}, EMB_{a2}, .., EMB_{am}, EMB_{I[SEP]}]$
$\forall x \forall y \in (\text{encoder}(x) \wedge (\text{emb}(y) \in \text{inputs})$ do
  $\forall z \in (\text{self-attention}(z))$ do
    $Q_i \leftarrow emb_j * W_{Qi}$
    $K_i \leftarrow emb_j * W_{Ki}$
    $V_i \leftarrow emb_j * W_{Vi}$
    $Z_i \leftarrow \text{softmax}\,(Q_i * K_i\text{^T}/\text{sqrt}(\dim(K_i))) * V_i$
  end for
  Multi-Head(Q, K, V) = Concat $(h_1, h_2, .. h_h)\, W_o$
  $emb_j^{new} = Z * W_F + b_F$
  inputs $\leftarrow emb_{1..sizeof(inputs)}^{new}$
end for

$$s_{i_1}^L = \sum_{j=1}^{d} H_{ij} * W_{sj}$$

$$E_{i_1}^L = \sum_{j=1}^{d} H_{ij} * W_{ej}$$

$S = HW_s + b_s$
$E = HW_e + b_e$
$(i^*, j^*) \leftarrow \text{argmax}(i, j)\,[\, s_i + e_j\,]$
retrieved_answer $\leftarrow D(\{p_k \mid i^* \le k \le j^*\})$
return retrieved_answer

Fig 2. Algorithm for our QA system

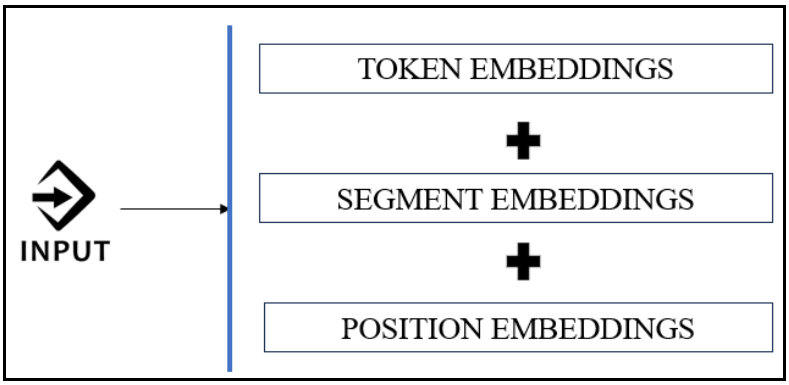

Fig 3. BERT model

learnable parameters W_F (weight matrix) and bF (bias vector) to transform the input vector. The output vector emb_j^{new} is computed by applying an activation function F as shown in equation (10).

Each input is taken from the input list, and a new embedding is created for it using the vector emb. The same is depicted in equation (11).

$$emb_j^{new} = Z * W_F + b_F \quad (10)$$

$$\text{inputs} \leftarrow emb_{1..sizeof\,(inputs)}^{new} \quad (11)$$

The equations (12) and (13) compute two separate output vectors, S and E, by taking the weighted sum of the elements in the input matrix H. The weights used for S are from the matrix Ws, while the weights used for E are from the matrix We.

$$s_{i_1}^{L} = \sum_{j=1}^{d} H_{ij} * W_{sj} \quad (12)$$

$$E_{i_1}^{L} = \sum_{j=1}^{d} H_{ij} * W_{ej} \quad (13)$$

As shown in equations (14) and (15), two linear transformations are applied to the input vector H, resulting in output vectors S and E. The transformations involve matrix multiplications with learnable weight matrices Ws and We, and additions of learnable bias vectors bs and be.

These transformations can be used to extract different features or representations from the input data, depending on the context.

$$S = HW_s + b_s \quad (14)$$

$$E = HW_e + b_e \quad (15)$$

Equation (16) finds the pair of indices (i*, j*) that corresponds to the maximum value of the sum of the elements from the vectors S and E for the optimal alignment between two sequences.

$$(i^*, j^*) \leftarrow \text{argmax}(i,j)\ [S_i + E_j] \quad (16)$$

Consequently, the retrieved answer from the dataset D is extracted, leveraging the optimal indices (i*, j*) to define the start and end points of the retrieved answer, the same is shown in equation (17)

$$\text{retrieved\_answer} \leftarrow D(\{p_k \mid i^* \le k \le j^*\}) \quad (17)$$

Three embedding totalling the input, are combined into one: token embedding (word representation), segmentation embedding (representing the segment or sentence), and position embedding (representing the word's position in the sequence). The same is shown in Fig. 3

V.EVALUATION METRICS

This section outlines the existing evaluation metrics used to assess performance, as well as the new metrics we have introduced to enhance evaluation effectiveness.

A. Jaccard Measure

The Jaccard measure, often called the Jaccard index or Jaccard similarity, is a technique for calculating the degree of similarity between two sets. The computation involves dividing the area where the two sets meet by the area where their union occurs. This can be mathematically expressed in equation (18).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (18)$$

In comparing two words, A and B, we look at the number of words that appear in both sets (A ∩ and B) and the number of words, in both sets (A ∪ and B).

The Jaccard measure ranges from 0 to 1, where 0 signifies no similarity (no shared words) and 1 denotes similarity (in words). Table VI shows the Jaccard measure of the model across datasets. Our dataset outperforms SQuAD by ~40% and CHAII by ~20%, setting a new benchmark. Our analysis revealed that the underlying reason for this disparity was attributed to the linguistic quality of the dataset.

B. Exact Match (EM)

The Exact Match (EM) score, widely utilized in natural language processing in question-answering tasks, gauges the precision of predictions. It analyses whether the reference answer and the anticipated response match.

Here is how equation (19) represents this. Table VI contains the model's EM scores across each dataset. Our dataset outperforms SQuAD by ~50% and CHAII by ~25%. Our analysis revealed that the underlying reason for this improvement was attributed to the domain specificity of the dataset.

$$EM\ Score = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Exact\ Matches}\ X\ 100 \quad (19)$$

C. *F1 Score*

A measure of a test's accuracy in binary classification issues is the F1 score. Equation (20) shows the same issue.

To calculate the score, it takes recall and precision as factors. The F1 score, a harmonic mean of precision and recall, provides a balance between the two.
Precision measures the proportion of correctly identified positives among all predicted positives. (Positive Predictive Value). Table VI lists the model's precision over several datasets.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (20)$$

Recall (True Positive Rate): The proportion of all observations made in the actual class to all positively anticipated observations. The same is depicted in equation (21). The Recall of the model over various datasets is listed in Table VI.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (21)$$

The F1 score represents the average of precision and recall as illustrated in equation (22). The models' F1 Score, across datasets, can be found in Table VI.

$$F1\ Score = 2\ X\ \frac{Precision\ X\ Recall}{Precision + Recall} \quad (22)$$

Although our dataset performs better than others on the existing evaluation metrics, we found that these metrics themselves are limited in their ability to improve the model's performance. This realisation has led us to develop conceptual evaluation metrics that can better guide and improve the performance of the models. By introducing these new metrics, we aim to achieve further advances in model capabilities. The metrics we propose should provide actionable insights and enable more effective model refinement and optimisation.

D. Fidelity Fit

Conventional metrics like precision, accuracy, recall, and F1-score give an overview of a model's performance, but they might not always fully reflect the subtleties of a given issue or situation. In these situations, the model's performance can be more accurately evaluated by using extra or different evaluation metrics.
Several important findings emerged from the analysis of the SQuAD and CHAII datasets. These included the existence of foreign languages (apart from Tamil) in certain contexts, which might impair model performance, and inconsistent question-answer and context-question pairs, in which some questions had no relation to the contexts in which they were asked, and some answers did not directly address the questions.
These issues highlight the significance of fine-tuning data and putting quality control measures in place to ensure contextual relevance and linguistic coherence. Optimizing the performance of the model may also require preprocessing and filtering the data. These findings must be taken into consideration to improve the model's performance and accuracy, as well as generate more consistent and helpful results.

Recognizing the significance of these parameters in influencing model performance, our dataset was meticulously curated to ensure optimal quality and relevance. Consequently, our dataset demonstrated superior performance compared to two other benchmark datasets, as evident in Table VII. The careful crafting of our dataset paid off, yielding impressive results that surpass those of existing datasets.

VI. DISCUSSION

To evaluate our dataset's quality, IMQuAD, we employed the MuRIL model, a BERT variant proficient in the Tamil language. We ran MuRIL [46] on three datasets: SQuAD, CHAII, and IMQuAD.

This approach allows us to evaluate our dataset's effectiveness by comparing MuRIL's performance across the three datasets. By doing so, we can identify potential biases, weaknesses, or areas for improvement in IMQuAD. Ultimately, this evaluation will help us refine our dataset and ensure its quality for future use in Tamil language understanding tasks. During our dataset quality assessment, we conducted a thorough evaluation of MuRIL's performance on three datasets. The datasets used for this evaluation were SQuAD, CHAII, and our own IMQuAD dataset. We evaluated the model's performance using an extensive set of benchmark evaluation indicators. As shown in Fig. 4, the Jaccard measure, Precision, Recall, F1 Score, and Exact Match (EM) were among these measurements.
The MuRIL model was trained and tested on each dataset to ensure a fair comparison.

Our analysis revealed that the MuRIL model achieved superior performance on our IMQuAD dataset. The results showed that IMQuAD outperformed the accuracy and efficacy of the other two datasets. This suggests that our dataset is of high quality and well-suited for Tamil language understanding tasks. The superior performance of the MuRIL model on IMQuAD demonstrates the value of our dataset in supporting NLP applications.
Overall, our evaluation confirms that IMQuAD is a reliable and effective dataset for Tamil language understanding research. To delve deeper into the dataset's quality, we scrutinized its fundamental components: Context, Questions, and Answers.

A. Linguistic Homogeneity Index (LHI)

We have found that the linguistic consistency of the context significantly affects the performance of the model. In particular, the fact that the context is entirely in Tamil improves performance, which was a limitation in other datasets. Figure 5a highlights the presence of non-Tamil text in the context and reveals striking differences between the datasets. CHAII had a notable 87.56% of contexts containing foreign languages, corresponding to approximately 322 out of 368 context-question-answer sets. This means that only 12% of the QA pair consists entirely of Tamil text.
Similarly, the SQuAD dataset had 58.78% of contexts with foreign languages, corresponding to approximately 707 instances. Conversely, 41.22% of the QA pair is made up entirely of Tamil text. In stark contrast, our dataset has a remarkable 100% of contexts with Tamil text, highlighting its linguistic homogeneity and potential for improved model performance.

TABLE VI.
PERFORMANCE OF THE MODEL OVER ALL DATASETS

Dataset	Jaccard Measure	Precision	Recall	F1 Score	Exact Match
SQuAD	0.38	0.37	0.39	0.3	30%
CHAII	0.622	0.64	0.67	0.62	54%
IMQuAD	0.779	0.78	0.79	0.784	78%

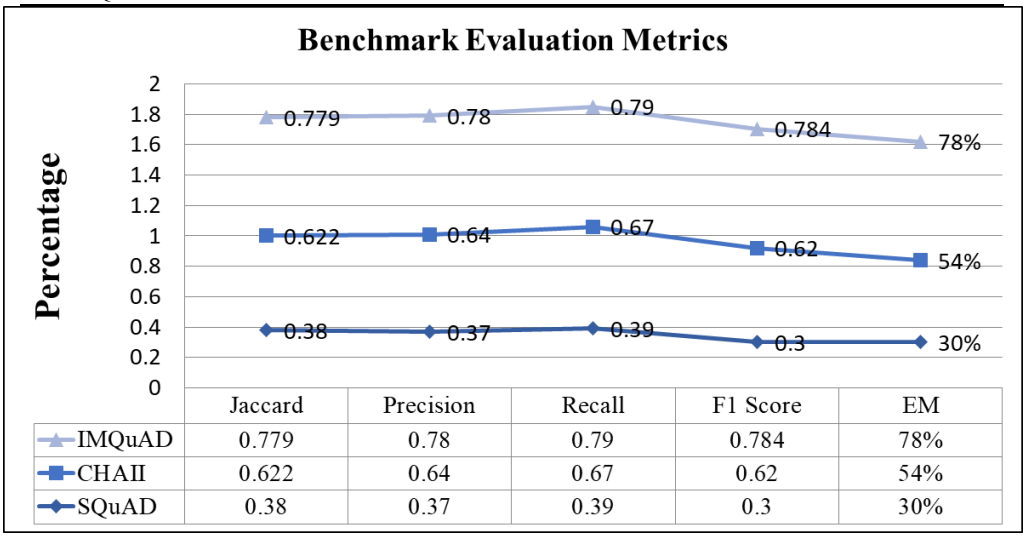

Fig 4. Benchmark Evaluation Metrics

B. Question-Context Focused

It has been shown that question-context (QC) fit is an essential criterion for assessing the quality of a QA dataset. This metric measures the match between questions and the corresponding context. A high-quality dataset depends heavily on this aspect. When analysing the SQuAD dataset, we found that 95 questions did not match their context.

This corresponds to 8% of the total questions in the dataset. Consequently, the SQuAD dataset scored 92% in the QC alignment metric. Figure 5b: In contrast, the CHAII dataset did not exhibit this problem. We took note of this discrepancy and ensured that our dataset maintained a high standard of QC alignment.

Our dataset was carefully curated to avoid questions that did not fully rely on the context provided. In this way, we were able to guarantee a high-quality dataset with excellent QC alignment.

C. Context – Answer Focused

We evaluated the Context-Answer-Alignment (CA) to assess the quality of our dataset. This evaluation checks whether the answer is present in the specified context. The metric for CA alignment is shown in Fig. 5c. In the SQuAD dataset, we found that 48 responses did not match their context. This discrepancy accounts for 4% of the total responses in the SQuAD dataset, which includes 1201 responses. In the CHAII dataset, on the other hand, 7 responses did not match their context.

This represents 2% of the total responses in the CHAII dataset, which comprises 368 responses. Consequently, the SQuAD dataset achieved 96% CA matching, while the CHAII dataset achieved 98% CA matching. We recognised the importance of CA matching in maintaining a high-quality dataset. Therefore, we have ensured that our dataset achieves 100% full CA matching. This means that every response in our dataset is present in its appropriate context. In this way, we have ensured the accuracy and reliability of our dataset. Due to the flawless CA alignment, our dataset stands out from others and is an invaluable tool for future studies and progress.

In addition to evaluating the dataset's internal quality metrics, we also sought to assess its effectiveness in a real-world application. Specifically, we wanted to investigate how our dataset performs when utilized to fine-tune the popular BERT model, MuRIL. This evaluation would provide insights into our dataset's ability to enhance the model's language understanding capabilities, particularly in the context of Tamil language processing.This assessment would shed light on how well our dataset can improve the model's language comprehension abilities, especially when it comes to processing Tamil language input. We aimed to identify the model's strengths and weaknesses to enhance our dataset for future NLP tasks. We did this by evaluating the model's performance on our dataset. This allows us to fully appreciate the worth and potential for improvement of our dataset by looking at both model performance and dataset quality.

We conducted a thorough evaluation of our model's performance by comparing the actual answers present in the dataset with the responses generated by our model. Our analysis revealed that the overall performance of our model can be broadly categorised into three different response types, as summarised in Table VIII.

Correct answers: These are cases where our model found exactly the answer available in the dataset, demonstrating a high degree of precision and reliability. There are 509

TABLE VIII.
OVERALL PERFORMANCE OF ALL THE DATASETS

	Correct answers	Incorrect answers	Potential Lead	Total
SQuAD	461	674	66	1201
CHAII	229	65	74	368
IMQuAD	397	25	87	509

questions in our dataset. Of these, 397 of the answers perfectly matched the answers available in the dataset. Similarly, 461 out of 1201 questions from the SQuAD dataset and 229 out of 368 questions from the CHAII dataset were correctly retrieved.

Incorrect answers: Conversely, this category includes cases where our model's answer differed from the correct answer and indicated areas for improvement and refinement. There were 25 instances of this type of QA in our dataset. Similarly, 65 for CHAII and 674 questions for the SQuAD dataset.

Potential Lead: This intriguing category encompasses answers that, while not exact matches, showed promise and relevance to the query, indicating opportunities for further

TABLE VII.
FEDELITY FIT

Dataset	No. of Questions	Linguistic Homogeneity Index (in %)	Question-context alignment (in %)	Context – Answer alignment (in %)
SQuAD	1201	58.78%	92%	96%
CHAII	368	87.56%	100%	98%
IMQuAD	509	0%	100%	100%

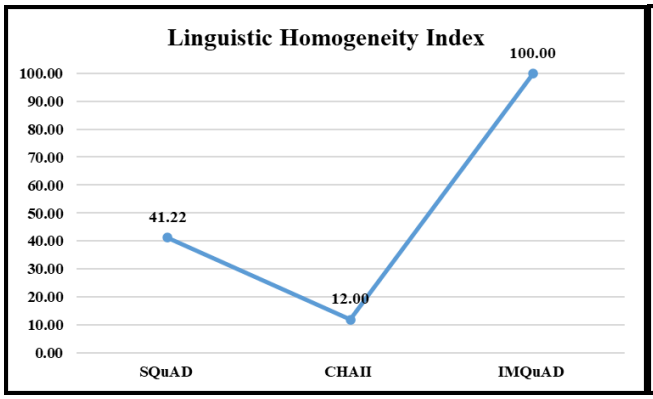

Fig 5a. Linguistic Homogeneity Index

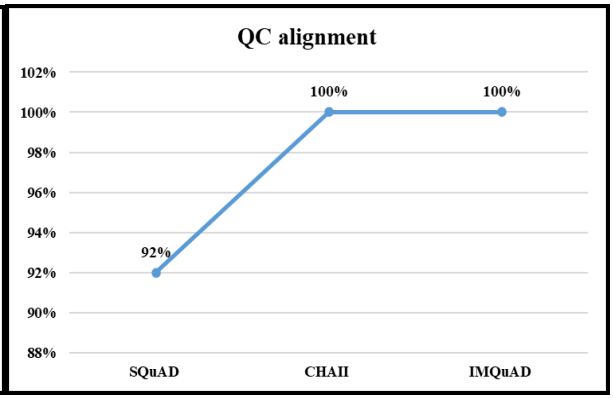

Fig 5b. QC alignment

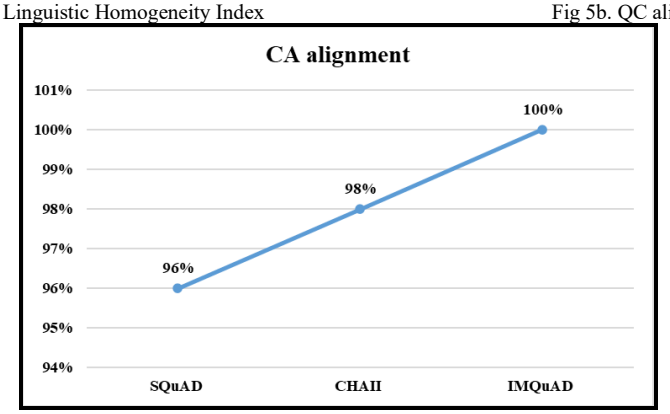

Fig 5c. CA alignment

TABLE IX.
NUMERICAL SPACING

Question	Answer in the dataset	Answer Retrieved by our Model
நன்கு தேர்ச்சி அடைந்த பின் எந்த அளவில் மாற்றிக்கொள்ளலாம்? At what level can you change after mastering well? Nan̲ku tērcci aṭainta pin̲ enta aḷavil mārrikkoḷḷalām?	01:04:02	01¸04¸02
எந்த அளவு தேன் சேர்த்துத் தினமும் காலையில் வெறும் வயிற்றில் அருந்தினால் நோய் குறையும்? How much honey and drinking it every morning on an empty stomach will reduce the disease? Enta aḷavu tēn̲ cērttut tin̲amum kālaiyil ver̲um vayirril aruntin̲āl nōy kur̲aiyum?	25 மி.லி. 25 ml. 25 ml.	25 மி. ¸லி. 25 ml. 25 ml.
எந்த சிந்தாந்தப்படி எல்லா வகையான நோய்களையும் இயற்கை உணவே மருந்தாகிக் குணப்படுத்துகிறது? According to which idea, natural food cures all kinds of diseases? Enta cintāntappaṭi ellā vakaiyāna nōykaḷaiyum iyar̲kai uṇavē maruntākik kuṇappaṭuttukir̲atu?	உணவே மருந்து: மருந்தே உணவு	உணவே மருந்து¸: மருந்தே உணவு

TABLE X.
N-GRAM OMISSION AND COMMISSION

Question	Answer in the dataset	Answer Retrieved by our Model	Type of Issue
இயற்கை உணவுகளை உண்ணும்போது ஜீரண உறுப்புக்கள் எப்படி இயங்கும்? How does the digestive system work when eating natural foods? Iyaṟkai uṇavukaḷai uṇṇumpōtu jīraṇa uṟuppukkaḷ eppaṭi iyaṅkum?	செம்மையாக இயங்கி Runs perfectly Cem'maiyāka iyaṅki	செம்மையாக Perfectly Cem'maiyāka	Omission
சிலந்தி நாயகம் என்றல் என்ன? What is silanthi nayagam? Cilanti nāyakam enṟāl enna?	ஒரு வகை வெடிக்காய்ச் செடி An Firecracker plant Oru vakai veṭikkāyc ceṭi	வெடிக்காய்ச் செடி Firecracker plant Veṭikkāyc ceṭi	Omission
யோகாசனப் பயிற்சிகள் எப்பொழுது செய்ய வேண்டும்? When to do yoga exercises? Yōkācaṇap payiṟcikaḷ eppoḻutu ceyya vēṇṭum?	காலையில் in the morning Kālaiyil	தினம் காலையில் Every morning Tiṇam kālaiyil	Commission

TABLE XI.
ERROR IN MORPHOLOGY

Question	Answer in the dataset	Answer Retrieved by our Model
எது ஆல்ஃபா நிலையாகும்? What is the alpha position? Etu ālḥpā nilaiyākum?	தூக்கத்திற்கும் விழிப்பிற்கும் இடைப்பட்ட நிலை The state between sleep and wakefulness Tūkkattiṟkum viḻippiṟkum iṭaippaṭṭa nilai	தூக்கத்திற்கும் விழிப்பிற்கும் இடைப்பட்ட நிலையே The state between sleep and wakefulness Tūkkattiṟkum viḻippiṟkum iṭaippaṭṭa nilaiyē
எண்ணெய் கொப்பளிப்பிற்கு சிறந்த நேரம் எது? What is the best time for oil extraction? Eṇṇey koppaḷippiṟku ciṟanta nēram etu?	காலை நேரம் Morning time Kālai nēram	காலை நேரமே Morning time Kālai nēramē

exploration and refinement to tap into the model's latent capabilities. Our analysis revealed three categories that required refinement.

A comprehensive comparison of the three metrics across the three datasets is vividly illustrated in Figure 6 and provides valuable insights into their performance and trends. A detailed discussion of the model's performance and possible enhancements is presented below.

C.1. Numerical Spacing: In our analysis, we encountered 3 specific questions that had numerical answers. The same is shown in Table IX (spaces are depicted using "␣"). However, due to a subtle issue with spacing between words, our model incorrectly identified these numerical answers as distinct entities. This meant that the model failed to recognize the identical numbers as the same answer, leading to inaccuracies in its response. To improve our model's performance, we need to address this spacing issue and enhance its ability to recognize numerical answers with varying formats. A similar problem was also observed with 6 questions from the CHAII dataset and 4 questions from the SQuAD dataset.

C.2. N-gram Omission and Commission: Our analysis revealed that approximately 45 questions had challenges in accurately identifying the correct answers, stemming from two distinct types of N-gram issues: Commission and Omission. The same is shown in Table X (displayed prominently in bold, italics, and underlined for emphasis).

- Commission issues occurred when the model incorrectly identified additional words that were not present in the correct answer.
- Omission issues arose when the model failed to identify certain words that were part of the correct answer.

In essence, Commission errors involved "false positives" (extra words), while Omission errors involved "false negatives" (missed words).

Addressing these N-gram issues will help refine our model's accuracy in identifying correct answers. The same problem also occurred with 44 questions from the CHAII dataset and 51 questions from the SQuAD dataset.

C.3. Error in Morphology: Tamil is an agglutinative language, characterized by its complex system of affixes that attach to root words to form new words.

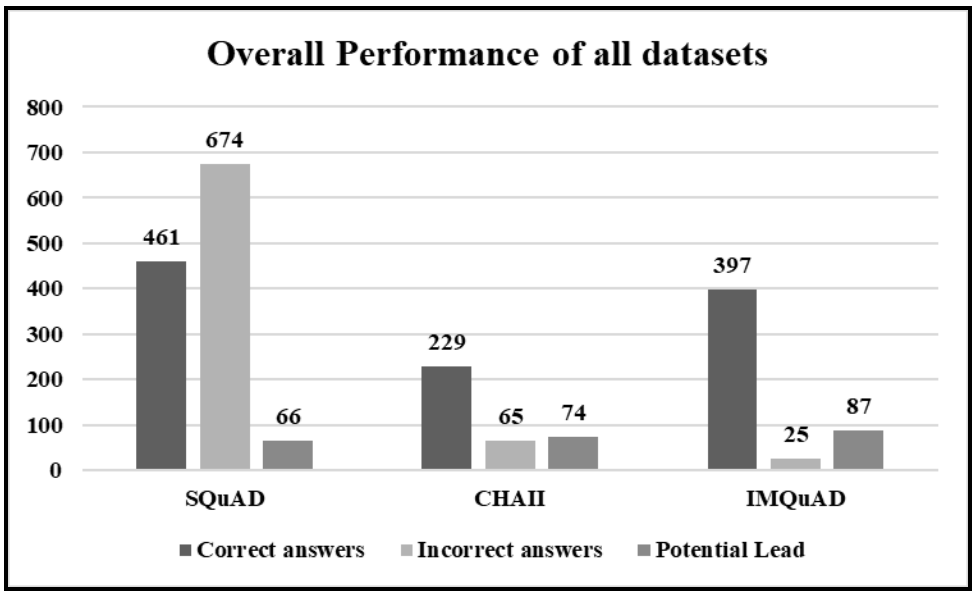

Fig 6. Overall Performance of all Datasets

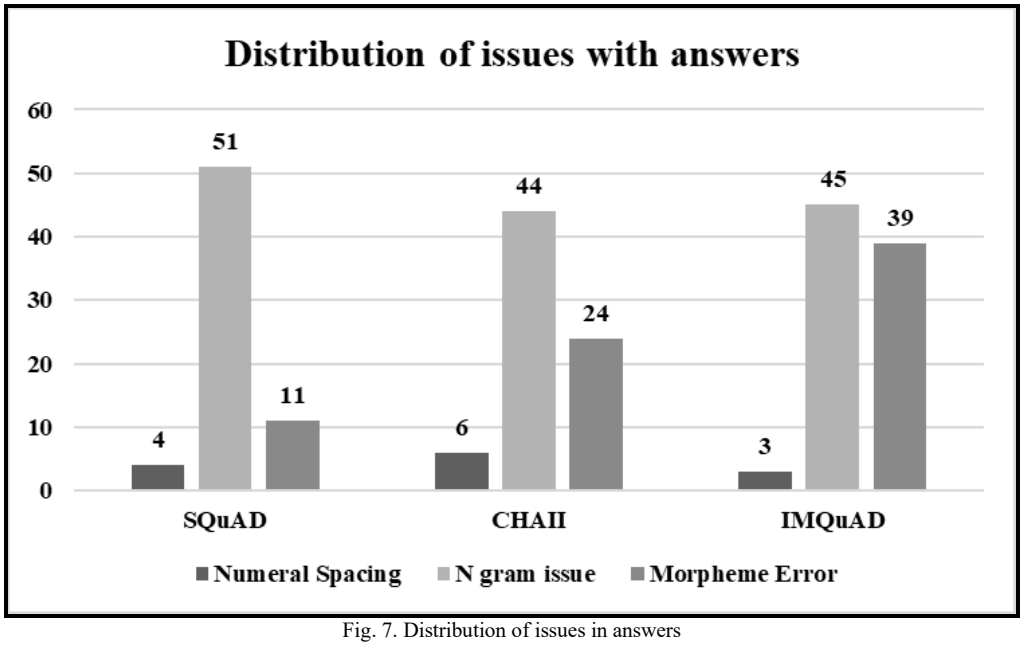

Fig. 7. Distribution of issues in answers

With approximately 2000 affixes in the Tamil language, words can have multiple forms, making it challenging for language models to understand their relationships. Our model, not being trained on the Tamil language, struggled to identify morphologically equivalent words.

Due to its lack of training, our model failed to recognize these nuances, leading to inaccuracies in its responses. The same is depicted in Table XI.

Agglutinative languages like Tamil pose unique challenges for natural language processing tasks, highlighting the need for language-specific training and adaptation. By acknowledging these limitations, we can work towards developing more effective language models for Tamil and other agglutinative languages. Our analysis of the 509-question dataset revealed promising results, with correct answers retrieved for 397 questions (77.99%). However, 25 questions (4.9%) yielded incorrect answers. Notably, a grey area emerged with 87 questions (17.09%),where accurate answers could potentially be retrieved if specific issues were adequately addressed, highlighting opportunities for further improvement. The same is depicted in Fig. 6. Upon further investigation into the grey area, we identified that the 17.09% performance lag was primarily due to inadequate handling of specific attributes, which, if addressed, could potentially bridge this gap. It was observed from the analysis that the model's performance can be significantly enhanced by optimizing three key parameters: numeral spacing, morphological variations in words, and precise N-gram selection for answer retrieval. Notably, despite three answers being contextually accurate, incorrect spacing between words led to their incorrect labeling. Furthermore, approximately 45 answers could be improved by refining the word count in the retrieved answers. Moreover, considering Tamil's agglutinative nature, accurate morphological analysis of words could lead to a significant increase of 39 correct

retrievals. By overcoming these challenges, we expect to significantly improve the performance of the model, which could achieve an
impressive accuracy rate of 90.2%. The preceding discussion focuses on our dataset. However, we have also analysed several other benchmark datasets in detail. The results of this comprehensive analysis are shown in Figures 6 and 7.

VII. CONCLUSION AND FUTURE WORK

We have introduced IMQuAD, a pioneering Tamil question-answering dataset carefully compiled with over 509 expert-annotated questions from a published book.[44]. IMQuAD was inspired by the renowned SQuAD and CHAII datasets and is capable of significantly advancing research in the field of multilingual natural language processing. Our ground-breaking model, MuRIL, achieves remarkable success, outperforming human ability with an impressive F1 score of 0.784 and an exact match of 78%. This breakthrough highlights IMQuAD's potential to revolutionise Tamil language processing and paves the way for future innovations in multilingual AI.

Through rigorous analysis of fidelity fit, we identified key areas for improvement and found that accounting for number spacing, morphological variations in words, and precise selection of N-grammes for answer retrieval could further boost performance to 95.08%. We applied this not only to our dataset, but also to the benchmark dataset.
The performance improved from 38.38% to 43.88% for SQuAD and from 62.28% to 82.33% for the CHAII dataset. To achieve even higher accuracy, our model also requires extensive training on Tamil grammar and uses Tamil Part-Of-Speech (POS) to refine its understanding. By addressing these aspects, we can realise the full potential of IMQuAD and MuRIL and make significant advances in multilingual natural language processing.

DATA AVAILABILITY STATEMENT

Data is available to access in the following drive link: https: //drive.google.com/file/d/1xvIreuWf3dM8Ao6BvQR5Sew935Aa3h0p/view?usp=sharing

ACKNOWLEDGMENT

We thank Dr. A. Bagiyalakshmi, who is an Assistant Professor (Ret.) of Jaya College of Arts and Science, for meticulously reviewing the dataset for semantic inaccuracies, syntactic errors, and inconsistencies.

REFERENCES

[1] Krishnan et al., “An Extractive Question Answering System for the Tamil Language,” presented at the Advances in Science and Technology, 2023, pp. 312–319.

[2] Murugathas et al., “Domain specific Question & Answer generation in Tamil,” presented at the International Conference on Asian Language Processing (IALP), 2022, pp. 323–328.

[3] Namasivayam et al., “Answer Prediction for Questions from Tamil and Hindi Passages,” presented at the Procedia Computer Science, 2023, pp. 1985–1993.

[4] Kolhatkar et al., “Indic Language Question Answering: A Survey,” presented at the Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2023, pp. 697–703.

[5] Srivatsun et al., “Machine Comprehension System in Tamil and English based on BERT,” presented at the 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), 2022, pp. 847–854.

[6] Pandian et al., “Tamil Question Classification Using Morpheme Features,” in *Advances in Natural Language Processing*, Gothenburg, Sweden, Aug. 2008.

[7] S et al., “Code Mixed Question Answering Challenge using Deep Learning Methods,” presented at the th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 1331–1337.

[8] Ilyas et al., “Review and Analysis of Different Approaches to Semantic Level Question Answering and Information Retrieval,” presented at the International Journal of Science and Research (IJSR), 2021, pp. 1238–1244.

[9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” Oct. 10, 2016, *arXiv*: arXiv:1606.05250. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1606.05250

[10] E. M. Voorhees and D. M. Tice, “The TREC-8 Question Answering Track,” in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00)*, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer, Eds., Athens, Greece: European Language Resources Association (ELRA), May 2000. Accessed: Sep. 04, 2024. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2000/pdf/26.pdf

[11] M. Wang, N. A. Smith, and T. Mitamura, “What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA”.

[12] Y. Yang, W. Yih, and C. Meek, “WikiQA: A Challenge Dataset for Open-Domain Question Answering,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 2013–2018. doi: 10.18653/v1/D15-1237.

[13] P. Bajaj *et al.*, “MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset,” Oct. 31, 2018, *arXiv*: arXiv:1611.09268. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1611.09268

[14] T. Kwiatkowski *et al.*, “Natural Questions: A Benchmark for Question Answering Research,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, Nov. 2019, doi: 10.1162/tacl_a_00276.

[15] E. Choi *et al.*, “QuAC : Question Answering in Context,” Aug. 27, 2018, *arXiv*: arXiv:1808.07036. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1808.07036

[16] S. Reddy, D. Chen, and C. D. Manning, “CoQA: A Conversational Question Answering Challenge,” Mar. 29, 2019, *arXiv*: arXiv:1808.07042. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1808.07042

[17] A. Trischler *et al.*, “NewsQA: A Machine Comprehension Dataset,” Feb. 07, 2017, *arXiv*: arXiv:1611.09830. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1611.09830

[18] Z. Yang *et al.*, “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering,” Sep. 25, 2018, *arXiv*: arXiv:1809.09600. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1809.09600

[19] M. Richardson, C. J. C. Burges, and E. Renshaw, “MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text”.

[20] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, “TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension,” May 13, 2017, *arXiv*: arXiv:1705.03551. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1705.03551

[21] M. Dunn, L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho, “SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine,” Jun. 11, 2017, *arXiv*: arXiv:1704.05179. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1704.05179

[22] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, “RACE: Large-scale ReAding Comprehension Dataset From Examinations,” Dec. 05, 2017, *arXiv*: arXiv:1704.04683. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1704.04683

[23] P. Rajpurkar, R. Jia, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD,” Jun. 11, 2018, *arXiv*: arXiv:1806.03822. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1806.03822

[24] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, “DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs,” Apr. 16, 2019, *arXiv*: arXiv:1903.00161.

Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1903.00161
[25] P. Efimov, A. Chertok, L. Boytsov, and P. Braslavski, "SberQuAD -- Russian Reading Comprehension Dataset: Description and Analysis," vol. 12260, 2020, pp. 3–15. doi: 10.1007/978-3-030-58219-7_1.
[26] S. Lim, M. Kim, J. Lee, and L. Cns, "KorQuAD1.0 Korean QA Dataset for Machine Reading Comprehension".
[27] M. d'Hoffschmidt, W. Belblidia, Q. Heinrich, T. Brendlé, and M. Vidal, "FQuAD: French Question Answering Dataset," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, 2020, pp. 1193–1208. doi: 10.18653/v1/2020.findings-emnlp.107.
[28] W. He *et al.*, "DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications," in *Proceedings of the Workshop on Machine Reading for Question Answering*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 37–46. doi: 10.18653/v1/W18-2605.
[29] C. C. Shao, T. Liu, Y. Lai, Y. Tseng, and S. Tsai, "DRCD: a Chinese Machine Reading Comprehension Dataset".
[30] C. P. Carrino, M. R. Costa-jussà, and J. A. R. Fonollosa, "Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering," Dec. 12, 2019, *arXiv*: arXiv:1912.05200. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1912.05200
[31] H. Mozannar, K. E. Hajal, E. Maamary, and H. Hajj, "Neural Arabic Question Answering," Jun. 12, 2019, *arXiv*: arXiv:1906.05394. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1906.05394
[32] A. Kazemi, J. Mozafari, and M. A. Nematbakhsh, "PersianQuAD: The Native Question Answering Dataset for the Persian Language," *IEEE Access*, vol. 10, pp. 26045–26057, 2022, doi: 10.1109/ACCESS.2022.3157289.
[33] S. Kazi and S. Khoja, "UQuAD1.0: Development of an Urdu Question Answering Training Data for Machine Reading Comprehension".
[34] S. Arif, S. Farid, A. Athar, and A. A. Raza, "UQA: Corpus for Urdu Question Answering," Jul. 22, 2024, *arXiv*: arXiv:2405.01458. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/2405.01458
[35] N. Abadani, J. Mozafari, A. Fatemi, M. Nematbakhsh, and A. Kazemi, "ParSQuAD: Persian Question Answering Dataset based on Machine Translation of SQuAD 2.0," *International Journal of Web Research*, vol. 4, no. 1, Jun. 2021, doi: 10.22133/ijwr.2021.293313.1101.
[36] P. Lewis, B. Oğuz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating Cross-lingual Extractive Question Answering," May 03, 2020, *arXiv*: arXiv:1910.07475. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/1910.07475
[37] D. Gupta, S. Kumari, A. Ekbal, and P. Bhattacharyya, "MMQA: A Multi-domain Multi-lingual Question-Answering Framework for English and Hindi".
[38] J. H. Clark *et al.*, "T Y D I QA: A Benchmark for Information-Seeking Question Answering in *Ty* pologically *Di* verse Languages," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454–470, Dec. 2020, doi: 10.1162/tacl_a_00317.
[39] J. Liu, Y. Lin, Z. Liu, and M. Sun, "XQA: A Cross-lingual Open-domain Question Answering Dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 2358–2368. doi: 10.18653/v1/P19-1227.
[40] A. Asai, J. Kasai, J. H. Clark, K. Lee, E. Choi, and H. Hajishirzi, "XOR QA: Cross-lingual Open-Retrieval Question Answering," Apr. 13, 2021, *arXiv*: arXiv:2010.11856. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/2010.11856
[41] S. Longpre, Y. Lu, and J. Daiber, "MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering," Aug. 16, 2021, *arXiv*: arXiv:2007.15207. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/2007.15207
[42] M. Sabane, O. Litake, and A. Chadha, "Breaking Language Barriers: A Question Answering Dataset for Hindi and Marathi," Feb. 17, 2024, *arXiv*: arXiv:2308.09862. Accessed: Sep. 04, 2024. [Online]. Available: http://arxiv.org/abs/2308.09862
[43] Niveditha S, "IMQuAD." Year. [Online]. Available: https://drive.google.com/file/d/1xvIreuWf3dM8Ao6BvQR5Sew935Aa3h0p/view?usp=sharing
[44] M. A. Appan, *Iyarkai Unavin Athisayam Arockya Vaazhvin Ragasiyam (Tamil)*, 1st ed., vol. 1. Chennai: Popular Publications, 2015.
[45] Dongnguyen, "Attention is all you need summary implementation of transformer model." Accessed: Oct. 08, 2024. [Online]. Available: https://medium.com/@minhanh.dongnguyen/attention-is-all-you-need-summary-implementation-of-transformer-model-b424b03c2728
[46] Khanuja et al., "MuRIL: Multilingual Representations for Indian Languages," *ArXiv, abs/2103.10730*, 2021.

Niveditha S holds a Bachelor of Engineering (B.E.) in Computer Science and Engineering (2007) and a Master of Engineering (M.E.) in Computer Science and Engineering (2009), both from Sri Krishna College of Engineering and Technology, India. With nearly 13 years of teaching experience, she is currently pursuing her Ph.D. in Natural Language Processing at SRM Institute of Science and Technology, which is expected to be completed in 2024.

Presently, she serves as an Assistant Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Vadapalani Campus, where she imparts her extensive knowledge and expertise to students. Her research interests include Natural Language Processing and Computational Linguistics, with a particular focus on sentiment analysis, machine translation, and sign language recognition, among other related areas.

Dr. Paavai Anand G is currently serving as an Assistant Professor (Senior Grade) in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Vadapalani Campus. She obtained her B.E. degree from Periyar University, Salem, and subsequently earned her M.E. and Ph.D. degrees from the College of Engineering, Guindy (CEG), Anna University, Chennai. With over 15 years of teaching experience, she has established expertise in both academia and research.

Her research work has involved developing and applying Machine Learning algorithms, including Probabilistic methods, Bootstrapping, and Genetic Algorithms, to enhance the performance of search engines operating across both the surface web and the deep web. She has published extensively in international conferences and peer-reviewed journals, including contributions to the prestigious ACM Computing Surveys. Some of her conference papers have also earned Best Paper Awards. At present, her research interests include Automated Machine Translation, Information Extraction using Machine Learning, and the classification of web pages.
.