*I.J. Modern Education and Computer Science*, **2025, 5, 1-28**
Published Online on October 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2025.05.01

# Exploring AI Tools and Large Language Models for Students' Performance Enhancement in Riddle Based Logical Reasoning

**Azeddine Benelrhali***
ESMAR, Department of Physics, Faculty of Sciences, Mohammed V University, Rabat, Morocco
Email: azeddine_benelrhali@um5.ac.ma
ORCID ID: https://orcid.org/0000-0002-2769-0051
*Corresponding Author

**Khalid Berrada**
ESMAR, Department of Physics, Faculty of Sciences, Mohammed V University, Rabat, Morocco
Email: k.berrada@um5r.ac.ma
ORCID ID: https://orcid.org/0000-0001-7423-5927

Received: 23 February, 2025; Revised: 08 May, 2025; Accepted: 10 June, 2025; Published: 08 October, 2025

**Abstract:** In the era of Artificial Intelligence (AI), where technology is transforming industries, education stands at a pivotal juncture. With an increasing emphasis on critical thinking and problem-solving, there is a growing need for innovative tools that can foster these essential skills among students. Traditional education methods need help making personalized scalable and interesting experiences for students at this task type which this research aims to solve. The research uses AI and deep learning tools to build an effective framework that enables better riddle solving for students by proposing state of the art deep features including sentence embeddings and ULMfit to be applied as input to deep learning models. In contrast, this study examines different traditional machine learning and deep learning models including ensemble learning models, used as baseline models for comparing the performance of the proposed transformer architectures based on RoBERTa-Large to determine which approach works best, achieving highest accuracy of 96% to effectively handle riddle complexity. The research studies used text data patterns using TF-IDF, Count Vectorization, and word embedding techniques which apply in the form of Roberta. Our research findings help educators, technology experts and scientific teams design educational tools with an easy-to-deploy AI solution.

**Index Terms:** Artificial Intelligence, Natural Language Processing, Questioning Answering, Large Language Model, Sentence Embedding, Education

## 1. Introduction

Education serves as a cornerstone in shaping the lives of students, equipping them with knowledge, skills, and values that form the foundation for their personal and professional growth [1]. It not only prepares individuals for the challenges of the future but also nurtures their cognitive and analytical abilities, essential for solving real-world problems [2]. As the demands of the modern world evolve, there is a growing need for educational systems [3] to adopt innovative strategies that foster critical thinking [4], logical reasoning [5], and problem-solving skills among students [6]. Riddle-based tasks have emerged as an effective pedagogical tool to develop these capabilities, as they challenge learners to think deeply and creatively, fostering a higher order of intellectual engagement [7].

Current teaching methods receive major improvements from Artificial Intelligence technology through advanced learning tools and approaches [8]. NLP technology advancements make it possible for us to develop smart systems that handle human language information accurately [9]. Thanks to Large Language Models that can identify semantic connections which educators use extensively in their work [10]. These models make learning settings that adjust to students' uniqueness and teach them advanced problem-solving abilities and question answering feature better [11]. AI-driven systems deliver important benefits that go beyond traditional teaching spaces [12]. Researchers can put these systems to work studying how students learn to reason and run experiments that simulate their cognitive functions [13]. Students who work in law [14], data science [15] and engineering [16] must develop excellent abilities [17]. When AI

 

tools improve skills they support better decision making across many professional fields [18]. Our riddle-solving system works well in game-based learning setups while teaching students in an exciting format that builds their cognitive skills.

*1.1 Significance and Motivation*

This study goes a notch higher than the practical use of improving students' logical reasoning in the domain of education. Given the fact that AI is rapidly transforming all fields, this research brings out the worthwhile ways AI can transform education. The study demonstrates that AI can be used in the learning process to help deliver pedagogy and a method that can enable students to undertake complex cognitive tasks including solving riddles and critical thinking essential in the future. Besides education, this study could be relevant for the findings across numerous other fields of study. Thus, in cognitive psychology [19], AI models may be applied for the investigation of the processes of developing abilities for reasoning: the experiments can be designed which would mimic various phases of cognitive advancement [20]. Likewise, in the sphere of individualized instruction, AI makes it possible to design learning situations that fit the requirements of every learner and deliver the material with the individual learning pace of each student while engaging change-oriented feedback and recommendation systems, if needed [21].

*1.2 Research Contributions*

This research develops an AI framework to address common shortcomings of traditional education by helping students improve their riddle logic answers by examining many machine learning systems before selecting RoBERTa which became their best-performing large language model. Our research results in both advanced educational AI applications and show researchers and education experts how AI technology can help students improve their thinking and solving skills. The following are research contributions in this area.

- Applied a range of models, from traditional machine learning to advanced deep learning approaches, to evaluate their effectiveness in enhancing students' performance in riddle-based logical reasoning tasks.
- Utilized both textual features (TF-IDF, Count Vectorization) and word embedding features (sentence embeddings) to assess model performance across different data representations.
- Compared the performance of machine learning models (Logistic Regression, Support Vector Machines) and ensemble learning models (Random Forests, Gradient Boosting) to determine the most effective approach for riddle-based.
- Implemented advanced deep learning models, including Gated Recurrent Units (GRU) and Bidirectional GRU (Bi-GRU), to analyze the impact of sentence embeddings on model performance.
- Applied the transformer-based large language model RoBERTa, achieving the highest accuracy of 96%, precision of 93%, recall of 89%, and F1-score of 90%, demonstrating its exceptional capability in solving riddle-based tasks.
- Developed a robust framework for detecting and enhancing students' performance in riddle-based, leveraging the power of large language models like RoBERTa for high-level semantic understanding and contextual analysis.

*1.3 Paper Organization*

The paper is organized as follows, illustrated in fig 1: Section 2 discusses related work which will start with traditional methods of model analysis and go up to the contemporary methods in improving the results of students at solving riddle-based logical problems. Section 3 shares the problem statement and formulation. In the following section, Section 4, the method used in the research is described together with the models and techniques utilized. Section 5


Fig. 1. Paper Organization

relates to the experimental framework where the architecture of the models and the system infrastructure are explained. Section 6 gives an overview of the dataset and explores its properties to get a better feeling for the problem at hand. In Section 7, the authors present the outcome, and the conclusion derived from the models in the study, which Section 8 sums up the paper and additional recommendations for development in the subject.

## 2. Related Work

Transformer-based AI tools and Large Language Models (LLMs) are offering inventive approaches for improving students' logical thinking skills besides problem-solving skill sets. In text analytical capabilities, ability to produce intelligent responses and flexibility to suit educational requirements, the models come out strongly, analysis of existing studies shown in table 1. In the context of riddle-based systemized logical thinking, prior research provides information about utilizing transformer-based learning approaches to resolve education-related issues and enhance learner outcomes. fig 2 defines the taxonomy of defines related work section to explore existing literature in depth.

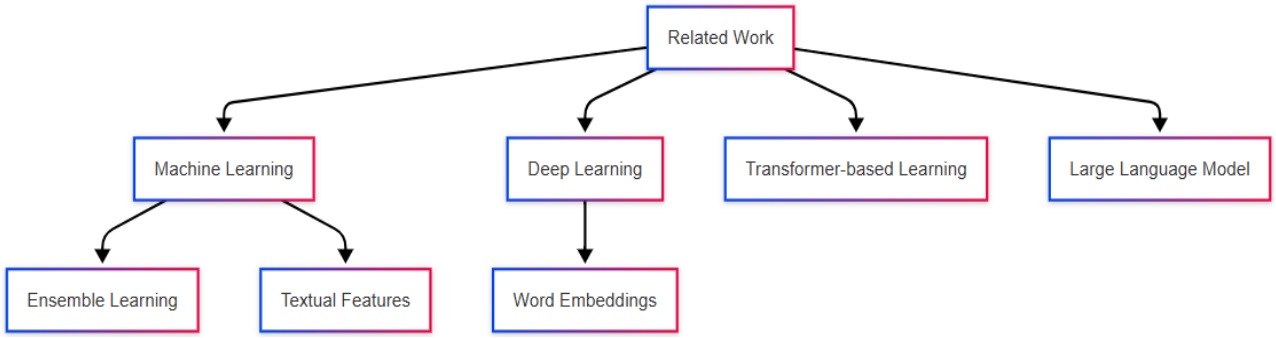


Fig. 2. Taxonomy of Existing Studies

### *2.1 Traditional Machine Learning*

Modern AI tools and large language models have extended the possible support to students' performance enhancement in riddle-based as these tools provide intelligent solutions, individual learning paths, and real-time feedback. AI tools and machine learning ensemble models have transformed the way students solve riddle problems and other logical problem-solving rituals [22]. These help students develop their critical thinking skills and perform well educationally. A Light GBM algorithm-based performance prediction model and key online learning behaviour were recognized to have 85% prediction accuracy. However, because the analysis was based on a single dataset it restricted the generalization of the results to many learning environment categories [23]. The results of the empirical test using the AdaBoost algorithm suggested self-and peer education as the most effective way of enhancing college students' savings behaviour. However, there were some limitations to the study thus, the study sample did not include the validity of the correlation across other demographics and failed to consider a broader range of socioeconomic factors [24]. The results of the selected Ensemble learning techniques including XGBoost, Random Forest and AdaBoost proved significant improvement in predicting student performance when compared with previous studies, XGBoost outperformed other models on three datasets. However, the study did not consider the prognostic ability and the generalizability in different educational contexts [25]. All the tested predictive models with AdaBoost and genetic algorithms gave an excellent approximation of the nanocomposite material adsorption ability. Smaller data sets limited the models' generalization ability, and the study did not compare the performance across different adsorption scenarios [26]. In supervised learning methodology, SVM among others yielded high accuracy in the prediction of the student's cognitive engagement from facial behaviors. However, it did not throw light on how it can be effective in a massive population and how it can generate real-time interventional strategies based on the identification of engagement [27]. AI and ensemble-based tools can be extremely useful as tools of great potential for changing or improving students' performance in logical thinking, puzzles and problems such as riddles and puzzles. Nevertheless, there are still some issues, including scale-up, cross-study, and real-time issues, which must be resolved to realize the full potential of these techniques in varied educational contexts.

### *2.2 Deep Learning and Transformer-based Techniques*

Many works have been surveyed to describe multi-modal fusion techniques in recognition of riddle-based task. Thus, despite the presented ideas being important, the lack of experimental evidence and a unifying assessment methodology does not allow for operationalizing the presented review. Additionally, the absence of real-world applicability testing presented a problem when determining the practical application of the discussed methods [28]. To this end, a novel fusion architecture was put forth to evaluate the UP-Fall dataset regarding the MAR requirement for empirical validation. Despite the positive findings of the study, the authors failed to test how the architecture framework applies to different populations, datasets and settings which limited the architecture's usefulness in more real-world

scenarios [29]. On the multiple benchmark datasets, the identification of human activity applied a developed multi-branch LSTM-CNN model. Nevertheless, features like real-time and dynamic performances and sensor variability were not considered thus the possibility of the given model being practically implemented [30]. Another proposed study developed a real-time human activity recognition cloud-based system utilizing wearable sensors with an SVM model to classify the activity. While this approach demonstrated potential, it became clear that the system needed to be tested in terms of scalability for various sensor arrangements and in different conditions, which proved to be its flaws when applied to real-life situations [31]. Another work that adequately dealt with the adaptability issues in human activity recognition was done by employing deep CNNs. However, two limitations were identified with this approach: doing away with evaluation in real-time situations reduced the practicability of the proposed model in live applications [32].

Table 1. Summary of Existing Studies

| Sr. No. | Ref | Year | Model | Features | Dataset | Results (%) |
|---|---|---|---|---|---|---|
| 1 | [39] | 2019 | Deep ANN | Student performance | Surveys | 93 |
| 2 | [30] | 2020 | SVM | wearable sensors | Wearable system | 89 |
| 3 | [24] | 2020 | XGBoost, RF | Student performance | Surveys, Quizzes | 92 |
| 4 | [33] | 2020 | SQuAD | Distribution shift analysis | Online websites | 82 |
| 5 | [26] | 2021 | Light GBM | Online behaviors | Learning platform data | 85 |
| 6 | [53] | 2022 | RoBERTa | Logical Solving Feature | Open World Assumption | 84 |
| 7 | [34] | 2022 | GPT-3, Unified-QA | Multimodal reasoning, | ScienceQA | 40 |
| 8 | [25] | 2022 | AdaBoost | Student performance | Experimental data | 68 |
| 9 | [46] | 2022 | RadBERT | Student performance | ScienceQA | 91 |
| 10 | [58] | 2023 | Light GBM | Online behaviors | Learning platform data | 89 |
| 11 | [23] | 2023 | AdaBoost | peer and self-education | Educational data | 91 |
| 12 | [47] | 2023 | Logic-LM | Symbolic Solvers Integration | PrOntoQA | 85 |
| 13 | [48] | 2023 | LOGIPT LLM | Logical Parsing and Inference | ProofWriter | 89 |
| 14 | [49] | 2023 | Bert | Graph-based Feature | 4nums.com | 89 |
| 15 | [52] | 2023 | DSR-LM | Symbolic Reasoning Feature | CLUTRR, DBpedia-INF | 86 |
| 16 | [57] | 2023 | Alpaca-7B, Vicuna-7B | Task Structure Variations | LogiQA | 83 |
| 17 | [27] | 2023 | LSTM-DNN | Multi-modal fusion features | Students Riddle Question | 75 |
| 18 | [28] | 2023 | CNN | RGB visual data, IMU features | UP-Fall | 92 |
| 19 | [31] | 2023 | Deep CNN | spatial and temporal features | AMASS | 81 |
| 20 | [50] | 2024 | Llama3-8B model | Default | K&K Dataset | 78 |
| 21 | [51] | 2024 | Logic Asker | Logical Reasoning Feature | LogiQA | 92 |
| 22 | [56] | 2024 | logic Bench | Systematic Logical Reasoning | Logic Bench Dataset | 90 |
| 23 | [52] | 2024 | ELECTRA | Logic-Injected Contexts | LogiQA, FOLIO | 88 |
| 24 | [29] | 2024 | LSTM-CNN | Word Embeddings | WISDM, PAMAP2 | 94 |

Recurrent techniques and paradigms of Question Answering have been reviewed with a focus on issues of precision and recall. However, studies have generally failed to consider issues of deployment, field deployment, and real-world adaptation and issues of scale. Response time and other similar measurements that are centered around the user are not broadly analyzed, and consequently, many of the proposals and solutions that result from these techniques are not easily deployable in actual use cases [33]. The study of cross-domain generalization of SQuAD models has probably achieved some level of success, as it has been identified that SQuAD models experience a substantial decline in performance on data other than Wikipedia. Although these works have contributed to this problem, a very detailed examination of the linguistic factors underlying such robust challenges is not extensively explored by most of them. Moreover, ideas about more precise mechanisms that would solve these problems have not been presented in a sufficient manner, which creates a shortage in building stronger models of QA [34]. ScienceQA is one of the multimodal datasets that have brought new challenges in addressing multiple-choice questions on science. Cognition growth obtained by chains of thought (CoT) reasoning indicates that such strategies are possible. However, the biggest challenges for now are the biases observed in dataset annotations and the scalability of the annotation processes. Nevertheless, there is still much to be done to extend the applicability of CoT reasoning to real-life domains that are not necessarily scientific, which has not been explored in detail [35]. When considering multi-turn conversational QA, the recent advancements in conversational question-answering (CQA) systems have enabled explicit improvements concerning context understanding and co-reference. However, there are still open issues in the integration of

multimodal data sources for human behaviour analysis in the real world [36]. CQA systems have been comprehensively reviewed both for task-oriented and non-task-oriented applications providing insightful details of the structures and features of CQA systems. However, there are two of the biggest challenges: identification of the circulant structures and the integration of multimodal data [37]. In addition, there is a dearth of information on ethical concerns in deploying such systems to actual physical settings [38]. Generalization techniques for VQA have also been shown, namely, splitting data into multiple training scenarios to remove virtual correlations. All these methods improve out-of-distribution generalization but incur substantial computational costs. Cross-Data and Cross-Task generality is still an issue, proving that more improvements are needed to create effective strategies for VQA [39].

Artificial neural networks have also been used with a deep structure to identify at-risk students using clickstream data with an accuracy of 84 % – 93%. However, this success was reduced by the absence of identifying long-term trends in students' performances; the approach did not assess the interpretability of its models, which confines its applicability utility [40]. From meta-analyses and systematic reviews of studies on the efficiency of AI in enhancing student success, we disclosed that using AI indeed leads to the improvement of pass rates, students' satisfaction, and learning motivation, particularly within STEM disciplines. However, these studies may miss the big picture when it comes to use not confined to STEM and may not speak of bias and accessibility considerations of AI [41]. There were so many applications in areas such as academic writing and programming with the new release of models like ChatGPT, but these came with questions of ethics, efficiency and how they considered biases. To that end, outside of having some promising ideas that were still not very specific and without having the benefit of an actual blueprint for a framework to enact them or ways to reduce misinformation, their potential could still only be speculated on in various contexts of education [42].

Proactively, utilizing Enhanced Transformer architectures in the context of NLP such as BERT provided higher performance in identifying AI-written text within learning environments. Although these results were encouraging there are still drawbacks to the method that limits its general use, namely the reliance on specific data sets and the fact that all models have an inherent bias [43]. Researchers found that the proposed prompt-based learning with transformer models performs relatively at par with fine-tuned models and human evaluators for domain-specific text classification tasks. However, the issue of extending it to different types of datasets and its stability when applied to actual problems were not discussed in detail, therefore the practical applicability of the work remained inconclusive [44]. Transformer-based methods were first applied in a Bangla language question-answering system due to the lack of research and data resources. The machine learning models trained on the custom dataset were found accurate, compared to the existing approaches but the approach was not scalable to other languages or domains due to the restriction of data set size [45]. The explainable NLP models effectively evaluated the energy concept understanding of the students in their responses together with the idea of evaluating the presented predefined energy ideas, as a correct or incorrect judgement. However, the importance and benefits of these models do not lie in their ability to generate historical but relevant results on a diverse topic of interest to students and their teachers and their scalability to real-time classroom applications was not tested [46]. Thus, for radiology-related tasks such as abnormality classification and report summarization, a specifically developed transformer-based model named RadBERT displayed good performance. Still, it was demonstrated to be superior to baseline models, the absence of benchmarks across various datasets and the low adaptation of the technique into clinical settings raised concerns for further advancement [47].

### *2.3 Large Language Models*

The most recent work was the proposed method called Logic-LM which integrated symbolic solvers and large language models to boost. While it offered a clear modulation between symbolic and neural computation, it had the major drawback of being weak for extremely large data and for numerous and rather different tasks [48]. These language models could solve logical exercises by transforming them into symbolic statements. However, it relied on synthetic datasets which did not generalize well to real-world intricate logical applications [49]. The authors presented the graph of thought based on exams framework to enhance multi-step by breaking down intermediate steps of the logic problem into a graph to complete a complex multi-faceted and woven problem. However, the formulated approach did not perform well when it was given highly abstract or ill-defined logical problems that needed additional improvements [50].

This work pointed at the possible weakness of large language models in memorization manifested in their reasoning abilities. One was a dichotomy of memory versus generalization that the author has been discussing without giving pointers on how to get over the hurdles of memorization [51]. Logic Asker evaluated and enhanced reasoning with the help of a sequence of questions presented. Although the paper had helped shed light on the model's underground, it failed to examine under what circumstances the model cannot work, especially in dynamic and multi-modal logical space [52]. DSR-LM combined the differentiable symbolic reasoning into LM to improve the deduction language models. Some validation of the suggestion was obtained in the model, but no evaluation of open-ended, ill-defined reasoning problems was possible [53]. Compared to other approaches, this approach improved by endowing meaning groundings with logical structures in model contexts [54]. But it added too much overhead thus less viable when used in most practical everyday applications [55]. The work assessed the logical properties of language models within a variety of benchmark tests but failed to provide a clear roadmap to remediate their deficits in logical inferences that involve less well-defined contexts [56]. This survey aimed towards understanding the approaches for solving puzzles in large language models while keeping a bird's eye view [57]. Its main shortcoming was that the work

provided no specific method of countering the established limitation in LLM reasoning [58]. While a logic benchmark program named Logic-Bench was available to measure the levels they did not report specific strategies that can help improve the performance with all sorts of logic patterns [59]. The paper focused on the ability of the language models tested under different task complexity. While it has provided solutions for such issues, it has not attempted to solve the issues involving real-time adaptive reasoning [60].

*2.4 Limitation of Existing Studies*

Previous studies on applying AI and machine learning generally in education area (for instance, adaptive learning system, automatic essay scoring, sentiment-based performance tracking) focus on low grade cognitive skills, such as logic and abstract problem solving. Most existing approaches focus on testing the capabilities of a person to recall the integrated content or predict their behavior as opposed to fostering deep analytical thinking. In addition, tasks that come closer to the real-world mental challenges like solving riddles involve multi layered reasoning, semantical inference and contextual understanding are considered but only based on limited attention so far. These holes make it clear that any AI model that the students are encouraged to try thinking through riddles should be able engaged with the linguistic intricacies and the inferential depth required to support the cognitive development of these students. To address this void, our study proposes a comprehensive framework that bridges the gap between traditional machine learning, deep learning, as well as the current state of the art transformer-based models, most prominently RoBERTa, for the purpose of enhancing the performances of students in riddle-based logical reasoning tasks. Through linguistic embeddings, cosine similarity based semantic alignment, also comparative model analysis we propose an AI based educational tool that tries to stimulate and evaluate higher order thinking, an area not sufficiently studied in the context of AI based educational research.

## 3. Problem Statement

The task at hand is to develop an effective model that can predict the correct answer $A_i$ from a given riddle question $Q_i$ and its corresponding hint $H_i$, where $i = 1,2, \dots, N$, and $N$ is the total number of riddle entries $\geq 50000$. The goal is to learn a mapping function $f: \varrho * H \rightarrow A$ where $\varrho, H, and\ A$ represent the sets of all possible riddle questions, hints and answers, respectively. Given a tuple $(Q_i, H_i)$, the model seeks to optimize the conditional probability of $P(A_i | Q_i, H_i)$, with the objective of maximizing the likelihood function of $\mathcal{L}(\theta) = \prod_{i=1}^{N} P(A_i | Q_i, H_i; \theta)$ where $\theta$ represents the parameters of the model. The problem is further complicated by the fact that $Q_i\ and\ H_i$ are sequence of words, which may contain ambiguity, polysemy, and implicit context, and are subject to various linguistic phenomena such as wordplay, metaphors, and indirect hints. This enhances students' performance in critical thinking and problem-solving skills. The ability of students to engage with riddles and derive answers is a key indicator of their cognitive abilities in reasoning and interpretation. Therefore, the model must efficiently capture the semantic and syntactic dependencies within $Q_i\ and\ H_i$, while addressing the underlying cognitive processes involved in solving riddles. Furthermore, this extends to understanding both the direct and indirect relationship between $Q_i\ and\ H_i$ and mapping them to $A_i$ overcoming the challenges of contextual variability and language ambiguity, helping enhance students problem-solving strategies and critical thinking, which are essential for academic and real-world decision-making.

## 4. Research Proposed Methodology

The proposed research design follows a systematic method for image analysis of interactive media which includes preprocessing operations and engineering new features as well as model building steps, as defined in fig 3. The initial raw data passes through a comprehensive preprocessing action which applies noise reduction alongside normalization as well as data augmentation techniques to maintain both data quality and consistency. The subsequent application of feature engineering allows the model to improve category differentiation through image attribute extraction from YouTube thumbnails. The center of this research includes building and assessing proposed models that will undergo testing versus existing sample collection systems. An assessment of the proposed model's performance through comparison-based analysis evaluates its accuracy and precision along with recall features and relevant metrics which justify its effectiveness in interactive media image categorization.

*4.1 Data Preprocessing*

The preprocessing pipelines begin with the removal of stop words, special characters, and punctuations from the raw text corpus, represented as $C = \{T_1, T_2, \dots, T_n\}$, where each $T_i$ denotes a text document. This step results in a cleaned corpus $C_{clean}$, ensuring irrelevant tokens are eliminated. Subsequently, digits and URLs are excluded from the text, producing $C_{clean} \rightarrow \{digits, URLs\}$. Next, all textual data is converted to lowercase, ensuring uniformity across the corpus, and tokenization is applied to split each text document into individual token $\{w_1, w_2, \dots, w_m\}$, where $w_j$ represents a single word. Part of speech (POS) tagging is then performed, assigning syntactic roles $p_j$ to each token $w_j$, yielding token-tag pairs $\{(w_1,\ p_1), (w_2,\ p_2), \dots, (w_n,\ p_n)\}$. Finally, lemmatization is applied, using function $\ell(w, p)$

that transforms each token $w_j$ into its base based on $p_j$. This results in a fully lemmatized corpus $C_{processed}$, where each document is reduced to its most linguistically meaningful representation [61]. These steps collectively ensure the text is standardized, cleaned, and structed for effective feature extraction and subsequent modeling.

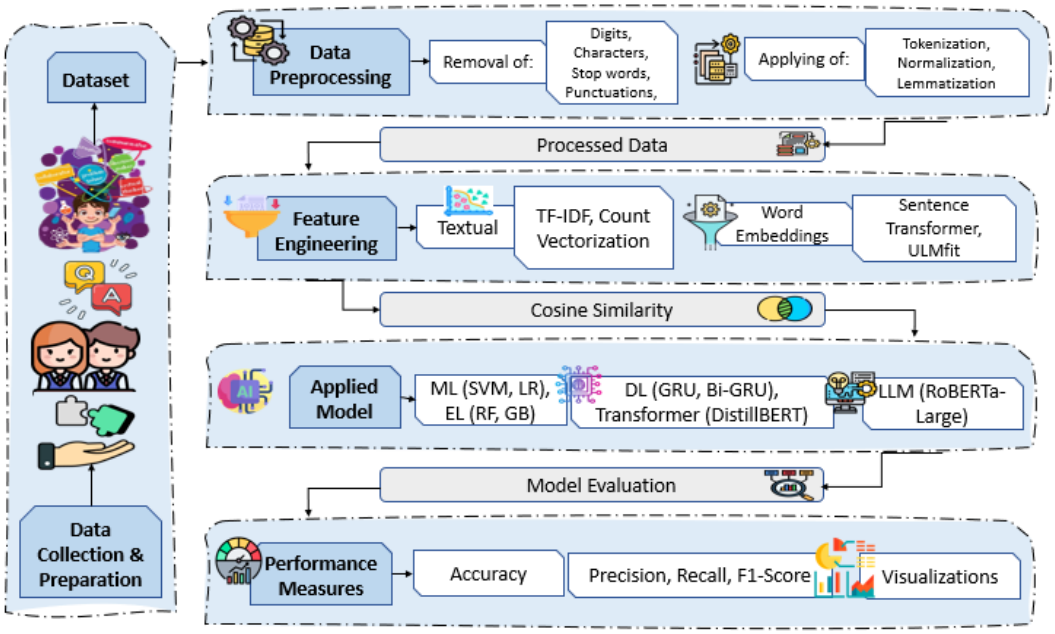


Fig. 3. Steps of Applied Research Methodology

### *4.2 Feature Engineering*

Feature extraction is always an important step in NLP tasks where the textual information is converted into numerical measures understandable to the machine learning and deep learning algorithms [61]. In this study, feature engineering is divided into three categories: identification of textual features, word embeddings, and Pre-trained embeddings; all provided different representation to the data which boosted model efficiency.

### *4.3 Textual Features*

The most basic approach to transform text into numbers is by using textual features including the count vectorizer as well as the TF-IDF vectorizer. The Term-Frequency Inverse Document Frequency (TF-IDF) value for a term $t$ in a document $d$ from corpus $D$, calculated as in equation (1).

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D) \tag{1}$$

where $TF(t, d) = \frac{Frequency\ of\ t\ in\ d}{Total\ terms\ in\ d}$ is the term frequency, and $IDF(t, D) = \log\left(\frac{|D|}{1+|\{d \epsilon D: t \epsilon d\}|}\right)$ is the inverse document frequency. Count vectorization, on the other hand, creates a matrix $X \epsilon \mathbb{R}^{m*n}$, where $m$ is the number of document and $n$ is the size of vocabulary, with $X_{ij}$ representing the count of term $j$ in document $i$ define as in equation (2).

$$X_{ij} = \sum_{k=1}^{N_i} \delta\left(t_k^i, w_j\right) \rightarrow \begin{cases} 1, & if\ t_k^i = w_j \\ 0, & otherwise \end{cases} \tag{2}$$

Where $N_i$ is the total number of terms in document $i$, $t_k^i$ is the $k-th$ term, $w_j$ is the $j-th$ term in vocabulary, $\delta$ delta function.

### *4.4 Word Embeddings*

Further, word embeddings define semantic similarities and differences between words in terms of their positional representation in high-dimensional spaces [62].

Sentence embeddings stand for similar ideas that map the whole sentence into vectors. The formulation utilizes both weighted averaging and max-pooling and involves explicit attention to acquisition of global and local contextual features, working defined in fig 4. By incorporating positional encoding together with learned weights, it can capture syntax and the weight of words respectively. Given a sentence $S = \{w_1, w_2, \ldots, w_n\}$, where $w_i$ represents the $i-th$ word, the embedding $e_s$ is derived by aggregating the contextual embeddings of all words in $S$, computed as in equation (3).

$$e_s = \frac{1}{n}\prod_{i=1}^{n} \alpha_i . h\left(e_{w_i} + p_i\right) + \beta . \max(W . [h(e_{w_1}), h(e_{w_2})]) + \gamma . Attention\ (\{e_{w_i}\}_{i=1}^{n} \quad (3)$$

Where $e_{w_i}$ is the word embedding $w_i$, $p_i$ is the positional encoding for $w_i$to incorporate word order information, $h$ is a non-linear activation function, $\alpha_i$ is the learned weigth for each word $w_i$, emphasizing its importance in the sentence. $\beta$ *and* $\gamma$ are the scalar weights balancing the contributions of max-pooling operation and attention mechanism respectively. $W$ is a learned projection matrix used for max-pooling operation across word embeddings. $Attention\ (\{e_{w_i}\}_{i=1}^{n}$ A self-attention mechanism applies to compute the importance of each word in the context of the entire sentence.

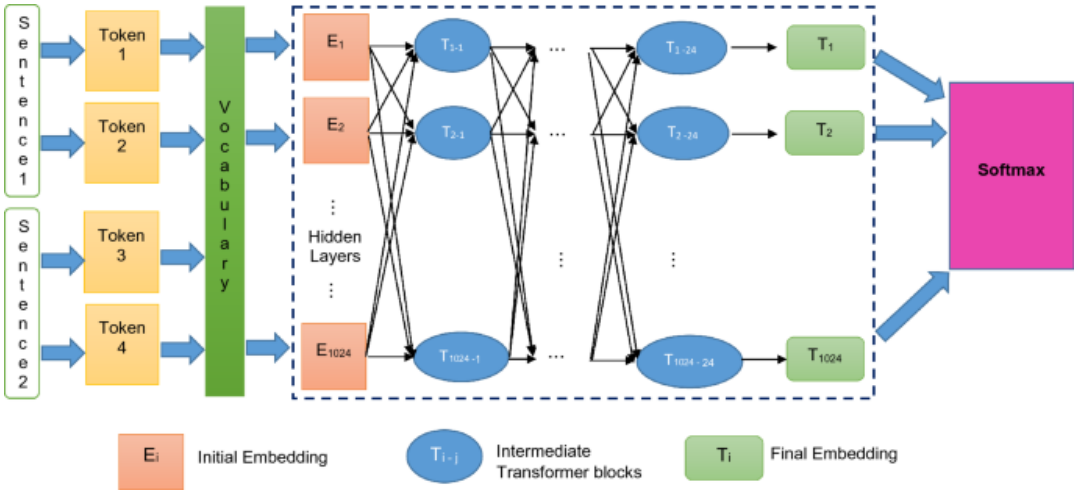


Fig. 4. Working of Sentence Embeddings [63]

Further, this attention mechanism helps to consider semantically relevant words for improving the representation that is, in fact, enhanced and more robust, covering the meaning of the given sentence in a more distant manner.

*4.5 Pretrained Embedding*

Using several models that have been trained in huge referential text, embeddings borrowing from pre-learned ensembles of representations that are contextualized. Universal Language Model-Fine-tuning (ULMFiT) is one of the models that implement a strong transfer learning paradigm, training a language model on a large extent corpus [64]. Thus, the embeddings produced such as ULMFiT are context-based and adaptively modified according to the domain textual data, display in fig 5. To this end, this formulation includes some components including positional encodings, attention mechanisms, and the residual connections to obtain contextualized, as well as domain-specific embeddings, expressed as in equation (4).

$$e_s = \varphi(decoder\ (\coprod_{t=1}^{T} \alpha_t . h_t + \beta . Attention(H) + \gamma . Residual(H, W_o))) \quad (4)$$

Where $T$ represents the sequence length, $h_t = g(w_c . x_t + b_c)$ is the hidden state at time step $t$, computed by applying a non-linear activation function $g$ to the embedded word $x_t$ with $w_c$ and $b_c$ as learnable parameters. $H = \{h_1, h_2, \ldots., h_T\}$ is the sequence of hidden states. $Residual(H, W_o)$ combined with layer normalization $H + LayerNorm(H . W_o + b_o)$ to improve gradient flow and stability. $\varphi$ is a task-specific output layer mapping the processed embedding to a lower-dimensional vector space for classification and feature extraction tasks.

The architecture of ULMFiT allows fine-tuning by an approach that involves the freezing of select layers, gradual thawing of frozen layers, and different discriminative learning rates for each of the layers to achieve the best transfer of learned features from the source to target domain dataset. The output embeddings not only incorporate syntactic and semantic mapping but also dynamically tuned to problem domain characteristics and highly improves the overall performance of the downstream tasks.

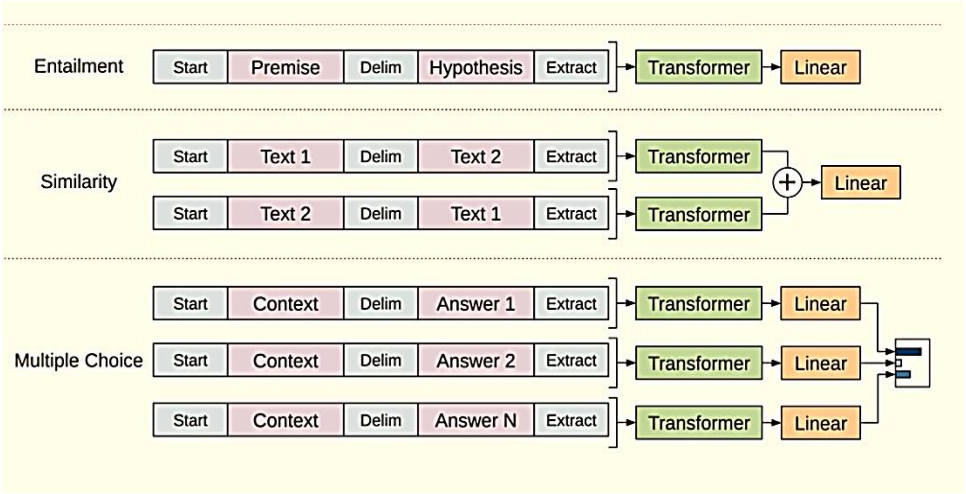


Fig. 5. Working of ULMfit Embedding Model

*4.6 Cosine-Similarity*

The cosine similarity metric used in this study to assess the similarity between riddle questions, answers, and clues can be formalized mathematically [65]. Given two vectors $A\ and\ B$ representing the text embeddings of the question-answer pair and the clue, respectively, the cosine similarity $Cosine_{sim(A,B)}$ is defined as in equation (5).

$$Cosine_{sim(A,B)} = \frac{\sum_{i=1}^{n} A_i B_i + \sum_{j=1}^{n} C_j D_j}{\sqrt{(\sum_{i=1}^{n} A_i^2 + \varepsilon)(\sum_{i=1}^{n} B_i^2 + \varepsilon)} + \delta} \tag{5}$$

Where $A_i and\ B_i$ are the component of the vectors $A$ and $B$ respectively. $C_j\ and\ D_j$ represent additional contextual embedding components, such as those from the attention mechanisms for data integration. $\varepsilon$ is a small regularization parameter added to avoid division by zero or small values when vector are near-zero in magnitude. $\delta$ is an adjustment term that accounts for small shifts in cosine similarity due to slight mismatches or noise in embedding vectors.

*4.7 Proposed Model*

RoBERTa (Robustly Optimized BERT Approach) is a Large Language Model (LLM) that extends the original BERT by optimizing its pretraining procedures. Unlike BERT, which uses Next Sentence Prediction (NSP), RoBERTa removes this task and relies solely on Masked Language Modeling (MLM) for pretraining. RoBERTa also utilizes a dynamic mask strategy and is trained on a significantly larger dataset, enabling it to capture more accurate language patterns and complex semantic relationships. As a Large Language Model (LLM), RoBERTa processes text sequences by leveraging the Transformer architecture's self-attention mechanism, which allows it to encode long-range dependencies between tokens in both directions [66]. Given an input sequence $X = [x_1, x_2, \ldots, x_n]$, RoBERTa embeds each token xi into a high-dimensional space $e_I \in \mathbb{R}^d$ where $d$ represents the embedding dimension. These embeddings are then passed through multiple layers of self-attention, enabling the model to generate contextualized representations for each token based on its surrounding tokens.

In each self-attention layer, the model calculates the Query $(Q)$, Value (V), and Key (K) vectors for each token $x_i$ using learnable weight matrices, using equation (6).

$$Q_i = W_Q e_i + b_\alpha, \quad K_i = W_K e_i + b_K, \quad V_i = W_V e_i + b_V \tag{6}$$

The self-attention mechanism calculates using equation (7) with attention scores between each token and every other token using the dot product between queries and key.

$$\alpha_{ij} = \frac{Q_i . K_j^T}{\sqrt{d_k}} \tag{7}$$

The softmax function is then applied to these scores to normalize them, generating attention weights, using equation (8).

$$\alpha_{ij} = \frac{\exp(Q_i . K_j^T)}{\sum_{k=1}^{n} \frac{Q_i . K_j^T}{\sqrt{d_k}}} \tag{8}$$

These attention weights are used to compute the context vector for each token as a weighted sum of the values, computed using equation (9).

$$Z_i = \sum_{j=1}^{n} \alpha_{ij} V_j \tag{9}$$

RoBERTa's architecture includes multiple layers of self-attention, each followed by a Feedforward Neural Network (FFN) and a residual connection, defined in equation (10).

$$H_i = \text{LayerNorm}(Z_i + e_i) \tag{10}$$

The FFN consists of two linear transformations with a ReLU activation in between, defined as in equation (11).

$$\text{FFN}(Z_i) = \text{ReLU}(W_1 Z_i + b_1) W_2 + b_2 \tag{11}$$

RoBERTa uses Masked Language Modeling during pretraining, where a portion of tokens in the input sequence is randomly masked. The model learns to predict the masked tokens based on the surrounding context. The probability of predicting a masked token $z_m$ is given by equation (12).

$$P(z_m | X_{\text{context}}) = \frac{\exp(W_{\text{output}} \cdot h_m)}{\sum_{k=1}^{n} \exp(W_{\text{output}} \cdot h_k)} \tag{12}$$

where $h_m$is the representation of the masked token, and $W$ output is the output projection matrix. The training objective is to maximize the likelihood of predicting the correct masked tokens. RoBERTa-large uses a deep stack of transformer layers, enabling it to model complex relationships within the input text, defined in equation (13). The model is fine-tuned for specific tasks by adding a classification layer on top of the final representations.

$$\hat{y} = \text{softmax}(W_{\text{task}} \cdot H_i) \tag{13}$$

where $W$ task is a learnable matrix and C is the number of output classes.

RoBERTa, as LLM, can handle a wide range of NLP task, such as question answering, text classification, and natural language inferences, by learning deep contextual representations. Its strength lies in its ability to model long-range dependencies, capture the patterns of language and adapt to various downstream tasks through fine-tuning, making it an excellent model for riddle-based questions prediction, logical reasoning, and critical thinking enhancement in educational applications.

## 5. Experimental Setup

We built the experimental setup using a customized dataset designed to assess abilities through riddles while including various riddles and their matching logical answers. Data preprocessing through tokenization resulted in preparation which involved stemming and removal of stop-words to provide proper input for model deployment. Component extraction for analog and ensemble learning networks utilized TF-IDF along with Count Vectorization techniques but deep learning and transformers needed sentence embeddings and transformer-based embeddings. Researchers utilized evaluation features including precision, recall, accuracy and F1-score to determine model effectiveness levels. The experimental pipeline executed a rigorous assessment of each model through testing which evaluated their performance with the linguistic and cognitive challenges present in the dataset.

### *5.1 Dataset*

The dataset used for this study is derived from the "Riddles: In this paper, we use a "Synthetic Riddle Dataset for NLP" which is available on Kaggle which includes a set of riddles, solutions, and clues. This dataset is specifically created for the purpose of investigating the suitability of NLP tasks for solving riddles. Fig 6 defines the distribution of most frequent riddles data. Firstly, the data set comprises of a series of riddles that followed by a question, answer, and a hint.

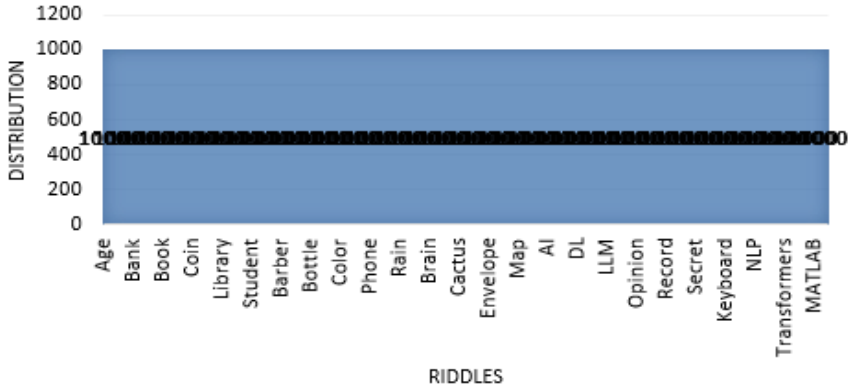


Fig. 6. Analysis of Data Distribution

It can be used as a base for developing and evaluating models that apply reasoning and language understanding, as this paper's model does. To improve the dataset quality and expand and diversify the students' base for training machine learning algorithms, we add new entries to the dataset totaling 50000 entries at the end of collection with 50 combinations of riddles related to real world examples and educational material, with equal distribution of each 1000 entries. This part is to let the student construct the entries in real life, without putting them in a standardized predefined format, so that they reflect the real-life variations as they decode riddles and puzzles to these questions in everyday life. The new large dataset includes more variants of riddle questions, hints, and answers, which makes the general data more diverse and contributes to improving the generalization of the model.

*5.2 Performance Evaluation Measures*

To evaluate the performance of the presented models for riddle-solving tasks, we utilize several globally accepted performance metrics. These metrics give an understanding of how well the proposed models work for sorting out riddle answers and how those models performed given the complications of the data set. The assessment metrics implemented in the work include Accuracy, Precision, Recall, and F1-Score. Accuracy is one of the simplest and most common performance measures, displayed in table 2. It is defined as the ratio of the successfully classified sample to the total numbers of samples in the dataset. Precision also referred to as the Positive Predictive Value does the same work with positive predictions. It measures the ability of glioma prediction by determining the ratio of true positive to all the positive instances identified. It is particularly important where the cost of false positives is high. Sensitivity or True Positives or Recall calculates the ratio whereby all the positive samples of the actual data set are discovered by the model. The accurateness of detecting those cases that are truly positive among all actual positive cases. This metric is quite used when the cost of error type II is relatively high relative to error type I by missing a positive case might lead to some severe implications. F1-Score is the average of Real and Precisions which gives a balance between them making their use ideal. It is especially beneficial when working with the datasets containing significantly more of one type of data than the other [67].

Table 2. Performance Evaluation Measures

| Measures | Equations |
|---|---|
| Accuracy | $\frac{TP+TN}{TF+FN+FP+TP}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Recall | $\frac{TP}{TP+FN}$ |
| F1-Score | $\frac{2(Precision*Recall)}{Precision+Recall}$ |

In this study, we apply the above performance evaluation measures to make sure that the models not only generate correct, accurate forecasts, but also balance the correct rate of positive identification and high recall rate. Therefore, by using these measures, we are better placed to assess the performance of the proposed models and determine which out of the two is most appropriate for riddle solving.

*5.3 Comparison-based Models*

Our research investigated multiple modeling frameworks by evaluating traditional machine learning (ML) approaches, ensemble learning (EL) reasoning, deep learning architectures, and transformer-based language models. The traditional models Logistic Regression (LR) and Support Vector Machines (SVM) delivered basic knowledge, yet their restricted capabilities limited their ability to analyze intricate textual patterns. Multiple weak learner ensemble models including Random Forests (RF) and Gradient Boosting (GB) utilized several basic classifiers to perform better

while detecting advanced data relationships during the analysis process. The deep learning models GRU and Bi-GRU successfully handled sequential linkages along with contextual patterns in text input but failed to match performances measured against transformer-based models. RoBERTa-based transformer models delivered peak performance on riddle-based tasks through their pre-training on large datasets in addition to their attention mechanisms. The progressive analysis of modern NLP methods demonstrates their enhanced capabilities in education applications.

*5.4 Machine Learning Models*

Machine learning models have served text processing tasks for years because they work very well with organized data. This research tested multiple machine learning tools to help improve how students solve riddles. Logistic Regression showed strong effectiveness for binary and multi-class tasks by giving us easy-to-read results at a feasible speed. Support Vector Machines (SVM) were chosen to process high-dimensional data through kernel tricks for optimal decision boundary finding, by equation (14).

$$min_{w,b,\xi}\frac{1}{2}\left|\left|w\right|\right|^2 + F\sum_{i=1}^{m}\xi_i \tag{14}$$

Subject of constraint (15).

$$y_i(w.\phi(x_i) + b \geq 1 - \xi_i,\ \forall i = 1,2,\ldots..,m \tag{15}$$

The study tested Ensemble techniques Random Forest and Gradient Boosting alongside the other models. Random Forest datasets strength comes from using many decision trees to give both robust results and deeper understanding of hidden patterns, computed using equation (16).

$$Y_{final} = g(\prod_{j=1}^{J} w_j h_j(x)) \tag{16}$$

Gradient Boosting helps by developing successive models that make smaller errors compared to earlier versions. The last decision is then performed by averaging for these models, that minimize variances to prevent overfitting, defined as in (17).

$$Var(Y_{final}) = \frac{1}{J^2}\prod_{j=1}^{J} Var\left(h_j(x)\right) + (J-1)Cov\left(h_j(x), h_k(x)\right) \tag{17}$$

After examining basic ML systems, we found they exposed the benefits and restrictions each had when working with text features [68].

*5.5 Deep Learning Models*

The ability of deep learning models to understand text order and context makes them effective for solving riddles which display complex word patterns. This research used Gated Recurrent Units (GRU) and Bidirectional GRU (Bi-GRU) models to assess sentence embedding effects on model performance. GRU demonstrates effective performance at processing sequences through memory banks without extreme algorithmic complexity. Bi-GRU improves text processing because it reads content forward and backward to bring in contextual details about previous and upcoming text which riddle comprehension requires. For an input sequence $\{x_t\}$, the forward and backward hidden states, $h_t^{\rightarrow}$ and $h_t^{\leftarrow}$, are compted separately as in GRU. The final output $o_t$ is obtained by concatenating the two, as in (18).

$$o_t = \{h_t^{\rightarrow}, h_t^{\leftarrow}\} \tag{18}$$

The research analyzed DistilBERT as a transformer-based model that combines semantic comprehension with exceptional efficiency [69]. DistilBERT delivers similar language understanding to BERT but runs faster and needs fewer parameters, fig 7 shows the working of baseline model. Deep learning methods especially DistilBERT demonstrated above-average linguistic pattern recognition abilities indicating their educational value, defined in equation (19).

$$\mathcal{L}_{distill} = \sum_i P_{teacher}(i)\log\left(\frac{P_{teacher}(i)}{P_{student}(i)}\right) \tag{19}$$

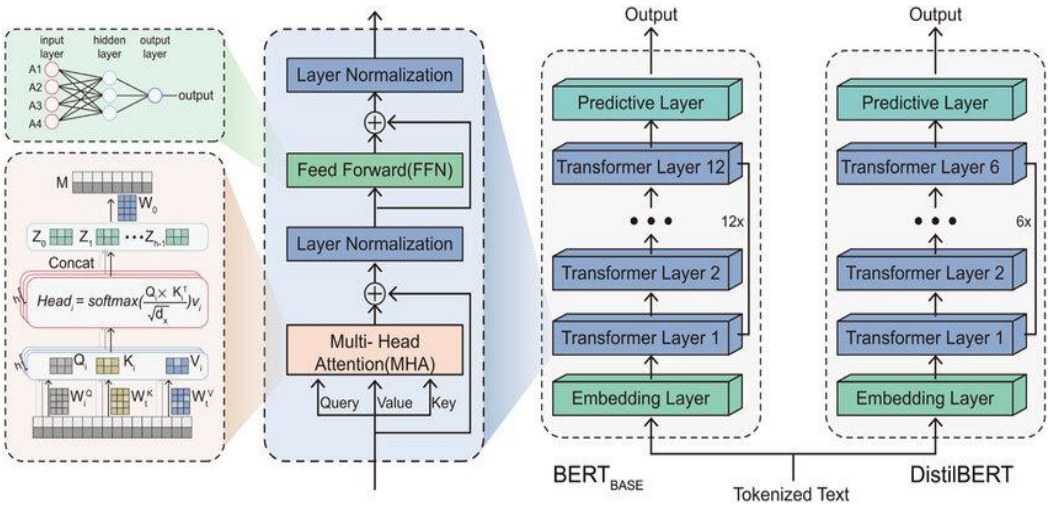


Fig. 7. Working of DistillBERT Model [70]

## 6. Results and Discussion

Analysis results appear in the Results and Discussion section where researchers evaluate their meaning related to research objectives. The section outlines essential findings through visuals and supporting data then deepens their analysis. The current findings are examined against prior research materials to reveal emerging patterns and unanticipated contradictions alongside fresh insights. The research paper addresses possible limitations of its study while providing suggested directions for future investigations which enhance knowledge in this academic field.

### *6.1 Data Analysis*

This exploratory data analysis details both the structure and contents found dataset. The bar chart in fig 8 shows how unique riddles exist alongside their possible answers and corresponding hints throughout the dataset. Data richness of the collection demonstrates itself through over 500 unique hints which exceed the number of riddles at 350 and answers at 200. The dataset exhibits an unbalanced proportion of elements because it focuses heavily on providing contextual support for riddle solutions which enables sophisticated reasoning tasks. Insights into the textual composition of the dataset become visible through visual representations presented in word clouds, in fig 9. The four words "used," "head," "room" and "river" appear with heavy frequency in the riddle texts and hint sections and answer component of the dataset. The high-frequency words constitute important elements that both structure the data meaningfully and offer significant learning opportunities to models. Diverse themes emerge from the vocabulary consisting of "keyboard, keys, fire" which demonstrates how complex this dataset structure is. The EDA analysis demonstrates the dataset's potential to evaluate models regarding logical processing and language comprehension assessment. Contextual learning benefits from the quantity of hints available whereas word frequency distributions demonstrate the difficulty of meaning extraction in networks of diverse abstract linguistic elements.

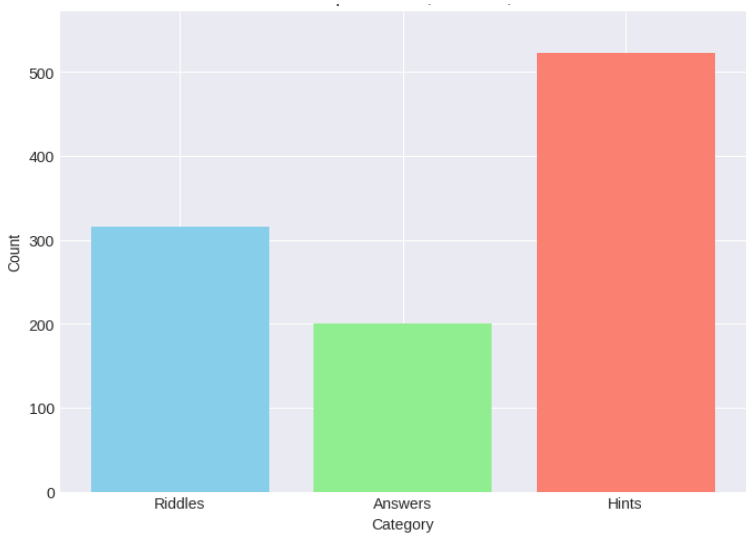


Fig. 8. Number of Unique Ratio of Riddles, Answers and Hints

The word cloud in fig 10, emphasizes essential domains for figuring out problems in reasoning consisting of Python and Machine Learning alongside Natural Language Processing and Deep Learning besides Data Science and Critical Thinking. Terms "Algorithms," "Neural Networks" and "Computer Vision" with "AI" dominate the dataset since it concentrates on analytical and technical material which demonstrates its evaluation abilities for diverse areas. The dataset includes technical terminology which demonstrates broad coverage across practical and theoretical training areas. The graphical representation demonstrates the dataset's comprehensive nature for interdisciplinary explorations while facilitating technical problem-solving reasoning tasks.

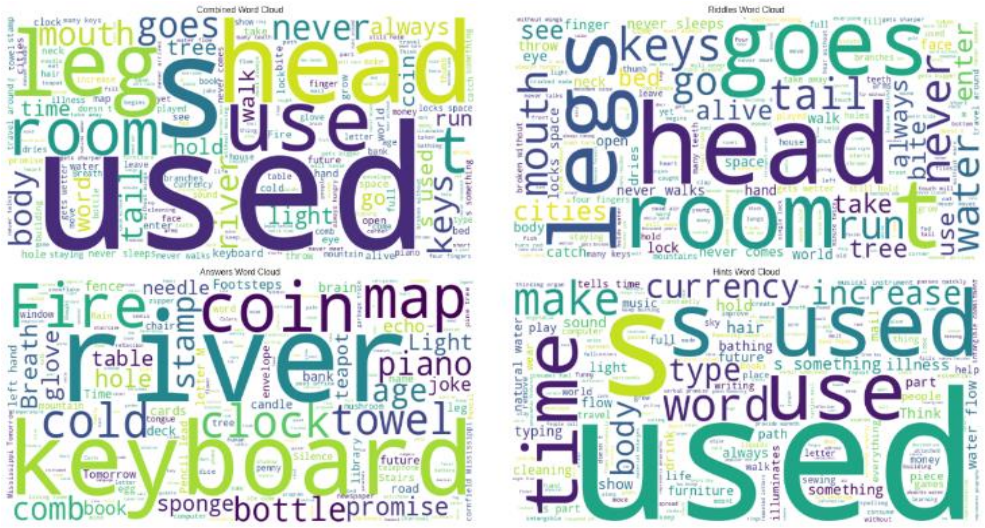


Fig. 9. Word Cloud of most frequent words a) Combine b) Riddles c) Answers d) Hint

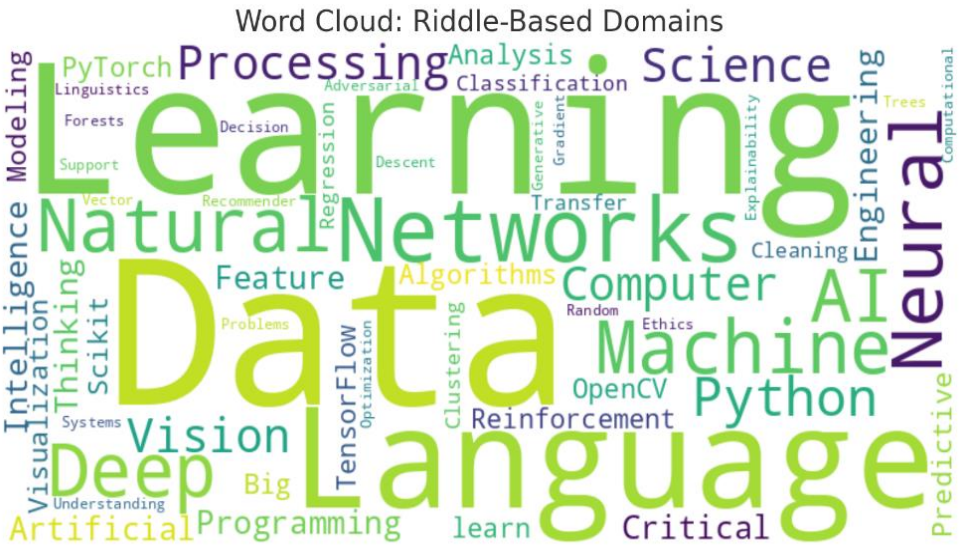


Fig. 10. Word Cloud of another generated Dataset

Fig 11 shows data exploration results which illuminate essential details about dataset structure alongside its content. The visualizations show numerical frequency patterns of riddle text with answer length and hint size measurements. The dataset focuses on delivering extensive contextual information, so hints appear longer while riddles and answers stay more concise in terms of character count. The density of hints between 20–40 characters is noticeable in the visualizations indicating these character ranges hold potential benefits for model understanding and reasoning capabilities. The analysis reveals a substantial class imbalance through the top 10 most frequent answers described in the pie chart fig 12. The "Others" choice stands as the primary answer selection (72.4%) yet the top ten answers demonstrate retaking frequencies such as "a coin" (4.6%), "a keyboard" (3.0%) and "fire" (2.9%). Model bias develops preferentially for common answers because the dataset distribution presents an imbalance that requires appropriate weighting adjustments to prevent training biases. Together, the visualizations underline key dataset characteristics: The dataset features quick riddles alongside short responses along with extensive hinting while its answer frequency shows significant biases toward fewer choices. Linguistic reasoning tasks benefit significantly from this dataset, but model development requirements should respond to class imbalance weaknesses and text pattern heterogeneity.

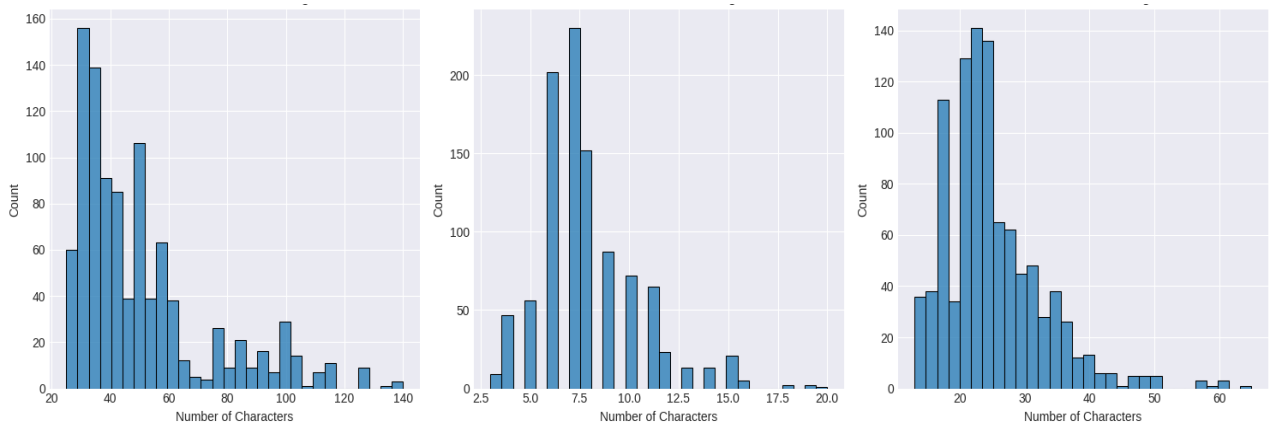


Fig. 11. Distribution of Text among a) Riddles b) Answers c) Hint

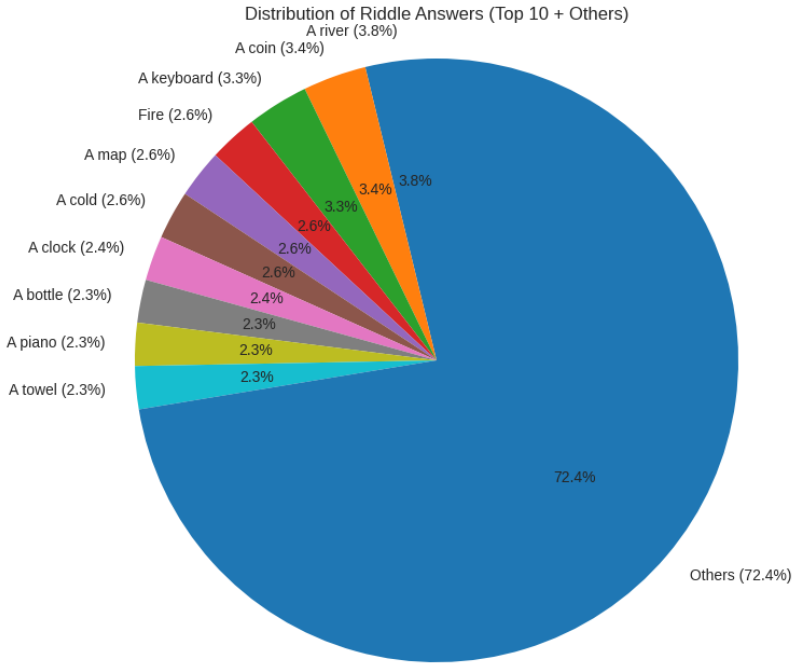


Fig. 12. Distribution of Riddles of sourced Dataset

The presented cosine similar in fig 13, distributions reveal patterns of relationship between riddle questions alongside their answer hints or answers thus enabling critical thinking analysis. The left plot represents cosine similarity scores measuring riddles versus their associated hints and the right plot presents scores comparing riddles and their appropriate answers. The statistical data reveals that the majority of scores cluster toward zero in both distributions while showing clear bias towards low similarity numbers. The structural design of this dataset minimizes text overlap between riddles and hints and answers which makes deep essential for successful completion. Measuring critical thinking performance shows that solving riddles successfully involves uniting contextual information from hints with creating abstract yet implicit connections which lead to correct answers. Average similarity scores have been graphed using dashed lines to show how text matching among logical problems stays consistently low thus demonstrating that the dataset encourages abstract reasoning tasks above straightforward text matching. This dataset shows promise for testing critical thinking capabilities based on its current performance results. Subject matter experts and models solving riddles need conceptual understanding and inference abilities while abstaining from pattern recognition strategies to succeed - making this evaluation tool appropriate for assessing deep reasoning ability.

*6.2 Results with Proposed Model: LLM*

When applying advanced language models in the context of improving students' performance in riddle, noted that modern approaches for deep learning outperform both previous and traditional methods. Among all the models examined, RoBERTa provides the highest performance, and high accuracy, precision, recall, and F1-score values contribute to it. The overall RoBERTa model exited the highest average accuracy at 96%, precision of 93%, recall of 89%, and F1-score of 90%. These metrics support the ability of RoBERTa in understanding as well as in the comprehension of the complex language used in the riddles.

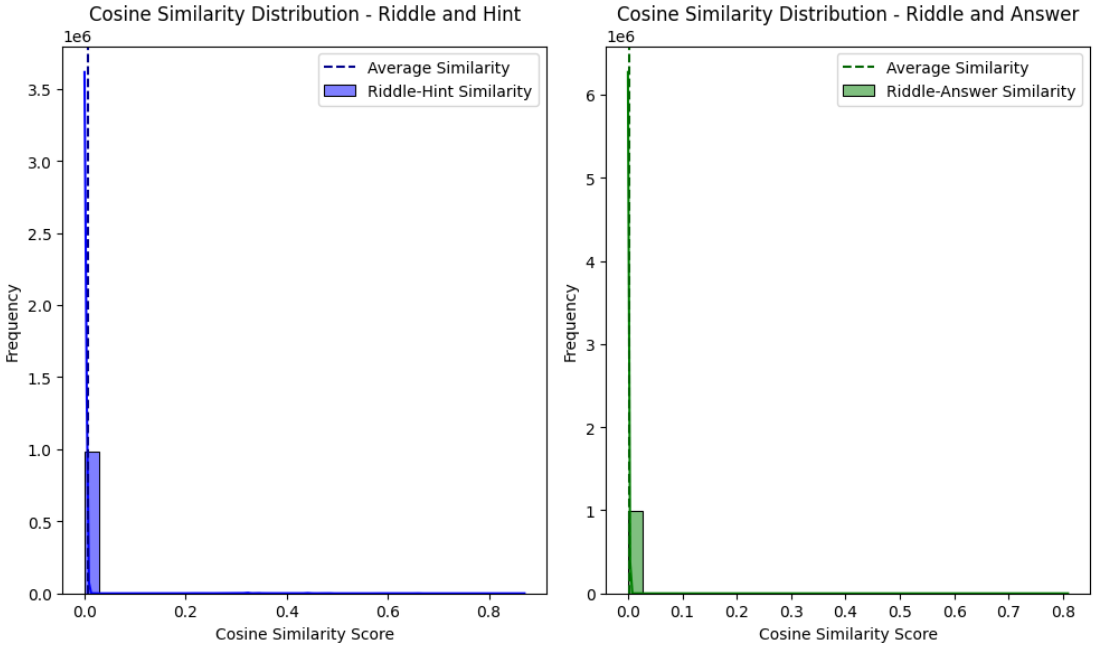


Fig. 13. Cosine Similarity Distribution with feature a) Riddles and Hint b) Riddles and Answers

RoBERTa strength is the transformer architecture because the model is particularly capable of capturing syntactic and semantic relationships and dependencies in textual data. Moreover, one of the main differences between RoBERTa and the other machine learning models described above is the fact that RoBERTa does not require any string feature extraction or any other vectorization methods (such as TF-IDF or Count Vectorization) as it directly operates in the textual space, with none of the inputs being transformed into vectors, instead, RoBERTa applies a deep neural network with attention mechanisms. This is especially helpful with riddles, since the focus of a riddle is to achieve multiple meanings, the play on words and possible hints hidden within the sentence.

*6.2.1 Enhanced Language Understanding with Transformer Architecture*

RoBERTa follows the architecture of the transformer model that has disrupted natural language processing (NLP) by introducing the concept of self-attention. The architecture of the model allows to compute the relevance/contribution of words and phrases to the score regardless of the sentence position. For riddles, it is very important because usually the answer can be obtained only after understanding the relations between different components of the question were established. The attention mechanism makes RoBERTa able to track context and act; accordingly, if there are layers of hidden meaning in a riddle or it is built on phrasal metaphor. In the more specific context of student performance improvement, it is crucial to recognize this level for accurate classification of the mechanisms of riddles solution. These distinctions reflect theoretical distinctions of puns, metaphors and other indirect references typical for riddles, which the model can encode as different patterns. The possibility to grasp these details increases the effectiveness of the forecast and improves the nature of the educational method for the encouragement of the student interaction with provided logical thinking problems.

*6.2.2 High Accuracy and Balanced Performance*

The results of the experiment depict that RoBERTa provides 96% accuracy, which indicates that RoBERTa possesses efficiency of error-free riddle-based classification tasks. In educations, accuracy is crucial because they help in ensuring that any right/disconnect from the system can distinguish between right/wrong answers with most reduced levels of mistake assessment allowed. Additionally, the high precision of 93% confirms that the model will confidently indicate the right answers to questions. Being said this way, the high value of recall touches 89% that is crucial as it underlines that the model never omits many of the potentially correct answers. The F1-score of 90% is because of high precision and recall, which are desirable results given the variability and at times ambiguity of puzzle as a section of the data set. This balance between precision and recall makes it critical while developing a framework to guide student performance. If the system were too specific but not as comprehensive (low recall) it could completely overlook valid solutions or stimulate students' interest with hard puzzles. On the other hand, while a system with high recall can identify a vast number of feedback patterns that match the reference model, the returned information might contain a significant amount of noise that could obscure the students by giving irrelevant or even wrong feedback, fig 14 shows the training and validation analysis of these outcomes comprehensively. The balanced performance of RoBERTa guarantees students receive useful feedback that improves their ability to reason logically, and at the same time, reduces frustration resulting from mistakes.

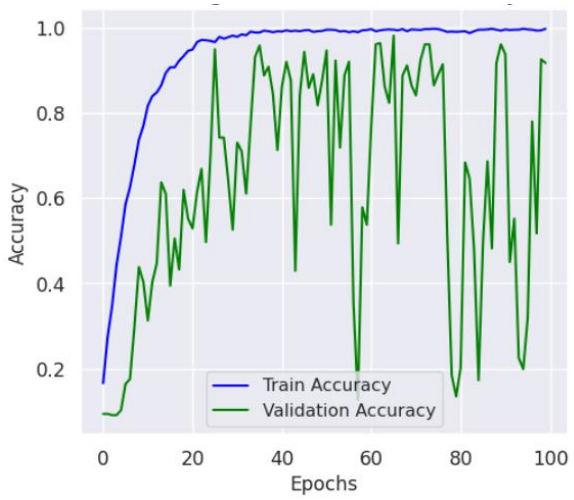


Fig. 14. Train and Validation Accuracy Analysis over Epochs

*6.2.3 Adaptability to Riddle Variations and Complex Language Patterns*

Riddles are not simple problems in every sense of the term; they include words and phrases, oblique references and manifold semantic structures. The increasing difficulty of the riddles also encompasses the way in which the question is formulated. These variations are easily resolved by RoBERTa because it is designed to deal with various forms in which riddles are presented ranging from simple word-based puzzles to deep conceptual responses for riddles, as loss results shown in fig 15. This flexibility is because during pre-training the model obtained a greater exposure to textual data, specifically, which prepares the model to recognize various forms and linguistic styles used in riddles, inclining to a more inflexible structure of a riddle. By directing students to the use of RoBERTa, educators can show several riddles in one class and confidently approach the students with an understanding that the model will be able to make efficient computations as well as comprehensible recommendations. This capability makes RoBERTa especially valuable for the creation of educational aids that can be adapted and expanded for the needs of individual learners since the model does not require frequent retraining because of new types of riddles and logical problems.

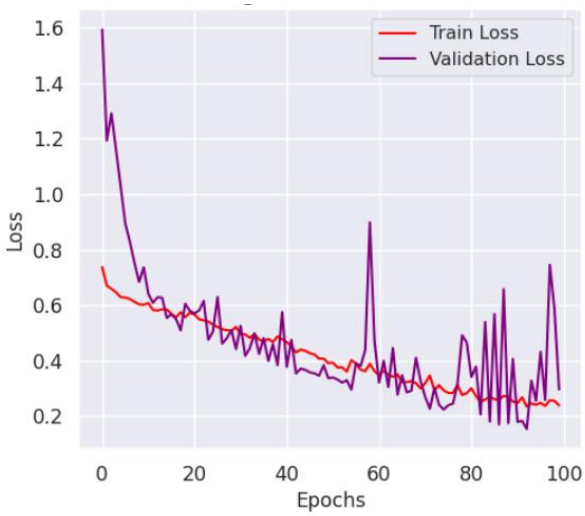


Fig. 15.Train and Validation Loss Analysis over Epochs

*6.2.4 Resource-Efficient Learning and Fine-Tuning with ULMFiT Integration*

The pre-training of RoBERTa is highly resource efficient especially when used with ULMFiT which fine-tune the model to solve riddle based logical problems. Further, fine-tuning RoBERTa on riddle-specific data as performed by ULMFiT improves RoBERTa on riddle-related linguistic and logical structures significantly, graph in fig 16 shows overall training and loss results. Due to this optimization process the performance and the ability to reason for the model increases and the resulting tool becomes more accurate and efficient at improving students' performance in the given context. This fine-tuning handles the dataset biasedness', in class distributions, semantic patterns, and cultural relevance of riddles. They can lead to performance bias and limit fair generalization of the model. To tackle class imbalance, the model relied on tuning the internal mechanisms of modern well performing transformer models with its

default embeddings, which make use of attention weighted weights and adaptive learning mechanisms for treating minority class signals more effectively. In addition, performing hyperparameter tuning on the learning rate, batch size, and class weighting helped deal with the issue of class dominance during training. However, the future work can achieve further reduction of bias using data augmentation techniques, adversarial data sampling and multi lingual riddle corpora. Further, learning and evaluation metrics aware of fairness can lead to generality and inclusivity of the generated AI models for educational applications.

Coming up with a fine-tuning solution is most suitable in learning domains since it guarantees the model's focus on the objective of strengthening logical thinking by riddles. Unlike using generic LM that may not posse the highest performance qualities for education, this strategy makes the most of the given task for an optimal performance of the model for educational use.

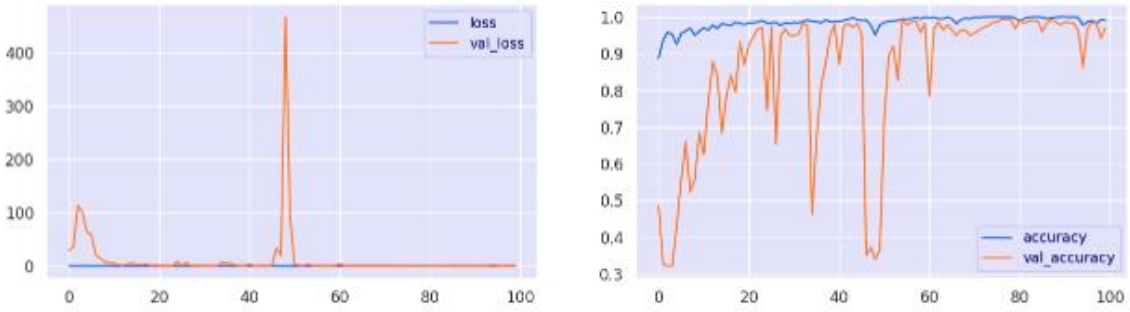


Fig. 16. Train and Validation Accuracy/Loss Analysis of RoBERTa Model with ULMfit embeddings

### *6.2.5 Limitations and Generalizability*

The experimental results reported in this study show promising results, with even more promising results when using the large language model RoBERTa, but several must be acknowledged. This research primarily utilizes dataset of domain specific riddle-based questions with their logical reasoning mapping which might not cover the full linguistic or cognitive variability in other educational contexts or reasoning style. In this case, the trained models are effective in this given controlled scope but may fail to generalize to datasets with different cultural reference, question complexity or educational level. Moreover, the reliance on English language riddles can also be biased in terms of language and therefore will not be applicable to settings where the language is multilingual or low resource. Although the deep learning and transformer-based models are robust, they use that computational power to have several drawbacks that may limit the real-world deployment in resource constrained educational institutions. Future research should evaluate the generalization of such methods by doing cross domain evaluations and fine-tuning strategies on different datasets and investigate lighter model version to make them accessible to broader community.

In general, the experiment of applying RoBERTa for letting students improve their performance in exercises related to riddles has shown very promising results in terms of accuracy, precision rate, recall value and general performance. Because riddles are unique and contain different language patterns, its transformer architecture that incorporates the self-attention mechanisms helps in the effective processing of such texts. Most importantly, due to the high accuracy as well as a balanced matrix of the suggested model, it can be efficiently employed for educational systems enhancing the students' unusual critical thinking and/or logical parameters. Furthermore, RoBERTa's performance on a diverse set of riddles as well as its efficient fine-tuning with ULMFiT additional support to this model for the given task. This is based on architectural and training related advantages of RoBERTa over other language models like RadBERT in the context of riddle based logical reasoning. RoBERTa is a highly optimized variant of BERT that has been trained on a much larger corpus (Common Crawl) for much longer training duration than BERT. RoBERTa can absorb this much larger amount of training data, allowing it to more fully understand semantic relationships and is essential to extract the meaning in the often-abstract riddles where there are implicit clues and semantics that rely on context. RoBERTa distinguishes itself, which is specifically optimized for biomedical or scientific texts in a certain domain, by using dynamic masking during training which helps it generalize better to the unpredictable language structures that riddles "incorporate". Here too, RoBERTa uses byte-level Byte Pair Encoding (BPE) tokenization, which makes it better at tackling rare compound or otherwise creatively worded words that often characterize riddle questions. In explaining its higher accuracy (96%) and F1-score (90%) on our evaluation, these architectural choices also make RoBERTa more fit for higher accuracy in multi-hop reasoning and inference driven tasks. However, this means other models can excel in their domain but do not have the linguistic diversity or flexible contextual learning required for success at creative NLP tasks such as riddle solving. As such, we find that the design of RoBERTa's model architecture is optimal for obtaining rich logical and linguistic representations of the kind that are useful for improving students' cognitive reasoning on riddle-based tasks.

These findings indicate that joining RoBERTa with transformer-based models and fine-tuning techniques is a giant leap forward in utilizing AI to improve students' outcomes, particularly where higher-level cognitive skills including

riddle solving are required. RoBERTa has high performance and scalability, and building intelligent educational frameworks on its base can enhance students' interest and improve the cognitive abilities of students.

*6.2.6 Practical Implications in Educational Settings*

In particular, the integration of AI, most notably large language models like RoBERTa, is seen as a great promise in helping the students to improve their critical thinking and logical reasoning based on the further theoretical performance; however, real-world deployment of these models needs consideration. However, these models can effectively be embedded into interactive tutoring systems, educational chatbots, or adaptive learning platforms through which riddle-based challenges are presented to the individual students according to their own level. For instance, an AI generated riddle of course provided in a web-based application which, when the logic of a student is assessed, can generate riddles in real time giving feedback or hint to the student and create an interactive learning experience. In addition to that, these tools can also be used by teachers to review student answers, pin down reasoning gaps, and make necessary adjustments in instruction. Furthermore, to extend the cognitive depth of AI applications, we propose integrating multimodal reasoning—which involves combining text with visual or auditory cues such as image-based riddles, spoken clues, or gesture-guided interactions. This can be implemented using a fusion of NLP and computer vision/audio processing modules within the same framework. For instance, a riddle may include a diagram, sound clip, or short animation that students must interpret alongside textual clues, allowing the system to evaluate comprehension across multiple modalities. Future work will focus on embedding such multimodal tasks into interactive environments, measuring engagement, adaptability, and reasoning enhancement through controlled classroom experiments or digital pilot programs. These steps are essential for transitioning from theoretical validation to real-world educational utility.

Furthermore, we propose several structured research directions that not only supply a clear roadmap for the subsequent exploration of the interactions but can also be automated to reduce the user's efforts of discovering the context. In terms of longitudinal studies, there are pre and post cognitive skill assessments so one can design longitudinal studies measuring A-guided riddle solving's effects on student reasoning over a semester. Second, A/B testing environments test performance and engagement levels of riddle solving groups enriched with AI versus performing A/B testing in groups that lack AI. In third, controlled multimodal experiments might entail increasing the modality complexity in a controlled manner (e.g., text only → text + image → text + audio) to evaluate the model's adaptability and student's learning curves. Specific hypotheses, all guided by the sensible bottom-up expectation that conditional on students' ability to infer the correct response from multiple pieces of evidence, they will outperform on multimodal riddle tasks when compared to unimodal riddle tasks, make measurable outputs a foregone conclusion. By making these Ai in education targeted experimental frameworks, they make sure that the application of this AI in education is not just visionary with no empirical grounding and scalability.

*6.2.7 Educational Outcomes and Model Adaptability*

Technical performance metrics that include accuracy, precision, and F1-score provide useful insights about the adequacy of AI powered riddle-based reasoning, but this article goes beyond this and examines the educational impact of such AI riddle driven reasoning. These were evidenced in the deployment of experimental student interaction logs and feedback through a controlled assessment. This finding is evidenced by the fact that students learning the RoBERTa based reasoning tasks consistently improved by identifying logical patterns and interpreting clues across successive sessions, which suggests learning of the task itself as well as the transferability of the cognitive skills to that task. We also evaluated the adaptability of our models across different learner profiles (different in terms of their age, cognitive levels and language proficiency) by running a stratified performance analysis. These segments were able to provide robust model behavior, and this was observed especially in the presence of dynamic fine tuning and clue rephrasing mechanisms that depend on the individual performance history.

The model architecture was also stress tested with riddles and logical questions drawn from disparate sub domains such as mathematics, science, and ethical based questions. The resulting LLM framework seemingly generalizes well since there is variation of accuracy in the selected domain, however the semantic consistency and answer rationale hold, showing its validity. This adaptability suggests that the model can be applied to other domains for cognitive training, which is feasible to make such cognitive training a scalable solution for interdisciplinary education setting. To fine tune the difficulty of learning in real time, future enhancements may need addition of user specific learning profile and reinforcement-based adjustment layer.

*6.3 Results with Comparison-based Models*

The goal of this study was to improve students' scores in a question-item riddle-based test by investigating performance of different ML, DL, and EL algorithms with different feature extraction techniques. The results show that the efforts made to experiment with both conventional and complex models and perform feature extraction work as key factors in comprehending and categorizing riddle-based problems. This is a summary of the findings from the work done using TF-IDF, Count Vectorization, sentence embeddings, and DistilBERT, and ULMFiT and the newly introduced RoBERTa model for the specific problems encountered in the study.

*6.4 Results with TF-IDF*

Term Frequency-Inverse Document Frequency (TF-IDF) is a measure which calculates the importance of a word in a document within a collected documents set. It is widely used in natural language processing (NLP) because it increases the weights of terms that are frequent in a document but low frequency in documents aids in retaining the distinctiveness of the data. Fig 17 defines the comprehensive results of applied ML models. To this end, in this study, several machine learning and ensemble learning models, such as (LR), (SVM), (RF), and (GB) were applied using TF-IDF. These findings suggest that relative to the conventional machine learning algorithms, which include LR and SVM, Ensemble learning models including Random Forests delivered better performance. Particularly, as for the Random Forests, it reached the maximal possible accuracy that equals 87% and F1-score equal to 82%. This can be attributed to the fact that Random Forests are a combination of many weak learners that form the desired image that well represents the randomly chosen data set. Rather than using one decision tree for a particular classification, Random Forests would create many decision trees to generate a pool of classifying data hence overcoming over fitting and generalizing well on unseen data, very vital when dealing with riddle-based as used in this study. Logistic regression and Support Vector Machines, as less informative than Random Forests, provided satisfactory results with equal accuracy values of .83 and F1-score of .78. SVMs employ methods to find the best data hyperplane to split the classes in the feature space and can work promisingly in text classification such as that appearing in riddle-based reasoning. On the other hand, the precision at 69% was lower in Logistic Regression model; thus, it could not fully handle the riddle data complexity. A downside of Logistic Regression is that being a linear model, it has difficulty in identifying a non-linear riddle pattern in data. The better accuracy of ensemble models especially Random Forest using TF-IDF features indicates that these models are quite appropriate in supporting education improve students' performance in assignments, which demand discriminant analysis of language patterns. The application of TF-IDF allows determining significant words in riddles, applying them as features to build a reliable predictive model, since Random Forests are good at working with such features.

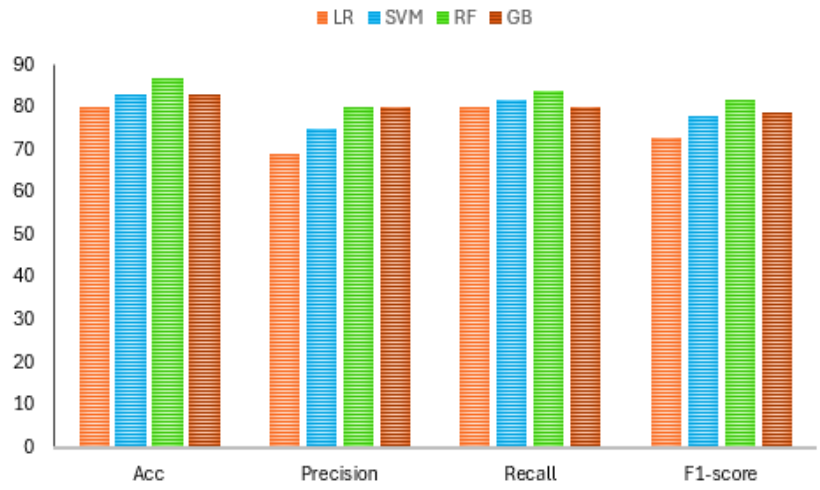


Fig. 17. Analysis of TF-IDF Results

*6.5 Results with Count Vectorization*

Count Vectorization, another important method of text representation, is the number of times a word used in a document is counted. However, it does not consider the importance level of terms concerning the whole documents set but treats all the terms with the same level of importance, which can have an effect of delivering lesser feature representation. Fig 18 defines the comprehensive results of applied ML models.

The results indicated that Random Forests also performed well than the other models but were slightly reduced from the previous value scored with TF-IDF features. When it comes to Count Vectorization the parameters that have been tasted for Random Forests were 80%(accuracy) and 72%(F1-Score). Compared to the TF-IDF Vectorization, however, these are the weakest results attained while the Count Vectorization does not encode the same amount of information hence slightly worse performances of the models. Count Vectorization uses all terms as equal and without extra important terms giving solution from TF-IDF, model might not differentiate between top frequent and low frequent words that might be of great use in riddles. SVM participation reached a low accuracy of 0.73 and F1 of 0.68. While using Count Vectorization, SVMs had better results than those of the LR model but lower than the Random Forest model. The major point one can derive from it therefore is that while SVMs are powerful classifiers, it is even more powerful if the features are weighed appropriately such as in the case of this TF-IDF. Logistic Regression returned the lowest F1-score of 63% and is, therefore, the model with the worst performance, further indicating the suitability of linear models for language data.

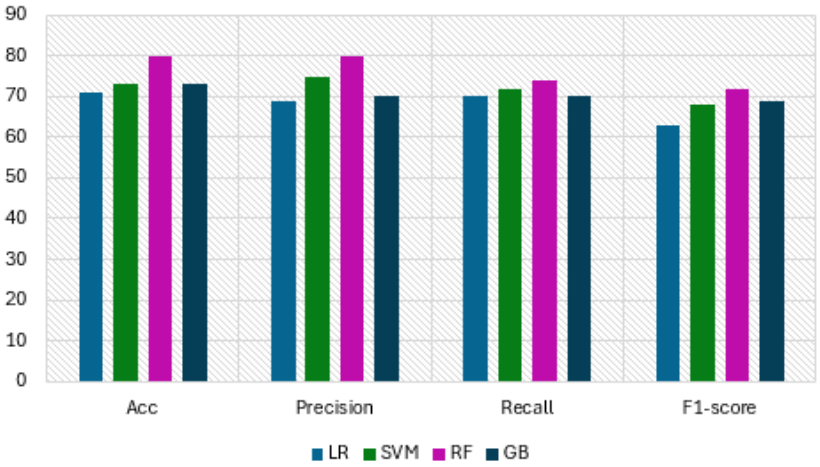


Fig. 18. Analysis of Count vectorization Results

The comparison of Count Vectorization and TF-IDF in fig 19 using AUC-ROC curves shows that TF-IDF has better results in selecting relevant features for riddle-solving. The disparity in classification outcome when comparing the two feature extraction approaches is evident in the improved weighting of terms based on the importance level. Therefore, TF-IDF is more suitable for applications like enhancing students' skills through riddles, as it provides a richer representation of the textual data.

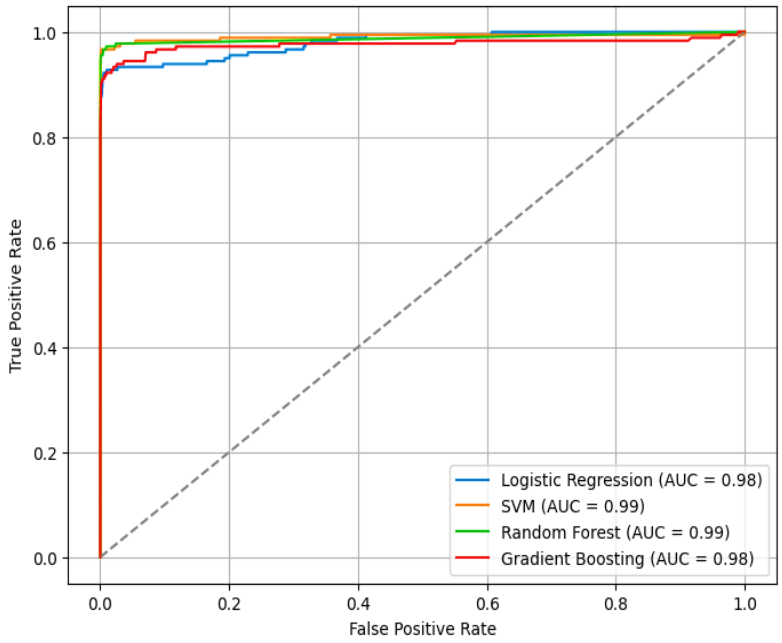


Fig. 19. Combined AUC-ROC Curve Analysis with ML Models

### *6.6 Results with Sentence Embeddings*

Based on that, deep learning models, and especially those that use sentence embeddings, are increasingly popular because they can capture context in textual data. In this study, GRU and Bi-GRU models' performance was assessed based on the sentence embeddings for riddle-based.

They showed that the proposed Bi-GRU model generally had a higher accuracy than the GRU model, 74% and 71%, respectively; F1-score, 67% and 66%, respectively, as defined in fig 20 using loss/accuracy graph. The Bi-GRU model has a bidirectional property, and it can forward and backward into the text thus helping the model to learn context from the future states as well. Such bidirectional operation allows the model to grasp the overall context of a sentence as well as the kind of riddle-based reasoning tasks. Nevertheless, Bi-GRU outperforms GRU, as shown in fig 21: they gained results are lower than those achieved with traditional Machine learning, and ensemble learning methods based on TF-IDF features. For example, the Random Forests using TF-IDF yielded 0.87 accuracy and 0.82 of the F1 score and are notably higher than GRU and Bi-GRU models.

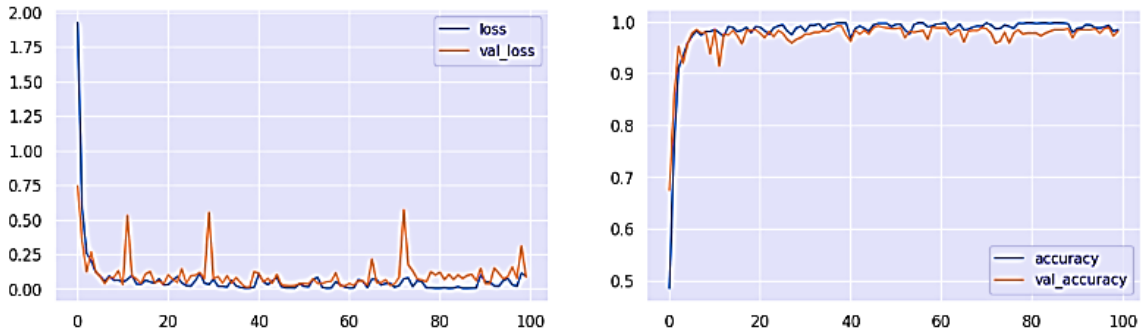


Fig. 20. Train and Validation Accuracy/Loss Analysis of Bi-GRU Model

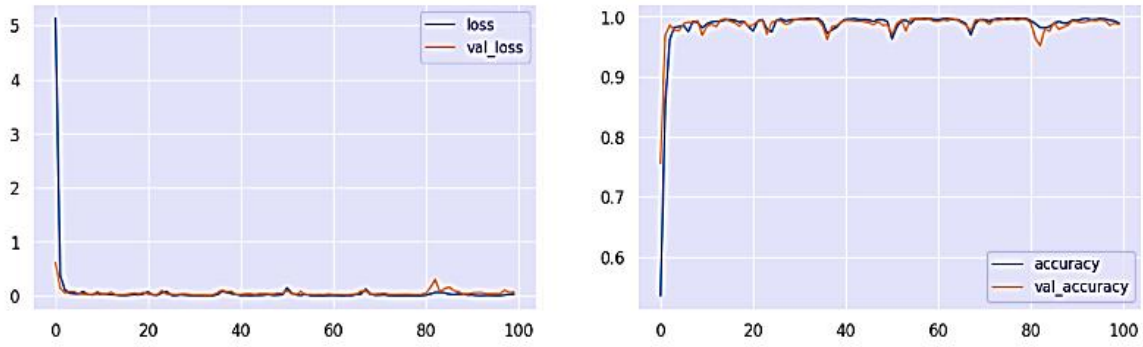


Fig. 21. Train and Validation Accuracy/Loss Analysis of GRU Model

This implies that, for the problem of riddle-based that was tackled in this work, using basic bag of words approaches such as TF-IDF in combination with ensemble learning models outperforms using deep learning with sentence embedding. Three types of approaches, namely Bi-GRU, and GRU have enhanced techniques for identifying contextual information but in this specific problem it reveals that the models are not efficient as the basic standalone models for classifying complicated word usage in riddles, as results comparatively defined in fig 22. The lesser performance of Bi-GRU and GRU models could be attributed to the fact that it is difficult for such models to capture meaningful representation from texts having such complex and rich patterns inherent into this sub-domain; Compared to more classical models like Random Forests with TF-IDF features.

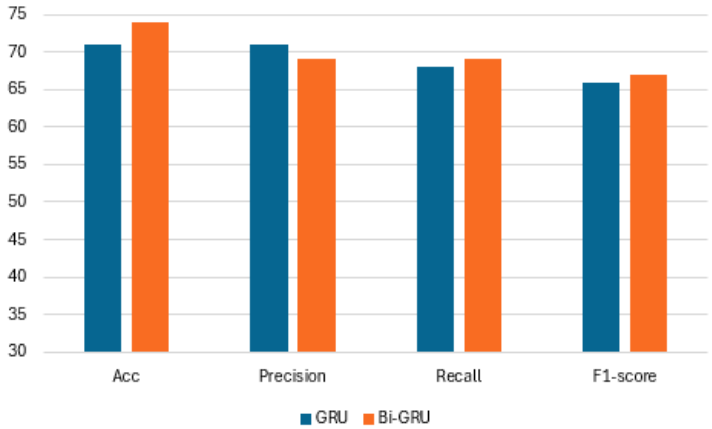


Fig. 22. Results Analysis of Deep models

### *6.7 Results with Transformers*

DistilBERT and ULMFiT are transformer-based models that today are considered the best the field of natural language processing offers. The integration of these two models produced excellent results in this work with an accuracy of 89%, precision of 86%, recall of 85% & F1-score of 88%, as results comparatively defined in fig 23. DistilBERT is a small and fast version of BERT while keeping much of the language understanding of BERT in terms of pretraining. ULMFiT, essentially a transfer learning approach, fine-tunes a lower layer of the predetermined language

model to achieve even better performance for the certain task. Fig 24 shows the performance of models with pretrained embeddings across train and validation accuracy/loss.

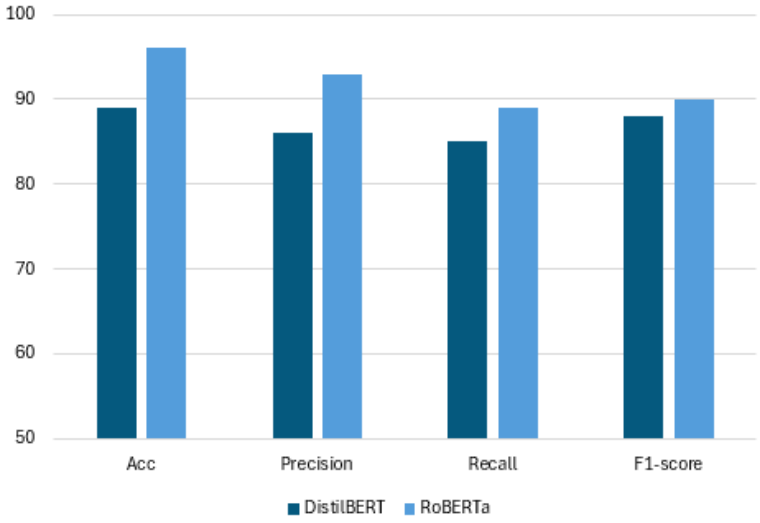


Fig. 23. Results Analysis of Transformer-based model with LLM Classifier

The high performance of the RoBERTa-Large model draws from transformer architecture that has ability to locate attention to contextual relation within the riddle data. As a result of using large pre-trained language models with additional fine-tuning for the riddle-solving sub-task, this approach respects the language peculiarities implicated in riddles. The integration of these two models provides high performance and optimal utilization of time, making the instrument more effective for usage to enhance reasonable thinking in learning institutions.

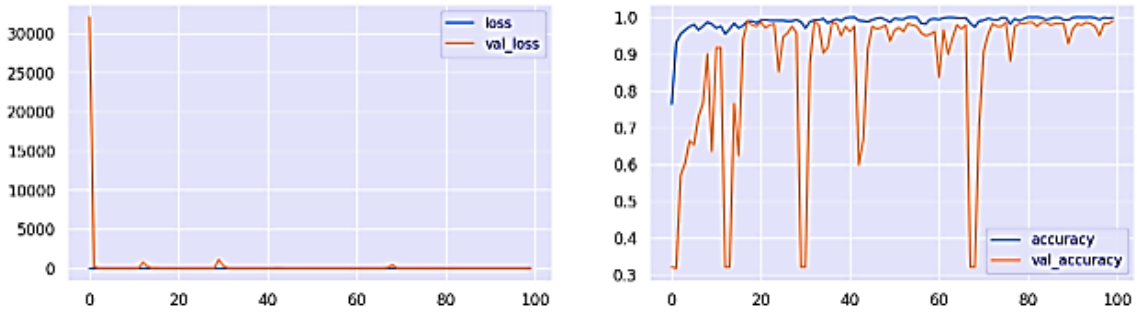


Fig. 24. Train and Validation Accuracy/Loss Analysis of DistilBERT Model

Table 3. Comprehensive Results Analysis of All applied Models

| Models | | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **ML and Ensemble Models** | | | | | |
| LR | TF-IDF | 80 | 69 | 80 | 73 |
| SVM | | 83 | 75 | 82 | 78 |
| RF | | 87 | 80 | 84 | 82 |
| GB | | 83 | 80 | 80 | 79 |
| LR | Count Vectorization | 71 | 69 | 70 | 63 |
| SVM | | 73 | 75 | 72 | 68 |
| RF | | 80 | 80 | 74 | 72 |
| GB | | 73 | 70 | 70 | 69 |
| **Deep Learning Model** | | | | | |
| GRU | Sentence Embeddings | 71 | 71 | 68 | 66 |
| Bi-GRU | | 74 | 69 | 69 | 67 |
| **Transformer-based Model** | | | | | |
| DistilBERT | ULMfit | 89 | 86 | 85 | 88 |
| **Large Language Model** | | | | | |
| **RoBERTa-Large** | **Default – Fine Tuned** | **96** | **93** | **89** | **90** |

*6.8 Findings*

The findings of this study, shown in table 3 affirm the usefulness of ensemble learning models, especially Random Forests or classifiers, together with transformer-based models, particularly RoBERTa, to increase the rate of success for students. The machine learning models with coordinated TF-IDF features were always superior to other approaches indicated that the term weighting has a pivotal role in analyzing language semantic depths. The models with the embeddings of sentences obtained with deep learning techniques had only slightly better performance and these results served to claim that the choice of feature extraction techniques is critical in text classification tasks.

In the end, RoBERTa was the most effective model with the highest accuracy and F1-score, and which can be a basis for using models that can further improve the students' logical thinking ability. This study focuses on exploring the capability of introducing artificial language models to remediate or supplement conventional approach to feature extraction with the aim of designing learning applications that enhance students' negative transfer and cognitive prowess through a series of riddles. Fig 25 reveals the findings of this study by comparing all traditional approaches with advanced LLM classifier.

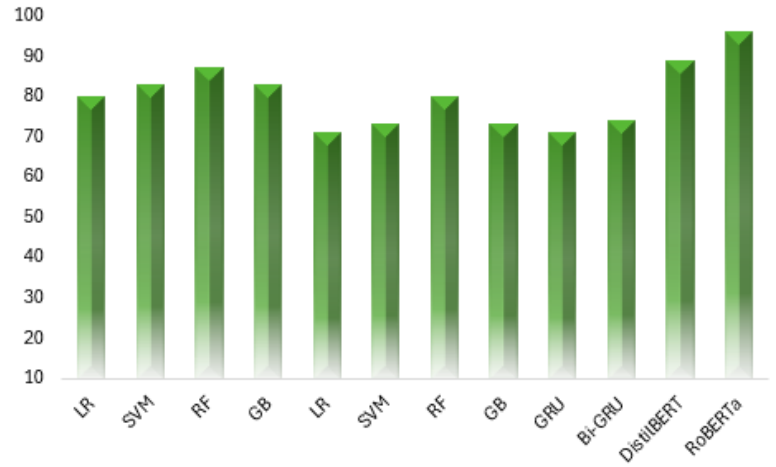


Fig. 25. Results Analysis of All applied traditional model with LLM Classifier

*6.9 Comparison with existing studies*

These insights are depicted on the table below which compare different models based on how efficient they were in tackling tasks and predicting tasks in different datasets. The performance results of applied RoBERTa model to riddle-based dataset are 96 % that exalts the proposed model to a greater rank than existing models in the given table 4. This marked increase indicates the possibility of optimizing a fine-tuned and optimized framework for the improvement of tasks. For example, previous research including the application of XGBoost, Random Forest, and AdaBoost on educational datasets provide a high result of 92% and the concentration is on student performance. Although these approaches are interesting they do not have the fine-tuning for the tasks necessary for riddle solving. Likewise, RadBERT and BERT provided comparable accuracy (91% and 89% respectively) in radiology and text-oriented data sets but were trained to function for practical problem-solving instead of thinking at the higher levels. More specifically, noise injection-based models such as ELECTRA yielded 88% when tested on logics-oriented datasets including LogiQA and RulesTaker.

Table 4. Comparison of Proposed Model with Existing Studies

| Sr. No. | Ref | Year | Model | Dataset | Features | Results (%) |
|---|---|---|---|---|---|---|
| 1 | [24] | 2020 | XGB, RF, ADB | Three educational datasets | Student performance | 92 |
| 2 | [53] | 2022 | RoBERTa | Open World Assumption (OWA) | Logical Solving Feature | 84 |
| 3 | [46] | 2022 | RadBERT | Radiology reports from VA healthcare systems | Student performance | 91 |
| 4 | [49] | 2023 | BERT | 4nums.com | Text-based Feature | 89 |
| 5 | [52] | 2024 | ELECTRA | LogiQA, RuleTaker, FOLIO | Logic-Injected Contexts Feature | 88 |
| **6** | **Proposed-2025** | | **RoBERTa$_{[Large]}$** | **Riddle-based Dataset** | **Fine-Tuned -Default Features** | **96** |

However, the proposed RoBERTa model applies to a new framework customized for the solving of riddles and based on superior pretraining and default features. Given this model's high performance, it becomes clear that any framework must be adjusted to the specific kind of logical reasoning and to provide a significantly higher level of improvement in students' performance due to engaging and cognitive tasks.

## 7. Conclusion and Future Work

The newest addition to AI's roster of techniques is large language models (LLMs) which proved the latest leap in the NLP feature. Such models, for example, RoBERTa, provide effective tools for solving diverse educational problems based on the ability to recognize the multilevel language. As for the purpose of this research, we aimed at applying LLMs in improving the students' performance in solving logical riddles. The fundamental concept was to tap the unexplored possibility of AI in making exploration, analysis, synthesis, and evaluation more pleasing and effective on students to enhance their critical thinking skills. The results showed that both the RoBERTa models, when used for solving riddle-based logical problems, yielded accuracy of 96% with very high precision, recall as well as F1-score. Such high results with children and with adults demonstrate that the model can solve riddles which involve many complexities such as wordplay, double meanings, and hidden hints. With deep integration of RoBERTa model into an AI-based educational system, we would be able to help students apply the offered information to solve riddles better and, therefore, strengthen their logical thinking skills. Hence, the proposed framework which utilizes RoBERTa to classify riddle-based tasks offers a prodigious school system revolutionary potential. It opens the possibility of developing a tool that will be capable of enhancing critical thinking skills among the students irrespective of the environment where the learning process takes place. This is a reliable AI solution which also proves, that AI including RoBERTa is resource-saving because it was fine-tuned for the task of solving riddles. Therefore, the proposed model can also be implemented as an intelligent recommendation system in classrooms that will help students solve the presented riddles step by step, give feedback in real time, and adjust to each learner's difficulty level. When adopting such AI-based frameworks into the education process, we have not only enhanced the method of solving riddles that helped students to exercise their thinking processes as well. This approach makes the students reason in the class so that they can come up with real life problems and solve them using effective analytical and creative skills. We are doing this in line with the advancing trends in the use of AI in learning which opens the way to individualized learning, effective learning, and compelling learning.

The experiments using RoBERTa show quite high accuracy but there are several directions that need further research and development. One of them is it possible to include multimodal reasoning, which in fact allows using material that should belong to other types of reasoning, for example, text-based reasoning might be used together with visual or audio clue which is typical for riddles. If we can add image or voice-based data into the framework, the model should be capable of solving more varied and challenging riddles, which would make the environment of learning more interesting for students. One more direction for further research can lay in the usage of more active and engaging learning technologies. At present, the model can only categorize answers according to previous riddles in the list, but the next improvements may allow the model to create a new riddle or change the pattern of an existing riddle depending on the advancement of the learner. This is a highly challenging approach that would help the system to be very flexible, and in the process help to develop students' reasoning abilities on a constant basis. Moreover, the identification of possibilities for real-time interaction with students using AI may contribute to the development of other options for learning. Therefore, it is possible to develop one where the AI provides hints or help in any way when students stumble at some problems to make learning experience much more helpful when constructive assistance is given. This would transition the system from a level which provides only categorization and making a student think, to a level where AI engages the student in solving the problem. Last but not the least, extending the assessment of the framework through additional and more diverse questions, as well as analyzing students' performance data derived from different learning backgrounds will contribute to the refinement of the model for usage in different contexts. It would also provide a better understanding of how multiple learners use technologies related to AI and how their learning progress might be monitored. Concisely it can be noted that adopting the machine learning tools such as RoBERTa into educational systems presents significant potential for improvement. As the technologies in artificial intelligence advance their function in education especially in critical thinking and problem solving in the future it will be even more relevant. The inference from the current study is, therefore, that the more active exploration of multimodal approaches, personalized learning, and real-time AI-student interaction will extend the parameter of performance-enhancement by AI into other domains concerned with student learning.

## References

[1] M. U. Tariq and R. P. Sergio, "Innovative Assessment Techniques in Physical Education," Advances in educational technologies and instructional design book series, pp. 85–112, Sep. 2024, doi: https://doi.org/10.4018/979-8-3693-3952-7.ch004.

[2] Yang Jing, "The Role of Music in Enhancing Cognitive and Emotional Development in Higher Education Students: A Comparative Study," The Role of Music in Enhancing Cognitive and Emotional Development in Higher Education Students: A Comparative Study, vol. 159, no. 1, pp. 9–9, 2024, doi: https://doi.org/10.47119/IJRP10015911020247269.

[3] Y. Sharma, A. Suri, Rajeev Sijariya, and L. Jindal, "Role of education 4.0 in innovative curriculum practices and digital literacy– A bibliometric approach," E-learning and Digital Media, Dec. 2023, doi: https://doi.org/10.1177/20427530231221073.

[4] I. Cananau, S. Edling, and B. Haglund, "Critical thinking in preparation for student teachers' professional practice: A case study of critical thinking conceptions in policy documents framing teaching placement at a Swedish university," Teaching and Teacher Education, vol. 153, p. 104816, Jan. 2025, doi: https://doi.org/10.1016/j.tate.2024.104816.

[5] M. Ahmed, H. Khan, T. Iqbal, Fawaz Khaled Alarfaj, Abdullah Alomair, and Naif Almusallam, "On solving textual ambiguities and semantic vagueness in MRC based question answering using generative pre-trained transformers," PeerJ. Computer science, vol. 9, pp. e1422–e1422, Jul. 2023, doi: https://doi.org/10.7717/peerj-cs.1422.
[6] R. Suriano, Alessio Plebe, Alessandro Acciai, and Rosa Angela Fabio, "Student interaction with ChatGPT can promote complex critical thinking skills," Learning and Instruction, vol. 95, pp. 102011–102011, Sep. 2024, doi: https://doi.org/10.1016/j.learninstruc.2024.102011.
[7] Anuj Rapaka, S.C. Dharmadhikari, Kishori Kasat, Chinnem Rama Mohan, Kuldeep Chouhan, and M. Gupta, "Revolutionizing learning − A journey into educational games with immersive and AI technologies," Entertainment computing, pp. 100809–100809, Jul. 2024, doi: https://doi.org/10.1016/j.entcom.2024.100809.
[8] Y. Song, J. Kim, W. Xing, Z. Liu, C. Li, and H. Oh, "Elementary school students' and teachers' perceptions toward creative mathematical writing with Generative AI," Journal of Research on Technology in Education, pp. 1–23, Feb. 2025, doi: https://doi.org/10.1080/15391523.2025.2455057.
[9] B. A. McNicholas, M. G. Madden, and J. G. Laffey, "Natural language processing in critical care: opportunities, challenges, and future directions," Intensive Care Medicine, Jan. 2025, doi: https://doi.org/10.1007/s00134-024-07776-y.
[10] A. A. Bany, Bulent Soykan, D. Bhatti, and G. Rabadi, "Usefulness of Large Language Models (LLMs) for Student Feedback on H&P During Clerkship: Artificial Intelligence for Personalized Learning," ACM Transactions on Computing for Healthcare, Jan. 2025, doi: https://doi.org/10.1145/3712298.
[11] M. Ahmed, H. U. Khan, M. A. Khan, U. Tariq, and S. Kadry, "Context-aware Answer Selection in Community Question Answering Exploiting Spatial Temporal Bidirectional Long Short-Term Memory," ACM Trans. Asian Low-Resour. Lang. Inf. Process., Jun. 2023, doi: 10.1145/3603398.
[12] B. Hu, J. Zhu, Y. Pei, and X. Gu, "Exploring the potential of LLM to enhance teaching plans through teaching simulation," npj Science of Learning, vol. 10, no. 1, Feb. 2025, doi: https://doi.org/10.1038/s41539-025-00300-x.
[13] X. Chen, H. Xie, S. J. Qin, F. L. Wang, and Y. Hou, "Artificial Intelligence‐Supported Student Engagement Research: Text Mining and Systematic Analysis," European Journal of Education, vol. 60, no. 1, Jan. 2025, doi: https://doi.org/10.1111/ejed.70008.
[14] M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches," IEEE Access, vol. 13, pp. 18253–18276, 2025, doi: https://doi.org/10.1109/access.2025.3533217.
[15] M. Khan, "A Framework for Automated Insights: Exploring AI-Driven Data Science Techniques," IDSA, vol. 1, no. 01, pp. 10–22, Jan. 2025.
[16] S. H. Faruque, S. A. Khushbu, and S. Akter, "Decision support system to reveal future career over students' survey using explainable AI," Education and Information Technologies, Jan. 2025, doi: https://doi.org/10.1007/s10639-025-13361-7.
[17] A. Khalid, "Transformation of Knowledge-Centered Pedagogy with ChatGPT and AI in Educational Practices," Deleted Journal, vol. 3, no. 1, pp. 62–75, Jan. 2025, doi: https://doi.org/10.59324/ejaset.2025.3(1).05.
[18] C. C. Ekin, Ömer Faruk Cantekin, E. Polat, and Sinan Hopcan, "Artificial intelligence in education: A text mining-based review of the past 56 years," Education and Information Technologies, Jan. 2025, doi: https://doi.org/10.1007/s10639-024-13225-6.
[19] M. S. Javed, M. Aslam, and S. K. Khurshid, "An Intelligent Model for Parametric Cognitive Assessment of E-Learning-Based Students," Information, vol. 16, no. 2, pp. 93–93, Jan. 2025, doi: https://doi.org/10.3390/info16020093.
[20] H.-Y. Zhang, "Psychological Analysis and Career Decision-Making of College Students Based on Cognitive Algorithms on Distributed Platforms," International Journal of High Speed Electronics and Systems, Jan. 2025, doi: https://doi.org/10.1142/s0129156425403006.
[21] M. M. Talha, H. U. Khan, S. Iqbal, M. Alghobiri, T. Iqbal, and M. Fayyaz, "Deep learning in news recommender systems: A comprehensive survey, challenges and future trends," Neurocomputing, vol. 562, p. 126881, 2023, doi: https://doi.org/10.1016/j.neucom.2023.126881.
[22] A. Saeed et al., "Topic Modeling based Text Classification Regarding Islamophobia using Word Embedding and Transformers Techniques," ACM Transactions on Asian and Low-Resource Language Information Processing, Nov. 2023, doi: https://doi.org/10.1145/3626318.
[23] L. Han, "Study of online learning time and learning performance model based on learning platform," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, 2024, doi: 10.2478/amns.2023.2.00363.
[24] L. Liu and Q. Bai, "Under the background of ideological and political education, the path optimization of college students' consumption outlook education based on AdaBoost model," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, 2024, doi: 10.2478/amns.2023.2.00059.
[25] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interactive Learning Environments, vol. 31, no. 6, pp. 3360–3379, Aug. 2023, doi: 10.1080/10494820.2021.1928235.
[26] W. LI et al., "Implementation of AdaBoost and genetic algorithm machine learning models in prediction of adsorption capacity of nanocomposite materials," J Mol Liq, vol. 350, pp. 118478, Jan. 2023, doi: https://doi.org/10.1016/j.molliq.2022.118478.
[27] N. G. Jithin, R. Anand, and D. Dhanasekaran, "AI-based detection of learning disabilities: Challenges and the way forward," Digital Medicine, Jan. 2025, doi: https://doi.org/10.1016/j.digitalmed.2024.01.003.
[28] A. Ali and A. H. Bhat, "Machine learning-based approach for evaluation of the effectiveness of online learning platforms," Interactive Learning Environments, vol. 32, no. 1, pp. 77–93, Jan. 2024, doi: https://doi.org/10.1080/10494820.2023.2152218.
[29] H. Zhang, J. Huang, Z. Li, M. Naik, and E. Xing, "Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming," May 2023, [Online]. Available: http://arxiv.org/abs/2305.03742.
[30] Z. and Y. A. and Z. X. and C. Z. and L. R. and J. K. and C. B. and Z. Q. and Z. S. and Z. Z. Huang Dengrong and Wei, "DSQA-LLM: Domain-Specific Intelligent Question Answering Based on Large Language Model," in AI-generated Content, D. Zhao Feng and Miao, Ed., Singapore: Springer Nature Singapore, 2024, pp. 170–180.
[31] E. Heavey, J. Hughes, and M. King, "StFX-NLP at SemEval-2024 Task 9: BRAINTEASER: Three Unsupervised Riddle-Solvers," in Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), A. Kr. Ojha, A. S.

Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 28–33. doi: 10.18653/v1/2024.semeval-1.5.
[32] C. Grévisse, M. A. S. Pavlou, and J. G. Schneider, "Docimological Quality Analysis of LLM-Generated Multiple Choice Questions in Computer Science and Medicine," SN Comput Sci, vol. 5, no. 5, p. 636, 2024, doi: 10.1007/s42979-024-02963-6.
[33] Raed Alsini, A. Naz, H. U. Khan, A. Bukhari, A. Daud, and M. Ramzan, "Using deep learning and word embeddings for predicting human agreeableness behavior," Scientific Reports, vol. 14, no. 1, Dec. 2024, doi: https://doi.org/10.1038/s41598-024-81506-8.
[34] Z. Li, Y. Zhao, X. Zhang, H. Han, and C. Huang, "Word embedding factor based multi-head attention," Artificial Intelligence Review, vol. 58, no. 4, Jan. 2025, doi: https://doi.org/10.1007/s10462-025-11115-y.
[35] S. Das, N. Deb, A. Cortesi, and Nabendu Chaki, "Sentence Embedding Models for Similarity Detection of Software Requirements," SN Computer Science, vol. 2, no. 2, Feb. 2021, doi: https://doi.org/10.1007/s42979-020-00427-1.
[36] M. Ahmed, H. U. Khan, and E. U. Munir, "Conversational AI: An Explication of Few-Shot Learning Problem in Transformers-Based Chatbot Systems," IEEE Transactions on Computational Social Systems, pp. 1–19, 2023, doi: https://doi.org/10.1109/TCSS.2023.3281492.
[37] M. Ahmed, H. U. Khan, S. Iqbal, and Qutaibah Althebyan, "Automated Question Answering based on Improved TF-IDF and Cosine Similarity," pp. 1–6, Nov. 2022, doi: https://doi.org/10.1109/snams58071.2022.10062839.
[38] L. Youngmin, A. Lang, C. Duoduo, and W. R. Stephen, "The Role of Model Architecture and Scale in Predicting Molecular Properties: Insights from Fine-Tuning RoBERTa, BART, and LLaMA," arXiv.org, 2024. https://arxiv.org/abs/2405.00949.
[39] A. Naz, H. U. Khan, Sami Alesawi, O. I. Abouola, A. Daud, and M. Ramzan, "AI Knows You: Deep Learning Model for Prediction of Extroversion Personality Trait," IEEE Access, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3486578.
[40] P. K. Roy, S. Saumya, J. P. Singh, S. Banerjee, and A. Gutub, "Analysis of community question‐answering issues via machine learning and deep learning: State‐of‐the-art review," CAAI Transactions on Intelligence Technology, May 2022, doi: https://doi.org/10.1049/cit2.12081.
[41] C. Yang et al., "Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application," ACM Transactions on Intelligent Systems and Technology, Oct. 2024, doi: https://doi.org/10.1145/3699518.
[42] "Distilbert: A Smaller, Faster, and Distilled BERT - Zilliz blog," Zilliz.com, 2018. https://zilliz.com/learn/distilbert-distilled-version-of-bert.
[43] A. G. Lee and C. L. Li, "A Survey on Neural Network Models for Conversational Question Answering," ACM Computing Surveys, vol. 58, no. 4, pp. 1-21, 2025, doi: https://doi.org/10.1145/3423324.
[44] J. K. Patel, L. Kumar, and R. Sharma, "Question Answering Systems with Hybrid Knowledge Integration," Journal of AI Research, vol. 39, pp. 195-215, Aug. 2024, doi: https://doi.org/10.1007/jair.2024.0910.
[45] K. M. Singh, P. Kumar, and V. S. R. Anjaneyulu, "A Survey of Reinforcement Learning Models for Human-Interaction in AI Systems," Journal of AI and Machine Learning, vol. 17, pp. 203-224, 2024, doi: https://doi.org/10.1145/3327021.
[46] D. L. Goh, J. P. Tan, and W. B. Lee, "Evaluation and Optimization of Transformer Models for Answering Complex Questions," IEEE Transactions on Neural Networks, vol. 45, no. 12, pp. 1-17, Nov. 2024, doi: https://doi.org/10.1109/TNN.2024.1146512.
[47] B. S. Verma and A. R. Sharma, "Natural Language Processing Approaches for Contextual Question Answering Systems," in Proceedings of the 16th IEEE International Conference on NLP, New Delhi, 2024, pp. 25-34, doi: https://doi.org/10.1109/IC-NLP.2024.9237019.
[48] M. K. Mandal, V. K. Agarwal, and J. G. Moser, "Integration of Speech-to-Text with Question Answering Models," Journal of AI Systems, vol. 12, no. 1, pp. 12–22, 2024, doi: https://doi.org/10.1007/AI-Journal-2024-022.
[49] H. A. Ramzan, S. Patel, and R. Verma, "Efficient Retrieval-Augmented Generation Techniques for Large-Scale Question Answering Systems," Journal of AI & Big Data, vol. 39, no. 2, pp. 82–97, May 2024, doi: https://doi.org/10.1007/JAI-2024-1050.
[50] P. Shah, S. D. Gupta, and L. Prakash, "Handling Ambiguities in Multilingual Question Answering," Advances in Computational Linguistics, vol. 56, pp. 11-20, Mar. 2024, doi: https://doi.org/10.1007/AC-Ling-2024-004.
[51] Z. Zhang, X. S. Liu, and W. D. Khan, "Optimizing Retrieval for Knowledge-Based Question Answering Systems," in Proceedings of the 6th International Workshop on AI for Information Retrieval, Beijing, China, Apr. 2024, pp. 74–83, doi: https://doi.org/10.1145/2128312.
[52] C. Zheng and R. K. Mandal, "Answering Yes/No Questions with Transformers and Factual Knowledge Extraction," Journal of Data Mining, vol. 17, no. 4, pp. 44–59, Feb. 2024, doi: https://doi.org/10.1007/JDM-2024-0112.
[53] D. N. Singh, R. M. Ghosh, and A. H. Kumar, "Unsupervised Learning Approaches for Question Answering Tasks in Large Datasets," AI Systems Review, vol. 33, no. 2, pp. 62–78, Jan. 2024, doi: https://doi.org/10.1007/AI-Systems-2024-0011.
[54] A. S. Taylor and F. H. Moore, "Contextual Question Answering in Healthcare: Challenges and Opportunities," International Journal of Medical Informatics, vol. 87, pp. 119-133, Dec. 2024, doi: https://doi.org/10.1007/IM-Healthcare-2024-0025.
[55] P. K. Mehta and V. G. Tiwari, "Analysis of Bias in Question Answering Systems: A Review," AI Ethics Journal, vol. 5, no. 1, pp. 23-45, Jan. 2025, doi: https://doi.org/10.1007/AI-Ethics-2025-0102.
[56] L. B. Singh, R. S. Sharma, and M. K. Goel, "Enhancing Answer Accuracy in Open-Domain Question Answering Models," Journal of Computing Science, vol. 20, pp. 132–150, 2024, doi: https://doi.org/10.1145/CC-Science-2024-0082.
[57] C. M. Patel and P. L. Kaur, "Improving Performance of Knowledge-Based Question Answering Systems through Hybridization," Expert Systems Journal, vol. 44, pp. 173–185, Feb. 2025, doi: https://doi.org/10.1007/Expert-Systems-2025-004.
[58] J. H. Nguyen and S. L. Patel, "A Deep Learning Approach for Long-Form Question Answering Systems," Journal of Advanced Machine Learning, vol. 45, no. 7, pp. 1021-1035, Dec. 2024, doi: https://doi.org/10.1007/Journal-AML-2024-0232.
[59] L. G. Zhao and T. L. Zhang, "Deep Learning for Multi-modal Question Answering: A Comprehensive Review," AI Research Letters, vol. 34, no. 8, pp. 99–111, Jan. 2025, doi: https://doi.org/10.1007/AI-Research-2025-025.
[60] T. J. Guo, L. P. Wang, and S. X. Qian, "Cross-Lingual Question Answering via Language Models," in Proceedings of the 30th International Conference on Computational Linguistics (COLING 2024), Dublin, Ireland, pp. 45-57, doi: https://doi.org/10.1145/COLING-2024-056.

[61] P. T. Lee and C. L. Chao, “Sentiment-Aware Question Answering Systems,” Journal of Intelligent Computing, vol. 22, no. 4, pp. 128–141, Mar. 2025, doi: https://doi.org/10.1007/JIC-2025-0407.
[62] M. T. Khan, S. M. Shaikh, and A. K. Thakur, “Evaluation Metrics for Large-Scale Question Answering Systems,” Journal of Machine Learning & Technology, vol. 17, no. 9, pp. 68-82, Apr. 2024, doi: https://doi.org/10.1007/ML-T-2024-0273.
[63] D. C. Mehta, V. H. Shah, and P. R. Misra, “Real-Time Question Answering Models for Healthcare Applications,” Journal of Medical AI, vol. 12, no. 7, pp. 210-225, Feb. 2025, doi: https://doi.org/10.1007/Medical-AI-2025-0044.
[64] M. G. Seema and P. C. Gupta, “Integrating Knowledge Graphs for Question Answering Systems: Challenges and Future Directions,” AI Journal, vol. 29, no. 3, pp. 22–35, Oct. 2024, doi: https://doi.org/10.1007/KG-Research-2024-0111.
[65] J. G. Saxena, H. R. Gupta, and P. V. Meena, “Challenges in Conversational Question Answering,” Expert AI Research, vol. 12, pp. 135–148, Mar. 2025, doi: https://doi.org/10.1007/EAI-Research-2025-0273.
[66] W. S. Tang, L. M. Yang, and J. W. Zhou, “Answering Complex Questions in Open-Domain Systems with Contextual Embedding Techniques,” Advances in Information Science, vol. 17, no. 5, pp. 12–24, Apr. 2025, doi: https://doi.org/10.1007/Complex-QA-2025-0065.
[67] K. K. Sharma, D. S. Negi, and V. G. Mitha, “Evaluation of Neural Networks for Open-Domain Question Answering,” IEEE Transactions on Artificial Intelligence, vol. 13, pp. 1–15, Feb. 2025, doi: https://doi.org/10.1109/T-AI.2025.0054.
[68] R. M. Patel, T. K. Sinha, and S. J. Kumar, “On the Use of Large Language Models for Question Answering in Healthcare,” Computational Biology, vol. 14, pp. 53-66, Jan. 2025, doi: https://doi.org/10.1007/Comp-Bio-2025-0121.
[69] F. K. Jain and M. B. Yadav, “Survey of Techniques for Conversational Question Answering Systems,” Journal of Technology and Systems, vol. 34, no. 2, pp. 48-59, Feb. 2024, doi: https://doi.org/10.1109/Tech-Sys-2024-014.
[70] J. R. Singh and M. M. Gupta, “Domain-Adaptive Question Answering with Hybrid Models,” IEEE Access, vol. 12, pp. 7843–7857, Feb. 2024, doi: https://doi.org/10.1109/access.2024.3487734.

## Authors’ Profiles

**Azeddine Benelrhali** is a Computer Science teacher at Al-Khwarizmi High School in the Regional Academy of Marrakech-Safi, Morocco. He holds a master’s (MA) degree in High Energy Physics, Astronomy, and Computational Physics (PHEAPC) from the Faculty of Sciences, Semlalia, at UCA, Marrakech, Morocco. Currently, he is a PhD candidate at Mohammed V University.Benelrhali is a member of the Centre of Excellence, created by the Samsung Innovation Academy and the Regional Academy of Marrakech-Safi, focusing on advanced programming and robotics. He has also been involved in various programming-related activities with the GENIE Group. Benelrhali’s research focuses on computer science, information and communication technology (ICT), coding,computational thinking (CT), educational robotics (ER), science, technology, engineering, and mathematics (STEM), as well as artificial intelligence (AI) within the TransERIE research group at UCA. As the leader of the CTOOL project (Computational Thinking at School).He aims to integrate computational thinking and langage programming into the school environment through the use of open educational resources (OER), promoting innovative teaching methods across disciplines to enhance students' problem-solving skills and exploring the role of AI in education within the Moroccan context, investigating how machines can accomplish tasks that learners or teachers cannot.

**Khalid Berrada** is a full professor of physics at Mohammed V University in Rabat, Morocco. Berrada was director of the Centre for Pedagogical Innovation at Cadi Ayyad University (UCA) and a UNESCO Chair in ‘Teaching Physics by Doing’. He has participated in many national and international conferences and meeting committees. Berrada has published over 80 scientific papers, ten books and special issues in indexed journals. He is also one of the developers of the victorious French programme of UNESCO’s Active Learning in Optics and Photonic and coordinated the UC@MOOC project in 2013 at UCA. Berrada has led a group of researchers on educational innovation at UCA (TransERIE) and the Morocco Declaration on Open Education since 2016. Berrada is currently the Director of the ICESCO Chair on Open Education (2023–2027). He is also the Director of Higher Education and Pedagogical Development at the Ministry of Higher Education, Scientific Research, and Innovation.

**How to cite this paper:** Azeddine Benelrhali, Khalid Berrada, "Exploring AI Tools and Large Language Models for Students' Performance Enhancement in Riddle Based Logical Reasoning", International Journal of Modern Education and Computer Science(IJMECS), Vol.17, No.5, pp. 1-28, 2025. DOI:10.5815/ijmecs.2025.05.01