Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Transforming Applied Behavior Analysis Therapy: An Internet of Things-Guided, Retrieval-Augmented Large Language Model Framework

HAOMIN QI[1], CHUNG HO, SIN[1], ROSANNA Y.-Y. CHAN[1,2] (Fellow, IEEE), C. M. V. WONG [3]

[1]Department of Information Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR
[2]Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR
[3]Department of Special Education and Counselling, The Education University of Hong Kong, Tai Po, N.T., Hong Kong SAR

Corresponding author: Rosanna Y.-Y. Chan (E-mail: ychan@ie.cuhk.edu.hk).

This research was partially funded by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the HKSAR Government's Innovation and Technology Commission (ITC)'s InnoHK Scheme.

ABSTRACT We propose **ABA-RAG**, a retrieval-augmented generation (RAG) framework specifically tailored for applied behavior analysis (ABA) interventions, which integrates real-time emotional and behavioral data from Internet-of-Things (IoT) wearable devices. In ABA-RAG, we systematically retrieve semantically relevant prompts from a structured ABA task repository using embedding-based semantic search and dynamically construct context-aware prompts by incorporating IoT-derived predictions of learners' emotional and behavioral states. To efficiently adapt large language model (LLM) to ABA contexts, we evaluate and compare several low-rank fine-tuning methods—including Low-Rank Adaptation (LoRA), Butterfly Orthogonal Fine-Tuning (BOFT), and LoRA with Gradient Approximation (LoRA-GA). Among these techniques, LoRA-GA demonstrates the best balance between computational efficiency and generation quality, making it particularly suitable for resource-constrained environments. Comprehensive experiments, validated through rigorous quantitative metrics and expert ABA practitioner evaluations, demonstrate that ABA-RAG significantly reduces manual workload while enhancing the precision, contextual relevance, and clinical utility of generated ABA tasks. Integrated into a practical web-based system, ABA-RAG provides ABA professionals with a scalable and real-time platform to generate individualized interventions.

INDEX TERMS Retrieval-augmented generation (RAG), large language model (LLM), applied behavior analysis (ABA), IoT-based emotional analysis, efficient parameter fine-tuning (PEFT), autism spectrum disorders (ASD), special educational needs (SEN)

I. INTRODUCTION

Applied behavior analysis (ABA) is an evidence-based framework aimed at improving socially significant behaviors in learners with special educational needs (SEN), leveraging reinforcement, prompting, and shaping techniques [1]–[3]. Although effective, traditional ABA interventions demand continual monitoring and manual adjustments to address learners' evolving emotional and behavioral states, placing significant burdens on practitioners and limiting scalability [4], [5]. To overcome these challenges, we propose **ABA-RAG**, a novel retrieval-augmented generation (RAG) framework specifically tailored for ABA interventions. **ABA-RAG** first utilizes physiological sensor data (blood volume pulse, galvanic skin response, wrist temperature, and acceleration) collected from wearable Internet-of-Things (IoT) devices, classifying real-time emotional and behavioral states via a Transformer-based deep learning model. These predicted states then guide a semantic retrieval module, employing sentence embeddings and cosine similarity to automatically identify and extract relevant ABA task prompts from a structured repository. Finally, a large language model (LLM), fine-tuned through low-rank adaptation methods, including Low-Rank adaptation (LoRA), butterfly orthogonal Fine-Tuning (BOFT), and Low-Rank adaptation with gradient approximation (LoRA-GA), is used to generate individualized, context-sensitive ABA intervention tasks.

TABLE 1: ABA-RAG Framework Components: Inputs, Outputs, and Challenges

Component	Inputs	Outputs	Challenges Addressed
Learners	Real-time physiological signals; learner profile (developmental level, history, targets)	Behavioral responses	Natural variability in engagement; ensuring signal reliability
Practitioners	RAG-generated prompts; live sensor feedback	Structured behavioral annotations; fidelity metrics; prompt adjustments	Balancing real-time decisions with AI recommendations; maintaining ABA fidelity
Tasks	Repository entries (domain, task, sub-task); learner profile; history	Semantically retrieved, context-aware instructional cues	Semantic alignment across diverse learner needs; task variability
Interventions	Context-aware LLM prompts; practitioner selections	Individualized teaching episodes; recorded outcomes	Ensuring clinical relevance; integrating IoT predictions seamlessly
IoT Integration	Continuous multimodal sensor streams; session event markers	Synchronized data streams; monitoring dashboards	Robust multimodal fusion; sub-ms synchronization; real-time quality control

FIGURE 1: Illustration of the ABA training scenario conducted in this work.

A. CURRENT CHALLENGE AND THE PROPOSED FRAMEWORK

In today's ABA therapy practice, it is often challenging to personalize interventions in real time due to reliance on manual observation and static task repositories. Here we introduce our proposed solution framework and provide a high-level example to illustrate how each component is connected.

1) Framework Components and Role Definitions

The proposed ABA-RAG framework comprises five interconnected components:

- **Learners**: Children with SEN who receive ABA therapies designed to increase adaptive behaviors and decrease problematic behaviors through systematic application of behavior-analytic principles [1].
- **Practitioners**: Board Certified Behavior Analysts (BCBAs) who design, implement, and supervise evidence-based behavioral interventions.
- **Tasks**: Structured ABA learning activities derived from comprehensive behavioral assessments that target specific skill acquisition or behavior reduction goals.
- **Interventions**: Systematic behavior-analytic procedures including discrete trial instruction, naturalistic teaching strategies, and functional communication training that are implemented to promote meaningful behavior change in learners.
- **IoT Integration**: Wearable devices and edge computing systems that continuously monitor learner physiological responses during intervention delivery.

Table 1 lists the inputs, outputs, and challenges addressed by each component. These components work together to overcome key limitations in current ABA practice.

2) A High-Level Example

Figure 1 illustrates our approach in a real-world ABA therapy session. A learner wears an IoT-enabled wristband that streams physiological signals (e.g., heart rate, motion) to an edge platform, while a practitioner uses a web interface to retrieve and refine ABA tasks based on the learner's profile and real-time sensor insights. By integrating RAG-LLMs with sensor input, our framework dynamically recommends tasks that maximize the learner's success probability, addressing limitations of traditional labor-intensive ABA interventions.

B. OUR CONTRIBUTION

To our knowledge, this is the first attempt to unite RAG-LLM methods with IoT-driven data analysis in ABA therapy, thereby offering an end-to-end framework optimized for SEN education. We detail four main contributions:

(1) **IoT-Based Data Pipeline:** We develop a transformer-based deep model that classifies emotional and behavioral states from multimodal sensor data, enabling context-aware task generation.
(2) **ABA-RAG Design:** We design a specialized retrieval pipeline that automatically provides relevant ABA prompts, reducing the workload on practitioners.
(3) **LLM Fine-Tuning:** Low-rank adaptation techniques, including LoRA, BOFT, and LoRA-GA, optimize LLM fine-tuning for high-quality, computationally efficient task generation.
(4) **Web System Integration and Deployment:** We integrate our framework into a user-friendly web interface that enables practitioners to generate, test, and refine tasks. Empirical trials were also conducted in real-world settings to validate its practical effectiveness.

All training scripts, source code, and the integrated web-based system implementation are publicly available at https://github.com/1314spb/IoT-RAG-ABA.

II. RELATED WORK

A. AI-ENHANCED ABA SYSTEM AND REAL-WORLD TRIALS

ABA is a foundational approach for promoting socially meaningful behavior change in learners with Special Educational Needs (SEN) [6]. Early work aimed to reduce practitioner workload by streamlining interventions and aligning them with each learner's core goals [7], [8]. Later efforts introduced data-driven methods to better manage session flow and track behavioral improvements [9].

Recent large-scale or longitudinal trials have demonstrated the real-world efficacy of digital ABA systems. For instance, a five-month retrospective study reported significant improvements in participants' target behaviors, supporting ABA's effectiveness in naturalistic settings [10]. Beside, the Stanford Superpower Glass study [11] found that a Google Glass-based wearable significantly improved socialization skills in 71 children, matching standard care results. Similarly, a 12-month evaluation of the AI-powered CognitiveBotics platform with 43 children showed developmental gains when used alongside standard therapy [12]. While these advances laid the groundwork for automated ABA practice, most lack real-time adaptation to emotional or behavioral cues. Our framework addresses this by embedding data-driven adjustments into every stage of task design, bridging the gap between theory and application in SEN education.

B. RELATED EMERGING TECHNOLOGIES

Technologies like gamification, software, and robotics have been used to augment ABA interventions [13]. Recent reviews of AI-assisted and digital ABA interventions demonstrate the real-world effectiveness of AI and related technologies in enhancing ABA outcomes [14]–[16]. Building on this evidence, we review emerging AI-enhanced ABA technologies that directly inform our current work.

1) LLMs in Digital Health and Special Education

LLMs are transforming healthcare, particularly in clinical decision support, with domain-specific models outperforming traditional methods on knowledge-intensive tasks [17]. Parameter-efficient fine-tuning (e.g., LoRA) has become the preferred strategy for medical adaptation [18]. RAG further boosts reliability by grounding responses in authoritative medical sources [19]. RAG-based LLMs further show strong promise in educational and therapeutic contexts. By combining generative capabilities with targeted retrieval, they enable personalized, context-aware content generation [20]. Studies further demonstrate improved precision and relevance in educational materials [21], [22], supporting adaptive learning, timely feedback, and reduced manual workload for practitioners.

Our system retrieves ABA tasks enriched with IoT-based emotional cues to generate context-aware prompts, enabling dynamic, personalized content and bridging the gap between LLM potential and real-world ABA practice.

2) IoT-Enabled Therapy and Recent Advances

IoT-enhanced observation has been used in special education for enhancing behavioral assessment [23]. IoT systems that track variables like skin conductance and classroom temperature offer deeper insights into learners' affective states [24]. Studies also show that sensor-driven feedback improves responsiveness in both educational and clinical contexts, particularly for monitoring stress and engagement [25]. Advanced IoT-based emotion recognition systems further achieve high accuracy [26], and real-time vibro-tactile feedback has helped adults with ASD identify up to seven distinct emotions [27]. Recent work has advanced IoT-enabled therapeutic platforms. For instance, a system combining wearables, noise-reduction, and a Particle Swarm Optimization-Support Vector Machine (PSO-SVM) classifier supports real-time health monitoring in adolescent rehabilitation [28]. A recent review also highlights Artificial Intelligence (AI), IoT, and sensor technologies in ASD diagnosis [29].

Our work extends these ideas by capturing emotional and behavioral signals from connected IoT devices, then incorporating these signals into the retrieval-augmented generation pipeline. This approach offers immediate contextualization for ABA tasks, allowing interventions to be more directly aligned with each learner's current needs.

3) Low-Rank Adaptation for Domain-Specific Fine-Tuning

Fine-tuning large language model for specialized tasks can be prohibitively resource-intensive, which has prompted the introduction of leaner parameter update strategies [30], [31]. LoRA addresses this challenge by constraining modifications

to compact matrices while preserving the core parameters of the pre-trained model. Extensions such as LoRA-GA and BOFT further incorporate gradient-aligned initialization or orthogonal transformations to improve training stability and accelerate convergence [32]. These techniques have proven effective in domains that require ongoing adaptation, such as personalized learning or dynamic content generation [33]. We adopt these methods to manage model complexity within our IoT-driven framework, ensuring that the retrieval-augmented LLM can handle evolving SEN contexts without incurring high computational overhead.

C. HUMAN-IN-THE-LOOP IN AI-ENHANCED SYSTEMS

Human-in-the-loop (HITL) systems that integrate human expertise with AI or machine learning algorithms are becoming increasingly prevalent across various industries [34]. HITL approaches are vital for ensuring safety, accountability, and clinical or educational effectiveness in advanced AI-driven systems, particularly in sensitive domains such as ABA for individuals with SEN [35]–[37]. HITL design ensures that practitioners remain central to the intervention process [38]. By embedding human oversight at critical decision points such as validating AI-generated recommendations, the system combines automated intelligence with clinical expertise to support informed, context-sensitive decisions [39]. In the context of ABA, this human-AI synergy enables the dynamic adaptation of intervention tasks to a learner's current emotional and behavioral state, while preserving practitioner control and ensuring adherence to ethical standards.

HITL design plays a central role in our framework. By maintaining practitioner control over intervention decisions, our system promotes the ethical and responsible use of AI in sensitive SEN contexts.

III. METHODOLOGY

A. PARTICIPANTS AND ETHICAL CONSIDERATIONS

A total of $N = 43$ participants diagnosed with ASD (13 females, 30 males) were recruited. Data from $N = 33$ learners (aged 6 to 15) enrolled in ABA therapy for at least three months were used, with 10 learners (5 females, 5 males) included in the empirical evaluation. Two expert evaluators (1 female, 1 male) independently assessed the quality of LLM-generated task descriptions. All study procedures were approved by the Survey and Behavioural Research Ethics Committee of The Chinese University of Hong Kong (Ref. SBRE-24-0806). Written informed consent was obtained from parents or legal guardians, and all data were anonymized and securely stored in a private network.

B. DATASET OVERVIEW AND PREPROCESSING

1) IoT Data Collection Protocol

We used a three-phase IoT data collection protocol to ensure systematic capture of physiological and behavioral data:

1) **Phase 1: Setup and Initialization** The learner wears the E4 wristband on the non-dominant wrist. The practitioner powers on the wristband and edge device to begin data collection, display tasks, and record assessments.
2) **Phase 2: Training and Data Collection** During ABA interventions, the system presents tasks based on instructional cues. The practitioner observes responses and records outcomes using: "+" (Positive), "–" (Negative), "P" (Prompt), and "OT" (Off Task), while real-time data are captured.
3) **Phase 3: Completion and Transfer** The practitioner ends recording, removes the wristband, powers down devices, and securely transfers the session data to a local server for analysis.

2) Dataset Composition

Our framework utilizes (1) the ABA Task Dataset and (2) the SEN Multimodal Dataset, as described below.

1) The **ABA Task Dataset** contains approximately 10,000 structured ABA intervention records. The preprocessing pipeline includes:
   - Data Structuring: Tasks are uniformly reformatted into structured records consisting of three explicit fields: domain, task, and sub-task, that include areas such as social emotion and daily living skills to ensure comprehensive coverage for ABA interventions.
   - Textual Normalization: To standardize the textual data for accurate retrieval, preprocessing includes converting to lowercase, removing punctuation, and lemmatizing terms to their base lexical forms to ensure consistency in semantic interpretation and retrieval accuracy.
2) The **SEN Multimodal Dataset** consists of synchronized physiological and behavioral data collected through Empatica E4 wristbands during ABA therapy sessions. Each student was uniquely identified by a distinct student ID, and the dataset includes:
   - Physiological Signals: four sensors on E4 wristband, namely a photoplethysmography (PPG) sensor, electrodermal activity (EDA) sensor, temperature sensor, and a 3-axis accelerometer respectively produces four continuous data streams namely blood volume pulse (BVP), galvanic skin response (GSR), wrist temperature, and 3-axis acceleration data.
   - Behavioral Annotations: Each recorded instance is labeled with behavioral outcome classifications using the four-category system: "+" (Positive), "-" (Negative), "P" (Prompt), and "OT" (Off Task).

These physiological signals have been shown to predict ABA outcomes, as demonstrated in our previous work [9]. Table 2 summarizes the key parameters, input-output formats, and selection justifications for both the ABA Task Dataset and the SEN Multimodal Dataset.

3) Data Synchronization and Quality Assurance

To align physiological signals with behavioral annotations, all IoT devices streamed sensor data with Universal Time

TABLE 2: Dataset Parameters, Input–Output Formats, and Selection Justifications

Dataset	Parameters	Input–Output Format	Justification
ABA Task Dataset	• Domain (categorical) • Task (text) • Sub-task (text)	• Input: query embedding + domain filter • Output: structured record with domain, task, sub-task	• Covers key ABA skill domains • Enables precise semantic retrieval • Text normalization ensures consistency
SEN Multimodal Dataset	• BVP (continuous, 64 Hz) • GSR (continuous, 4 Hz) • Temperature (continuous, 1 Hz) • Acceleration (continuous, 32 Hz) • Trial label (categorical: +, –, P, OT)	• Input: synchronized time-series streams + timestamped labels • Output: feature vectors (windowed sensor statistics) + label	• Multimodal signals capture emotional/behavioral state • High sampling rates balance resolution and noise • Categorical labels align with ABA coding schema

Coordinated (UTC)-based timestamps. Trained ABA therapists used a tablet app to log behavioral events with precise timestamps. Post-session, sensor and annotation data were merged automatically, achieving sub-second synchronization (±50 ms). Practitioners followed a standardized annotation guideline specifying event definitions, temporal precision, and behavioral tags to ensure consistency and reliability across sessions.

4) Data Preprocessing Pipeline

The preprocessing pipeline for the SEN multimodal dataset follows these systematic steps:

- Individual Student-based Data Structuring: Data is structured individually for each student based on their unique identifiers, ensuring personalized preprocessing and training rather than merging data across multiple students indiscriminately.
- Data Quality Filtering: Each student's data is filtered to maintain only records that contain complete sensor data, valid timestamps, and clearly defined behavioral outcome labels. This step ensures the high relevance and quality of each student's dataset.
- Outlier Detection and Removal: Data entries with extreme readings, defined as those exceeding three standard deviations from the mean, are systematically removed to ensure robustness in model training and prevent distortion caused by anomalous data points.

C. TRANSFORMER-BASED BEHAVIOR OUTCOME CLASSIFIER

1) IoT Data Encoding

After preprocessing, the SEN multimodal dataset is encoded into numerical formats for deep learning classification. Raw physiological signals and behavioral annotations are transformed into structured inputs to predict emotional and behavioral states, which inform the ABA-RAG framework. Behavioral outcomes ("+", "–", "P", "OT") are label-encoded and one-hot encoded. Physiological signals—BVP, GSR, temperature, and 3-axis acceleration—are standardized. The dataset is then stratified into training and test sets to ensure balanced class representation, as shown in Figure 2.

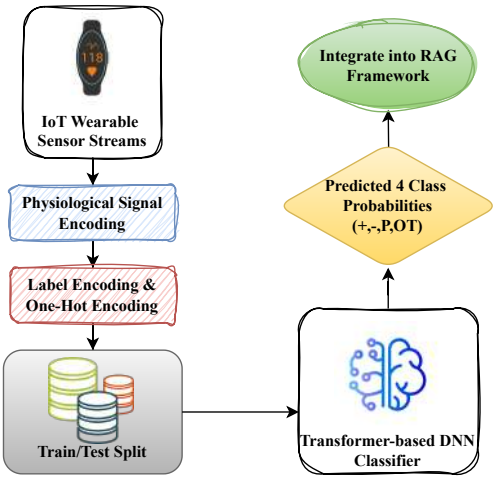

FIGURE 2: IoT data encoding and deep learning model training pipeline. Key stages include IoT physiological signals numerical encoding, label encoding of behavioral outcomes, Transformer-based classification training, and generation of predicted class probabilities for RAG integration.

2) Classifier Selection Rationale

Given the complexity and temporal nature of multimodal IoT signals, we adopted a transformer-based architecture for classification [40]. Prior evaluations of SVMs, Random Forests, and MLPs showed limited ability to capture temporal dependencies and subtle patterns. In contrast, the transformer's self-attention mechanism [41] effectively modeled long-range relationships in sensor data, yielding significantly higher accuracy. Despite added complexity, the transformer's

superior performance justified its use for accurate emotion and behavior prediction in personalized ABA interventions.

3) Transformer Architecture

The transformer-based model employed in this study comprises the following components:

- **Input Layer:** Matches the dimension of numerically encoded features.
- **Transformer Encoder Layers:** Two stacked encoder blocks, each consisting of:
  - Multi-head self-attention with 4 attention heads, enabling the model to attend to multiple segments of IoT sensor data simultaneously.
  - Feed-forward neural network with two fully connected layers containing 128 and 64 neurons respectively, activated by Rectified Linear Units (ReLU) [42].
  - Layer normalization and dropout (rate = 0.1) after each sub-layer to stabilize training and mitigate overfitting.
- **Output Layer:** A softmax classification layer producing probability distributions across the four behavioral outcome classes ("+", "−", "P", "OT").

4) Training Algorithm and Hyperparameters

The model training process is outlined explicitly in Algorithm 1. The transformer parameters are optimized using the cross-entropy loss function:

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{1}$$

where $\hat{y}$ denotes the predicted class probabilities, and y represents the true class labels.

The following hyperparameters were used for model optimization: learning rate $= 2\times10^{-5}$, weight decay $= 1\times10^{-4}$, batch size $B = 64$, MLP dropout $= 0.4$, dropout $= 0.1$, and number of training epochs $E = 70$. Optimization was performed using the AdamW algorithm, with all values empirically determined through grid search to ensure effective convergence and robust generalization during training.

Upon post-training, the transformer outputs class probabilities for behavioral outcomes, quantifying the learner's likely emotional and behavioral states. These probabilities are fed into the ABA-RAG framework to dynamically retrieve tasks aligned with the learner's current state and anticipated reinforcement needs.

Algorithm 1 Transformer-based IoT Classification Model Training

Require: Encoded feature matrix X, labels y, epochs E, batch size B
Ensure: Optimized neural network parameters θ

```
Initialize neural network parameters θ randomly
Define loss function ℒ as Cross-Entropy Loss
Initialize AdamW optimizer for parameters θ
Split dataset into stratified training (X_train, y_train) and test sets (X_test, y_test)
for epoch = 1 to E do
    Set model to training mode
    for each training batch (X_b, y_b) do
        Forward propagate X_b to obtain predictions ŷ_b
        Compute loss: loss = ℒ(ŷ_b, y_b)
        Backward propagate loss
        Update model parameters θ using AdamW optimizer
    end for
end for
Evaluate the trained model on test set to generate predicted class probabilities
return trained neural network parameters θ
```

D. ABA-RAG AUTOMATION FRAMEWORK

The ABA-RAG automation framework consists of three structured layers, aligned with the components detailed earlier.

- **Predicted probabilities (P_e):** These provide real-time, learner-specific emotional context, enabling the model to adjust tone, content difficulty, or reinforcement mechanisms appropriately [43].
- **Retrieved tasks ($T_{\text{retrieved}}$):** They supply grounded, semantically relevant ABA content that ensures generated responses are not only syntactically correct but clinically actionable.
- **Domain knowledge (D):** It acts as a theoretical scaffold to constrain the model's behavior, ensuring generated content conforms to accepted ABA guidelines and preserves task fidelity.

Each parameter involved in the ABA-RAG framework is carefully chosen to support key functional roles in ABA task generation:

1) **IoT Behavioral Outcome Prediction Layer:** Uses transformer-generated probability distributions across four behavioral states ("+", "−", "P", "OT") to adjust task difficulty, reinforcement, and prompting based on real-time emotional cues (e.g., high "P" triggers supportive prompts).
2) **Semantic Retrieval Layer:** Retrieves relevant tasks from the structured ABA dataset using dense embeddings generated by a Sentence-Transformer (M_{ST}). User queries—explicit or inferred—are matched to task embeddings via cosine similarity. Tasks exceeding similarity threshold $\alpha = 0.75$ are selected, with the top $k = 5$ forming $T_{\text{retrieved}}$, grounding LLM outputs in real ABA practices.
3) **Domain Knowledge Layer:** Embeds curated ABA domain knowledge—including six key intervention areas—into prompts to maintain alignment with professional standards and ensure theoretical rigor.

The integrated interaction of these layers is illustrated in

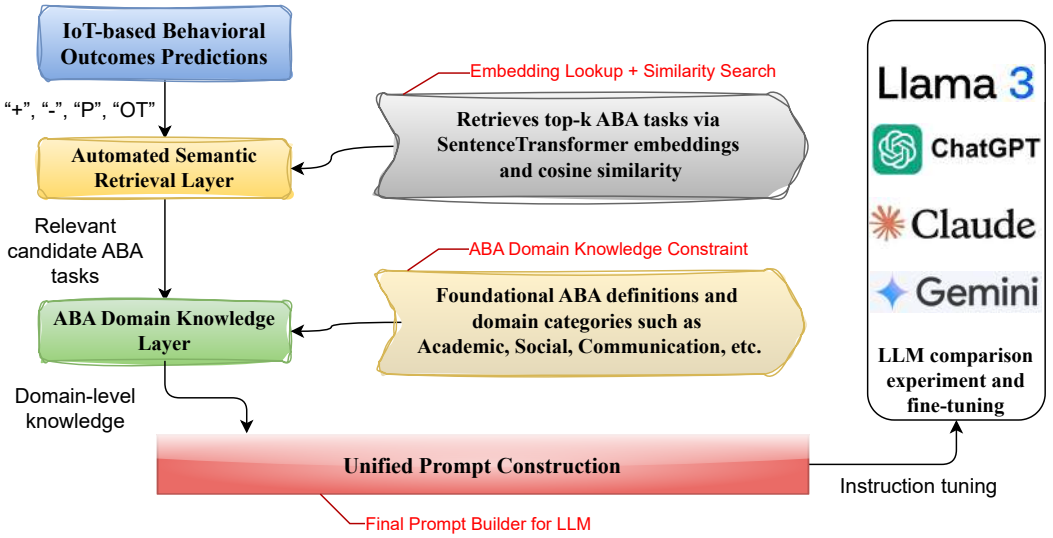

FIGURE 3: The ABA-RAG framework illustrating the integration of IoT-derived behavioral outcome predictions, embedding-based semantic task retrieval, and structured ABA domain knowledge into enriched prompts guiding personalized ABA task generation by the LLM.

Figure 3, highlighting the flow from real-time predictions and retrieval to enriched prompting. The process of prompt construction is detailed in Algorithm 2. It describes how semantic search results, behavioral outcome predictions, and domain-level knowledge are combined to form an informative prompt for the large language model.

Algorithm 2 Prompt Construction for ABA-RAG Task Generation

Require: IoT predicted emotional probabilities P_e, user query Q, ABA task dataset embeddings E_{task}, Sentence-Transformer model M_{ST}, similarity threshold α, number of retrieved tasks k, ABA domain knowledge D

Ensure: Constructed prompt $Prompt_{\text{final}}$ for LLM

Encode user query Q into embedding vector $\mathbf{q}$ using M_{ST}
Initialize empty result set $\mathcal{S} \leftarrow \emptyset$
for each embedding $\mathbf{t}_i$ in E_{task} **do**
  Compute cosine similarity $s_i = \frac{\mathbf{q} \cdot \mathbf{t}_i}{\|\mathbf{q}\| \|\mathbf{t}_i\|}$
  if $s_i > \alpha$ **then**
    Add (s_i, t_i) to $\mathcal{S}$
  end if
end for
Sort $\mathcal{S}$ in descending order by similarity
Select top k tasks and aggregate into formatted text $T_{\text{retrieved}}$
Format P_e into readable summary T_{emotion} (e.g., "Predicted: 70% Prompt, 20% Positive")
Construct full prompt: $Prompt_{\text{final}} = [T_{\text{emotion}}; T_{\text{retrieved}}; D]$
return $Prompt_{\text{final}}$

E. EFFICIENT LLM FINE-TUNING

To improve efficacy and efficiency within the RAG framework, we evaluated several advanced fine-tuning techniques [44]. These methods reduce computational overhead by restricting weight updates to a low-dimensional space while preserving performance to varying degrees.

1) Low-Rank Adaptation (LoRA)

LoRA decomposes the weight updates into two smaller matrices, dramatically reducing trainable parameters [45]. Mathematically, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, the low-rank update is approximated as:

$$W = W_0 + \Delta W = W_0 + A \cdot B, \tag{2}$$

where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ contain significantly fewer parameters than the original W_0, with $r \ll \min(d, k)$. During training, gradients with respect to these matrices are computed by:

$$\frac{\partial L}{\partial A} = \frac{\partial L}{\partial W} \cdot B^\top, \quad \frac{\partial L}{\partial B} = A^\top \cdot \frac{\partial L}{\partial W}, \tag{3}$$

and the corresponding updates are subsequently applied via gradient descent:

$$A^{t+1} = A^t - \eta \frac{\partial L}{\partial A^t}, \quad B^{t+1} = B^t - \eta \frac{\partial L}{\partial B^t}. \tag{4}$$

Although LoRA significantly reduces the number of trainable parameters, basic LoRA can face limitations in handling complex semantic tasks required by ABA interventions.

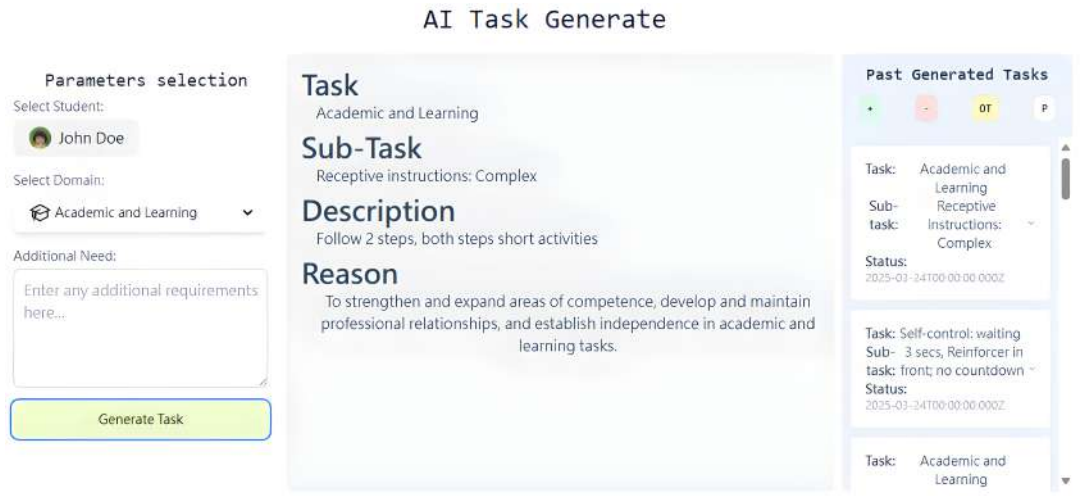

FIGURE 4: Web interface for generating ABA tasks. Users select a student profile and specific domain to generate tailored interventions based on real-time IoT-derived emotional and behavioral states.

2) Butterfly Orthogonal Fine-Tuning (BOFT)

BOFT uses orthogonal butterfly matrices for efficient parameterization and stable training. The weight matrix is given by:

$$W = B_k B_{k-1} \dots B_1. \tag{5}$$

Each B_i is an orthogonal butterfly matrix that preserves gradient magnitudes, preventing explosion or vanishing and stabilizing training. Gradient updates for each B_i follow the structured form

$$B_i^{t+1} = \big(I + \epsilon Q_i\big)^{-1}\big(I - \epsilon Q_i\big)B_i^t, \tag{6}$$

where Q_i Q_i is a skew-symmetric matrix ensuring orthogonality, and ϵ is a step size. This structure keeps the spectral norm bounded [32], stabilizing gradient flow:

$$\|W\|_2 = \big\|B_k B_{k-1} \dots B_1\big\|_2 = 1. \tag{7}$$

3) LoRA with Gradient Approximation (LoRA-GA)

LoRA-GA further refines initialization and convergence speed by aligning low-rank matrix updates with the principal directions of the full gradient matrix [46]. Specifically, let G_W denote this gradient matrix. Its singular value decomposition (SVD) is given by

$$G_W = U\,\Sigma\,V^\top,$$

where U and V are orthonormal matrices, and Σ is a diagonal matrix of singular values. LoRA-GA employs this decomposition to initialize the low-rank factors:

$$A_0 = U\,\Sigma^{\frac{1}{2}}, \quad B_0 = V\,\Sigma^{\frac{1}{2}},$$

providing optimal starting directions for weight updates.

We define $\Delta W = A\,B$ as the low-rank update to the original weight matrix. Then LoRA-GA seeks to minimize the difference between ΔW and $\eta\,G_W$, formulated as:

$$\min_{A,B}\Big\|\Delta W - \eta\,G_W\Big\|_F = \Big\|\eta\,(A\,B - G_W)\Big\|_F,$$

where η is a learning rate factor, and $\|\cdot\|_F$ denotes the Frobenius norm. This strategy ensures that ΔW is well aligned with the principal gradient directions, thereby accelerating convergence and supporting rapid, stable adaptation. Such efficiency is especially critical in highly dynamic ABA contexts [47].

Compared to LoRA and BOFT, LoRA-GA offers greater efficiency and faster convergence by aligning low-rank updates with the gradient's principal directions, allowing more informative updates with fewer parameters. While BOFT provides enhanced stability through orthogonal transformations, it incurs higher computational cost. LoRA-GA strikes a balanced trade-off between expressiveness, efficiency, and overhead, making it well-suited for real-time, resource-constrained ABA task generation.

IV. WEB SYSTEM INTEGRATION

We developed the ABA-RAG Framework into a web-based platform (Figure 4), which seamlessly integrating IoT-driven emotional and behavioral classification with state-of-the-art LLMs and advanced fine-tuning methods. This platform enables ABA professionals to generate personalized tasks, review prior interventions, and refine learner-specific ABA programs efficiently. The platform consists of three primary interface components:

1) **Task Generation Interface:** As shown in Figure 4, users select a student and domain, then click "Generate." The system retrieves relevant tasks, integrates IoT-derived emotional states, and generates personalized suggestions via the fine-tuned LLM—enabling rapid, tailored intervention.
2) **Therapy Overview and Editing:** Figure 5 displays a dynamic therapy dashboard listing generated tasks per learner. Users can edit task details, adjust reinforcement

strategies, and track completion, supporting real-time intervention management.

3) **Performance Analytics:** Figures 6 and 7 present visual analytics on task outcomes, session duration, and emotional indicators. Practitioners can quickly assess progress and target areas needing attention.

Each component realizes the ABA-RAG framework, and ensure the alignment across task generation, sensor insights, and ABA practice.

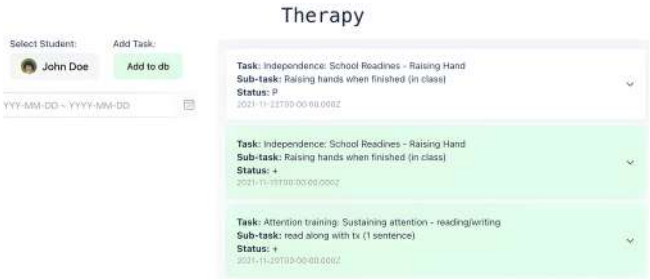

FIGURE 5: Therapy management interface allowing ABA professionals to review, edit, and manage generated tasks directly, facilitating responsive adjustments to learner interventions.

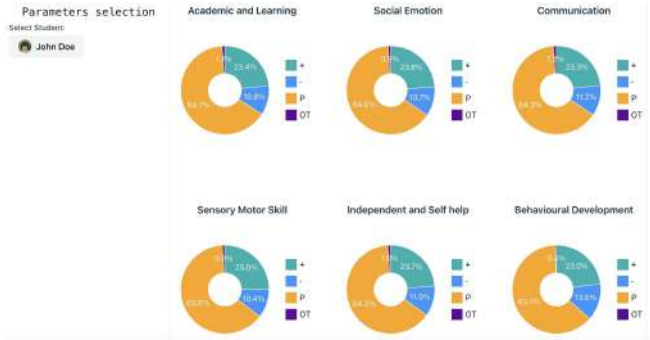

FIGURE 6: Performance analytics interface providing detailed summaries of learner outcomes and sensor-based emotional states, enabling practitioners to track progress and adjust intervention strategies accordingly.

V. EXPERIMENTS

We implemented a four-stage experimental framework to validate our system for personalized ABA task delivery.

- First, we assessed the accuracy of a Transformer-based classifier in inferring emotional and behavioral states from IoT sensor streams, establishing the reliability of the IoT component for generating context-relevant labels (Experiment 1; see Section V-A).
- Second, we introduced the ABA-RAG framework for context-sensitive task generation, comparing it against baseline configurations (No RAG, Average RAG) to evaluate the performance gains achieved through retrieval augmentation (Experiment 2; see Section V-B).

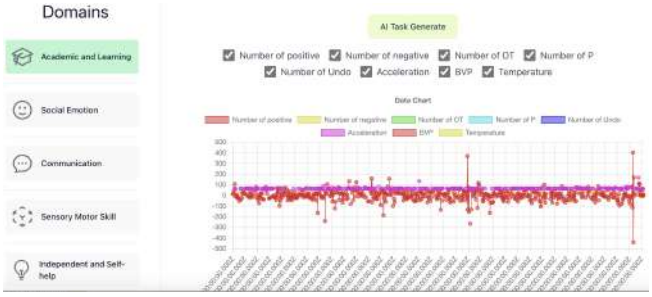

FIGURE 7: Historical data visualization page displaying longitudinal records of learner performance and IoT-derived emotional states, supporting informed, data-driven intervention planning.

- Third, we enhanced the best-performing RAG configuration using low-rank adaptation techniques to improve generation quality in resource-constrained settings (Experiment 3; see Section V-C).
- Finally, we conducted a real-world evaluation of the best RAG-LLM outputs with actual learners and ABA practitioners to assess practical effectiveness and usability (Experiment 3; see Section V-D).

Collectively, these experiments demonstrate the feasibility and impact of integrating IoT-driven classification, semantic retrieval, and fine-tuned LLMs into a unified, adaptive system for delivering timely and personalized ABA interventions.

A. EXPERIMENT 1: IOT CLASSIFICATION FOR EMOTIONAL STATES

This experiment evaluates the reliability of a deep learning model designed to classify a learner's behavioral outcome into four categories: `+`, `−`, `P`, and `OT`. The full dataset consists of around 2,400 minutes of sensor data collected from 33 students. For clarity, the results in this section focus on the dataset from a single student, containing 20,000 labeled samples. Each sample includes physiological signals such as BVP, GSR, wrist temperature, and tri-axis acceleration. Table 3 presents the distribution of the four ABA outcome classes in our dataset. The "Prompt Required" (P) and "Positive" (+) categories are most frequent, while "Off-Task" (OT) is underrepresented.

TABLE 3: Class-wise Distribution of ABA Outcome Labels

Class	Count	Percentage (%)
Positive (+)	6,100	30.5
Negative (−)	4,900	24.5
Prompt (P)	7,200	36.0
Off-Task (OT)	1,800	9.0
Total	20,000	100

Because the class distribution reflects real-world ABA session frequencies and practitioner coding, we did not apply class balancing. Preserving this natural imbalance allows the model to learn from authentic patterns and supports deployment where certain outcomes are inherently rare. Preprocess-

ing included student-based partitioning, outlier removal, and feature scaling, followed by an 80/20 train-test split.

The classification model adopts a transformer-inspired architecture with fully connected layers, ReLU activations, and dropout regularization. It is trained using cross-entropy loss optimized with AdamW. To prioritize recall for the + (Positive) and P (Prompt) classes—critical for adapting task difficulty and reinforcement—threshold-based post-processing is applied. Based on ABA expert input, if the predicted probability for + or P exceeds a set threshold, the model directly assigns that label to avoid missing signs of engagement or partial compliance.

Tables 4 and 5 show that increasing the threshold from 0.05 to 0.15 improves recall while maintaining acceptable precision. ABA professionals confirm that maximizing recall for these categories is key to capturing teachable moments and delivering timely support.

Approach	Precision(+)	Recall(+)
Baseline (Argmax)	0.74	0.64
Threshold (0.05)	0.76	0.82
Threshold (0.10)	0.78	0.87
Threshold (0.15)	0.80	0.90

TABLE 4: Classification results for the "+" category before and after threshold tuning.

Approach	Precision(P)	Recall(P)
Baseline (Argmax)	0.68	0.60
Threshold (0.05)	0.70	0.74
Threshold (0.10)	0.73	0.80
Threshold (0.15)	0.75	0.86

TABLE 5: Classification metrics for the "P" category under different threshold values.

The model's ability to handle − and OT is also examined. Table 6 reports final precision, recall, and F1-scores for all four categories when threshold tuning is enabled for + and P. Table 7 then details the confusion matrix for 20,000 test samples. Our result shows that P occasionally overlaps with − or OT, though recall for + and P remains high. Despite lower precision on − and OT, these classes have lower support, indicating fewer occurrences.

Class	Precision	Recall	F1-Score	Support
+	0.79	0.90	0.84	6,100
−	0.48	0.45	0.46	4,900
P	0.74	0.85	0.79	7,200
OT	0.42	0.38	0.40	1,800
Accuracy = 0.73, Macro-F1 = 0.62, Weighted-F1 = 0.68				

TABLE 6: Final four-class metrics with threshold-based tuning for "+" and "P".

The lower F1-scores for the − (0.46) and OT (0.40) classes are due to their limited representation and semantic overlap that introduces classification ambiguity. However, we deliberately avoided rebalancing techniques to preserve the real-world class distribution, ensuring that evaluation reflects practical deployment conditions despite lower performance on minority classes.

Actual	Predicted				2*Total
	+	−	P	OT	
+	5500	200	340	60	6100
−	330	2200	1240	1130	4900
P	800	700	6100	600	7200
OT	200	500	360	740	1800
Total	6830	3600	8040	2530	21,000*

*Slight discrepancy due to rounding or data withholding.

TABLE 7: Confusion matrix (threshold-based) across four categories. Rows are actual labels, columns are predictions.

B. EXPERIMENT 2: RAG EVALUATION AND ABLATION COMPARISON

This experiment compares three RAG configurations across four leading LLMs, namely Llama-3.1-70B, Claude-3.5-Sonnet, GPT-4o-Latest, and Gemini-1.5-Pro, to identify the most effective setup for generating high-quality ABA tasks. The RAG variants offer different levels of contextual support. These widely used, state-of-the-art LLMs were chosen for their accessibility, relevance, and prevalence in benchmarking literature, ensuring both rigorous comparison and real-world applicability.

1) Candidate Configurations

No RAG: The LLM is given only a minimal prompt comprising an ABA definition and the six primary domains:

- Academic and Learning
- Social Emotion
- Communication
- Sensory Motor Skill
- Independent and Self-help
- Behavioral Development

Under this baseline, the LLM receives no external references or historical data.

Average RAG: The prompt includes two illustrative tasks from a cleaned ABA dataset, each aligned with the user-specified domain and follows a standardized structure:

- *Domain*: Same as domains in No RAG.
- *Task*: Specific intervention goal (e.g., turn-taking).
- *Sub-task*: A more granular activity refinement.
- *Description*: The intended therapy objective.
- *Materials*: Required resources for the session.
- *Procedure*: Step-by-step instructions for implementing the task.
- *Data Collection*: Methods to record behavioral outcomes.
- *Variations*: Possible task modifications or expansions.
- *Reinforcement*: Recommended incentives or feedback approaches.
- *Skills Developed*: Targeted behavioral skills.

Two ABA task examples serve as partial guidance, allowing the LLM to reference prior designs and domain knowledge more effectively than the No RAG setup. This variant excludes IoT emotional inputs and automated retrieval, relying solely on exemplars to ground ABA outputs.

Best RAG (ABA-RAG Framework) integrates three distinct information layers to maximize relevance:

1) *IoT Emotional Predictions*: Real-time probabilities for the learner's emotional or behavioral state (`+`, `−`, `P`, and `OT`) derived from the deep learning classifier.
2) *Automated Semantic Retrieval*: The system automatically retrieves multiple relevant tasks from the preprocessed ABA dataset, guided by user keywords and emotional cues from IoT data.
3) *ABA Domain Knowledge*: A more comprehensive definition of ABA principles, encompassing the core six domains, curated professional guidelines, and literature-based references.

These layers are combined into a structured prompt, giving the LLM detailed context that includes both sensor-derived emotional states and curated domain-specific tasks.

2) Evaluation Metrics and Results:

Each LLM is tested under No RAG, Average RAG, and Best RAG across six ABA domains. The focus is on:

- **Text Length**: The average number of tokens in generated outputs.
- **Response Time**: The total time required to generate each response.
- **Human Evaluation**: Human evaluators assign 10-point scores based on each generated task's clarity, adherence to ABA principles, and suitability for therapeutic use.

For human evaluation, two experts (a certified ABA therapist and an educational psychologist), were independently recruited to rate the LLM-generated tasks. Both experts were selected based on their professional certification and extensive experience in ABA practice, ensuring the reliability and relevance of the evaluation. They independently rated each task based on clarity, clinical appropriateness, and therapeutic suitability, following standardized evaluation guidelines adapted from established ABA assessment frameworks. Inter-rater reliability was high, ICC(2,1) = 0.76, F(59, 59) = 7.38, $p < .001$, 95% CI [0.66, 0.84].

We summarize the evaluation results for each LLM across the three RAG configurations. Figures 8 and 9 report the text length (in tokens) and response time (in seconds) across six ABA domains, while Figure 10 present human ratings from the ABA therapy expert.

C. EXPERIMENT 3: FINE-TUNING UNDER ABA-RAG FRAMEWORK

1) Dataset and Hardware

To further optimize the best-performing RAG model (Llama-3.1-70B), we conducted specialized fine-tuning aimed at aligning the model with complex ABA therapeutic principles while minimizing training costs. The fine-tuning dataset comprised over 10,000 lines from the first 20 chapters of *Applied Behavior Analysis: Pearson New International Edition* by Cooper et al. [1], supplemented with ABA-specific medical terminology and definitions. All data were preprocessed (normalization, tokenization, formatting) for model compatibility. Experiments were performed on a dual NVIDIA RTX 3090 GPU system (24 GB VRAM each), with an AMD Ryzen 9 5950X CPU and 128 GB RAM, enabling efficient parallel processing and stable fine-tuning.

2) Results

We evaluated the three low-rank adaptation methods introduced in Section III E. Table 8 compares these methods across BLEU score, average output length, response time, and training time (averaged over six trials). Table 9 presents expert ratings of ABA task generation quality under three RAG configurations (No RAG, Average RAG, Best RAG) and the three fine-tuning methods. Experts assessed relevance, clarity, and adherence to ABA principles on a 10-point scale, with higher scores indicating better alignment with ABA practice standards.

D. EXPERIMENT 4: COMPARISON OF RAG-ABA AND TRADITIONAL HUMAN-EXPERT APPROACHES

1) Experimental Design

To evaluate the effectiveness of our RAG-ABA system relative to the traditional human-expert task generation approach, we conducted a controlled within-subject experiment involving $N = 10$ participants (5 females, 5 males) aged 6 to 15 years. Each participant completed parallel ABA training sessions using both the RAG-ABA system and the conventional practitioner-led approach, allowing direct comparison of performance within the same individuals. Performance scores were recorded for each student under both conditions.

2) Statistical Analysis

We used the paired t-test to assess whether there was a statistically significant difference in scores between the RAG-ABA and Traditional approaches. To ensure robustness against non-normality due to the small sample size, we also performed the Wilcoxon signed-rank test.

3) Results

The mean score for the RAG-ABA condition was 48.0 (SD = 19.89), while the mean score for the Traditional condition was 33.9 (SD = 23.74). Statistical analysis using the paired t-test ($p = 0.156$) and the Wilcoxon signed-rank test ($p = 0.212$) indicated no within-subject significant difference between the two methods, demonstrating that the RAG-ABA system achieves comparable effectiveness to the traditional human-expert approach in ABA training.

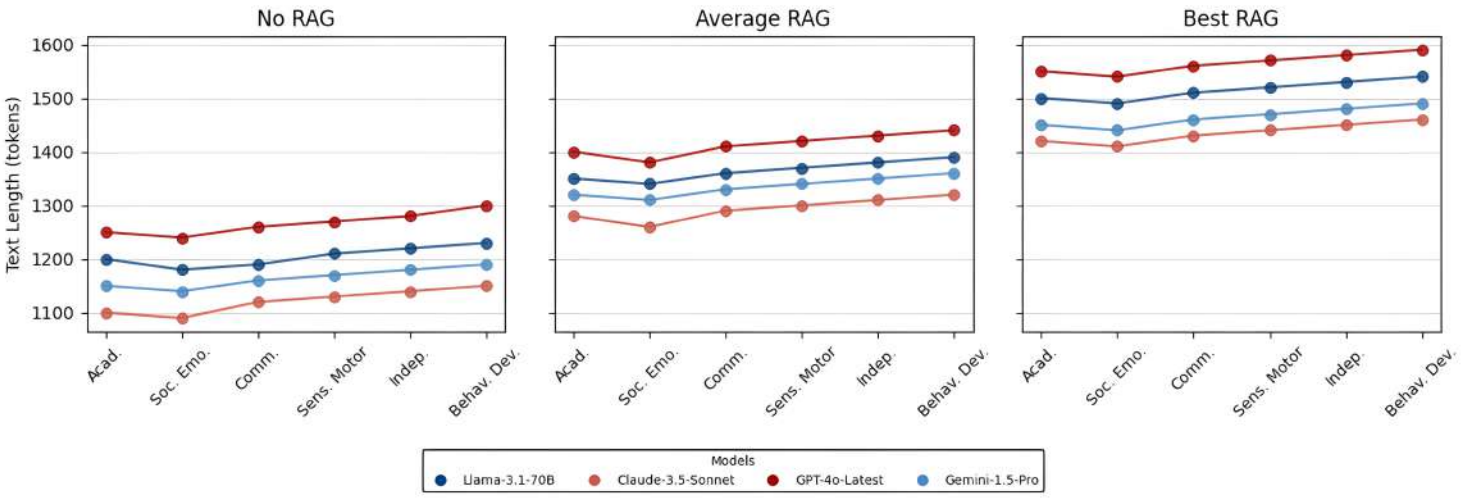

FIGURE 8: Text Length for four LLMs across different task domains and RAG configurations. (Task Abbreviations: Acad. refers to Academic; Soc. Emo. refers to Social Emotion; Comm. refers to Communication; Sens. Motor refers to Sensory Motor; Indep. refers to Independent; Behav. Dev. refers to Behavioral Development).

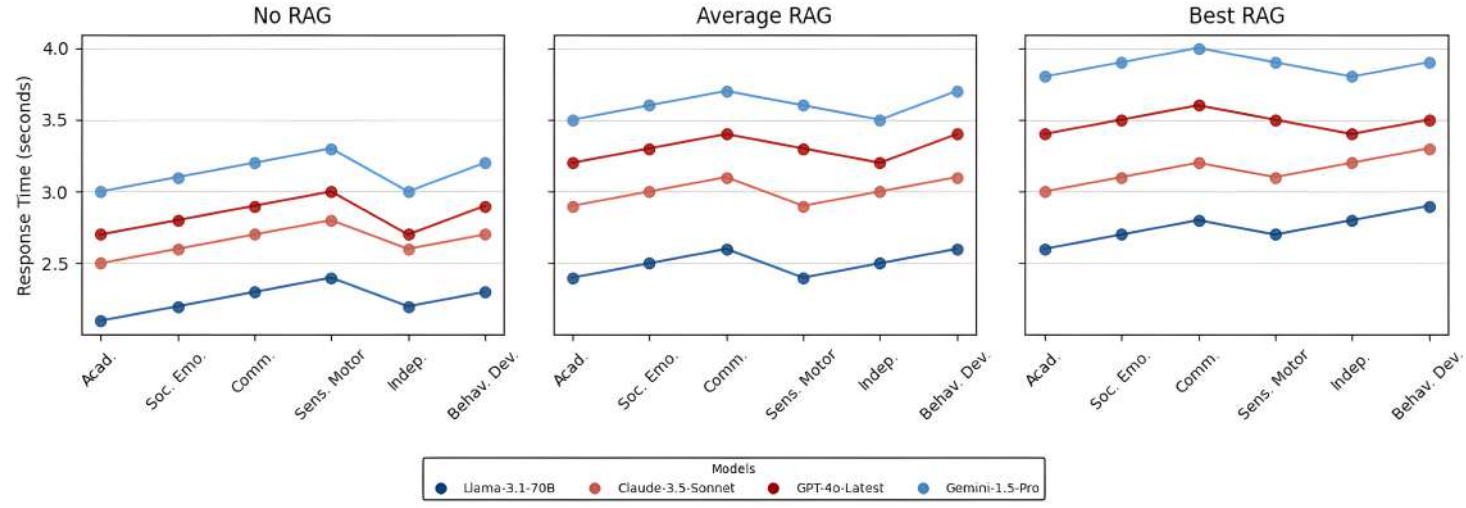

FIGURE 9: Response Time for four LLMs across different task domains and RAG configurations. (Task Abbreviations: Acad. refers to Academic; Soc. Emo. refers to Social Emotion; Comm. refers to Communication; Sens. Motor refers to Sensory Motor; Indep. refers to Independent; Behav. Dev. refers to Behavioral Development).

4) Interpretation

Both statistical tests indicate that the RAG-ABA system achieves comparable effectiveness to the traditional human-expert approach in ABA training. These results suggest that our automated framework can deliver ABA training outcomes comparable to those achieved by experienced practitioners, supporting its potential for scalable and efficient intervention delivery.

VI. ANALYSIS AND DISCUSSION

A. ENHANCING RAG PIPELINE ROBUSTNESS THROUGH IOT CLASSIFICATION ACCURACY

In therapeutic contexts, missing opportunities for reinforcement (+) or prompting (P) is more detrimental than generating occasional false positives. Feedback from ABA professionals highlights the importance of maximizing recall in these two categories to better support behavior acquisition and maintenance. Prioritizing recall for "+" and "P" aligns with the practical needs of real-world ABA interventions, where the observed trade-off between precision and recall is acceptable. This supports the adoption of threshold tuning in

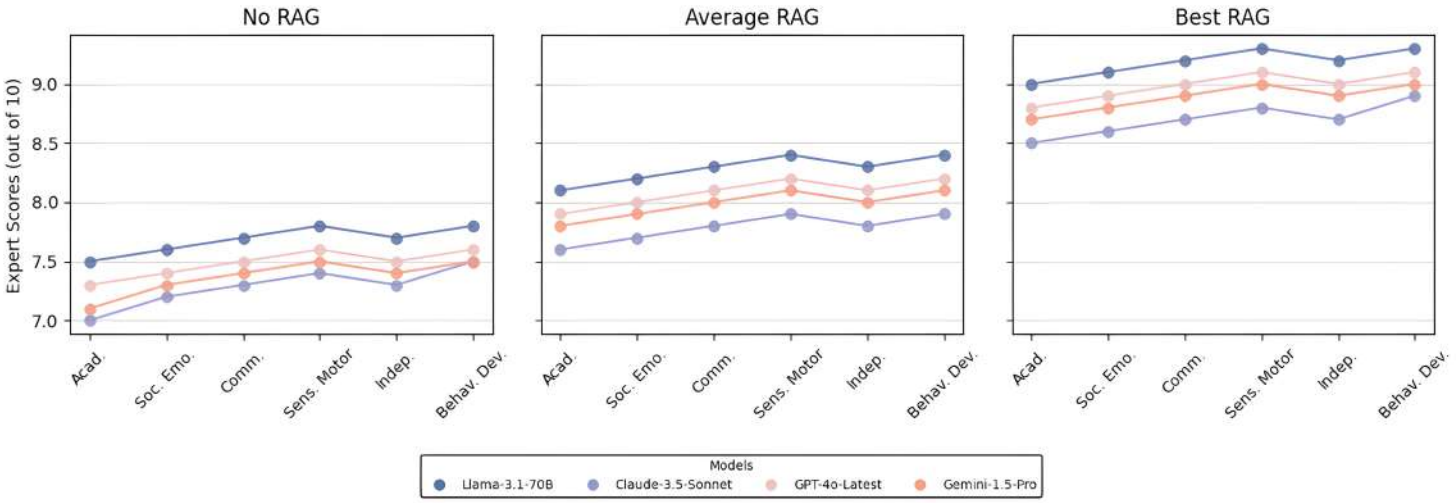

FIGURE 10: Expert Evaluation Scores for four LLMs across different task domains and RAG configurations. (Task Abbreviations: Acad. refers to Academic; Soc. Emo. refers to Social Emotion; Comm. refers to Communication; Sens. Motor refers to Sensory Motor; Indep. refers to Independent; Behav. Dev. refers to Behavioral Development).

Technique	BLEU Score	Avg. Length (tokens)	Response Time (ms)	Training Time (hours)	Trainable Parameters (%)
LoRA (Trial 1)	78	125	520	4.0	0.2%
LoRA (Trial 2)	79	128	515	4.1	0.2%
LoRA (Trial 3)	77	124	522	4.0	0.2%
LoRA (Trial 4)	78	126	518	4.0	0.2%
LoRA (Trial 5)	76	123	525	4.2	0.2%
LoRA (Trial 6)	79	127	519	4.1	0.2%
BOFT (Trial 1)	81	130	535	5.5	0.5%
BOFT (Trial 2)	82	132	540	5.6	0.5%
BOFT (Trial 3)	80	129	538	5.5	0.5%
BOFT (Trial 4)	81	131	536	5.6	0.5%
BOFT (Trial 5)	82	133	542	5.5	0.5%
BOFT (Trial 6)	81	130	537	5.6	0.5%
LoRA-GA (Trial 1)	85	140	500	4.8	0.2%
LoRA-GA (Trial 2)	86	142	495	4.9	0.2%
LoRA-GA (Trial 3)	84	138	502	4.8	0.2%
LoRA-GA (Trial 4)	85	141	498	4.8	0.2%
LoRA-GA (Trial 5)	86	143	496	4.9	0.2%
LoRA-GA (Trial 6)	85	139	501	4.8	0.2%

TABLE 8: Objective metrics and parameter efficiency for fine-tuning techniques.

Configuration	Technique	Trial 1	Trial 2	Trial 3
No RAG	LoRA	6.8	7.0	6.5
Average RAG	LoRA	7.8	7.9	7.6
Best RAG	LoRA	8.5	8.7	8.6
Best RAG	BOFT	8.9	9.0	8.8
Best RAG	LoRA-GA	9.5	9.8	9.6

TABLE 9: Expert evaluation scores for ABA task generation quality under different RAG configurations and fine-tuning techniques.

operational deployments to optimize task generation.

Achieving robust recall for the "+" and "P" categories significantly enhances the RAG pipeline's adaptive capacity, improving task retrieval and generation, particularly when integrating domain-specific knowledge and real-time sensor feedback. Misclassifications in "−" and "OT" categories have a smaller impact on system performance, and future dataset expansion and rebalancing are expected to further improve model accuracy across all categories.

B. LLMS PERFORMANCE AND RAG ABLATION

During the RAG ablation experiments, all four candidate LLMs were accessed uniformly via their official APIs to ensure consistency and fairness in evaluation. This API-based setup was adopted solely for benchmarking different RAG configurations and LLM performances. Training time, trainable parameters, and hardware configuration are thus not

applicable at this stage. These metrics were only assessed during the fine-tuning experiments, where local training was performed exclusively on Llama-3.1-70B.

In terms of text length and response time, GPT-4o-Latest consistently produced the longest outputs, averaging 1,580 tokens under the Best RAG configuration, followed closely by Llama-3.1-70B at 1,550 tokens. Claude-3.5-Sonnet, on the other hand, generated the shortest outputs, averaging 1,300 tokens across all configurations. Regarding response time, Llama-3.1-70B demonstrated the highest efficiency, averaging 2.6 seconds under the No RAG setting and 3.4 to 3.6 seconds under the Best RAG configuration. In contrast, GPT-4o-Latest and Gemini-1.5-Pro had slower response times, often exceeding 3.8 seconds. These results reveal a trade-off: while GPT-4o-Latest excels in generating highly detailed content, Llama-3.1-70B strikes a better balance between efficiency and quality.

Human evaluation shows Llama-3.1-70B achieved the highest score (9.3), surpassing GPT-4o (8.8) and Claude-3.5 (8.7), indicating superior alignment with expert expectations. Its score improved from 8.5 to 9.3 under the Best RAG framework, demonstrating the benefit of integrating IoT-driven emotional cues, semantic retrieval, and ABA knowledge. These results support Llama-3.1-70B as the optimal model for web-based, high-quality ABA text generation.

C. EVALUATION OF FINE-TUNING TECHNIQUES AND RAG CONFIGURATIONS

Results from our comparison of LoRA, BOFT, and LoRA-GA across key metrics (Table 8) indicate that LoRA-GA is best suited for ABA-RAG tasks. It achieves the highest BLEU score (85.5 vs. 81.2 for BOFT and 77.8 for LoRA), reflecting stronger alignment with reference tasks. LoRA-GA also produces the longest outputs (140.5 tokens), offering richer content beneficial for ABA. It delivers the fastest response time (498 ms) and maintains a minimal parameter footprint (0.2%, same as LoRA, compared to 0.5% for BOFT). Although its training time (4.83 hours) is slightly longer than LoRA's (4.07 hours), LoRA-GA strikes a better balance between quality and efficiency than BOFT (5.55 hours), making it ideal for resource-constrained deployment.

Results in Table 9 show that Best RAG consistently yields the highest scores across all fine-tuning methods. LoRA improves from 6.8 (No RAG) to 7.8 (Average RAG) and 8.6 (Best RAG), while BOFT and LoRA-GA achieve 9.0 and 9.6, respectively, under Best RAG. These findings underscore the importance of retrieval—particularly expert-curated prompts—in enhancing contextual relevance, clarity, and alignment with ABA principles.

All three fine-tuning methods show stable performance across runs. LoRA-GA consistently achieves BLEU scores between 84 - 86 and expert ratings of 9.5 - 9.8. While BOFT and LoRA are also reliable, LoRA-GA stands out for combining high-quality output with faster response times, making it the most suitable for ABA-RAG tasks.

D. ON DATA PRIVACY AND ETHICAL CONSIDERATIONS IN AUTOMATED ABA INTERVENTIONS

Ensuring robust data privacy and addressing the ethical implications of automating therapeutic decisions are central to the responsible deployment of AI-driven systems in healthcare and education [48]. In our work, all collected data were anonymized and securely stored, with access strictly limited to authorized research personnel. No personally identifiable information was used in model training or evaluation, and all results were reported in aggregate to further mitigate re-identification risk.

Beyond technical safeguards, the automation of therapeutic recommendations using LLMs may introduce important ethical considerations [49]. While our results demonstrate that the RAG-ABA system achieves comparable effectiveness to traditional human-expert approaches, it is essential to recognize that algorithmic decisions must remain transparent, interpretable, and subject to practitioner oversight. Automated systems risk amplifying biases present in training data and may lack the nuanced judgment of experienced clinicians, particularly in complex or ambiguous cases [50]. To address these concerns, our framework is designed to augment, not replace, human expertise. We adopt HITL principles [34] where practitioners always review, approve, or modify AI-generated prompts before delivery, ensuring that final intervention decisions are grounded in professional judgment and individualized to each learner.

E. LIMITATIONS

Despite the promising results of the proposed ABA-RAG framework, several limitations remain.

First, the retrieval pipeline introduces a latency of approximately 300–400 milliseconds due to embedding computations and similarity searches. While acceptable in current settings, this delay may impact scalability in real-time deployments. Future work will explore optimizations such as approximate nearest neighbor indexing and caching mechanisms to reduce retrieval overhead.

Second, the SEN Multimodal Dataset suffers from class imbalance, particularly for underrepresented behavioral states such as "–" and "OT". This imbalance may hinder the system's ability to generate accurate tasks for less frequent states. Preliminary oversampling methods yielded limited improvements at the current dataset scale. Future efforts will focus on dataset expansion and advanced augmentation techniques, including synthetic data generation, to enhance robustness across all behavioral categories.

A further limitation is our reliance on BLEU scores as the primary quantitative metric for evaluating generated ABA tasks. While BLEU provides a measure of lexical similarity, it does not fully capture clinical, contextual, or therapeutic appropriateness. Future work should incorporate additional evaluation metrics and expert qualitative assessments to better reflect the practical and semantic relevance of generated interventions.

VII. CONCLUSION

We introduced **ABA-RAG**, a retrieval-augmented generation framework that integrates real-time IoT data, semantic retrieval, and efficient low-rank fine-tuning to personalize ABA tasks for learners with special educational needs. The best-performing setup—LoRA-GA fine-tuning with a comprehensive RAG prompt—achieved an expert rating of 9.63/10, highlighting strong clinical relevance. The system also demonstrated robust classification performance (F1 = 0.84, recall = 0.90 for the "+" class) and produced tasks comparable in quality to those created by human experts.

The deployed web platform reduces practitioner workload by dynamically adapting interventions based on learners' physiological and emotional states. LoRA-GA consistently outperformed other tuning methods, offering high responsiveness with low computational cost, which is ideal for real-world ABA environments. Among the tested models, Llama-3.1-70B provided the best balance of quality and efficiency.

In summary, ABA-RAG offers a scalable and resource-efficient solution for personalized ABA delivery. Above all, ethical considerations remain paramount. To this end, ongoing monitoring for bias, transparent model documentation, and regular audits are essential for building trust. We strongly advocate for stakeholder involvement—including families, practitioners, and ethics boards—to co-develop guidelines that ensure the equitable and privacy-preserving use of LLMs in educational and clinical settings.

VIII. ACKNOWLEDGMENT

Rosanna Yuen-Yan Chan is a Principal Investigator of the Centre for Perceptual and Interactive Intelligence (CPII) under the InnoHK. This work was partially supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK. The authors would like to acknowledge Chris Wong for his support in the RAG framework evaluation.

REFERENCES

[1] J. O. Cooper, T. E. Heron, and W. L. Heward, Applied behavior analysis. Pearson UK, 2020.

[2] H. S. Roane, W. W. Fisher, and J. E. Carr, "Applied behavior analysis as treatment for autism spectrum disorder," The Journal of pediatrics, vol. 175, pp. 27–32, 2016.

[3] T. Eckes, U. Buhlmann, H.-D. Holling, and A. M"ollmann, "Comprehensive aba-based interventions in the treatment of children with autism spectrum disorder–a meta-analysis," BMC psychiatry, vol. 23, no. 1, p. 133, 2023.

[4] R. M. Foxx, "Applied behavior analysis treatment of autism: The state of the art," Child and adolescent psychiatric clinics of North America, vol. 17, no. 4, pp. 821–834, 2008.

[5] M. Gitimoghaddam, N. Chichkine, L. McArthur, S. S. Sangha, and V. Symington, "Applied behavior analysis in children and youth with autism spectrum disorders: a scoping review," Perspectives on behavior science, vol. 45, no. 3, pp. 521–557, 2022.

[6] G. Du, Y. Guo, and W. Xu, "The effectiveness of applied behavior analysis program training on enhancing autistic children's emotional-social skills," BMC psychology, vol. 12, no. 1, p. 568, 2024.

[7] S. Eldevik, R. P. Hastings, J. C. Hughes, E. Jahr, S. Eikeseth, and S. Cross, "Meta-analysis of early intensive behavioral intervention for children with autism," Journal of Clinical Child & Adolescent Psychology, vol. 38, no. 3, pp. 439–450, 2009.

[8] C. Wong, S. L. Odom, K. A. Hume, A. W. Cox, A. Fettig, S. Kucharczyk, M. E. Brock, J. B. Plavnick, V. P. Fleury, and T. R. Schultz, "Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review," Journal of autism and developmental disorders, vol. 45, pp. 1951–1966, 2015.

[9] R. Y.-Y. Chan, C. M. V. Wong, and Y. N. Yum, "Predicting behavior change in students with special education needs using multimodal learning analytics," IEEE Access, vol. 11, pp. 63 238–63 251, 2023.

[10] T. Peterson, J. Dodson, and F. J. Strale, "Replicative study of the impacts of applied behavior analysis on target behaviors in individuals with autism using repeated measures," Cureus, vol. 16, no. 3, p. e56226, Mar 2024.

[11] C. Voss, J. Schwartz, J. Daniels et al., "Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: A randomized clinical trial," JAMA Pediatrics, vol. 173, no. 5, pp. 446–454, 2019.

[12] A. H, N. S, and R. BR, "Effectiveness of artificial intelligence–based platform in administering therapies for children with autism spectrum disorder: 12-month observational study," JMIR Neurotech, vol. 4, p. e70589, 2025.

[13] F. J. Alves, E. A. D. Carvalho, J. Aguilar, L. L. D. Brito, and G. S. Bastos, "Applied behavior analysis for the treatment of autism: A systematic review of assistive technologies," IEEE Access, vol. 8, pp. 118 664–118 672, 2020.

[14] A. C. Meneses do Rêgo and I. Araújo-Filho, "Leveraging artificial intelligence to enhance the quality of life for patients with autism spectrum disorder: A comprehensive review," European Journal of Clinical Medicine, vol. 5, no. 5, pp. 28–38, 2024.

[15] N. Perry, C. Sun, M. Munro et al., "Ai technology to support adaptive functioning in neurodevelopmental conditions in everyday environments: A systematic review," npj Digital Medicine, vol. 7, p. 370, 2024.

[16] O. P. Adako, O. C. Adeusi, and P. A. Alaba, "Enhancing education for children with asd: A review of evaluation and measurement in ai tool implementation," Disability and Rehabilitation: Assistive Technology, pp. 1–18, 2025, epub ahead of print.

[17] O. K. Gargari and G. Habibi, "Enhancing medical ai with retrieval-augmented generation: A mini narrative review," Digital Health, vol. 11, pp. 1–7, 2025.

[18] R. Dutt, L. Ericsson, P. Sanchez, S. A. Tsaftaris, and T. Hospedales, "Parameter-efficient fine-tuning for medical image analysis: The missed opportunity," MIDL, 2024.

[19] Y. Lu, X. Zhao, and J. Wang, "ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval," in Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), S. Li, M. Li, M. J. Zhang, E. Choi, M. Geva, P. Hase, and H. Ji, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 64–68.

[20] C. N. Hang, C. W. Tan, and P.-D. Yu, "Mcqgen: A large language model-driven mcq generator for personalized learning," IEEE Access, 2024.

[21] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," Computers and Education: Artificial Intelligence, p. 100417, 2025.

[22] J. Swacha and M. Gracel, "Retrieval-augmented generation (rag) chatbots for education: A survey of applications," Applied Sciences, vol. 15, no. 8, p. 4234, 2025.

[23] A. Kumar, K. Sharma, and A. Sharma, "Memor: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries," Image and Vision Computing, vol. 123, p. 104483, 2022.

[24] C. M. V. Wong, R. Y.-Y. Chan, Y. N. Yum, and K. Wang, "Internet of things (iot)-enhanced applied behavior analysis (aba) for special education needs," Sensors, vol. 21, no. 19, p. 6693, 2021.

[25] F. M. Talaat, "Real-time facial emotion recognition system among children with autism based on deep learning and iot," Neural Computing and Applications, vol. 35, no. 17, pp. 12 717–12 728, 2023.

[26] A. Nandi and F. Xhafa, "A federated learning method for real-time emotion state classification from multi-modal streaming," Methods, vol. 204, pp. 340–347, 2022.

[27] L. Levy, A. Ambaw, E. Ben-Itzchak et al., "A real-time environmental translator for emotion recognition in autism spectrum disorder," Scientific Reports, vol. 14, p. 31527, 2024.

[28] Q. Yang, H. Zuo, R. Su et al., "Dual retrieving and ranking medical large language model with retrieval augmented generation," Scientific Reports, vol. 15, p. 18062, 2025.

[29] G. Bouchouras and K. Kotis, "Integrating artificial intelligence, internet of things, and sensor-based technologies: A systematic review of methodologies in autism spectrum disorder detection," Algorithms, vol. 18, no. 1, p. 34, 2025. [Online]. Available: https://doi.org/10.3390/a18010034
[30] K. Bałazy, M. Banaei, K. Aberer, and J. Tabor, "Lora-xs: Low-rank adaptation with extremely small number of parameters," 2024. [Online]. Available: https://arxiv.org/abs/2405.17604
[31] K. Li, S. Han, Q. Su, W. Li, Z. Cai, and S. Ji, "Uni-lora: One vector is all you need," 2025. [Online]. Available: https://arxiv.org/abs/2506.00799
[32] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal, "On orthogonality and learning recurrent networks with long term dependencies," arXiv preprint arXiv:1702.00071, 2017.
[33] Y. Deng, A. Zhang, N. Wang, S. Gurses, Z. Yang, and P. Yin, "Cloq: Enhancing fine-tuning of quantized llms via calibrated lora initialization," arXiv preprint arXiv:2501.18475, 2025.
[34] S. Kumar, S. Datta, V. Singh, D. Datta, S. K. Singh, and R. Sharma, "Applications, challenges, and future directions of human-in-the-loop learning," IEEE Access, vol. 12, pp. 75 735–75 760, 2024.
[35] J. M. Jackson and M. D. Pinto, "Human near the loop: Implications for artificial intelligence in healthcare," Clinical Nursing Research, vol. 33, no. 2-3, pp. 135–137, mar 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38247381/
[36] M. Elhaddad and S. Hamam, "Ai-driven clinical decision support systems: An ongoing pursuit of potential," Cureus, vol. 16, no. 4, p. e57728, Apr 2024.
[37] A. Holzinger, K. Zatloukal, and H. Müller, "Is human oversight to ai systems still possible?" New Biotechnology, vol. 85, pp. 59–62, 2025.
[38] F. Kabata and D. Thaldar, "Human in the loop requirement and ai healthcare applications in low-resource settings: A narrative review," South African Journal of Bioethics and Law, vol. 17, no. 2, p. e1975, aug 2024.
[39] L. Steven, O. Babajide, and K. Joseph, "Toward a responsible future: Recommendations for ai-enabled clinical decision support," Journal of the American Medical Informatics Association, vol. 31, no. 11, pp. 2730–2739, Nov 2024.
[40] V. Bianchi, M. Bassoli, G. Lombardo, P. Fornacciari, M. Mordonini, and I. De Munari, "Iot wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment," IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8553–8562, 2019.
[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2019, nAACL-HLT 2019.
[42] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of machine learning research, vol. 21, no. 140, pp. 1–67, 2020.
[43] D. D. Olatinwo, A. Abu-Mahfouz, G. Hancke, and H. Myburgh, "Iot-enabled wban and machine learning for speech emotion recognition in patients," Sensors, vol. 23, no. 6, p. 2948, 2023.
[44] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," arXiv preprint arXiv:2104.08691, 2021.
[45] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., "Lora: Low-rank adaptation of large language models." ICLR, vol. 1, no. 2, p. 3, 2022.
[46] S. Wang, L. Yu, and J. Li, "Lora-ga: Low-rank adaptation with gradient approximation," arXiv preprint arXiv:2407.05000, 2024, submitted on 6 Jul 2024, last revised 16 Jul 2024.
[47] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in International conference on machine learning. PMLR, 2019, pp. 2790–2799.
[48] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," BMC Medical Ethics, vol. 22, no. 1, p. 122, 2021. [Online]. Available: https://doi.org/10.1186/s12910-021-00687-3
[49] T. Mirzaei, L. Amini, and P. Esmaeilzadeh, "Clinician voices on ethics of llm integration in healthcare: a thematic analysis of ethical concerns and implications," BMC Medical Informatics and Decision Making, vol. 24, no. 1, p. 250, 2024.
[50] M. Abdelwanis, H. K. Alarafati, M. M. S. Tammam, and M. C. E. Simsekler, "Exploring the risks of automation bias in healthcare artificial intelligence applications: A bowtie analysis," Journal of Safety Science and Resilience, vol. 5, no. 4, pp. 460–469, 2024.

HAOMIN QI received the B.S. degree in Mathematics and Information Engineering from The Chinese University of Hong Kong in 2025. He is currently a senior research assistant with the Advanced Wireless Systems Group at The Chinese University of Hong Kong, focusing on leveraging large language models for network communication and CAD-based design automation. In addition, he serves as the operations director of Intell-Pro Global startup and works as a machine learning engineer. He is now pursuing the M.S. degree in Electrical and Computer Engineering at the University of California San Diego, USA. His research interests include structure-aware retrieval-augmented large language models: methods, systems, and cross-domain applications and robust multimodal (language and vision) reasoning.

SIN CHUNG HO received the B.Eng degree in Information Engineering from The Chinese University of Hong Kong in 2025. He has worked at various Institution and company, and contributed to projects at the the period of duty where he focused on integrating advanced techniques like containerization and Web Integration. With interdisciplinary interests spanning artificial intelligence, edge technique of application development and network systems, Sin Chung Ho is dedicated to advancing the frontiers of technology while fostering innovative practices and inclusive environments in reality.

ROSANNA YUEN-YAN CHAN (M'06 – SM'07 – F'25) received her B.Eng., M.Phil., and Ph.D. degrees in information engineering and M.Ed. degree in education psychology from the Chinese University of Hong Kong. She is a Principal Investigator and Senior Research Scientist at the Centre for Perceptual and Interactive Intelligence (CPII) and an Adjunct Associate Professor at the Department of Information Engineering, CUHK. Rosanna has been a Member-at-Large in the Board of Governors of the IEEE Education Society in 2016 – 2022. She received the 2021 IEEE William E. Sayle II Award for Achievement in Education and is the 2024 – 2025 Distinguished Lecturer of the IEEE Education Society. Her research focuses on augmentative and alternative communication and artificial intelligence for special education needs.

CHUN MAN VICTOR WONG is an industry practitioner and an EdD candidate in special education at the Department of Special Education and Counselling of the Education University of Hong Kong. He founded Bridge Academy in 2014. Victor has also founded Bridge AI, a company that innovates in AI and machine learning for the e-learning of SEN students. He has received several funds from the Hong Kong Innovative and Technology Bureau (ITB) to develop educational systems that tailor to the individual needs of the SEN students.