I.J. Modern Education and Computer Science, **2026, 1, 104-125**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2026.01.07

EduAgent: A Multimodal Chatbot Framework for Enhancing Interactive Learning in Electrical and Electronics Engineering Education

Sayem Shahad*
Bangladesh University of Engineering and Technology/Computer Science and Engineering, Dhaka, 1000, Bangladesh
E-mail: 1905064@ugrad.cse.buet.ac.bd
ORCID iD: https://orcid.org/0009-0005-1965-0892
*Corresponding Author

Salman Sayeed
Bangladesh University of Engineering and Technology/Computer Science and Engineering, Dhaka, 1000, Bangladesh
E-mail: 1905079@ugrad.cse.buet.ac.bd
ORCID iD: https://orcid.org/0009-0004-2054-968X

Md Naimur Rahman Khan Sifat
University of Asia Pacific/Electrical and Electronic Engineering, Dhaka, 1205, Bangladesh
E-mail: 22208152@uap-bd.edu
ORCID iD: https://orcid.org/0009-0009-6934-3769

Shuvo Biswas
University of Asia Pacific/Electrical and Electronic Engineering, Dhaka, 1205, Bangladesh
E-mail: 22220004@uap-bd.edu
ORCID iD: https://orcid.org/0009-0003-8696-9014

Tishna Sabrina
University of Asia Pacific/Electrical and Electronic Engineering, Dhaka, 1205, Bangladesh
E-mail: tishna@uap-bd.edu
ORCID iD: https://orcid.org/0000-0002-6072-6386

Received: 28 January, 2025; Revised: 28 May, 2025; Accepted: 08 November, 2025; Published: 08 February, 2026

Abstract: In recent days, we have largely adopted Advanced Large Language Models (LLMs) in educational settings, where we use them as content creators, teaching assistants, and interactive conversation agents. However, the responses generated by these models are often monotonous, verbose, and ambiguous, which can hinder their effectiveness in educational contexts. Addressing these shortcomings, we introduce EduAgent, a multimodal chatbot framework specifically designed to enhance interactive learning in Electrical and Electronics Engineering (EEE) education. EduAgent can respond with pedagogically enhanced answers to electronics-related queries, complemented by relevant images and detailed explanations. It is designed to provide complete, concise, step-by-step responses, ensuring that foundational knowledge is clearly mentioned before diving deep. To develop EduAgent, we constructed a dataset comprising 596 four-turn conversations and a collection of 118 images covering a wide range of EEE concepts. The conversation dataset was used to fine-tune the open-source LLMs and facilitate in-context learning. Both images and their corresponding explanations were integrated into a knowledge base for efficient retrieval. Finally, we evaluated multiple text generation and image retrieval methods using both automatic metrics and human assessments, demonstrating the effectiveness and engagement of our approach.

Index Terms: LLM, Chat Agent, Retriever, Pedagogical Response, Prompt Engineering, Fine Tuning

1. Introduction

Artificial intelligence (AI) has revolutionized education from different angles such as learning, teaching, and management [1]. The emergence of large language models (LLMs) such as OpenAI's GPT (Generative Pretrained

Transformer) [2], Google's Gemini [3], and BERT (Bidirectional Encoder Representations from Transformers) [4] has made tasks such as multimodal content creation, AI-driven teaching assistance, collaboration, and personalized assessment more accessible and effective.

With the integration of AI and LLMs in the educational fields, different pedagogical approaches, such as learning through teaching [5], practicing interpretive thinking [6], emphasizing the teacher-student relationship in online learning [7], etc have been explored in many research studies. LLMs have contributed to making learning more interactive, dynamic, and personalized. Researchers are developing tools to enhance the performance, efficacy, and quality of LLM-generated content, often focusing on specific domains [8, 9].

LLMs have helped explore methods such as simulating classroom environments [10], simulating teacher and student bots to facilitate learning through teaching [11], enhancing collaborative learning among students [12], training teachers with student agents, and recommending personalized pedagogies [13]. However, LLM-driven teaching assistants still fall behind when they are used to replicate the critical thinking, creativity, and human emotion essential for effective teaching [14].

In this paper, we present EduAgent, a tool designed to respond to learners' queries with correct, concise, complete, and engaging responses, complemented by appropriate images. The primary challenges we address are twofold: the lack of interactive and engaging educational tools in the realm of Electrical and Electronics Engineering (EEE), and the difficulties learners and professionals face when seeking timely, accurate answers that are explained in short, simple steps, building on the learner's prior knowledge.

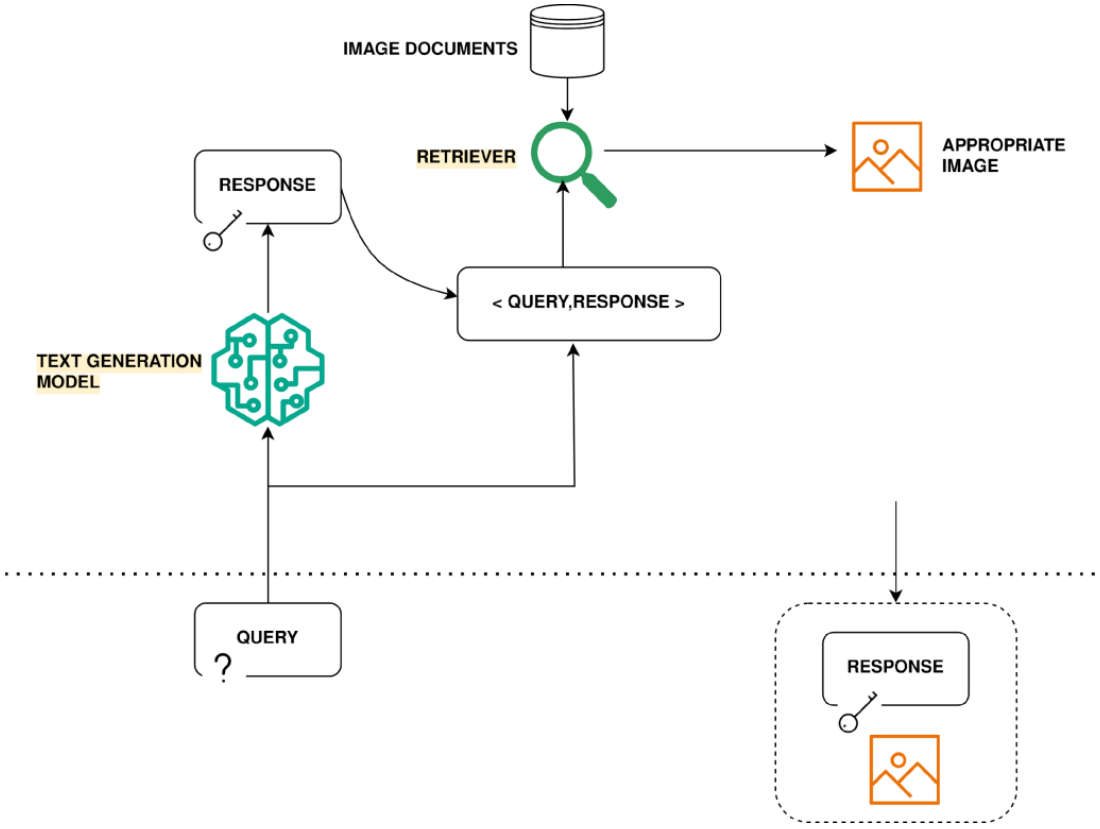


Fig. 1. The overall workflow of the chatbot. When a user asks a query, the chatbot invokes the text generation model to generate a pedagogically enhanced response. Later, the query-response pair is passed through the retriever which fetches the most relevant image and corresponding explanation from the image documents. The text response, image, and image explanation are returned to the user end.

In [15], authors suggested that beginning a lesson with a short preview of previous learning, and teaching in small steps are good practices. Moreover, learners want a response that mentions what they need to know in prior for efficient learning. To achieve these goals, we designed multiple text response generators using various prompting techniques and fine-tuned models. The images and their explanations are retrieved from a database containing 118 images we collected, covering a wide range of EEE concepts. Our evaluations, which included both automatic metrics and human assessments, revealed that in-context learning through prompting outperformed other methods in terms of response format understanding.

In Section 2, we review previous research related to our work. Section 3 details the process of generating the conversa- tion dataset, designing the text generators, and fine-tuning and prompting the models. Section 4 evaluates text generation and image retrieval through automated processes, while Section 5 discusses the evaluations involving learners and experts. Finally, in Section 6, we explore the potential and future directions for EduAgent.

2. Background and Related Work

In this section, we cover the theoretical background and discuss related work in areas including Pedagogical Agents, Large Language Models (LLMs), and prompting techniques. These concepts form the backbone of AI applications in educational settings, particularly in technical domains such as Electrical and Electronics Engineering (EEE).

2.1 Background

2.1.1 Large Language Models (LLMs)

As we entered the transformer era, parallel processing made it possible to train models having large numbers of parameters with large training corpus. BERT [16] by google was considered as the state of the art for question-answering and sentiment analysis. But as OpenAI's GPT models [2, 17] and Google's Gemini [3] started having more and more parameters, we got the true understanding of these models' ability to generate content. Some large language models with multimodal abilities can be used to diversify their usage. Large Language Models have come into prominence in health, education, content creation, because of the resemblance they have with human mechanisms in processing and generating. Large language models can also learn to align their generation according to the user intent. OpenAI's Codex [18] showcased LLMs' ability to write code. Google's PaLM [19] and DeepMind's Chinchilla [20] introduced scaling laws for optimal model training. Pre-trained models can be fine-tuned or prompted to return task specific responses. Open source models like Llama [21], mistral [22], and phi [23] not only made LLMs accessible to all but also made them lighter and power efficient for casual uses.

2.1.2 Prompting

Prompting facilitates downstream classification and domain-specific tasks where precision and clarity are required. It is based on methods such as zero-shot and few-shot prompting, which is essential in generalizing AI applications.

While zero-shot assumes no prior examples or instructions over any tasks, this becomes important in those applications when labeled data is non-existent since models can classify something or create an inference from general knowledge only [24]. Few-shot prompting, on the other hand, uses some examples to support the model in improving the generalization and precision if not much labeled data are available.

These prompting methodologies have demonstrated adaptability to tasks without requiring full datasets or fine-tuning. For instance, frameworks such as ProP (prompting as Probing) [25] enhance classification and information extraction through organized prompting methods. In addition, several other techniques, including Ask Me Anything Prompting (AMA) [26], that involve the combination of imperfect prompts on classification tasks, accumulate noisy but useful votes for making predictions with higher accuracy compared to regular few-shot techniques.

Advanced techniques like chain-of-thought [27] and self-consistency prompting [28] further improved model performance by encouraging orderly reasoning and iterative problem-solving. The methods found their highest value in educational applications since they can easily allow students to iteratively refine questions for deeper understanding [24].

2.1.3 Token Based Retrieval and Model Finetuning

A retrieval system retrieves documents that are similar to the given text or query. The similarity or alignment is calculated using different techniques. Tokenization transforms raw textual data into manageable units (tokens) such as words, subwords, or characters. Each token is mapped to a numerical ID. Methods like BM25 [29] rely on token-level statistics to calculate the similarity between the query and document corpus. In retrieval systems where the semantic similarity measures the query-document alignment, tokenized inputs are embedded using models (BERT [16], Sentence Transformers), and query-document similarity is computed in vector space.

Finetuning helps a pre-trained model adapt to a downstream task using a small, task-specific dataset. But training all the parameters of a model is expensive and often not necessary. That's why methods like PEFT are used. PEFT stands for Parameter-Efficient Fine-Tuning. It is a collection of techniques for fine-tuning large language models (LLMs) that allows a pretrained model to adapt to a new task using significantly fewer trainable parameters. LoRA [30] is a PEFT method that doesn't update the parameters of the main model while training. It adds very few new parameters to specific layers of the model and trains them.

2.2 Related Work

2.2.1 Pedagogical Agents (PAs)

According to a recent systematic review by [31], the identification of design features that actually improve learning is still at stake, especially for K-12[1] students. Knowledge gaps remain on issues concerning the effective impact of

[1] Students from kindergarten to 12th grade

design characteristics such as human-likeness and immersive environments on learning outcomes, amidst advances in agent technology. According to [32], with effective communication, PAs foster intrapersonal and interpersonal learning, self-regulation, and motivation of the users. However, their human-like capabilities in communication remain scarce and must be developed to enable adaptive and context-sensitive interactions.

To deal with the design issues, [33] proposed a taxonomy on the categorization of conversational agents for structure, technology, and user interaction. This framework outlined the design features influencing their effectiveness, hence setting the foundation for further research and practical implementation in education. In support, [34] also identified that 2D agents are more effective in improving learning outcomes when compared to 3D agents; thus, careful consideration is called for in designing multimedia PAs.

Embodied pedagogical agents were shown by [35] to greatly improve learning along many dimensions, although differential effectiveness by design and type of content supports the need for research into the optimization of specific design elements of these agents. Furthermore, [36] pointed out the need for adaptive PAs able to adapt interactions according to the particular needs of each learner to offer a more supportive and efficient learning environment.

2.2.2 Retrieval Augmented Generation (RAG) in Pedagogical Setting

In [37], authors compared various retrieval, indexing, and generation strategies of RAG (Retrieval Augmented Genera- tion). An assistant to help students in administrative queries was presented in [38] where the data was fetched and processed by the assistant from the university website. In [39], authors introduced CyberBOT, a domain-specific RAG-based chat- bot for cybersecurity education. CyberBOT integrated course content with a domain ontology to validate and constrain LLM-generated answers, ensuring accuracy and safety. Reference [40] explored whether LLMs can empower educa- tors to independently develop conversational agents for their teaching needs. It proposed a template-driven framework enabling educators to create domain-specific chatbots without deep technical knowledge. The tutoring chatbot system designed in [41] offers accurate, personalized, and contextually relevant assistance overcoming limitations of GPT-4o and other generic models. In [42], a modular AI chatbot framework was introduced that combined RAG with a recommender system for dynamic, real-time, and context-aware support.

2.2.3 Applications of Large Language Models

In health, such Large Models as Med-PaLM 2 [43] and BioGPT [44] currently have an expert level of answering medi- cal questions and can be fine-tuned in more specialized fields in order to support education and clinical practice [45]. Such applications are conversational models, which can make communication faster and more efficient in clinical practices. In [46], authors pointed out a gap in the evaluation of foundation models originally trained on electronic medical records. They developed a superior evaluation framework, which captured all the metrics relevant to health care significantly better and answered some fundamental questions on generalizability to medical practice. [45] discussed the strengths and weak- nesses of LLMs in healthcare and provided a summary of their role to the clinicians to appreciate how such models can further enhance clinical practice, education, and research. GatorTron [47] is a large clinical language model trained on the EMR narratives. It has advanced the state-of-the-art in EHR processing and evaluation of the performance of the system on clinical NLP tasks.

In[48], authors discussed the resultant paradigm shift by LLM-driven chatbots. They predicted that there is a likelihood of efficiencies in academic work but warned against bias and inaccuracies. They recommend discretion with the quantification of their limitations. Still, LLMs provide means to increase engagement, enable new technologies, and take on educational inequalities [49].

Systematic reviews highlight the competitiveness of newer LLMs, such as PolyCoder, across languages [50].Their deployment should be done responsibly, balancing the transformative potential of the models while putting ethical considerations and risk mitigation into perspective [51]. Several examples in telecom provide a vision into how LLMs will affect the industry by implementing anomaly detection and understanding of technical specifications [52]. In another breath, [53] has looked at NLP to recommend skills needed for assessment in industry sectors, focusing on the United Arab Emirates.

In [54], a critical overview of history, architecture, and applications of Large Language Models including their ap- plication in medicine, education, finance, and engineering is performed. Moreover, model bias, interpretability, ethical issues, and techniques related to enhancing robustness in the LLMs are also discussed. The authors in [55] discussed how much the LLMs will be able to revolutionize various fields of studies while keeping ethical considerations in mind and give useful insights into further research for enhancement in models' reliability.

Authors in [56] discussed a comprehensive survey of pretrained foundation models ranging from BERT to ChatGPT. They further discussed different architectures, various training methodologies, and LLM applications on different data modalities such as text, images, and graphs. Challenges on model efficiency, security, and privacy are discussed whereby future research directions are identified.

2.2.4 Applications of LLMs in Electrical and Electronic Engineering (EEE)

In [57], authors proposed a hybrid natural language processing (hybrid-NLP) algorithm for entity extraction from electrical equipment malfunction texts, which integrates a dictionary-based method, a language technology platform

(LTP), and a Bidirectional Encoder Representations from Transformers-Conditional Random Field (BERT-CRF) model. ChatGPT's ability to provide insights, personalized support, and interactive learning experiences in engineering education was assessed in [58] through different stakeholders like lecturers, students, and engineers. In [59], offered insights into the cur- rent usage patterns, perceived benefits, threats, and challenges, as well as recommendations for enhancing the adoption of LLMs among students and instructors. In [60], Multimodal Large Language Models' (MLLMs) ability to answer digital circuit questions was assessed using a benchmark dataset.

3. Methodology

EduAgent is designed to provide pedagogically enhanced responses to user queries, followed by appropriate images and explanations. In section 3.1, we have described the process we followed to create the conversational dataset that will be used to fine-tune and prompt open-source models to enhance their responses. Section 3.2 discusses image collection, formalization, and vectorization processes. We have described the prompting and fine-tuning processes in sections 3.3 and 3.4.

3.1 Conversational Dataset Creation

We planned to create a dataset consisting of multi-turn conversations on fundamental topics of the electronic courses. At the start, we collected 596 questions from 10 different topics mentioned in Appendix A. However, it is not feasible to manually simulate 596 conversations and ensure high-quality questions and responses.

So, we designed an answering agent that prompts GPT-4o to generate concise, easily understandable, and engaging responses. We ensured that the responses start by mentioning what prior knowledge is needed and divide explanations into short and simple steps. In Figure 2, we have shown how we enhanced the answering prompt iteratively.

We also needed a questioning agent that would ask the answering agent to simulate a multi-turn conversation between a teacher and a student. The query must not be repetitive, naive, and out of context. We also penalized the questioning agent for giving feedback instead of a question. Figure 3 shows how we enhanced the questioning prompt iteratively.

In the end, Every data point in the synthetic dataset is a 4-turn conversation between two agents. We checked the model outputs for harmful intent[2] and shallowness[3] bias, ensuring in-depth, precise, and meaningful responses. Educators evaluated randomly sampled conversations across different metrics to validate pedagogical value. Section 5.3 highlights the analysis of the evaluation.

Among the topics of the dataset, we kept questions from 8 topics for training and separated 2 topics (Miscellaneous Topics, and Logic Gates) for testing purposes. In the end, we were left with 547 training data points and 49 testing data points.

3.2 Image Dataset Creation

Images can greatly enhance the understandability of a response to a complex query. Keeping that in mind, we created an image vector database to help EduAgent respond with images that improve responses and introduce multi-modality. We can divide the database creation process into the following steps:

- **Image Collection:** At first, we collected 118 images from GeekForGeeks, Hackatronic, Researchgate, and other domains. We have listed the source for every individual image in our GitHub repository.
- **Image Explanation Generation:** We used InternLM [61], an open-source model to extract contextual information from the image and generate an explanation. This model supports multimodal input and uses its vision processing pipeline to extract the image contents and respond to the query based on that.
- **Topic Description Generation:** We also collected corresponding image topics along with the images. We asked Llama-3.1-8B [62] to generate a detailed text on each of these topics.

Topic description generation prompt:

```
"describe the topic {image_title} in 300
words. The topic should be described in the
context of electrical and electronic engineering.
An explanation of an image on that topic is also given
below for further context:\n {image_explanation}\n\n
```

[2]Responses that could encourage unsafe, unethical, or malicious behavior.

[3]LLMs' tendency to avoid complexity, providing oversimplified and generic answers.

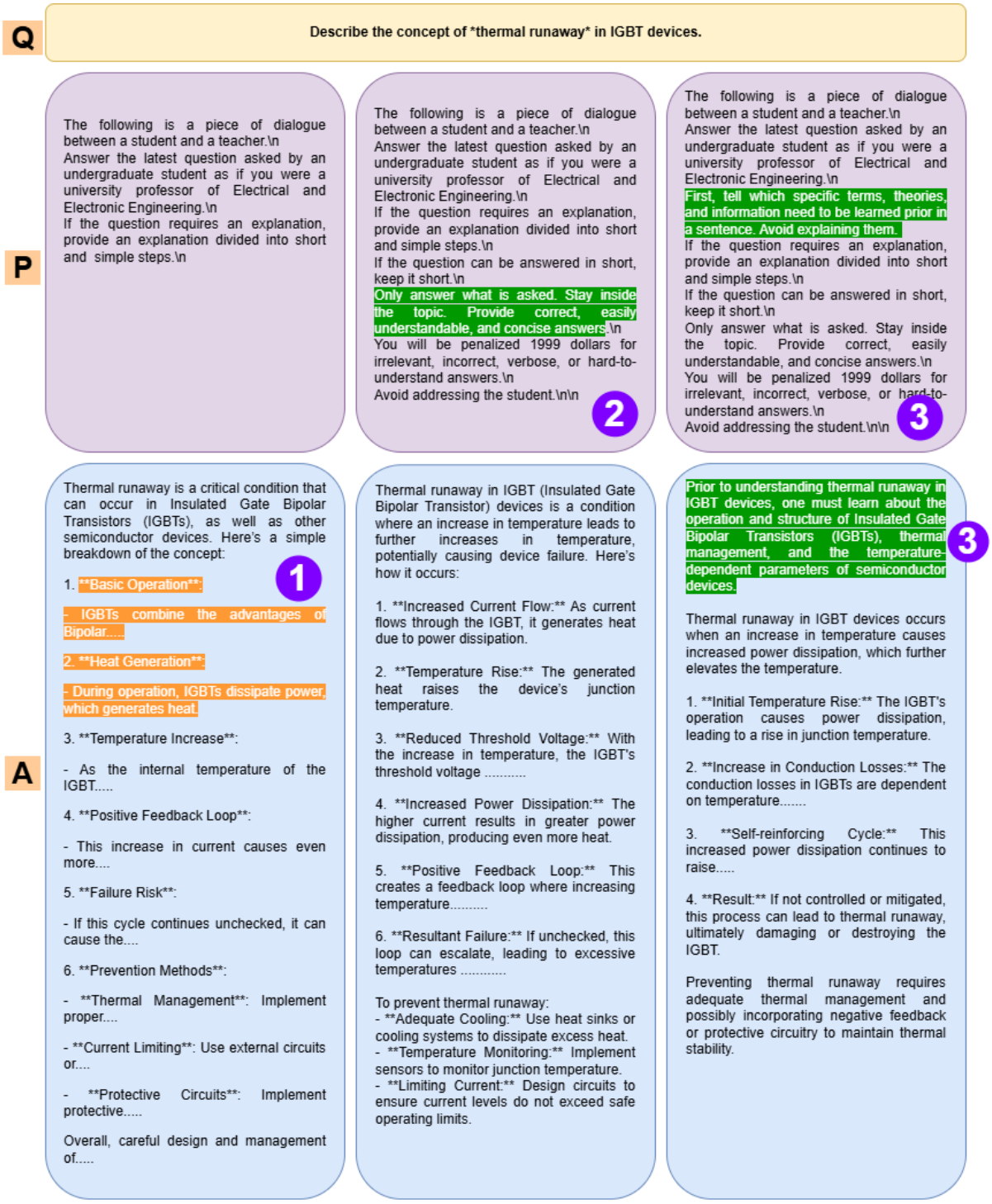


Fig. 2. Iterative prompt building process for the answering agent enhancing the response of GPT-4o (1) While describing thermal runaway, we don't need to explain what IGBT is. We can tell the learner that it is prior knowledge to have. (2) The response became shorter and unnecessary explanations were omitted from the previous iteration. (3) Mention what prior knowledge is needed. The response is shorter also.

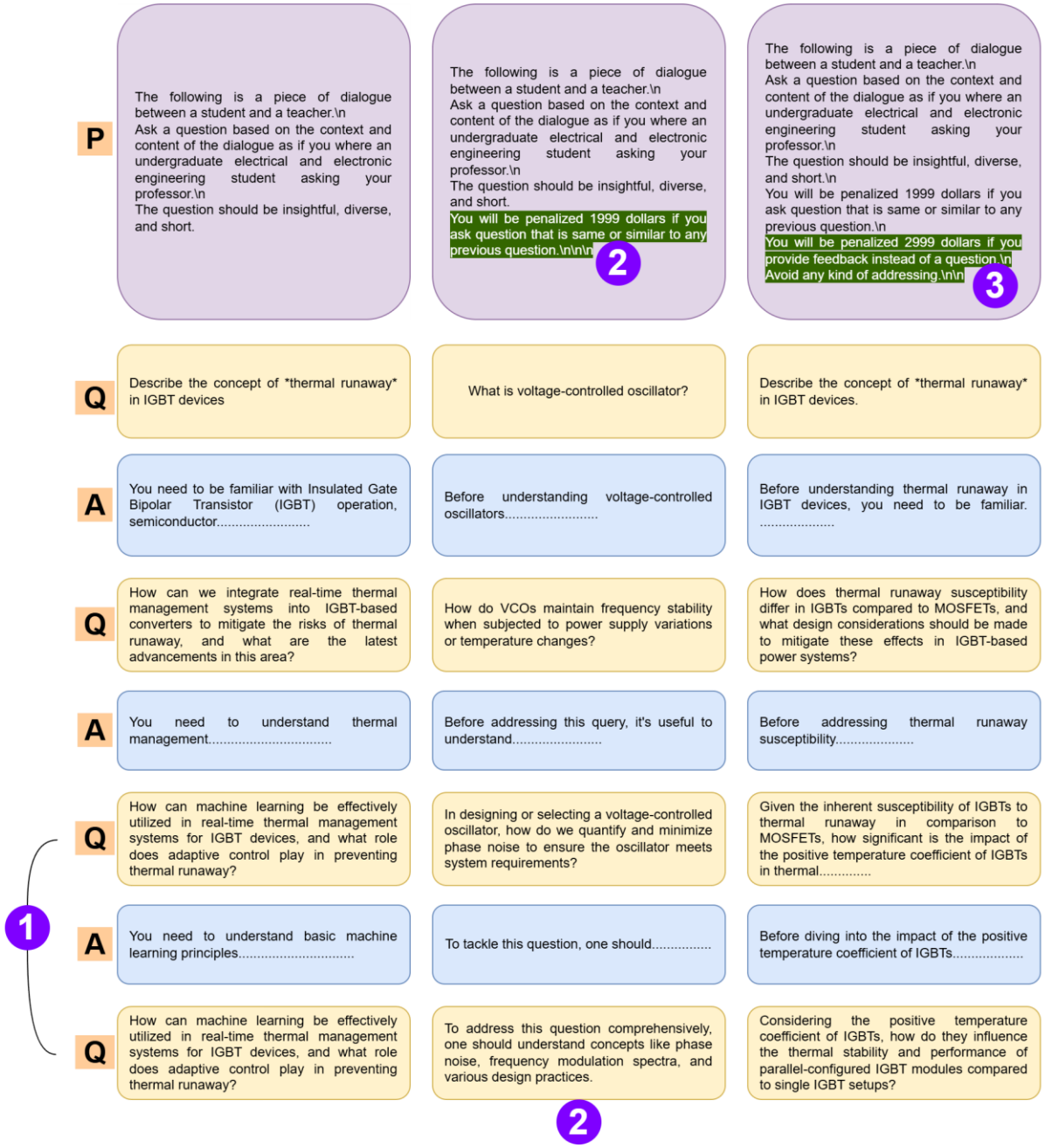


Fig. 3. Iterative prompt building process for the questioning agent enhancing the response of GPT-4o (1) Repetitive questions for turn 3 and 4 (2) In turn 4 of 2nd conversation, the questioning agent is replying with feedback instead of a question (3) Final prompt mitigating the issues

```
                The description should cover the usage of the algorithm,
        system or device described by the provided context and the topic
        name. Avoid including the image explanaton in the description. Avoid
        including anything outside the given context."}
```

- **Creation of Knowledge Base:** We tokenized the topic explanations as they can act as large texts to compare similarity with the query-answer pairs. As metadata, we had corresponding image URLs and explanations. Every data point in our knowledge base was in this format:

```
Document{
    page_content:{topic_description} metadata:{
        "topic", "image_url", "image_explanation"
    }
}
```

3.3 Few Shot Prompting Based Pedagogical Agent Design

We prompted open source models (Llama-3.1-8B [62] and Llama-3.2-3B [63]) to generate pedagogically enhanced responses. We have 2188 question-answer pairs from the training dataset (section 3.1). We tokenized these question- answer pairs using *o200k-base* tokenizer.

While retrieving example question-answer pairs (QA pairs) for a query, we tokenized[4] the query with the same tokenizer. Then we ranked the QA pairs based on how similar they are to the query using rank bm25 library in Python that ranked QA pairs based on term importance. Term importance is proportional to term frequency in one QA pair and inversely proportional to frequency in different QA pairs [29]. Lastly, we retrieve the top 5 similar QA pairs. For Zero Shot prompting, we didn't provide any example as the prompt. For One Shot prompting, one example was provided and increased incrementally for Three Shot and Five Shot prompting.

3.4 Fine Tuning Open Source Models with the Conversation Dataset

We fine-tuned two different models: Llama-3.1-8B [62] and Llama-3.2-3B [63] offered by Unsloth [64]. While fine- tuning, they would store and load their weights in a 4-bit quantized format. The internal calculations and representations will be carried out in 16-bit half-precision float values. In this way, we are ensuring that the models not only use less memory but also maintain respectable precision while computing. For resource constraints, we are using LoRA [30], a PEFT method that freezes the original model parameters, adds row rank matrices with some target layers, and updates the weights of the matrices as the training progresses. In this way, we are ensuring that the model generates task-specific responses without huge training overhead. As a result, the 3B model had 24 million trainable parameters (~1%) and the 8B model had 42 million trainable parameters (~0.5%).

Among 547 example conversations from the training dataset, 519 examples were used for training and 28 others were used for validating. After every 10 steps, a checkpoint was added to verify if the validation loss was decreasing. The training loss was 1.39 for the Llama-3.1-8B model at the first checkpoint which was reduced to 0.94. Similarly, the training loss was reduced to 1.00 from 1.55 for the Llama-3.2-3B model.

4. Automated Evaluation

In this section, we introduced different evaluation metrics provided by different libraries and tools. We conducted evaluations on text response generators comparing the generations with the ground truth, the test dataset from section 3.1. A retrieval test dataset was also designed to evaluate the image retrieval methods.

4.1 Evaluation of the Text Generated by the Open Source Models

In this section, we aim to evaluate if different prompting techniques and fine tuning have improved the quality of text responses generated by the open source models [62, 63].

4.1.1 Experimental Setup

In section 3.1, we have mentioned that there are 49 4-turn test conversations. For any turn of any conversation, given the previous turns as context, we can ask the improved open source models (through prompting or finetuning) current query and compare the generated responses to those generated by the answering agent mentioned in Figure 2.

However, we needed to make sure that the query for each turn remained the same across all models and techniques. So instead of simulating a conversation by feeding the previous responses as context for the current turn, we fetched previous turns from the test dataset. For one shot and few shots prompting, examples similar to that turn are fetched from the 2188 QA pairs of 547 training conversations using BM25 technique.

As human evaluation was not possible for such a huge evaluation set, BERT and BLEU scores were used to evaluate how much the responses from open-source models semantically aligned with the responses from the test dataset. A subset of conversation samples had been previously evaluated by expert educators, and the corresponding test responses were deemed pedagogically sound. Therefore, higher BERT and BLEU scores were interpreted as stronger alignment with these pedagogically aligned reference responses. Note that BERT and BLEU scores are positively correlated with human judgment. [65, 66].

We calculated the BERT-score using the Python bert-score library. As the model, we used bert-base-uncased.BERT generates precision and recall scores. As high precision results in low recall or vice versa, the F1 score works as a balance (harmonic mean of precision and recall score) between them. We will compare the text generation based on the f1 score provided by the library function. We used the sacrebleu [67] tool to calculate the BLEU score. A high BLEU score means the model has memorized the reference answers. A low BLEU score means the training loss is huge.

[4]Tokenization is the process of breaking input text into smaller units called tokens. These tokens are then mapped to numerical IDs. These IDs are the building blocks that language models use to understand and process text.

4.1.2 Evaluation Results

Table 1. BERT Score and BLEU score comparison between different prompting techniques for llama3.1-8B and llama3.2-3B base and fine-tuned model. Scores are organized for generations in different turns (1st to 4th turn) of the conversation. The Best scores for each model and each turn are highlighted in bold.

Generator		1st turn		2nd turn		3rd turn		4th turn	
Model	**Technique**	**BERT**	**BLEU**	**BERT**	**BLEU**	**BERT**	**BLEU**	**BERT**	**BLEU**
Llama-3.1-8B	Zero Shot	0.6482	19.93	0.6850	23.08	0.7202	27.21	0.7271	23.48
	One Shot	0.6571	**23.75**	0.6936	26.04	0.7249	**32.61**	**0.7453**	29.37
	Three Shots	**0.6684**	20.42	0.6994	**29.89**	**0.7372**	26.27	0.7426	29.23
	Five Shots	0.6649	20.91	**0.7067**	27.35	0.7328	27.38	0.7415	29.92
	Finetuning	0.6489	15.96	0.6865	21.24	0.7171	26.98	0.7293	**31.65**
Llama-3.2-3B	Zero Shot	0.6374	16.66	0.6866	23.16	0.6994	22.74	0.7176	25.99
	One Shot	0.6495	14.97	0.6902	23.21	0.7131	25.50	0.7245	**27.75**
	Three Shots	**0.6742**	**23.47**	**0.7024**	**28.13**	**0.7183**	**28.81**	**0.7357**	26.57
	Five Shots	0.6696	20.93	0.7001	28.03	0.7094	24.30	0.7290	26.52
	Finetuning	0.6240	18.54	0.6884	21.67	0.7056	23.93	0.7144	19.77

Table 1 shows the performance of two models, Llama-3.1-8B, and Llama-3.2-3B, in generating responses across four conversational turns using different prompting techniques: Zero Shot, One Shot, Three Shots, and Five Shots. These models were also finetuned and zero-shot prompted. Each method's effectiveness is measured using BERT and BLEU scores as mentioned in the previous section.

For both models, the base models with few-shot prompting outperformed their fine-tuned counterparts. In most cases, base models with zero shot prompting gave similar results to finetuned models. This could imply that the models benefit more from in-context learning (using examples in the prompt) than from task-specific fine-tuning in conversational scenarios.

For Llama-3.1-8B, both one-shot and three-shot prompting achieved higher BERT and BLEU scores across all techniques. However, for Llama-3.2-3B, results for three shots prompting appeared to be more prominent than one shot prompting. This indicates that Llama-3.2-3B requires more examples to learn the response format. Also, three-shots prompting scores were consistently higher than five-shots prompting scores for both models. Longer context lengths for five-shots prompting might be an issue here.

As the conversations progressed or turn increased, both BERT and BLEU scores across all techniques gradually increased. As the previous turns were from the test dataset instead of the models themselves, the models learned the response format from the previous turns even if no example was provided in case of zero shot prompting. In such cases, previous turns contributed to in-context learning.

4.2 Evaluation of Retrieved Image Using Synthetic Test Dataset

In section 3.2, we have described the process of creating the image database. In this section, we are going to evaluate different retrieval methods on the database. So, we need a test dataset: some question answer pairs and corresponding appropriate images as ground truths. The procedure to make the test dataset is described in Appendix C.

4.2.1 Experimental Setup

Given a query, we designed the retrieval process in 3 ways: BM25 Retriever removing the stopwords, BM25 Retriever without removing the stopwords, and TF-IDF Retriever.

TF-IDF [68], a core technique in information retrieval, calculates the importance of terms within documents based on their frequency in a document relative to their frequency across all documents. It is based on the idea that terms that appear frequently in a single document but infrequently in the entire document collection carry higher relevance. As we are working with domain-specific documents and queries, TF-IDF is a great option.

BM25 [29] is an improvement over the traditional vector space model, often favored for its ability to balance term frequency (how often a term appears in a document) with document frequency (how rare the term is across documents). BM25 includes adjustable parameters, *k*, and *b*, which control term saturation and document length normalization, respectively. In our case, we implemented the retriever using the python rank_bm25 library where we did the retrieval in two ways: with and without removing the stopwords such as is, are, the, etc. Words such as and, or, not, etc. were not considered stopwords because these words provide important contexts related to logic gate queries.

For each way, the document with the maximum score will be retrieved. Then, we compared each of the retrieval techniques on whether it could retrieve the same image as the ground truth. If the retriever was successful, it was considered a win and scored as 1. Otherwise, the retrieval was considered a loss. So, this evaluation represents a classification task with definitive correct/incorrect options. Win-rate evaluation appropriately measures the model's ability to select the most pedagogically relevant image from available options, similar to multiple-choice assessment.

4.2.2 Evaluation Result:

Table 2. Our test dataset contains the appropriate image response for every test data point. A correct image retrieval results in a score 1 and 0 otherwise. then the average of the total score is calculated as the success rate for every retrieval process.

Retrieval Process	Win Rate
BM25 Retriever without removing the stopwords	0.75
BM25 Retriever removing the stopwords	0.76
Langchain TF-IDF Retriever	0.82

The results from Table 2 indicate that removing stopwords in BM25 improved retrieval performance as removal of the stopwords can result in the domination of important terms on the score and reducing the noise.

In appendix D, retriever responses are shown for an example query. For that query, both BM25 retrievers failed, but the TF-IDF retriever was successful. While matching the query to the documents, TF-IDF retrievers give importance to rare terms in the document that are present in the query. BM25 retrievers can also downperform when the document lengths vary so much. On the other hand, if the query doesn't have an exact match to the document, BM25 can outperform sometimes.

Moreover, our image dataset is small. A larger pool of documents to retrieve can result in more successful retrieval.

5. Human Evaluation

In Section 5.1 and 5.2, we have described the evaluation process and the metrics. Dataset evaluation scores have been discussed in Section 5.3. Section 5.4 and 5.5 have described the text generation and image retrieval evaluation on pre-simulated conversations whereas Section 5.6 and 5.7 have discussed the live conversations in an interactive setting.

5.1 Design of Evaluation Process

The conversation dataset with 596 synthetic conversations created in section 3.1 was evaluated by 6 educators. We sampled 45 conversations randomly. Each of the conversations was evaluated independently by two experts.

Then we evaluated the text generated by the prompted or fine-tuned open-source models and images retrieved by the retrievers. It was done in two steps. At first, 20 evaluators evaluated the generated text, and 15 evaluators evaluated the retrieved images using evaluation scripts that contained pre-simulated conversations and retrieved image examples.

Then, 8 evaluators evaluated in an interactive environment using a live tool where learners can ask a query, and evaluate on the run. It was ensured that the learners had already completed courses related to the dataset topics.

To design the evaluation scripts, we chose three generators: GPT-4o, Llama-3.1-8B with Three Shots prompting, and Llama-3.1-8B fine-tuned model. Then we simulated multi-turn conversations between two agents where each generator acted as the answering bot and GPT-4o acted as the questioning bot. Thus, we had three conversations for three generators for one initial question. These three conversations created a conversation set. A snapshot of an example conversation set is shown in Appendix E. Then we created an image set where a QA pair with and without the retrieved image were shown side by side. We used BM25 Retriever removing the stopwords to retrieve the images and their corresponding explanations. A snapshot of an example image set is shown in Appendix F. Finally, we created an evaluation script for each of the evaluators, where the evaluators will rate multiple conversations and image sets.

To evaluate in an interactive environment, we chose the baseline method (GPT-4o) and our best-performing method (Llama-3.1-8B with Three Shots prompting and BM25 Retriever removing the stopwords for image retrieval). Responses from each method were displayed side by side so that learners could compare and rate them.

5.2 Explanation and Usage of the Evaluation Metrics

A response can be evaluated on the basis of the following metrics:

- **Correctness**: A response is correct when it aligns with the accepted facts. Evaluators will check if it conveys information in a way that leaves no room for ambiguity or misinterpretation in the given context.
- **Completeness**: A response is complete when it fully addresses the query in a thorough and meaningful way. Evaluators will ensure that all aspects are answered without leaving any gaps or without losing any context. The prior knowledge needed to understand this topic should also be mentioned.
- **Conciseness**: A response is concise when it dives into the main idea in a straightforward manner, avoiding unnec- essary complexity or ambiguity. Evaluators will check if it can be understood easily, the main idea is delivered in simple terms, and the explanation is done in small and simple steps. The coherence between subsequent steps should also be ensured.

- **Engagement**: A response that is natural, engaging, and easy to connect with for the learners. Overly formal or robotic phrasing should result in a lower score. The response should be approachable and enjoyable to read. It should convey the necessary information effectively. All important headings must be in bold text.

The synthetic dataset was evaluated on the basis of correctness, completeness, and conciseness by the educators. Learners evaluated the evaluation scripts and live responses in the interactive environment on the basis of all four metrics.

We evaluated image retrieval on the following criteria:

- **Relevance**: The retrieved image is consistent with the text response. Evaluators assess how well the image and corresponding explanation align with the query and the context.
- **Understandability**: The retrieved image increases the understandability of the response. Evaluators will check if the image and its explanation align with each other. Understandability is enhanced when the image provides clarifying information, breaks down complex concepts and terms, and illustrates abstract ideas of the text response more intuitively.

5.3 Evaluation of the Conversation Dataset

Figure 4 shows the average scores of the conversation dataset evaluation, where each sample conversation was evaluated by 2 educators independently. Correctness received the highest average score, with a median close to 4.75. The tightly clustered distribution indicated consistency for this case. Completeness scores showed a narrower spread, with a lower median around 4.0. The conciseness plot was similar to the completeness plot but with slightly more outliers. The variability of the conciseness scores was underestimated because of the outliers.

As each datapoint was evaluated by 2 evaluators on a scale from 1 to 5, we calculated Krippendorff's alpha score with ordinal level of measurement for three metrics: Correctness, Conciseness, and Completeness. For Correctness, the score was 0.671, indicating moderate agreement between two evaluators. The score was 0.536 for Completeness and 0.569 for Conciseness. As the metrics are highly subjective, different evaluators may have different expectations while judging how complete and concise the responses are.

5.4 Evaluation of the Text Generated by the Open Source Models

In Figure 5, the box plots illustrate the comparison of the responses generated by three models: GPT-4o, Llama-3.1-8B fine-tuned, and Llama-3.1-8B few-shot. The evaluation was performed on four different metrics: correctness, completeness, conciseness, and engagement. Overall, Evaluators rated responses from the Llama-3.1-8B few-shot model highly across all metrics. Higher medians and narrower interquartile ranges indicate responses appeared to be more reliable and higher-quality compared to the other models. Evaluators highly rated its responses because they were well- described, complete, and informative.

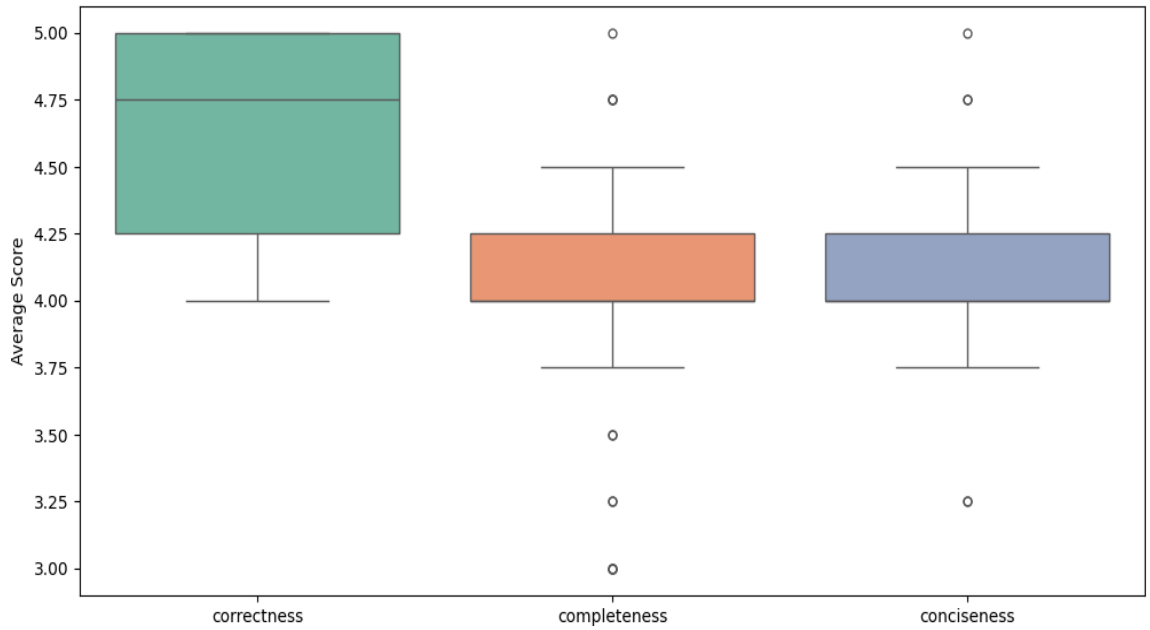
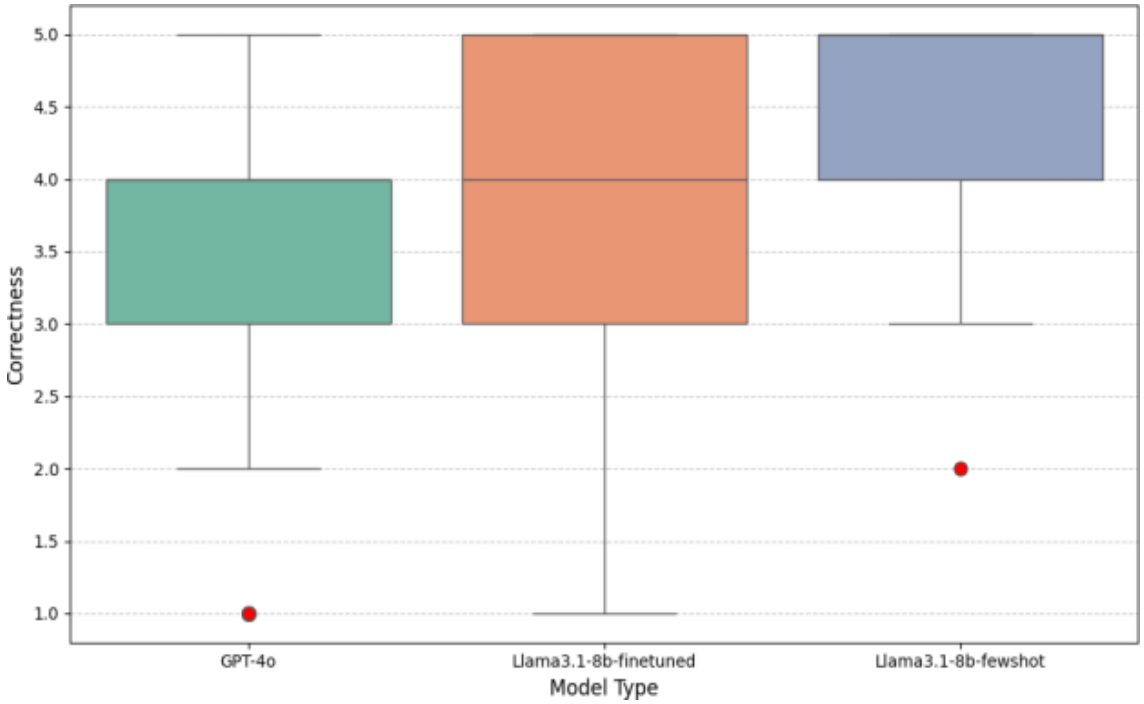
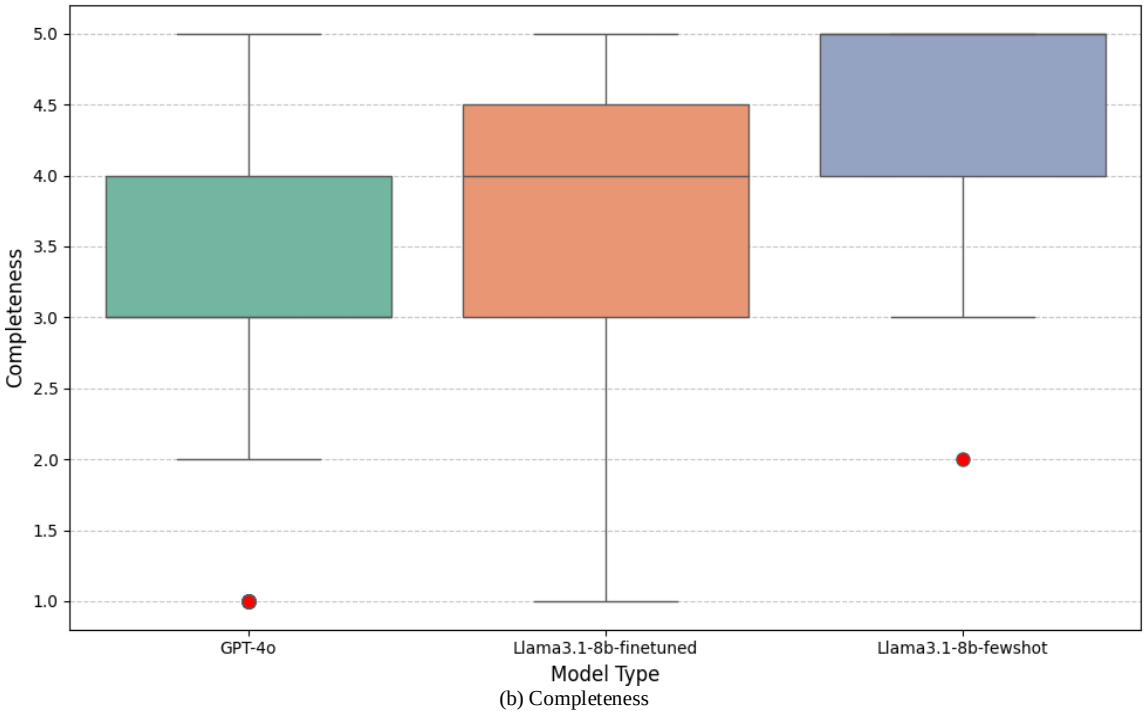


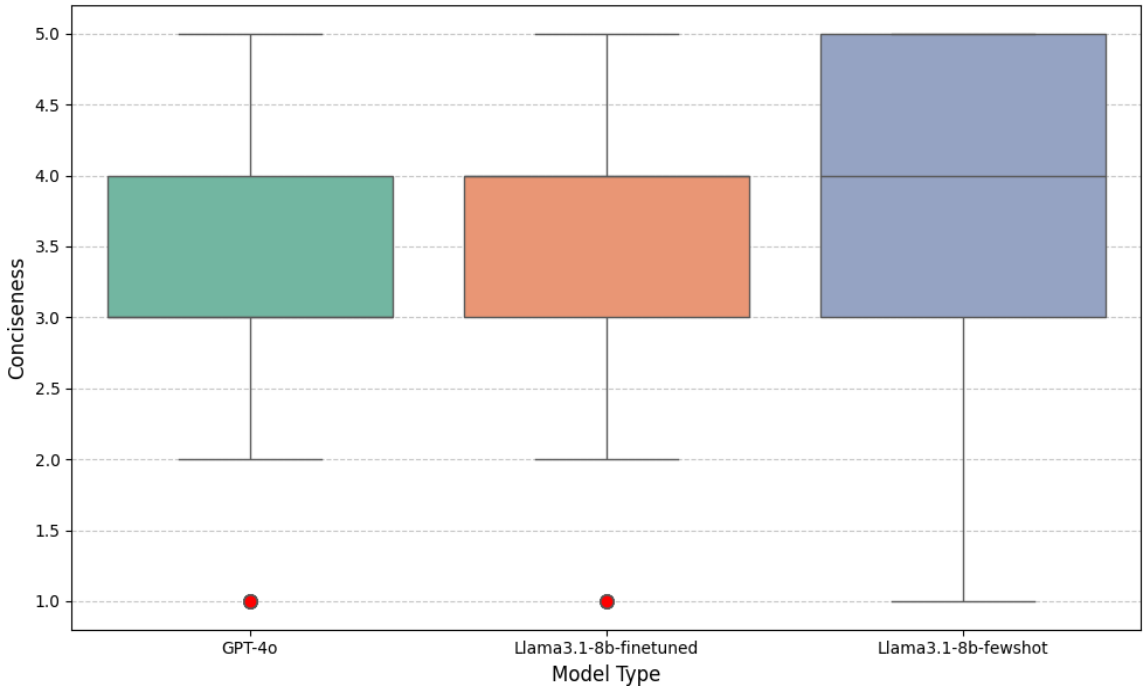
Fig. 4. 6 Evaluators rated 45 sample conversations from the conversation dataset on three matrices: correctness, completeness, and conciseness.



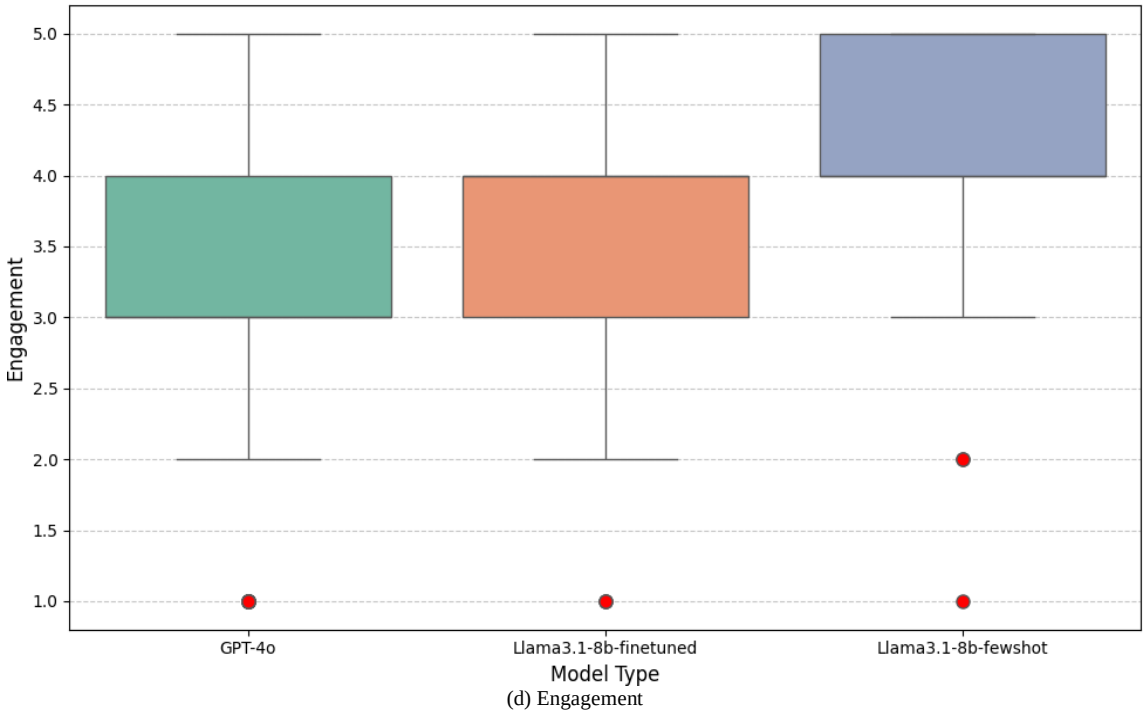
(a) Correctness



(b) Completeness



(c) Conciseness



(d) Engagement

Fig. 5. 20 Evaluators rated text generation of 3 models: GPT-4o, Llama-3.1-8B fine-tuned, and Llama3.1-8B-few shot on 4 different metrics: correctness, completeness, conciseness, and engagement.

For correctness, the Llama-3.1-8B few-shot model achieves the highest median score, suggesting it produced the most accurate responses. In contrast, GPT-4o shows lower scores with one outlier that indicates consistent poor performance. The Llama-3.1-8B fine-tuned model having a wide interquartile range performed better than GPT-4o but not as well as the few-shot version.

For completeness, the Llama-3.1-8B fine-tuned model shows significant variability demonstrating the widest interquartile range. The Llama-3.1-8B few-shot model achieved the highest score with a narrow interquartile range, while GPT-4o fell behind with a lower median.

Conciseness and engagement follow almost similar trends. However, the Llama-3.1-8B few-shot model achieved scores that have higher variability in the case of conciseness. While GPT-4o and Llama-3.1-8B fine-tuned models achieved the same scores, they fell short of the few-shot model's response quality based on these two metrics.

5.5 Evaluation of pedagogical understandability improvement with image

Figure 6 illustrates the evaluation scores of a BM25 retriever removing the stopwords, evaluated on two metrics: relevance and understandability of retrieved images. For relevance, the scores are distributed over a wide range (3.0 - 5.0), with a median close to 4.

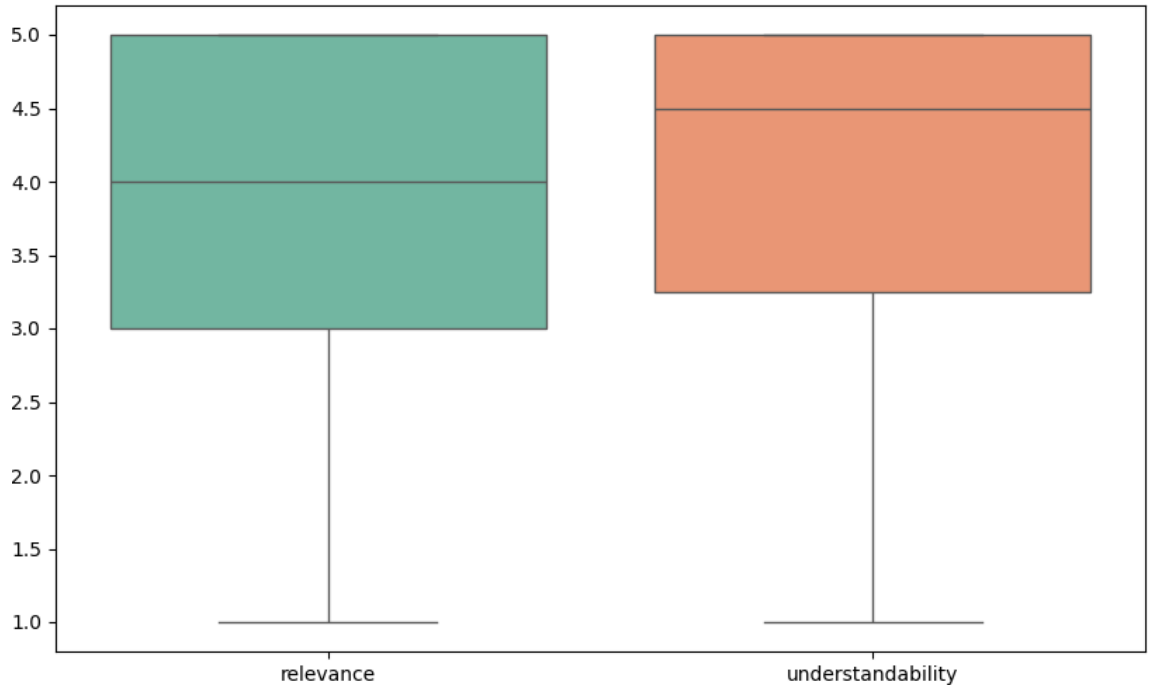


Fig. 6. 15 Evaluators rated the images retrieved by BM25 retriever removing stopwords on 2 different metrics: relevance, and understandability.

In the case of understandability, the median (4.5) is higher and the IQR is narrower (3.25 - 5.0) than those for relevance, reflecting that most images were moderately understandable.

5.6 Evaluation of Text Generation Through Live Interaction

Figure 7 illustrates the comparison of GPT-4o and Llama-3.1-8B three-shot prompting text responses in a live and interactive setting. For correctness, conciseness, and completeness, Llama-3.1-8B three-shot model consistently exhibited exceptionally high and remarkably consistent performance, with its Inter Quartile Range (IQR) appearing to be from 4.0 to 5.0. GPT-4o, on the other hand, consistently scored around the range of 3.0-4.0 for completeness and conciseness. But the range was narrower for correctness, indicating better agreement between raters. While GPT-4o provides good results, it consistently falls short of the scores achieved by the Llama-3.1-8B three shot model. The average for GPT-4o in these categories was approximately 4.0, and similar to Llama, its consistency was high, albeit at a lower performance level.

For engagement, however, the Llama-3.1-8B three-shot model showed a median score of approximately 4.25. For GPT-4o, the engagement scores show greater variability. The median is approximately 4.0, but the box extends further, suggesting a wider IQR. This indicates that while the typical engagement score for GPT-4o was around 4.0, there was more fluctuation in its performance, with some instances potentially scoring lower (as indicated by the whisker extending down towards 2.0).

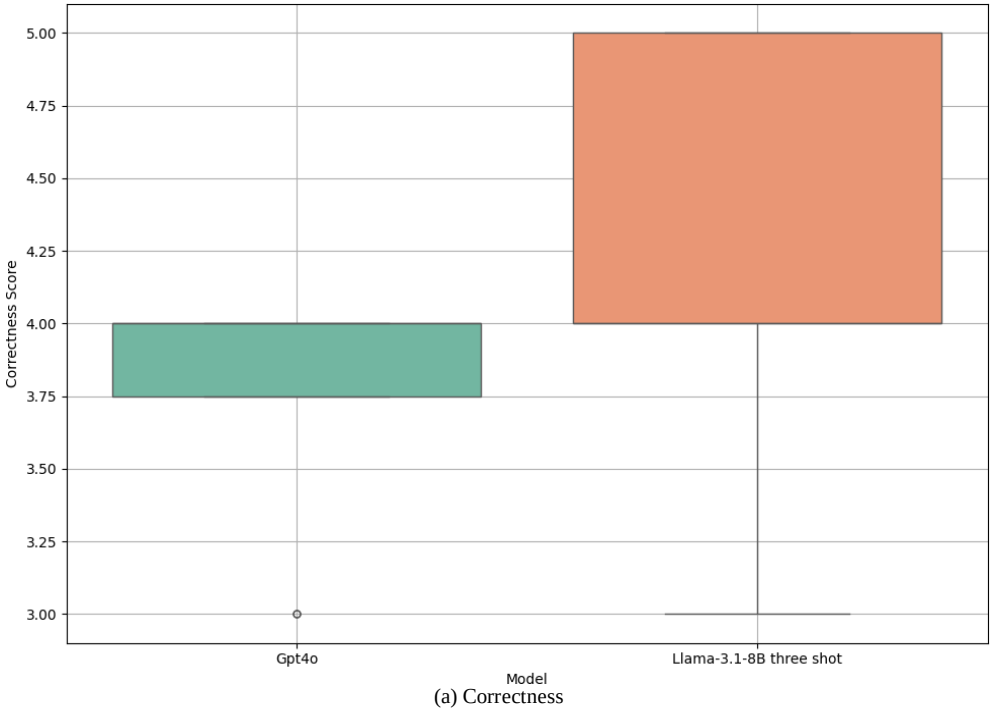
The presence of outliers underscored the variability for GPT-4o in correctness, and Llama-3.1-8B three-shot model for engagement. The Llama-3.1-8B three-shot model performed better on every metric, having scores between 4 and 5. Evaluators found responses from the Llama-3.1-8B three-shot model more learnable, easy to understand, and pedagogically sound.

5.7 Evaluation of Image Retrieval Through Live Interaction

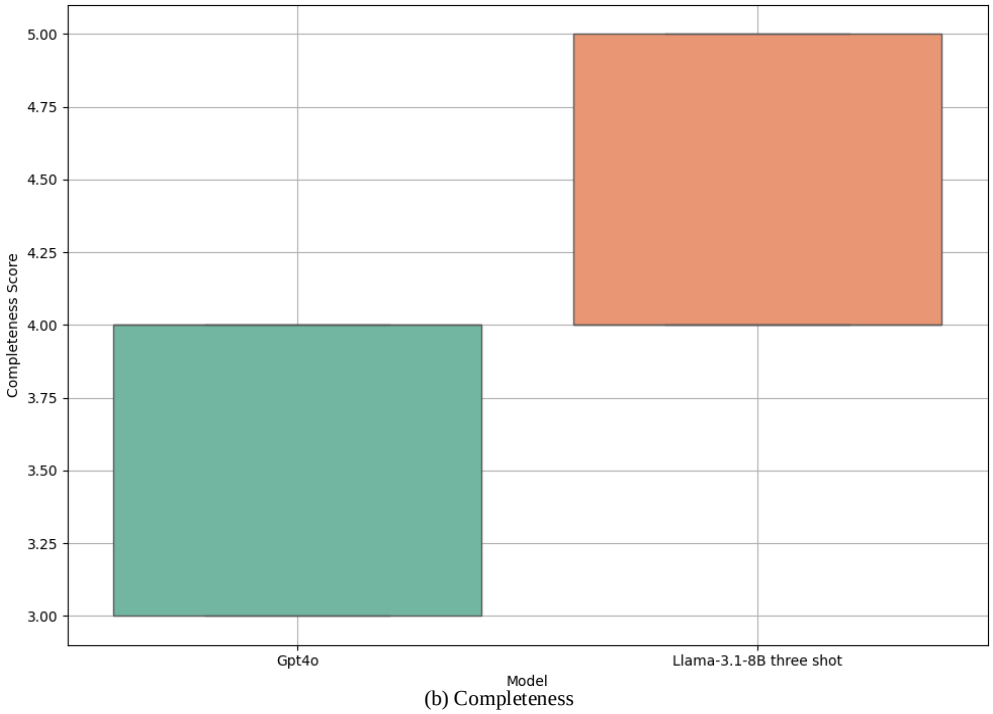
Figure 8 illustrates the evaluation scores of a BM25 retriever removing the stopwords in an interactive setting, evaluated on two metrics: relevance and understandability of retrieved images. For relevance, the scores are distributed over a range (4.0 - 5.0), with a median close to 4.75. In the case of understandability, the median (4.5) is lower than that for relevance, although the IQR range is the same.

6. Conclusion and Future Works

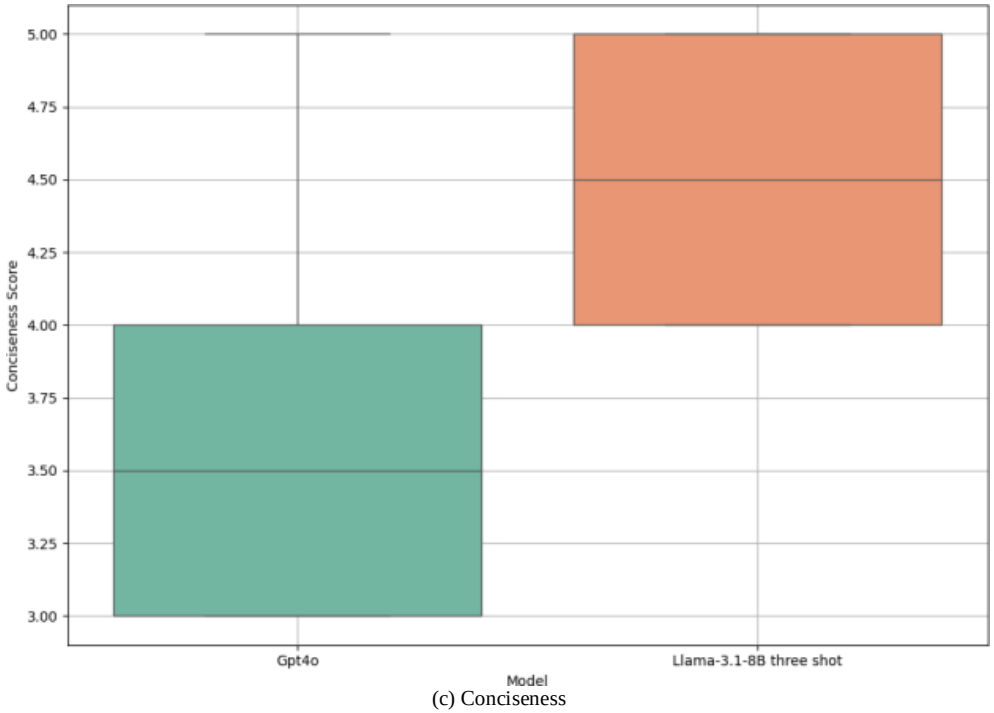
EduAgent linked with an image database generates pedagogically enhanced responses and retrieves appropriate im- ages from our image database. To enhance the responses, a dataset of synthetic conversations was created on 10 EEE topics. Those synthetic conversations were simulated between two LLM bots. Our selection of 10 foundational EEE topics was methodologically driven. Rather than attempting broad but shallow coverage, we deliberately focused on core areas to enable systematic evaluation of our pedagogical framework. We acknowledge that these topics do not comprehensively cover advanced areas. However, we note that our dataset creation framework is automated and topic-agnostic. As such, leveraging our pedagogical framework to expand the dataset into more advanced EEE areas is a trivial task, which we leave as future work.



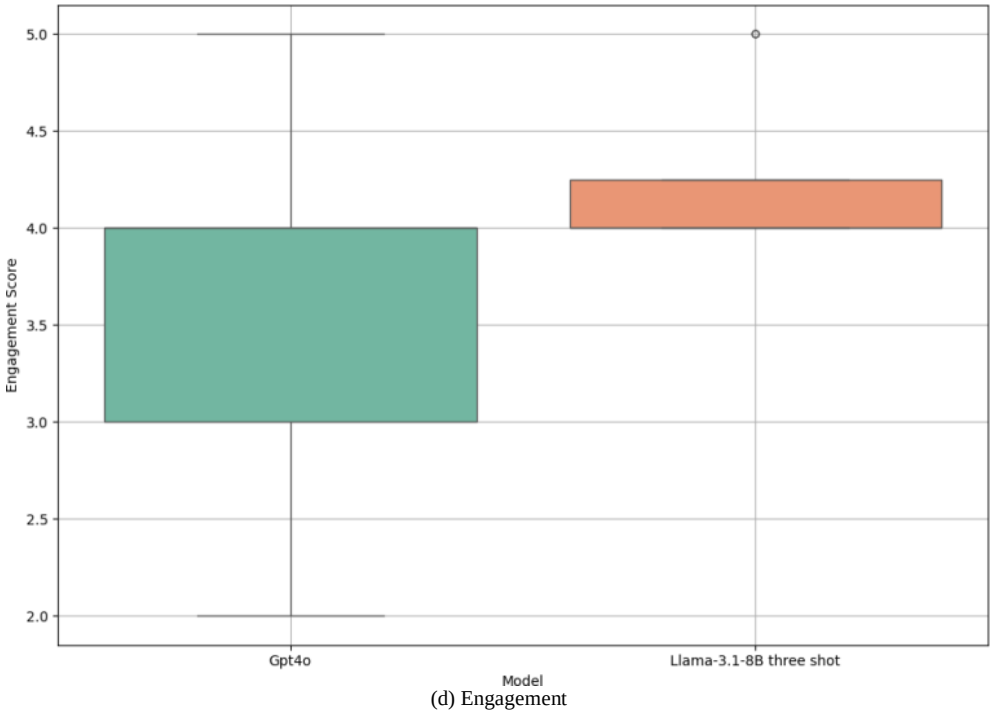
(a) Correctness



(b) Completeness



(c) Conciseness



(d) Engagement

Fig. 7. 8 Evaluators rated text generation of 2 models: GPT-4o and Llama3.1-8B-few shot on 4 different metrics: correctness, completeness, conciseness, and engagement.

LLMs inherently have bias. Incremental prompt engineering was done to mitigate harmful intent and shallowness bias from the responses. Some sample conversations were also evaluated to validate effectiveness in learning.

Then we compared different text generation and retrieval methods, and it turned out that fewshot prompting with appropriate examples can help the model learn the output format faster. All the fine-tuned models, datasets, notebooks, and evaluation data are in EduAgent github (https://github.com/hyadess/EduAgent).

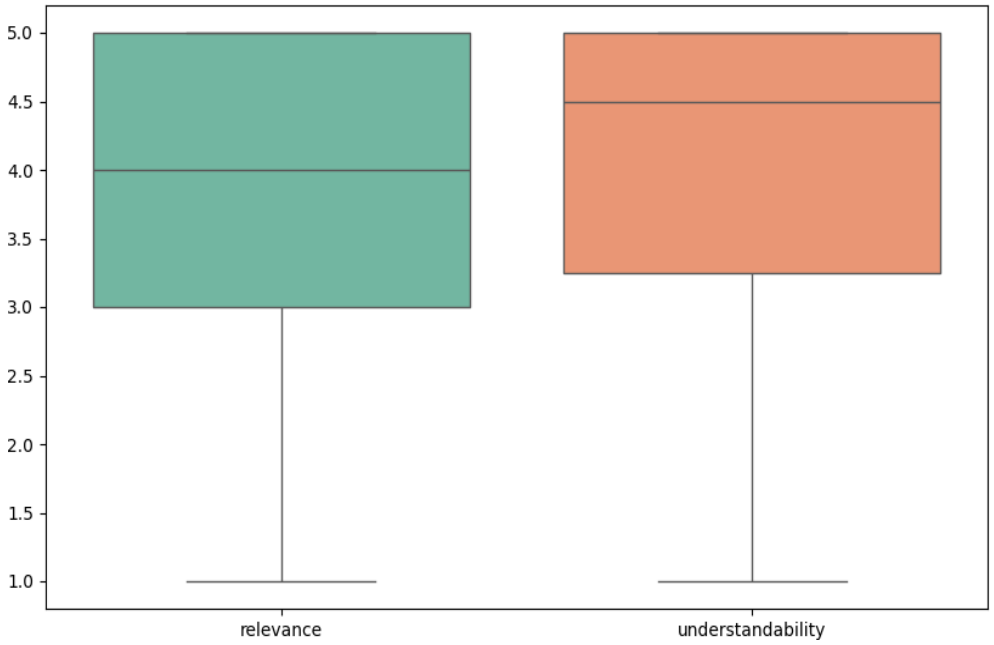


Fig. 8. 8 Evaluators rated the images retrieved by BM25 retriever removing stopwords on 2 different metrics: relevance, and understandability.

However, different people prefer to learn in different ways. EduAgent is designed on the presumption that a short, simple, and stepwise response that tells the learner what prerequisites are needed before it is pedagogically enhanced. Moreover, it still lacks the power to generate responses suitable for each individual. Our future work will focus on adding more individuality in generating responses so that EduAgent can respond with modified replies based on the user's interests, strengths, weaknesses, and learnability.

In the future, we will work on how EduAgent can be improved to help learners learn math more engagingly. Although LLMs still fall behind in the case of mathematical calculations, we will be focusing on mathematical understanding from different viewpoints.

Appendix

1. Appendix-A: Topics of the Dataset.

Table 3. Problem Count by Area

Area	Problem Count
Electronic Components and Applications	140
Power Electronics	120
Amplifiers and Oscillators	69
Digital Electronics	69
Transistors	69
Rectification and Diode Circuits	40
Electronic Instruments	34
Miscellaneous Topics	27
Logic Gates	22
Adders	6

2. Appendix B: Fine-tuning Parameters

Table 4. Fine-tuning parameters from Llama-3.2-3B model

Parameter Name	Value
per device train batch size	4
max steps	140
learning rate	2e-4
weight decay	0.01
optim	adamw 8bit
max seq length	2048
gradient accumulation steps	4

Table 5. Fine-tuning parameters from Llama-3.1-8B model

Parameter Name	Value
per device train batch size	2
max steps	120
learning rate	2e-4
weight decay	0.01
optim	adamw 8bit
max seq length	2048
gradient accumulation steps	4

3. Appendix C: Retriever Test Dataset Creation

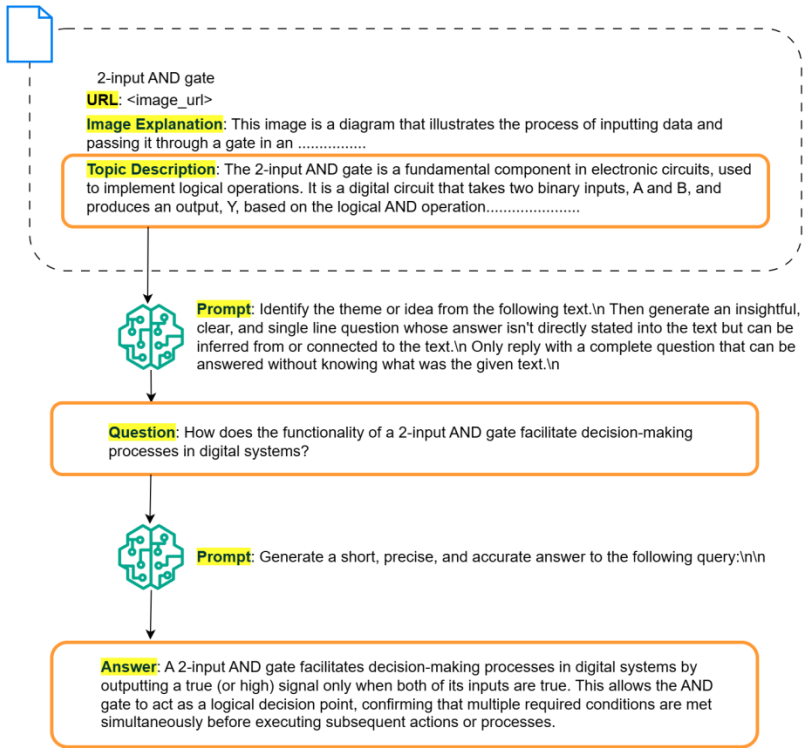


Fig. 9. Retriever Test Dataset Generation. From Topic Description, we generate a question using GPT-4o. Then using GPT-4o, we generate an answer for that question.

4. Appendix D: Retriever Comparison on an Example Query

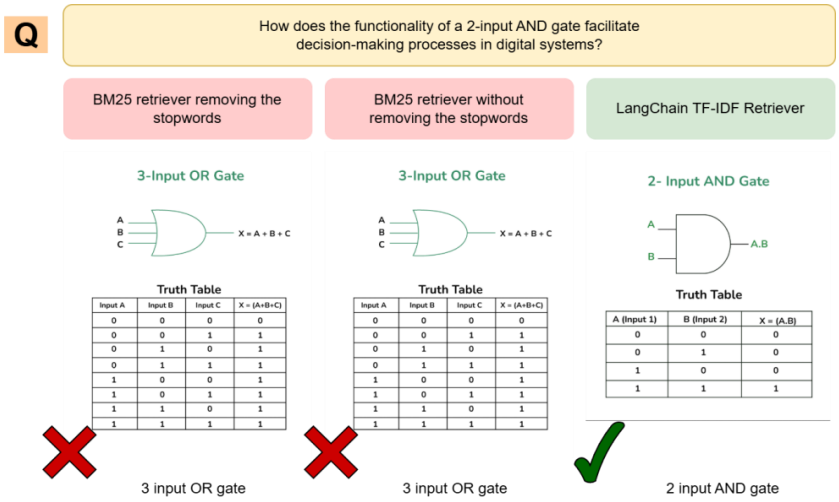


Fig. 10. Comparison of 3 types of retriever on a query. BM25 retriever failed to retrieve the appropriate image for the query for both cases. But, LangChain TF-IDF Retriever was successful as the image is appropriate given the context.

5. Appendix E: Snapshot of a Sample Conversation set

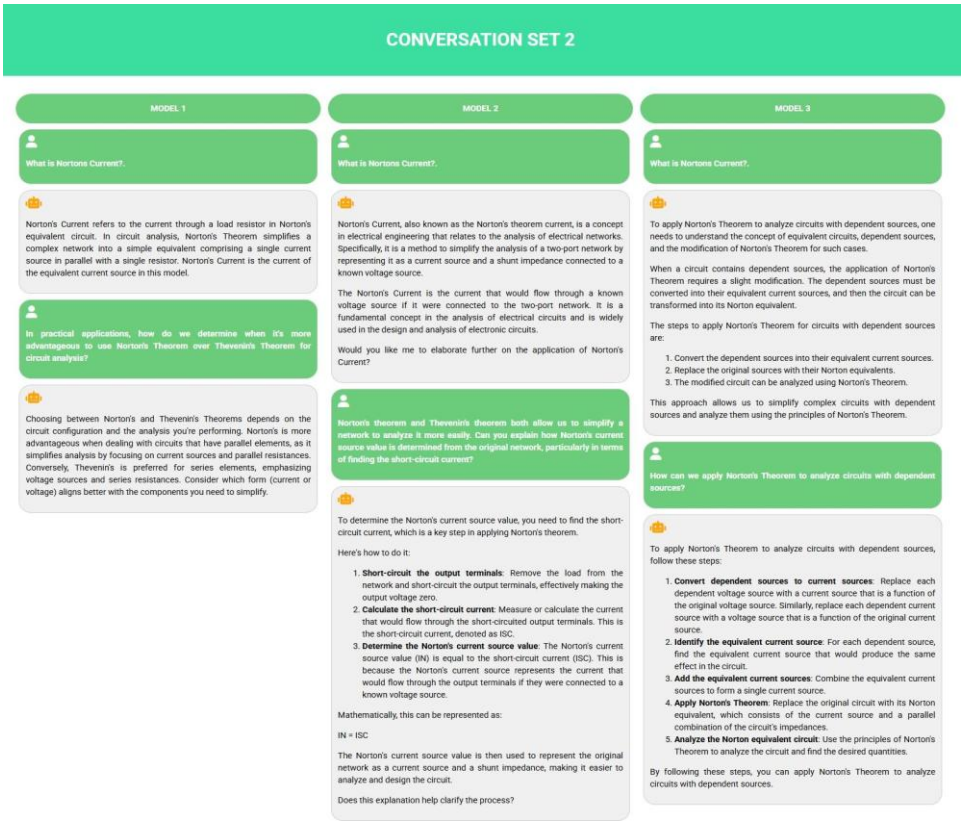


Fig. 11. A sample conversation set Where 3 conversations of 2 turn each are given side by side. Evaluators compare those conversations and rate them on the following metrics: correctness, conciseness, completeness, and naturalness.

6. Appendix F: Snapshot of a Sample Image set

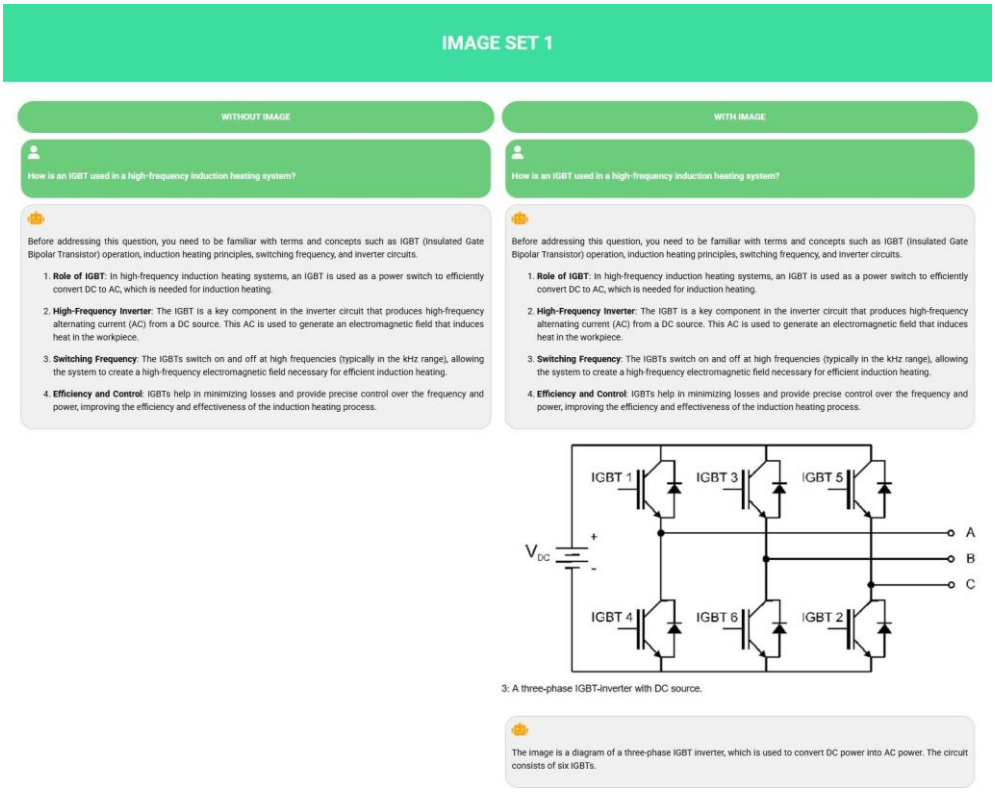


Fig. 12. A sample image set Where 2 conversations (with and without retrieved image) are given side by side. Evaluators compare those conversations and rate the retrieved image on the following metrics: relevance and understandability

References

[1] Lijia Chen, Pingping Chen, and Zhijian Lin. "Artificial Intelligence in Education: A Review". In: *IEEE Access* 8 (2020), pp. 75264–75278. DOI: 10.1109/ACCESS.2020.2988510.

[2] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[3] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2024. arXiv: 2312.11805 [cs.CL].URL: https://arxiv.org/abs/2312.11805.

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[5] David Duran. "Learning-by-teaching. Evidence and implications as a pedagogical mechanism". In: *Innovations in education and teaching international* 54.5 (2017), pp. 476–484.

[6] Pamela M Ironside. "Using narrative pedagogy: learning and practising interpretive thinking". In: *Journal of ad- vanced nursing* 55.4 (2006), pp. 478–486.

[7] Aarika Pardino et al. "The best pedagogical practices in graduate online learning: A systematic review". In: *Creative Education* 9.07 (2018), p. 1123.

[8] Bashaer Alsafari et al. "Towards effective teaching assistants: From intent-based chatbots to LLM-powered teach- ing assistants". In: *Natural Language Processing Journal* 8 (2024), p. 100101.

[9] Mahyar Abbasian et al. "Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Dia- betes Patients". In: *arXiv preprint arXiv:2402.10153* (2024).

[10] Zheyuan Zhang et al. *Simulating Classroom Education with LLM-Empowered Agents*. 2024. arXiv: 2406.19226 [cs.CL]. URL: https://arxiv.org/abs/2406.19226.

[11] Robin Schmucker et al. "Ruffle&riley: Towards the automated induction of conversational tutoring systems". In:*arXiv preprint arXiv:2310.01420* (2023).

[12] AN Sixu et al. "Developing an LLM-Empowered Agent to Enhance Student Collaborative Learning Through Group Discussion". In: *International Conference on Computers in Education*. 2024.

[13] Nasrin Dehbozorgi, Mourya Teja Kunuku, and Seyedamin Pouriyeh. "Personalized Pedagogy Through a LLM- Based Recommender System". In: *International Conference on Artificial Intelligence in Education*. Springer. 2024, pp. 63–70.

[14] Cecilia Ka Yuk Chan and Louisa HY Tsi. "The AI revolution in education: Will AI replace or assist teachers in higher education?" In: *arXiv preprint arXiv:2305.01185* (2023).

[15] Barak Rosenshine. "Principles of instruction: Research-based strategies that all teachers should know." In: *Ameri- can educator* 36.1 (2012), p. 12.

[16] Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[17] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[18] James Finnie-Ansley et al. "The robots are coming: Exploring the implications of openai codex on introductory programming". In: *Proceedings of the 24th Australasian Computing Education Conference*. 2022, pp. 10–19.

[19] Aakanksha Chowdhery et al. "Palm: Scaling language modeling with pathways". In: *Journal of Machine Learning Research* 24.240 (2023), pp. 1–113.

[20] Jordan Hoffmann et al. "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556*(2022).

[21] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971*(2023).

[22] Albert Q Jiang et al. "Mistral 7B". In: *arXiv preprint arXiv:2310.06825* (2023).

[23] Marah Abdin et al. "Phi-3 technical report: A highly capable language model locally on your phone". In: *arXiv preprint arXiv:2404.14219* (2024).

[24] H. Dang et al. "How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models". In: *CHI'22 Workshops, New Orleans, LA*. 2022.

[25] D. Alivanistos et al. *Prompting as Probing: Using Language Models for Knowledge Base Construction*. 2022. URL:https://github.com/HEmile/iswc-challenge.

[26] S. Arora et al. *ASK ME ANYTHING: A SIMPLE STRATEGY FOR PROMPTING LANGUAGE MODELS*. 2023. URL: https://github.com/HazyResearch/ama_prompting.

[27] B. Chen et al. "Unleashing the potential of prompt engineering in Large Language Models: a comprehensive re- view". In: (2024). eprint: arXiv:2310.14735.

[28] Xuezhi Wang et al. "Self-consistency improves chain of thought reasoning in language models". In: *arXiv preprint arXiv:2203.11171* (2022).

[29] Stephen Robertson, Hugo Zaragoza, et al. "The probabilistic relevance framework: BM25 and beyond". In: *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389.

[30] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685*(2021).

[31] L. Dai et al. "A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects". In: *Computers & Education* 190 (2022), p. 104607. DOI: 10.1016/j.compedu.2022.104607.

[32] P. Sikstro¨m et al. "How pedagogical agents communicate with students: A two-phase systematic review". In: *Com- puters & Education* 188 (2022), p. 104564. DOI: 10.1016/j.compedu.2022.104564.

[33] F. Weber et al. *Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education*. Completed Research Paper. 2022.

[34] J. C. Castro-Alonso et al. "Effectiveness of Multimedia Pedagogical Agents Predicted by Diverse Theories: a Meta-Analysis". In: *Educational Psychology Review* 33 (2021), pp. 989–1015. DOI: 10.1007/s10648-020- 09629-1.

[35] R. O. Davis, T. Park, and J. Vincent. "A Meta-Analytic Review on Embodied Pedagogical Agent Design and Testing Formats". In: *Journal of Educational Computing Research* 61.1 (2023), pp. 30–67. DOI: 10 . 1177 / 07356331221100556.

[36] R. Schlimbach et al. “A Literature Review on Pedagogical Conversational Agent Adaptation”. In: *Pacific Asia Conference on Information Systems 2022*. 2022.
[37] Zongxi Li et al. “Retrieval-augmented generation for educational application: A systematic survey”. In: *Computers and Education: Artificial Intelligence* (2025), p. 100417.
[38] Umair Hasan Khan, Muneeb Hasan Khan, and Rashid Ali. “Large Language Model based Educational Virtual Assistant using RAG Framework”. In: *Procedia Computer Science* 252 (2025), pp. 905–911.
[39] Chengshuai Zhao et al. “CyberBOT: Towards Reliable Cybersecurity Education via Ontology-Grounded Retrieval Augmented Generation”. In: *arXiv preprint arXiv:2504.00389* (2025).
[40] Andreas Martin et al. “ChEdBot: designing a domain-specific conversational agent in a simulational learning envi- ronment using LLMs”. In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1. 2024, pp. 180–187.
[41] Horia Alexandru Modran et al. “LLM intelligent agent tutoring in higher education courses using a RAG approach”. In: *International Conference on Interactive Collaborative Learning*. Springer. 2024, pp. 589–599.
[42] N Reddy. “Design and Implementation of an AI-Based Chatbot Framework with Retrieval-Augmented Generation and Integrated Recommender System for Interactive User Support”. In: *Available at SSRN 5250507* (2025).
[43] Karan Singhal et al. “Towards expert-level medical question answering with large language models”. In: *arXiv preprint arXiv:2305.09617* (2023).
[44] Renqian Luo et al. “BioGPT: generative pre-trained transformer for biomedical text generation and mining”. In:*Briefings in bioinformatics* 23.6 (2022), bbac409.
[45] A. J. Thirunavukarasu et al. “Large language models in medicine”. In: *Nature Medicine* 29.9 (2023), pp. 1740–1747. DOI: 10.1038/s41591-023-02448-8.
[46] M. Wornow et al. “The shaky foundations of large language models and foundation models for electronic health records”. In: *npj Digital Medicine* 6 (2023), p. 135. DOI: 10.1038/s41746-023-00725-2.
[47] X. Yang et al. “A large language model for electronic health records”. In: *npj Digital Medicine* 5 (2022), p. 194. DOI: 10.1038/s41746-022-00654-y.
[48] J. G. Meyer et al. “ChatGPT and large language models in academia: opportunities and challenges”. In: *BioData Mining* 16 (2023), p. 20. DOI: 10.1186/s13040-023-00270-9.
[49] E. Kasneci et al. “ChatGPT for good? On opportunities and challenges of large language models for education”. In: *Learning and Individual Differences* (2023), p. 102274. DOI: 10.1016/j.lindif.2023.102274.
[50] F. F. Xu et al. “A systematic evaluation of large language models of code”. In: *MAPS 2022: Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. 2022, pp. 1–10. DOI: 10.1145/3520312. 3534862.
[51] J. Huang et al. “Current state of large language models: Opportunities and challenges in artificial intelligence”. In:*Artificial Intelligence Review* 56.2 (2023), pp. 1377–1400. DOI: 10.1007/s10462-022-10207-0.
[52] A. Maatouk et al. “Large language models for telecom: forthcoming impact on the industry”. In: *IEEE Communi- cations Magazine* (2024). Early Access, pp. 1–7. DOI: 10.1109/MCOM.001.2300473.
[53] K. S. Unnithan. “An empirical study on Natural Language Processing in identifying industry demands to recom- mends required qualifications in the UAE”. Dissertation submitted in fulfillment of the requirement for the degree of MSc Informatics (Knowledge & Data Management). MA thesis. The British University in Dubai, 2021.
[54] M. U. Hadi et al. *A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage*. Computing and Processing, General Topics for Engineers. 2023. DOI: 10.1007/sXXXX-XXXX-XX.
[55] R. Ruffle and J. Riley. “The Future of Large Language Models: Opportunities and Ethical Considerations”. In:*Journal of Artificial Intelligence Research* 75 (2024), pp. 1–15. DOI: 10.1007/sXXXX-XXXX-XX.
[56] C. Zhou et al. “A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT”. In: (2023). arXiv preprint. DOI: 10.48550/arXiv.2302.09419.
[57] Z. Kong et al. “Entity Extraction of Electrical Equipment Malfunction Text by a Hybrid Natural Language Process- ing Algorithm”. In: *IEEE Access* 9 (2021), pp. 40954–40963. DOI: 10.1109/ACCESS.2021.3063354.
[58] Thanh Nguyen Ngoc et al. *AI-assisted Learning for Electronic Engineering Courses in High Education*. 2023. arXiv: 2311.01048 [cs.CY]. URL: https://arxiv.org/abs/2311.01048.
[59] Ishika Joshi et al. *”With Great Power Comes Great Responsibility!”: Student and Instructor Perspectives on the in- fluence of LLMs on Undergraduate Engineering Education*. 2023. arXiv: 2309.10694 [cs.HC]. URL: https://arxiv.org/abs/2309.10694.
[60] Pragati Shuddhodhan Meshram et al. *ElectroVizQA: How well do Multi-modal LLMs perform in Electronics Visual Question Answering?* 2024. arXiv: 2412.00102 [cs.CV]. URL: https://arxiv.org/abs/2412. 00102.
[61] Xiaoyi Dong et al. *InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model*. 2024. arXiv: 2401.16420 [cs.CV]. URL: https://arxiv.org/abs/ 2401.16420.
[62] *Llama3.1-8b-Instruct Unsloth model*. URL: https://huggingface.co/unsloth/Meta-Llama-3.1- 8B-Instruct-bnb-4bit.
[63] *Llama3.2-3b-Instruct Unsloth model*. URL: https://huggingface.co/unsloth/Llama- 3.2- 3B- Instruct.
[64] *Unsloth GitHub repository*. URL: https://github.com/unslothai/unsloth.
[65] Jiawei Shao, Yuyi Mao, and Jun Zhang. *Task-Oriented Communication for Multi-Device Cooperative Edge Infer- ence*. 2023. arXiv: 2109.00172 [eess.SP]. URL: https://arxiv.org/abs/2109.00172.
[66] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp.311-318. DOI:10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/.
[67] *Sacrebleu*. URL: https://github.com/mjpost/sacrebleu.
[68] Gerard Salton, Anita Wong, and Chung-Shu Yang. “A vector space model for automatic indexing”. In: *Communi- cations of the ACM* 18.11 (1975), pp. 613–620.

Authors' Profiles

Sayem Shahad is currently studying B.Sc. (hons.) degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), 1000, Dhaka Bangladesh. His major research interests are in the fields of Applied machine learning, Natural Language Processing, and Human Computer Interaction.

Salman Sayeed is currently studying B.Sc. (hons.) degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), 1000, Dhaka Bangladesh. His major research interests are in the fields of Applied machine learning, Natural Language Processing, and Human Computer Interaction.

Md Naimur Rahman Khan Sifat was born in Narayanganj, Dhaka, Bangladesh. He is currently in his 3rd year of Electrical and Electronic Engineering at University of Asia Pacific, Dhaka, Bangladesh. He began his undergraduate studies in 2022 and is expected to graduate in 2026. His interests are in the fields of electronics(IoT), natural language processing (NLP), and machine learning(ML).

Shuvo Biswas was born in Bangladesh on July 11, 1997. Earned a Bachelor of Science in electrical and electronics engineering from Faridpur Engineering College, Faridpur, Bangladesh, in 2021. He has gained professional experience in various roles, including Junior Engineer (Electrical), Elevated BIM/CAD Operator, Electrical Site Engineer, and Assistant Design Engineer. He is currently working as an Assistant Design Engineer in Mirpur, Bangladesh. His previous research interests included building electrical design and solar technology, while his current focus lies in electronics, semiconductors, and very-large-scale integration (VLSI).

Tishna Sabrina received B.Sc.Eng. (hons.) degree in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2005. She received a PhD degree from Monash University, Australia in 2014. She is an Assistant Professor in the department of Electrical and Electronic Engineering, University of Asia Pacific (UAP). Her major research interests are in the fields of Applied machine learning, Security privacy, and Communication Engineering.

How to cite this paper: Sayem Shahad, Salman Sayeed, Md Naimur Rahman Khan Sifat, Shuvo Biswas, Tishna Sabrina, "EduAgent: A Multimodal Chatbot Framework for Enhancing Interactive Learning in Electrical and Electronics Engineering Education", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.1, pp. 104-125, 2026. DOI:10.5815/ijmecs.2026.01.07