ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230388

T-VAKS: A Tutoring-Based Multimodal Dialog System via Knowledge Selection

Raghav Jain[a;*], **Tulika Saha**[b] **and Sriparna Saha**[a]

[a]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[b]University of Liverpool, UK
ORCiD ID: Raghav Jain https://orcid.org/0000-0002-5422-9498

Abstract. Advancements in Conversational Natural Language Processing (NLP) have the potential to address critical social challenges, particularly in achieving the United Nations' Sustainable Development Goal of quality education. However, the application of NLP in the educational domain, especially language learning, has been limited due to the inherent complexities of the field and the scarcity of available datasets. In this paper, we introduce *T-VAKS* (Tutoring Virtual Agent with Knowledge Selection), a novel language tutoring multimodal Virtual Agent (VA) designed to assist students in learning a new language, thereby promoting AI for Social Good. *T-VAKS* aims to bridge the gap between NLP and the educational domain, enabling more effective language tutoring through intelligent virtual agents. Our approach employs an information theory-based knowledge selection module built on top of a multimodal seq2seq generative model, facilitating the generation of appropriate, informative, and contextually relevant tutor responses. The knowledge selection module in turn consists of two sub-modules: (i) knowledge relevance estimation, and (ii) knowledge focusing framework. We evaluate the performance of our proposed end-to-end dialog system against various baseline models and the most recent state-of-the-art models, using multiple evaluation metrics. The results demonstrate that *T-VAKS* outperforms competing models, highlighting the potential of our approach in enhancing language learning through the use of conversational NLP and virtual agents, ultimately contributing to addressing social challenges and promoting well-being.

1 Introduction

Artificial intelligence (AI), particularly deep learning (DL), has profoundly impacted the field of education, offering novel solutions and tools to address critical social challenges. By employing AI and DL techniques, educators can effectively analyze patterns in large datasets, enabling them to better understand and track student performance [10]. The emergence of personalized, data-driven instruction has led to the development of targeted educational material tailored to individual learning styles and abilities [1]. Moreover, AI has improved assessment efficiency through automated feedback on assignments and homework [8]. One of the most promising applications of AI in education is the use of virtual tutors and educational chatbots, which have the potential to enhance the quality of education and make it accessible to a broader range of students [22]. These technologies not only provide personalized, one-on-one tutoring but can also bridge the instructional gaps resulting from a shortage of qualified teachers in certain areas [7]. By offering equal access to educational resources regardless of socio-economic status, virtual tutors can help reduce educational disparities among different demographic groups [2]. Additionally, they can prove particularly beneficial for students with disabilities or those facing language barriers [6]. Virtual automated tutors also play a key role in enhancing student engagement and motivation by delivering personalized and engaging learning experiences [22].

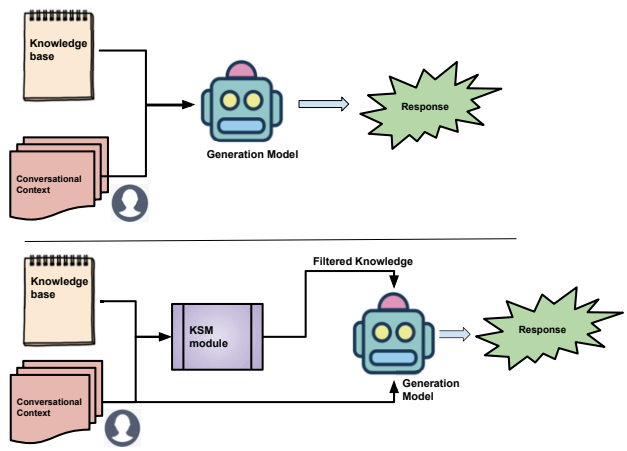


Figure 1: Existing models for Educational Tutoring Vs Our Proposed Approach, ***T-VAKS (Tutoring Virtual Agent with Knowledge Selection)***

Knowledge base (KB) serves as a vital resource for automated tutoring virtual agents (VAs) by enabling them to store and retrieve information related to the topics they teach [4]. Knowledge selection (KS) is a process that identifies one or more relevant pieces of information from a large pool of knowledge stored in a KB to be used in the conversations. Efficient KS is crucial to numerous AI applications, including educational VAs, for several reasons:

- **Selecting relevant knowledge for generating appropriate responses :** Knowledge selection plays a critical role in enabling virtual agents to generate contextually appropriate and meaningful

* Corresponding Author. Email:raghavjain106@gmail.com
[1] https://hbr.org/2019/10/how-ai-and-data-could-personalize-higher-education
[2] https://tech.ed.gov/files/2017/01/NETP17.pdf

responses. By identifying the most relevant pieces of information from the knowledge base, VAs can maintain the coherence and relevance of the conversation [17], ensuring that the generated responses address the user's queries and contribute to their learning experience.

- **Reducing processing time and resources through knowledge selection :** Efficient knowledge selection enables VAs to filter out unnecessary or irrelevant information before it is processed, thus, conserving computational resources. By focusing only on the most relevant knowledge, VAs can reduce the processing time and computational demands [14], resulting in more responsive and efficient systems. This is particularly important in educational settings, where prompt and accurate responses are essential to maintaining user engagement and promoting effective learning.
- **Addressing context length limitations in generative pre-trained language models (PLMs) :** Generative PLMs, such as GPT-2 [20], have fixed input context sizes that limit the amount of information they can process in a single pass [5]. This constraint can lead to issues when dealing with long conversational contexts or complex knowledge bases. Knowledge selection can help mitigate this issue by identifying the most relevant and contextually appropriate information, reducing the input size to fit within the language model's constraints.

Furthermore, incorporating multimodal information, such as images related to the topic being discussed, can enhance the performance of educational VAs by providing additional contextual cues [30]. The combination of different modalities enables VAs to achieve their objectives more accurately and efficiently. By addressing the context length issue and integrating multimodal information, KS can help overcome the inherent limitations of language models, resulting in more effective and resource-efficient VAs for language tutoring and other educational applications. Figure 1 depicts a high-level comparison between the Knowledge Selection-based Dialogue System and conventional dialogue systems.

This paper introduces *T-VAKS* (Tutoring Virtual Agent with Knowledge Selection); a VA that offers personalized, one-on-one tutoring to learn Italian language based on the CIMA dataset [27]. The primary objective of *T-VAKS* is to provide guidance and assistance to students learning the language, helping them to develop their skills and reach their academic goals. The VA's goal is to create a comfortable and encouraging learning environment for the students by offering them competent and expert tutoring. To ensure that the VA is equipped to fulfill this purpose, the KB of the dataset must be utilized to its fullest potential. In addition to this KB, we also use the multimodal information (images) provided in the dataset about the content being discussed. Our end-to-end system employs an information theory inspired Knowledge Selection module (*KSM*) to select relevant information from the KB over the top of a multimodal seq2seq generative model (*MGen*), which is designed to fuse textual and multimodal information to generate an appropriate, informative, and relevant tutor response.

The key contributions of this paper are as follows :

- *We introduce T-VAKS, a Tutoring-based Multimodal Dialog System with Knowledge Selection, imitating the behavior of an ideal tutor. It is capable of conversing with a learner/user on a given subject, aiding them in their language learning experience.*
- *Our end-to-end system employs two sub-modules, viz.,* ***KSM*** *framework to select the relevant information from the KB corresponding to the context, and* ***MGen*** *framework to generate tutor response utilizing the current context, selected knowledge information, and multimodal information (images).*
- *Additionally, the KSM sub-module encompasses two mechanisms, viz.,* ***knowledge relevance estimation*** *to evaluate the significance of each knowledge sentence in the KB, and* ***knowledge focusing mechanism*** *to select most contributing knowledge sentences.*
- *Empirical results indicate that our proposed system outperforms several baseline and state of the art models.*

Social Impact of Our Research. Our work on *T-VAKS* has far-reaching implications in various aspects of society, including: (a) **Enhanced Language Learning Opportunities :** *T-VAKS*, as a language tutoring multimodal virtual agent, provides students with improved language learning experiences; (b) **Increased Access to Quality Education :** *T-VAKS*'s efficiency, utilizing only 40% of the parameters compared to other models, enables deployment on a wider range of devices, including those with limited computational power; (c) **Reduced Environmental Impact :** The resource-efficient nature of *T-VAKS* contributes to lowering the environmental footprint of training and deploying AI models. By requiring fewer parameters, the model promotes the development of more sustainable AI solutions that align with the broader goal of sustainable development; (d) **Encouraging AI for Social Good :** The success of *T-VAKS* highlights the potential of AI-driven educational solutions in addressing societal challenges.

2 Related Works

In this section, we provide a concise summary of the latest developments in the domain of conversational agents, with a special emphasis on the progress made in educational virtual assistants specifically.

2.1 Natural Language Generation and Dialogue Systems

The field of dialogue systems and chatbots has witnessed considerable progress in recent years, driven by advances in natural language processing, machine learning, and artificial intelligence. These developments have led to the creation of more sophisticated and versatile chatbots, which are now employed in various domains, including customer service, e-commerce, and healthcare.

Rule-based and Retrieval-based Chatbots. Early dialogue systems primarily relied on rule-based [29] and retrieval-based [18] approaches. Rule-based chatbots, such as ELIZA [33], used pattern matching and pre-defined rules to generate responses, while retrieval-based systems selected responses from a predefined set based on the similarity between user input and available responses. Although these methods provided a foundation for chatbot development, they were limited in their ability to handle complex and diverse user inputs.

Generative Chatbots. The advent of deep learning and neural networks paved the way for generative models, which allowed chatbots to create responses based on learned patterns rather than relying on predefined rules or templates. Sequence-to-sequence (seq2seq) models [28] and attention mechanisms [2] significantly improved the quality of generated responses, enabling more flexible and coherent dialogue generation. The introduction of pre-trained language models, such as GPT-2 [20], BART [13], and T5 [21], further enhanced

the capabilities of dialogue systems. These models, trained on large-scale text corpora, facilitated the development of more contextually aware and human-like chatbots [23, 31, 24].

2.2 Educational Dialogue Systems

The integration of natural language processing and artificial intelligence in educational settings has given rise to the development of automated tutoring systems and educational chatbots. These systems aim to provide personalized and engaging learning experiences to students, catering to their individual needs and learning styles.

Early Automated Tutoring Techniques. Prior to the advent of virtual agents, automated tutoring predominantly employed techniques like flashcard-based systems [12]. However, these methods offered limited adaptability and interactivity, constraining their effectiveness in addressing diverse learning requirements. The development of rule-based [1] and template-based educational dialogue systems sought to address these limitations. While these systems provided more personalized learning experiences, their reliance on predefined rules and templates restricted their ability to generate flexible, open-ended responses.

Educational Chatbots and Datasets. The advancements in natural language generation and the introduction of pre-trained language models have spurred the development of more sophisticated virtual tutors for various subjects. This has led to the creation of AI-related datasets, such as CIMA [27] and the Teacher-Student Chatroom Corpus [3]. Despite the surge of research on generative models for dialogue systems, the application of large pre-trained models in educational tutoring systems remains limited. Most existing solutions are rule-based and do not generate open-ended responses. A recent study [9] utilized the CIMA dataset to build a conversational agent based on DialoGPT, although it did not take advantage of the image data provided.

In summary, the field of educational dialogue systems has evolved from early flashcard-based techniques to rule-based and template-based systems, and more recently, to the adoption of pre-trained language models. The advancements in natural language generation and artificial intelligence have enabled the development of more sophisticated and interactive virtual tutors, offering personalized learning experiences that cater to diverse student needs.

3 Dataset

This study employs an extended version of the CIMA dataset [27], which focuses on tutoring dialogue in one-to-one student-tutor conversations. An instance of this dataset is shown in Figure 2. The dataset features conversations in which a tutor helps students learn the Italian translation of an object and its features. The dataset includes object descriptions, grammar rules, intent tags, and action labels. These object descriptions and grammar rules are referred to as the knowledge base (KB) in the dataset. Student intent tags consist of Guess, Question, Affirmation, and Other, while VA's actions are labeled as Question, Hint, Correction, Confirmation, and Other. In the original CIMA dataset, intent-action labels are provided only for the final student utterance, and the gold tutor response, in each data instance, resulting in 2983 response-context pairs. However, this limited number of instances hampers the training of neural response generation models, which typically require substantial amount of data. To address this limitation, the authors of the EDICA framework [9] adopt a semi-supervised approach for labeling every utterance in the conversational history with the corresponding intent-action tags, thus, increasing the dataset length suitable for training neural generation models. They train a student intent classifier (SIC) and a tutor intent/action classifier (TIC) on the original gold-standard CIMA dataset. These classifiers are then used to assign silver-standard intent-action tags to the respective student-tutor utterances in the conversational history of each data instance present in the original CIMA dataset. The modified CIMA dataset with silver-standard labeling is referred to as the *extended-CIMA* dataset. The *extended-CIMA* dataset now encompasses 4322 response-context pairs, offering a more comprehensive collection of student-tutor conversations for training purposes.

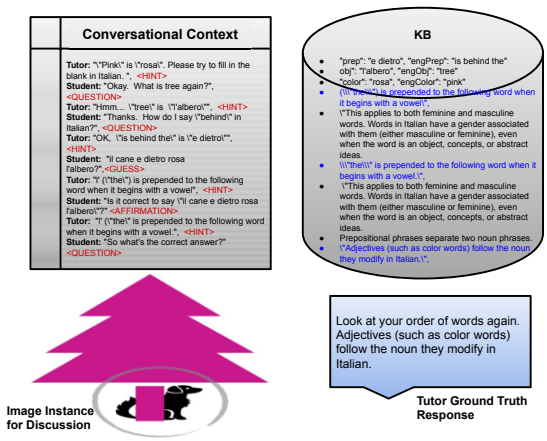


Figure 2: Illustration of a data sample from the CIMA dataset. Sentences highlighted in blue within the Knowledge Base signify that these sentences were utilized to produce the subsequent ground truth instructor response. In the conversational context, intent tags and action labels for utterances are displayed in red.

4 Proposed Methodology

This section discusses the problem statement, methodology and implementation details.

4.1 Problem Formulation

This paper addresses the challenge of generating appropriate and meaningful responses for a virtual agent (VA) in student-tutor conversations, considering various contextual factors. Given a student utterance, S_{n+1}, the conversation history, H, a knowledge base, K, the student intent I_{n+1}, the VA action a_{n+1}, and an image M, the objective is to produce a suitable tutor reply, $\hat{T_{n+1}}$. The primary goal is to develop a generative language model, G capable of generating such a response. Formally, the problem can be defined as follows: given the inputs, S_{n+1}, H, K, I_{n+1}, a_{n+1}, and M, the model G should produce a response $\hat{T_{n+1}}$ that adequately addresses the student's query and adheres to the context of the conversation. In other words, $\hat{T_{n+1}} = G(S_{n+1}|H, K, I_{n+1}, a_{n+1}, M)$.

4.2 Methodology

This section discusses the proposed approach in details.

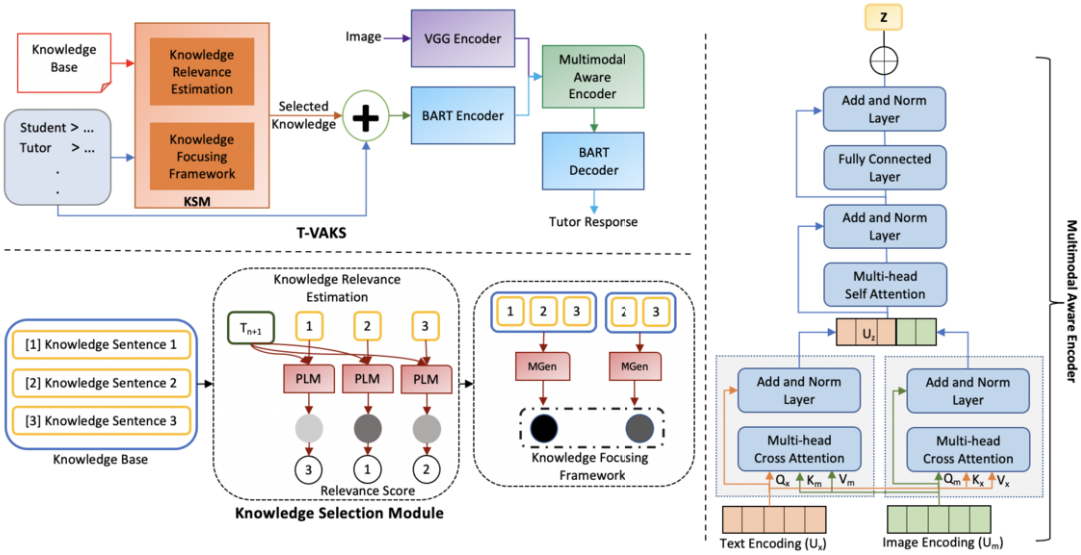


Figure 3: This figure presents the schematic design of the proposed *T-VAKS* framework. The top-left segment provides a comprehensive summary of the model, while the lower-left segment highlights the proposed knowledge selection module. On the right, the proposed multimodal generative encoder is demonstrated.

4.2.1 *Knowledge Selection Module (KSM)*

In previous research on tutor-student conversations, the primary focus has been on utilizing the entire domain knowledge base (KB) as an input to the system. As illustrated in Figure 2, it is evident that merely 3 out of the 9 sentences in the knowledge base (accentuated in blue) are employed to generate the ground truth response. However, this approach may not be optimal, as it can lead to processing large amount of irrelevant information, increasing the computational overhead and potentially diluting the context of the conversation. To address this challenge, this paper proposes a knowledge selection module (KSM) that is specifically designed to select pertinent and significant knowledge sentences from the entire KB, depending on the current context. The primary objective of KSM is to select relevant knowledge sentences k_i from the entire KB K, considering the prior n turns, $H = \{S_1, T_1, ..., S_n, T_n\}$, and the student's utterance, S_{n+1}. By selectively choosing the most relevant pieces of knowledge, KSM can enhance the efficiency and contextual relevance of the virtual agent's responses. KSM consists of the following two submodules that work in tandem to achieve this goal :

Knowledge Relevance Estimation. During training, the aim is to select the most relevant knowledge sentences based on the ground truth tutor response, T_{n+1}. The quality of a knowledge sentence can be evaluated by computing its information overlap with the tutor response, T_{n+1}. A high overlap of information guarantees that the knowledge sentence must be utilized or is pertinent in creating the tutor response; thus, helping us to evaluate the significance of each knowledge sentence in KB, K. To model this information overlap, a Shannon score based mechanism is employed to estimate the importance and relevance of each knowledge sentence k_i where $k_i \in K$ which works as follow : Given a language model, $P(k_i|T_{n+1})$ that outputs the probability distribution of a knowledge sentence, k_i associated with a given tutor response, T_{n+1}, we calculate the conditional information content, $InfoGain(k_i|T_{n+1})$ as the amount of knowledge gained from k_i through knowing the response T_{n+1}. The amount of knowledge gained is quantified as the shannon information content :

$$InfoGain(k_i|T_{n+1}) = -\log P(k_i|T_{n+1}) \quad (1)$$

To approximate $P(k_i|T_{n+1})$ and $InfoGain(k_i|T_{n+1})$, we compute the conditional probability of k_i being generated when T_{n+1} is provided as a prompt to a language model (GPT-2 in our case) [20]. We introduce a shannon score metric as :

$$SS = InfoGain(k_i) - InfoGain(k_i|T_{n+1}) \quad (2)$$

If k_i is of high importance and relevance, then the value of $InfoGain(k_i|T_{n+1})$ will be much smaller than $InfoGain(k_i)$, which implies that a large value for the *SS* score shows that sentence k_i is relevant. The Shannon score SS can be used to rank all knowledge sentences according to relevance, and then the Top-R most relevant knowledge sentences can be used as input for the model.

Knowledge Relevance during Inference : During inference, the ground-truth tutor response, T_{n+1} is unavailable to estimate the knowledge relevance. For this, we propose to use a Kullback-Leibler divergence loss [15] which works as follows: two conditional probability distributions are defined; Prior Distribution $P(k|S_{n+1}, H)$ and Posterior Distribution $P(k|S_{n+1}, H, T_{n+1})$ for a knowledge sentence k_j are defined as follows :

$$P(k_j|S_{n+1}, H) = \frac{\exp(k_j \cdot (S_{n+1} \oplus H))}{\sum_{k \in K} exp(k \cdot (S_{n+1} \oplus H))} \quad (3)$$

$$P(k_j|S_{n+1}, H, T_{n+1}) = \frac{\exp(k_j \cdot (MLP[S_{n+1} \oplus H; T_{n+1}]))}{\sum_{k \in K} exp(k \cdot (MLP[S_{n+1} \oplus H; T_{n+1}]))} \quad (4)$$

where MLP is a fully-connected feed-forward network. The KL-divergence loss, L_{kl} can be computed as follows :

$$L_{kl} = \sum_{k \in K} P(k_j|S_{n+1}, H, T_{n+1}) \log \frac{P(k_j|S_{n+1}, H, T_{n+1})}{P(k_j|S_{n+1}, H)} \quad (5)$$

The KL-Divergence loss function assists the model in selecting relevant knowledge sentences, even when the ground-truth tutor response is unknown (during inference), by approximating the posterior distribution using the prior distribution.

Knowledge Focusing Framework. Additionally, we propose a knowledge focus loss to encourage the model to produce the same output distribution for tutor response when the entire knowledge base K is fed as input and when only top ranked knowledge sentences are fed. Formally, knowledge focus loss can be defined as follows :

$$L_{kf} = (P \log P^{'\{k\}} + (1-P)\log(1-P^{'\{k\}})) \quad (6)$$

where P represents probability distribution of model $G(S_{n+1}|H, K, I_{n+1}, a_{n+1}, M)$ with complete KB, K. k represent the top ranked knowledge sentences from K, and $P^{'\{k\}}$ represents probability distribution of model G with only top ranked sentences k in the input instead of K. Primarily, the knowledge focus loss guarantees that P is nearly identical to $P^{'\{k\}}$ so that excluding non-important knowledge sentences will not effect the result of our generative model. This penalization will take into account the discrepancy in predicted probabilities when non-important sentences are removed from the knowledge base.

4.2.2 *Multimodal Generative Framework (MGen)*

Firstly, the current student utterance S_{n+1}, history H, student's intent, I_{n+1}, and VA's action, a_{n+1}, are concatenated to form an input string C_i. Additionally, KB K is fed into the *KSM* module that outputs the top-ranked knowledge sentences $\{k\}$. Finally, C_i and $\{k\}$ are concatenated to form a final input string X_i. Next, both the input string X_i and corresponding image M are correspondingly fed into a pre-trained BART encoder and VGG-19 encoder [26] to obtain encoded representations, U_x and U_m, respectively. To fuse the information between these two representations, we propose a Multimodal-aware encoder (shown in Fig. 3), an extension of the original transformer encoder [32]. We create two triplets of query, key and value matrices corresponding to U_x and U_m, respectively: (Q_x,K_x,V_x) and (Q_m,K_m,V_m). Unlike the original transformer encoder where the same input is projected as query, key, and value, in *MGen*, we propose a cross-attention layer with two sublayers of multi-head-cross attention and normalization. This layer exchanges the key and value by taking (Q_x,K_m,V_m) and (Q_m,K_x,V_x) as inputs to compute a cross infused vector representation defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (7)$$

where (Q,K,V) represents the set of query, key, and value and d_k represents the dimension of the query and key. This cross-attention layer facilitates the exchange of information between U_x and U_m. Now, these multi-head cross attention outputs ($U_{x->m}$ and $U_{m->x}$) contain information about each other. Following this, $U_{x->m}$ and $U_{m->x}$ are concatenated and the final concatenated output U_z is passed through a self-attention layer, normalization layers, and fully connected layers with residual connections to obtain final multimodal-aware input representation vector, Z. Finally, Z is fed to an autoregressive decoder following the standard decoder computations as defined in the original transformer network [32].

4.2.3 *Loss Function*

We initialize our model's weights θ using the weights of the pre-trained sequence-to-sequence generative model (BART-base). Our framework, *T-VAKS*, is then fine-tuned by optimizing a combined loss function defined in Equation 8, which is the average of the following three loss components:

- **Maximum Likelihood Estimation (MLE) Objective Function:** This loss (L_{mle}) component operates in a supervised manner to optimize the weights, θ, by minimizing the difference between the predicted and ground truth responses, encouraging the model to generate accurate and coherent responses during training.
- **Kullback-Leibler (KL) Divergence Loss:** As defined in Equation 5, the KL divergence loss (L_{kl}) helps to estimate the knowledge relevance during inference. This loss component measures the divergence between the predicted knowledge relevance distribution and the ground truth distribution, promoting the selection of relevant knowledge sentences from the knowledge base, which contributes to generating more contextually appropriate responses.
- **Knowledge Focusing Loss:** As defined in Equation 6, the knowledge focusing loss (L_{kf}) penalizes the model for attending to irrelevant knowledge sentences. By applying this loss component, the model is encouraged to focus on the most relevant parts of the knowledge base, which leads to improved response generation and a more effective language tutoring experience.

$$L_{final} = \frac{L_{mle} + L_{kl} + L_{kf}}{3} \quad (8)$$

4.3 *Experimental Setup*

We trained our models on a Tyrone machine equipped with an Intel Xeon W-2155 processor and a 11 GB Nvidia 1080Ti GPU. The training was conducted for 20 epochs with a learning rate of 5 x 10^{-5}, a batch size of 16, using the Adam optimizer, and an Adam epsilon value of 1 x 10^{-8}. Our proposed model, as well as all the ablated models, are built on top of the BART-Base architecture. After thorough investigation, we set the Top-R parameter for Knowledge Relevance Estimation to 3. All models were implemented using Scikit-Learn[3] and PyTorch[4]. The performance of generative models was evaluated using several metrics, including BLEU score [19], ROUGE-L score [16], BERT F1-score [35], and Embedding-based Metric [25]. To assess the performance of these models in terms of human evaluation, three independent human users from the authors' affiliation were asked to rate 100 simulated dialogues on a scale of 1 (worst), 3 (moderate), and 5 (best) based on two criteria: *fluency* (grammatical correctness) and *relevance* (response conditioning based on the conversation's trajectory). The final reported score is the average of the human-rated scores.

5 Results and Discussion

This section discusses the results and the experimental findings in detail.

Comparison with the Baselines. The automatic evaluation results of different baselines and the proposed *T-VAKS* model are shown in Table 1. A clear observation from the table is that GPT-based models significantly outperform text-based BART and T5 models. However, the absence of an encoder in GPT models limits their ability to integrate multiple information sources. This limitation inspired the development of a multimodal and knowledge selection framework based on the BART-base model, which achieved results comparable to GPT-2.

[3] https://scikit-learn.org/
[4] https://pytorch.org/

Table 1: Comparison of *T-VAKS* with other baselines on automated metrics; T:Text, I:Image, KRE: Knowledge Relevance Estimation, SS: Shannon Score, KFF: Knowledge Focusing Framework. The maximum scores attained are represented by bold-faced values. The † denotes statistically significant findings.

Model	BLEU	BERT F1	ROUGE-L	Embedding Metric		
				Average	Extrema	Greedy
GPT-1_T	0.33	0.49	0.53	0.67	0.41	0.60
GPT-2_T (Base)	0.34	0.54	0.55	0.69	0.41	0.61
$BART_T$ (Base)	0.33	0.52	0.55	0.67	0.40	0.60
$T5_T$ (Base)	0.32	0.52	0.54	0.67	0.39	0.60
$BART_T$+ROUGE (Base)	0.34	0.56	0.55	0.69	0.41	0.63
$BART_T$+WMD (Base)	0.34	0.57	0.54	0.69	0.42	0.63
$BART_{T+I}$ (Base)	0.34	0.53	0.55	0.68	0.41	0.62
$T5_{T+I}$ (Base)	0.34	0.52	0.55	0.69	0.40	0.62
$BART_{T+I}$ (Base)+KRE	0.36	0.57	0.56	0.70	0.43	0.65
$BART_{T+I}$ (Base)+(KRE-SS)	0.35	0.57	0.54	0.70	0.42	0.64
$BART_{T+I}$ (Base)+(KFF+KRE-SS)	0.36	0.58	0.57	0.70	0.43	0.66
PostKS [15]	0.34	0.50	0.53	0.70	0.41	0.65
T-VAKS	**0.39**†	**0.60**†	**0.58**†	**0.72**†	**0.45**†	**0.67**†

Incorporating images as multimodal information in the model ($BART_{T+I}$) enhances its performance compared to the vanilla BART model, as demonstrated in Table 1. We created two vanilla knowledge selection models ($BART_T$+ROUGE and $BART_T$+WMD) by selecting the top-2 knowledge sentences based on ROUGE/WMD [11] similarity of the context and the KB. The performance of both models surpasses that of all text-based GPT models and other generative models (BART and T5), indicating the effectiveness of knowledge selection in improving performance.

Table 1 reveals that the *T-VAKS* framework significantly outperforms its predecessor models across all evaluation metrics. This improvement can be attributed to two factors: (1) the addition of an image as an extra information source broadens the conversational context of the model, allowing for better predictions, and (2) selecting only relevant knowledge instead of the complete KB enhances the model's generation capabilities by focusing on pertinent parts and reducing the context length, thereby eliminating the need for truncation.

We also present a comparison of our model, *T-VAKS*, with the widely recognized Knowledge Selection-based Dialogue generation model, PostKS [15]. PostKS employs a prior distribution derived solely from utterances to approximate the posterior distribution, facilitating the selection of relevant knowledge. Nevertheless, *T-VAKS* surpasses PostKS by a considerable margin across all assessment metrics, which can be ascribed to three main factors: (1) In addition to the KL-Divergence loss used by PostKS to approximate the posterior distribution, our model incorporates Shannon score-based estimation during training, further enhancing its performance; (2) Unlike PostKS, our model takes advantage of additional modalities, such as images; and (3) *T-VAKS* employs pre-trained language models in its architecture, whereas PostKS utilizes standard GRU layers as encoder-decoder components.

We demonstrate the impact of each *T-VAKS* component by observing how adding each component over $BART_{T+I}$ affects the model's performance. Integrating the KRE module into the multimodal BART results in improvements compared to using Multimodal BART alone. The KRE module filters pertinent knowledge sentences for processing by the model, effectively reducing irrelevant information and making the input more efficient. The performance of the KRE module declines when the Shannon Score (SS) component is omitted, suggesting that SS assists the KRE module in identifying accurate and relevant knowledge sentences, given its utilization of the ground truth responses for knowledge relevance estimation. Layering the KFF on top of this model further bolsters its performance, as the knowledge focus loss helps the model concentrate on the most relevant knowledge areas, making the model more robust.

Comparison with the SOTA. In this section, we compare the performance of our proposed model, *T-VAKS*, with state-of-the-art models from the literature, specifically [27] and [9]. The comparison, as shown in Table 2a, is based on the same experimental settings. Our model, *T-VAKS*, demonstrates superior performance compared to the generation-based model [27], outperforming it by a significant margin on both BLEU and BERT-F1 metrics. Specifically, *T-VAKS* surpasses the generation-based model by 8% and 11% in terms of BLEU and BERT-F1 scores, respectively. Furthermore, *T-VAKS* outperforms all variants of EDICA [9] in terms of BERT F1 and BLEU scores. In comparison to $EDICA_{GPT-2BASE}$, *T-VAKS* outperforms it by a margin of 3% and 5% in BLEU and BERT-F1 scores, respectively. When compared to $EDICA_{DialoGPT}$, *T-VAKS* exhibits almost similar performance in terms of BLEU score while showing a 2% improvement in BERT-F1 measure. *All of the results are statistically significant*[5] *[34].*

Comparison of Parameters. The findings presented in Table 2b highlight the remarkable efficiency of *T-VAKS*, as it achieves comparable performance to both $EDICA_{GPT-2Medium}$ and $EDICA_{DialoGPT}$ while utilizing only 40% of the parameters. This significant reduction in the number of parameters showcases the ability of *T-VAKS* to deliver high-quality language tutoring without the need for extensive computational resources, setting it apart from other models in the field. The efficiency and effectiveness of *T-VAKS* have substantial implications for the field of NLP, particularly in addressing the growing demand for more efficient and accessible language technologies. By requiring fewer parameters, *T-VAKS* can be deployed on a wider range of devices, including those with limited computational power, thereby increasing its potential reach and impact on language learning across diverse communities. Furthermore, the reduced parameter count contributes to lowering the environmental footprint of training and deploying AI models, aligning with the

[5] We used the Student's t-test (p-value < 0.04).

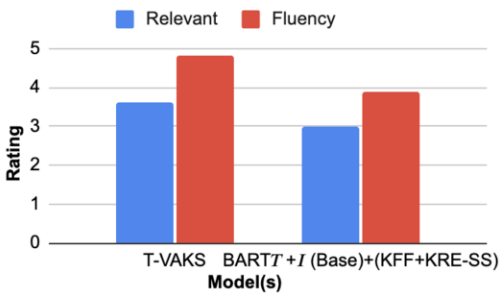
Table 2: (a) Comparison of *T-VAKS* with the State of the Art Models. The maximum scores attained are represented by bold-faced values. The † denotes statistically significant findings., (b) The trainable parameters for the proposed model, baselines and SOTA models

(a)

Model	BLEU	BERT F1
Generation based [27]	0.30	0.49
EDICA$_{GPT-1}$ [9]	0.34	0.57
EDICA$_{GPT-2_{Base}}$ [9]	0.36	0.55
EDICA$_{GPT-2_{Medium}}$ [9]	0.38	0.57
EDICA$_{DialoGPT}$ [9]	0.38	0.58
T-VAKS	**0.39**†	**0.60**†

(b)

Model	Total Parameters
EDICA$_{GPT-1}$ [9]	110000000
EDICA$_{GPT-2_{Base}}$ [9]	117000000
EDICA$_{GPT-2_{Medium}}$ [9]	345000000
EDICA$_{DialoGPT}$ [9]	345000000
T5	222903552
BART	139001589
T-VAKS	139420416



(a)

Example-1
Context : ['please try to fill in the blank in italian. ', 'how do you say in front of?', 'okay i'll give you a hint. "Is in front of the" is "ei di fronte"']
Student : and how do you say purple?
Ground Truth : viola is the word for purple
T-VAKS : hmm…"purple" is "viola"
BART$_{T+I}$(Base)+(KRE-SS+KFF) : "purple is voila"
Example-2
Context : ['please try to fill in the blank in italian.', 'how do you say blue in italian?', 'ok, "blue" is "blu"']
Student : how do you say under in italian?
Ground Truth : under is sotto
T-VAKS : remember that "is under the" is "e sotto"
BART$_{T+I}$(Base)+(KRE-SS+KFF) : "under" is "e sotta"

(b)

Figure 4: (a) Human evaluation scores of different models, (b) Sample response generated by the proposed model

broader goal of sustainable development.

Human Evaluation. We conducted a human evaluation to further assess the performance of *T-VAKS* compared to the best-performing baseline model (*BART$_{T+I}$ (Base)+(KFF+KRE-SS)*). As depicted in Figure 4a, *T-VAKS* demonstrates significant improvement over the baseline model. On average, *T-VAKS* scored 4.8 in fluency and 3.6 in relevance, outperforming the baseline model, which received an average score of 3.9 and 3.0 for fluency and relevance, respectively. These results indicate that *T-VAKS* generates outputs that are not only more contextually relevant but also exhibit superior syntactic accuracy. The human evaluation scores corroborate the quantitative findings, providing strong evidence that *T-VAKS* is an effective language tutoring tool.

We present sample-generated responses from *T-VAKS* in Figure 4b to provide qualitative insight into the system's performance. These examples demonstrate the model's ability to generate contextually relevant, informative, and syntactically accurate responses that effectively address students' inquiries or concerns. By examining the sample outputs in Figure 4b, we can observe that both *T-VAKS* and baselines successfully adapt to various language learning scenarios, ranging from vocabulary clarification to generating grammatical hints. However, it can be observed that *T-VAKS* responses are more fluent and complete as compared to the baseline. By generating more pertinent and syntactically accurate responses, *T-VAKS* has the potential to significantly enhance language learning experiences for students.

6 Conclusion

In conclusion, we have introduced *T-VAKS*, a cutting-edge language tutoring multimodal Virtual Agent (VA) designed to support students in learning new languages, effectively addressing critical societal challenges in education. This groundbreaking VA leverages an information theory-based knowledge selection approach to extract relevant information from an educational knowledge base, seamlessly integrated with a multimodal generative model. Our empirical results demonstrate the superior performance of *T-VAKS*, outshining baseline models and state-of-the-art alternatives across a diverse array of evaluation metrics. By enhancing language learning experiences through the utilization of conversational NLP and virtual agents, *T-VAKS* significantly contributes to addressing social challenges in education and promoting well-being, in accordance with the United Nations' Sustainable Development Goal of quality education. The success of *T-VAKS* in improving educational outcomes also encourages further research and development of AI-driven solutions that target social issues and contribute to a more equitable society.

Looking ahead, future work will explore extending educational chatbots like *T-VAKS* to additional tasks, such as summarization and personalized learning, further amplifying their impact on solving societal challenges in education and fostering more inclusive and accessible learning opportunities for all.

7 Acknowledgement

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

[1] Vincent Aleven, Octav Popescu, and Kenneth R Koedinger, 'A tutorial dialogue system with knowledge-based understanding and classification of student explanations', in *Working Notes of 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Citeseer, (2001).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', *arXiv preprint arXiv:1409.0473*, (2014).

[3] Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery, 'The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts', in *Swedish Language Technology Conference and NLP4CALL*, pp. 23–35, (2022).

[4] J.J. Castro-Schez, J. Gallardo, R. Miguel, and D. Vallejo, 'Knowledge-based systems to enhance learning', *Know.-Based Syst.*, **122**(C), 180–198, (apr 2017).

[5] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov, 'Transformer-xl: Attentive language models beyond a fixed-length context', *arXiv preprint arXiv:1901.02860*, (2019).

[6] Reham El Shazly, 'Effects of artificial intelligence on english speaking anxiety and speaking performance: A case study', *Expert Systems*, **38**, (05 2021).

[7] Harry Barton Essel, Dimitrios Vlachopoulos, Akosua Tachie-Menson, Esi Eduafua Johnson, and Papa Kwame Baah, 'The impact of a virtual teaching assistant (chatbot) on students' learning in ghanaian higher education', *International Journal of Educational Technology in Higher Education*, **19**(1), 1–19, (2022).

[8] Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs, 'Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8577–8591, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[9] Raghav Jain, Tulika Saha, Souhitya Chakraborty, and Sriparna Saha, 'Domain infused conversational response generation for tutoring based virtual agent', in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, (2022).

[10] Ijaz Khan, Abdul Ahmad, Nafaa Jabeur, and Mohammed Mahdi, 'An artificial intelligence approach to monitor student performance and devise preventive measures', *Smart Learning Environments*, **8**, (09 2021).

[11] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger, 'From word embeddings to document distances', in *Proceedings of the 32nd International Conference on Machine Learning*, eds., Francis Bach and David Blei, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966, Lille, France, (07–09 Jul 2015). PMLR.

[12] S. Leitner and R. Totter, *So lernt man lernen*, Angewandte Lernpsychologie ein Weg zum Erfolg, Herder, 1972.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, 'Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension', pp. 7871–7880, (01 2020).

[14] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. Learning to select knowledge for response generation in dialog systems, 2019.

[15] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu, 'Learning to select knowledge for response generation in dialog systems', *arXiv preprint arXiv:1902.04911*, (2019).

[16] Chin-Yew Lin, 'Rouge: A package for automatic evaluation of summaries', p. 10, (01 2004).

[17] Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, and Dilek Hakkani-Tur. Knowledge-grounded conversational data augmentation with generative conversational networks, 2022.

[18] Kristen Moore, Shenjun Zhong, Zhen He, Torsten Rudolf, Nils Fisher, Brandon Victor, and Neha Jindal. A comprehensive solution to retrieval-based chatbot construction, 2021.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 'Bleu: A method for automatic evaluation of machine translation', in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 311–318, USA, (2002). Association for Computational Linguistics.

[20] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 'Language models are unsupervised multitask learners', (2019).

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, 'Exploring the limits of transfer learning with a unified text-to-text transformer', *Journal of Machine Learning Research*, **21**(140), 1–67, (2020).

[22] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay, 'Quizbot: A dialogue-based adaptive learning system for factual knowledge', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–13, New York, NY, USA, (2019). Association for Computing Machinery.

[23] Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha, 'Towards motivational and empathetic response generation in online mental health support', in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 2650–2656, New York, NY, USA, (2022). Association for Computing Machinery.

[24] Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya, 'A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony', in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2436–2449, Seattle, United States, (July 2022). Association for Computational Linguistics.

[25] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio, 'A hierarchical latent variable encoder-decoder model for generating dialogues', in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 3295–3301. AAAI Press, (2017).

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[27] Katherine Stasaski, Kimberly Kao, and Marti A. Hearst, 'CIMA: A large open access dialogue dataset for tutoring', in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 52–64, Seattle, WA, USA → Online, (July 2020). Association for Computational Linguistics.

[28] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, 'Sequence to sequence learning with neural networks', in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, p. 3104–3112, Cambridge, MA, USA, (2014). MIT Press.

[29] Sandeep A Thorat and Vishakha Jadhav, 'A review on implementation issues of rule-based chatbot systems', in *Proceedings of the international conference on innovative computing & communications (ICICC)*, (2020).

[30] Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari, 'Dr. can see: Towards a multi-modal disease diagnosis virtual assistant', in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, p. 1935–1944, New York, NY, USA, (2022). Association for Computing Machinery.

[31] Abhisek Tiwari, Tulika Saha, Sriparna Saha, Shubhashis Sengupta, Anutosh Maitra, Roshni Ramnani, and Pushpak Bhattacharyya, 'A persona aware persuasive dialogue policy for dynamic and co-operative goal setting', *Expert Systems with Applications*, **195**, 116303, (2022).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in Neural Information Processing Systems*, eds., I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, volume 30. Curran Associates, Inc., (2017).

[33] Joseph Weizenbaum, 'Eliza—a computer program for the study of natural language communication between man and machine', *Commun. ACM*, **9**(1), 36–45, (jan 1966).

[34] Bernard L Welch, 'The generalization of 'student's'problem when several different population varlances are involved', *Biometrika*, **34**(1-2), 28–35, (1947).

[35] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi, 'Bertscore: Evaluating text generation with bert', in *International Conference on Learning Representations*, (2020).