A Multimedia Interactive Presentation System Based on AIoT and RAG-Enabled Large Language Models

Ming-Shun Wang[1,‡], Ming-Che Chen[1,*], *Member*, *IEEE*
[1]Department of Electronic Engineering, Southern Taiwan University of Science and Technology,
Tainan, Taiwan
Email: [‡]*db230201@stust.edu.tw*; [*]*jerryhata@stust.edu.tw*

2025 IEEE International Conference on Consumer Electronics (ICCE) | 979-8-3315-2116-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCE63647.2025.10930122

Abstract—**This paper proposes a multimedia interactive presentation system, Gen-Presenter, that leverages Artificial Intelligence of Things (AIoT) and Retrieval Augmented Generation (RAG)-enabled Large Language Models (LLM). The Gen-Presenter system integrates an edge-AI computing device with peripherals such as a camera, microphone, speakers, and display screen, in combination with a natural language processing (NLP) server. The system detects visitor activity through the camera, infers their age group, and uses this information to select appropriate slides and generate corresponding voice narration for interactive presentations. Experimental results show that Gen-Presenter, when responding to queries from users of different age groups, can select appropriate presentation slides and generating corresponding explanations, with performance closely matching human decisions. In terms of slide selection, the overall precision, recall, and accuracy were 0.87, 0.67, and 0.76, respectively. Additionally, the suitability rates for the generated age-appropriate text content across the three age groups exceeded 0.7. This demonstrates the system's success in delivering human-like slide explanations and interactive Q&A sessions.**

Keywords—Large Language Models, Retrieval-Augmented Generation, Slide Presentation, Edge computing.

I. Introduction

With the rapid development of artificial intelligence (AI) and natural language processing (NLP) technologies, the way information is presented and communicated is undergoing revolutionary changes. Particularly in human-computer interaction applications such as guided tours and explanations, the integration of generative AI technology is becoming increasingly prevalent, driving a growing demand for personalized and human-like interactions, and further promoting the widespread adoption and advancement of these technologies.

Traditional presentation systems offer robust functionality for presenters, such as flexible slide switching and viewing explanation prompts. However, in automatic playback scenarios, these systems typically rely on pre-arranged slide sequences and fixed content, lacking the flexibility to adjust based on real-time needs. This limitation is particularly evident in exhibition environments, where visitors' ages and interests vary greatly, requiring more targeted and responsive explanations. To address this issue, many studies have proposed interactive systems that dynamically generate content based on visitor inquiries and needs. For example, humanoid guide robots have been introduced in museums and exhibitions, enabling interaction with visitors, and providing detailed explanations of exhibits [1][2]. Nevertheless, despite their strong performance in interactivity, the high cost remains a major obstacle to their widespread adoption.

This paper proposes a low-cost multimedia interactive presentation system called Gen-Presenter. The system integrates AI edge computing technologies, large language models (LLM), and the Retrieval-Augmented Generation (RAG) framework. Gen-Presenter can detect visitor attention and recognize characteristics (such as age) to enable personalized interactions and proactively engage with visitors through voice communication. Additionally, the system automatically selects the most relevant slides based on visitor inquiries and generates corresponding explanation content, significantly enhancing interactivity and visitor experience, while providing more real-time and intelligent presentation capabilities.

II. The Proposed System

Fig. 1 illustrates the architecture and workflow of the Gen-Presenter system. The system integrates an edge-AI computing device (ECD) and a natural language processing (NLP) server. The ECD device includes several modules: automatic speech recognition (ASR), text-to-speech (TTS), gaze detection with age estimation (GD-AE), and data processing (DP). These modules interface with peripherals such as a microphone, speakers, a camera, and a display screen. The NLP server operates two large language models (LLM1 and LLM2) within a retrieval-augmented generation (RAG) framework, which powers the system's interactive capabilities.

As the interface for interacting with visitors, the ECD device operates a control process, dynamically connecting to the LLM models on the NLP server based on visitor attention. This allows for diverse interactive experiences and the display of presentation content. The ECD device is pre-loaded with slides and posters that can be shown on the screen. As shown in Fig. 2, when the system is activated, the ECD utilizes LLM1 to generate attention-grabbing advertisement text, which is then transmitted through the DP module to the TTS module for speech synthesis. The content is subsequently played through the speakers while simultaneously displaying a poster on the screen.

Next, the ECD uses the camera to capture images and detect visitors through the GD-AE module. If no visitors are detected, the system automatically plays the full presentation. During the presentation, LLM2 generates corresponding explanation text for each slide, which is processed by the DP module and sent to the TTS module for speech synthesis, and the output is broadcasted through the speakers. After the presentation ends, the ECD restarts the process by having LLM1 generate a new attention-grabbing advertisement text for the next round of engagement.

This work was supported in part by National Science and Technology Council (NSTC), Taiwan, under the following grants: NSTC 113-2221-E-218-010-.

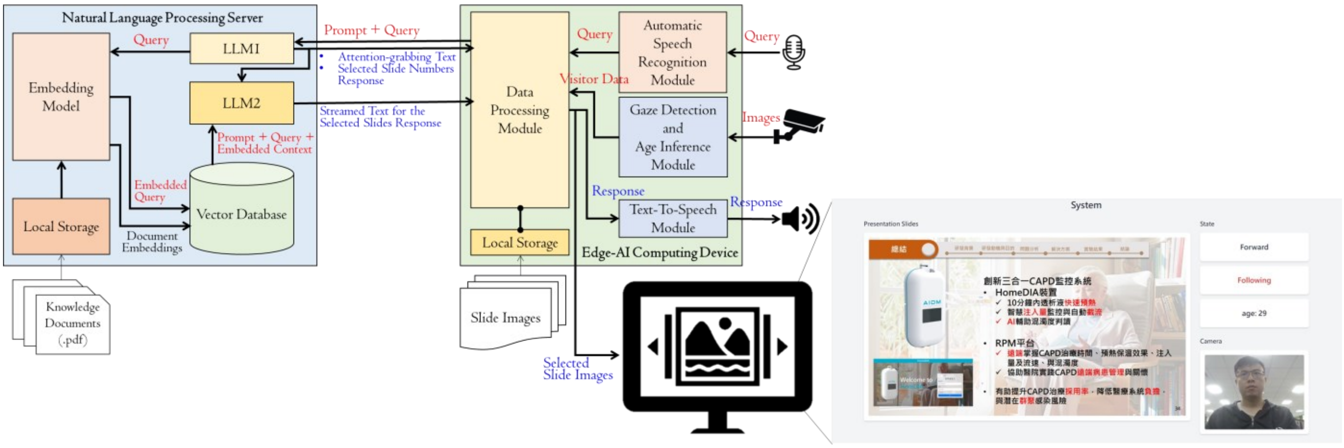
Natural Language Processing Server
Embedding Model
Query
LLM1
LLM2
Prompt + Query + Embedded Context
Embedded Query
Vector Database
Local Storage
Document Embeddings
Knowledge Documents (.pdf)
Prompt + Query
• Attention-grabbing Text
• Selected Slide Numbers Response
Streamed Text for the Selected Slides Response
Data Processing Module
Query
Visitor Data
Automatic Speech Recognition Module
Gaze Detection and Age Inference Module
Response
Text-To-Speech Module
Local Storage
Edge-AI Computing Device
Query
Images
Response
Slide Images
Selected Slide Images
System
Presentation Slides
總結
創新三合一CAPD監控系統
• HomeDIA裝置
✓ 10分鐘內透析液快速預熱
✓ 智慧注入量監控與自動截流
✓ AI輔助混濁度判讀
• RPM平台
✓ 遠端掌握CAPD治療時間、預熱保溫效果、注入量及流速、與混濁度
✓ 協助醫院實踐CAPD遠端病患管理與關懷
• 有助提升CAPD治療採用率，降低醫療系統負擔，與潛在群聚感染風險
State
Forward
Following
age: 29
Camera

Fig. 1 System architecture of the proposed Gen-Presenter system.

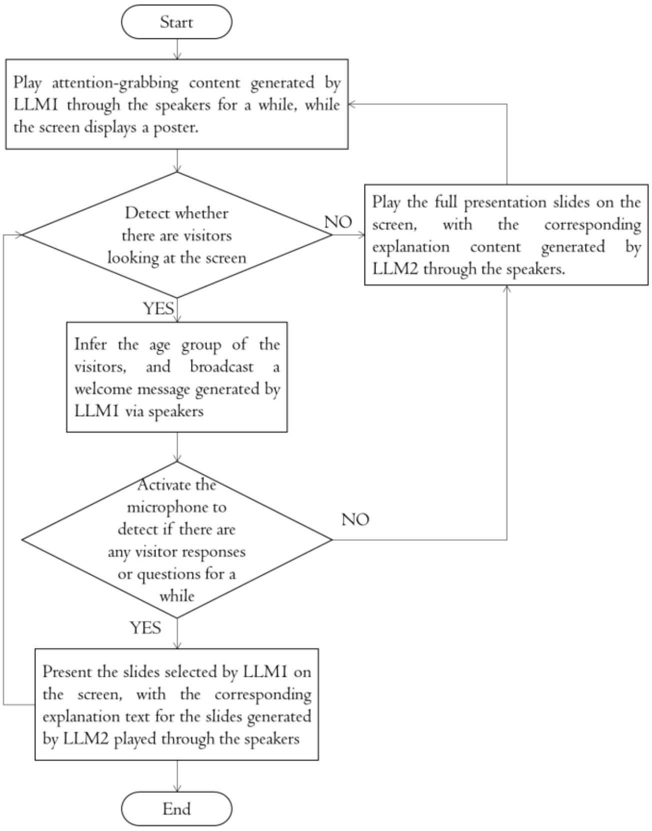
Start
Play attention-grabbing content generated by LLM1 through the speakers for a while, while the screen displays a poster.
Detect whether there are visitors looking at the screen
NO
Play the full presentation slides on the screen, with the corresponding explanation content generated by LLM2 through the speakers.
YES
Infer the age group of the visitors, and broadcast a welcome message generated by LLM1 via speakers
Activate the microphone to detect if there are any visitor responses or questions for a while
NO
YES
Present the slides selected by LLM1 on the screen, with the corresponding explanation text for the slides generated by LLM2 played through the speakers
End

Fig. 2 Operational flow of the ECD device in Gen-Presenter for visitor engagement

When visitor attention is detected (i.e., the visitor is gazing at the screen), the ECD uses the GD-AE module to infer the visitor's age and plays a welcome message generated by LLM1, initiating an interactive dialogue by proactively asking questions. During this process, the ECD activates the microphone to record the visitor's queries, converting the audio into text via the ASR module, which is then transmitted to the DP module along with the visitor's age information. The DP module uses the visitor's age as a key prompt parameter, combining it with the query text and forwarding it to LLM1.

LLM1 then selects the most relevant slide numbers based on the visitor's age and query, sending these slide numbers back to the DP module and forwarding them to LLM2 along with the age-related prompt information. Additionally, LLM1 sends the visitor's query to the embedding model, which converts high-dimensional data (such as the visitor's query and related presentation knowledge) into low-dimensional continuous vectors. These vectors are output in a fixed-dimension real number format and stored in a vector database. The vector database manages and retrieves these embedded vectors, providing LLM2 with the most similar vector lists along with their corresponding identifiers or metadata.

LLM2, receiving the selected slide titles and descriptions, uses the visitor's age-related prompt information to extract the most similar vectors and their expanded data from the vector database. It generates real-time streaming text explanations for each selected slide and sends them back to the DP module on the EC device. The DP module then displays the corresponding slides on the screen based on the selected slide numbers and transmits the streaming text to the TTS module, which ultimately broadcasts the corresponding voice explanations through the speakers. After the explanation is completed, the process loops back to detect if there are visitors, entering the next round of interaction.

To evaluate the operational performance of the Gen-Presenter system, this study used an Intel i7-9750H CPU paired with a NVidia Geforce GTX 1660 Ti GPU (with 6GB VRAM) as the hardware computing core for the ECD device, and a DGX V100 [3] as the NLP server. The ASR module on the ECD device was implemented using the OpenAI Whisper toolkit [4], while the TTS module was completed using the GPT-SoVITS toolkit [5][6]. The GD-AE module combined two open-source tools, Python-Gaze-Face-Tracker [7] and deepface [8], to enable facial tracking and age inference functionalities. The LLM model running on the NLP server was Llama-3.1 [9].

In the initial phase of the experiment, we referenced the authors' previous research [10] and designed two versions of the presentation, one for visitors under the age of 15 and the other for those above 15. Each version consisted of 10 slides (Slides #1~#10), with each slide accompanied by 4W1H question classifications, forming the core of the prompts. Additionally, the document [10] and its related files were incorporated into the embedding model of the NLP server, converted into fixed-dimension real number vectors, and stored in the vector database as a source for the RAG framework.

Table I presents the comparison results between the Gen-Presenter system and 10 test participants in selecting appropriate slides under five different question types (4W1H) scenarios. The selection strategy of the Gen-Presenter system is divided into two methods: one where LLM1 directly selects the slides, and another where LLM2 selects the slides with RAG support. These two methods limit the selection to a maximum of 5 slides per question as a response. The experimental results show that the version of LLM1 without RAG support had selections that were more consistent with human choices, achieving an overall precision of 0.87, a recall rate of approximately 0.67, and an accuracy of 0.76. Additionally, this version also outperformed the LLM2+RAG version in terms of processing time, with an average processing time of approximately 0.30 seconds. Table II displays the suitability rates of the slide explanations generated by LLM2 with RAG support under different age group conditions. The suitability rates were evaluated by the same group of 10 test participants, who assessed whether the slide explanations generated by LLM2 with RAG support for five different questions were appropriate for each of the three age groups (under 15-year-old, 15 to 65 years old, and over 65-year-old). The results indicate that the suitability rate for the explanation content across all three age groups exceeded 0.7.

III. Conclusion

This paper has proposed a low-cost Gen-Presenter system that integrates existing Edge-AI technology with LLM and RAG frameworks to enhance the effectiveness of interactive slide presentations. The system demonstrated satisfactory accuracy and efficiency in selecting and displaying relevant slides, with overall precision, recall, and accuracy reaching 0.87, 0.67, and 0.76, respectively. The average response processing time was 0.30 seconds, and the suitability rate for descriptive content across three age groups exceeded 0.7. Future work will focus on improving and optimizing prompt engineering and task collaboration between LLMs, as well as leveraging more visitor feature detection data to generate content explanations. This will allow the system to be applied to a broader range of presentation scenarios, further enhancing its interactivity and personalization capabilities.

References

[1] N. Ogata, et al., “Human-Like Guide Robot that Proactively Explains Exhibits,” *International Journal of Social Robotics*, vol. 12, no. 1, pp. 45-58, 2024. [Online]. Available: https://link.springer.com. Accessed: Oct. 7, 2024.

[2] Y. Tong, et al., “Advancements in humanoid robots: A comprehensive review and future prospects,” *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 2, pp. 301–328, Feb. 2024. doi: 10.1109/JAS.2023.124140

[3] "Volta V100 Datasheet," Nvidia. [Online]. Available: https://images.nvidia.com/content/technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf. [Accessed: Aug. 24, 2024].

[4] "OpenAI Whisper," GitHub. [Online]. Available: https://github.com/openai/whisper. [Accessed: Aug. 24, 2024].

[5] J. Xue et al., "Retrieval Augmented Generation in Prompt-based Text-to-Speech Synthesis with Context-Aware Contrastive Language-Audio Pretraining," *ArXiv*, 2024. [Online]. Available: https://arxiv.org/html/2406.03714v1

[6] "GPT-SoVITS," GitHub. [Online]. Available: https://github.com/RVC-Boss/GPT-SoVITS. [Accessed: Aug. 24, 2024].

[7] A. Alireza, “Python Gaze Face Tracker,” GitHub repository, https://github.com/alireza787b/Python-Gaze-Face-Tracker, accessed Oct. 7, 2024.

[8] S. Serengil, “DeepFace: A Lightweight Face Recognition and Facial Attribute Analysis Framework for Python,” GitHub repository, https://github.com/serengil/deepface, accessed Oct. 7, 2024.

[9] "Meta Llama 3.1 8B Instruct," Hugging Face. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct. [Accessed: Aug. 24, 2024].

[10] Ming-Che Chen et al., "iCAPD: A Deep Learning-Based Monitoring System for Continuous Ambulatory Peritoneal Dialysis," in *Proc. of 2023 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW 2023)*, Pingtung, Taiwan (R.O.C.), Jul. 17–19, 2023.

Table I: Experimental Results on Slide Selection Performance

Queries	Human Selection (Union)	LLM1 Selection	LLM2 Selection	Precision (LLM1/ LLM2)	Recall (LLM1/ LLM2)	Accuracy (LLM1/ LLM2)	Processing Time (s) (LLM1/ LLM2)
Query 1: What is this product?	Slides #1, #2, #6, #8, #9	Slides #1, #6, #8	Slides #1, #2, #3	**1.0**/0.67	**0.6**/0.4	**0.8**/0.6	**0.26**/7.26
Query 2: Who will use this product?	Slides #1, #2, #3, #4, #5, #7, #8	Slides #2, #3, #4, #8	Slides #4, #5	**1.0**/**1.0**	**0.57**/0.29	**0.7**/0.5	**0.28**/8.05
Query 3: Where is this product used?	Slides #2, #5, #6, #7, #8	Slides #2, #4, #5, #7, #8	Slides #1, #2	**0.8**/0.5	**0.8**/0.2	**0.8**/0.5	**0.22**/8.12
Query 4: When will this product be used?	Slides #1, #6, #7, #8, #9	Slides #5, #6, #7, #8	Slides #1, #8	0.75/**1.0**	**0.6**/0.4	**0.7**/**0.7**	**0.32**/8.04
Query 5: How is this product used?	Slides #2, #3, #4, #5, #6	Slides #3, #4, #5, #6, #10	Slides #3, #5, #7, #9	**0.8**/0.5	**0.8**/0.4	**0.8**/0.5	**0.41**/8.40
			Overall	**0.87**/0.73	**0.67**/0.38	**0.76**/0.56	**0.30**/7.97

Table II: Suitability Rates of Slide Explanations for Different Age Groups

Queries	under 15-year-old	15 to 65 years old	over 65-year-old
Query 1: What is this product?	0.9	1.0	0.8
Query 2: Who will use this product?	0.5	1.0	0.7
Query 3: Where is this product used?	1.0	1.0	0.6
Query 4: When will this product be used?	0.6	1.0	0.9
Query 5: How is this product used?	0.8	1.0	0.7
Overall	0.76	1.0	0.74