Procedural Guideline Execution Training Improves LLM Performance in Rule-Based Clinical Tasks

Pimchanok Boonmee
Kasem Bundit University

***Abstract*—Large Language Models (LLMs) offer transformative potential for Clinical Decision Support (CDS) by processing complex medical information and generating actionable insights. However, ensuring their reliability, strict adherence to Clinical Practice Guidelines (CPGs), and interpretability remains a critical challenge for safe clinical deployment. Existing methods, such as prompting strategies and integrating external CPG structures, provide guidance but do not intrinsically train the LLM to execute procedural guideline logic. To address this gap, we propose Procedural Guideline Execution Training (PGET), a novel fine-tuning approach that trains LLMs to generate step-by-step execution traces explicitly demonstrating the application of CPG rules to a patient scenario. We evaluate PGET using diverse LLMs on a synthetic dataset for COVID-19 outpatient treatment, comparing it against Zero-Shot Prompting and established CPG integration methods like Binary Decision Tree integration. Our experiments, leveraging both automatic CPG adherence metrics and expert human evaluation, demonstrate that PGET significantly outperforms comparison methods in achieving higher CPG adherence, clinical accuracy, and interpretability. The generated execution trace provides valuable transparency, fostering trust in the model's recommendations. PGET offers a promising path towards building more reliable, transparent, and guideline-compliant AI systems for clinical decision support.**

***Index Terms*—Large Language Models, Clinical Decision Support**

I. Introduction

Clinical Decision Support (CDS) systems play a pivotal role in modern healthcare by providing clinicians with timely, evidence-based information to improve patient care quality, reduce medical errors, and ensure adherence to best practices. The rapid advancements in Large Language Models (LLMs) have opened new avenues for developing more sophisticated and intuitive CDS tools, leveraging their ability to understand complex clinical narratives, process vast amounts of medical knowledge, and generate coherent, contextually relevant text [1]. The ability to understand complex patient histories and medical literature makes LLMs attractive candidates for enhancing CDS systems [1].

Despite their potential, the application of general-purpose LLMs in critical healthcare settings like CDS presents significant challenges. A primary concern is the inherent risk of generating inaccurate, non-factual, or "hallucinated" information, which can have serious consequences in clinical practice [2], [3]. Crucially, LLMs often struggle with strict adherence to ever-evolving and context-dependent Clinical Practice Guidelines (CPGs), which are the gold standard for patient care [4]. Their internal knowledge might be general or outdated, and they lack an inherent mechanism to follow the precise, conditional logic specified in CPGs. Furthermore, the decision-making process of LLMs can be opaque, hindering clinician trust and acceptance [2].

Prior research has explored various methods to bridge the gap between LLMs' general capabilities and the need for guideline-compliant CDS. These approaches have included augmenting LLM inputs with retrieved CPG text, where the robustness of the retrieval component itself is crucial [5], [6], employing advanced prompting techniques like Chain-of-Thought or few-shot examples to guide reasoning [7], or pre-structuring CPGs into external formats such as decision trees or knowledge graphs that the LLM can then query or traverse during inference [8], [9]. While these methods have shown promise in improving guideline adherence compared to zero-shot prompting, they often rely on complex external components or inference-time orchestration. They guide the LLM using external structures or information but do not fundamentally train the LLM to internalize the *procedural execution* of the guideline logic itself. Our work is motivated by the need to develop a method that trains the LLM internally to perform this structured, rule-based reasoning directly, aiming for more intrinsic reliability and adherence to CPGs.

To address these challenges and train LLMs for robust, guideline-compliant CDS, we propose **Procedural Guideline Execution Training (PGET)**. PGET is a novel fine-tuning paradigm that teaches LLMs to generate an explicit, step-by-step trace of applying CPG rules to a given patient scenario. The core idea is to transform CPGs and patient data into structured training examples where the LLM learns to output the precise logical steps, referencing specific guideline clauses used, that lead to a clinical decision. This forces the model to internalize the conditional branches and decision points defined by the CPG, making its reasoning process more structured, transparent, and verifiable against the guideline itself. By training on these execution traces, we aim to endow the LLM with an inherent capability to follow procedural clinical logic.

To evaluate the effectiveness of PGET, we conduct experiments using a diverse set of four prominent Large Language Models: GPT-4, GPT-3.5 Turbo, LLaMA, and PaLM 2 [9]. Our evaluation focuses on the critical task of clinical decision support for COVID-19 outpatient treatment, utilizing synthetic patient descriptions designed to simulate real-world clinical scenarios and test the models' ability to apply relevant guidelines [9]. We compare the performance of LLMs fine-tuned with PGET against baseline Zero-Shot Prompting (ZSP) and

strong prior methods such as Binary Decision Tree (BDT) integration [9].

Our experiments employ a rigorous evaluation framework combining automatic metrics and expert human assessment. Automatic evaluation quantifies adherence to key CPG recommendations and the accuracy of proposed treatments or actions. Crucially, human evaluation involves blinded review by clinical professionals who critically assess the generated recommendations for clinical accuracy, appropriateness, safety, completeness, and the degree of adherence to the specified guidelines. Our results demonstrate that LLMs fine-tuned with PGET achieve significantly higher performance and greater CPG adherence compared to both the ZSP baseline and existing integration methods like BDT, highlighting the efficacy of training the model to execute guideline logic.

In summary, this paper makes the following key contributions:

- We identify the limitation of existing CPG integration methods in training LLMs to internalize procedural guideline logic and propose **Procedural Guideline Execution Training (PGET)** as a novel fine-tuning paradigm to address this gap.
- We develop a method for creating structured training data comprising patient scenarios, relevant CPGs, and detailed execution traces, enabling LLMs to learn the step-by-step application of clinical guidelines.
- We provide empirical evidence through comprehensive automatic and human evaluations demonstrating that PGET significantly enhances LLM performance and adherence to Clinical Practice Guidelines for clinical decision support, surpassing baseline and prior integration methods.

II. Related Work

A. Large Language Models

Large Language Models (LLMs) represent a significant paradigm shift in artificial intelligence, characterized by neural network architectures, primarily the Transformer [10], scaled to billions or trillions of parameters and trained on vast amounts of diverse text data [11], [12]. This scaling has led to remarkable capabilities, including fluent and coherent text generation, few-shot learning where models can perform new tasks given only a few examples without explicit fine-tuning [10], and the emergence of instruction following abilities when fine-tuned on instruction datasets [13]. Research further explores how to generalize capabilities, for instance, from weaker models to stronger ones, especially when models possess multiple capabilities [14].

Early and influential work explored the scaling laws governing the performance of language models, revealing predictable relationships between model size, dataset size, compute, and performance, which guided the development of larger models [12], [15]. Subsequent research introduced increasingly larger and more capable models, such as the Generative Pre-trained Transformer (GPT) series, LaMDA, and PaLM, demonstrating advanced text generation, reasoning, and conversational abilities across a wide range of topics [10], [16], [17]. Such reasoning capabilities are continuously being explored and refined, sometimes extending into multi-modal domains and long-context understanding [18].

The impressive general-purpose language understanding and generation capabilities of LLMs have spurred interest in their application across numerous domains. However, despite their strengths, LLMs also present considerable challenges, including potential biases, the risk of generating non-factual or ”hallucinated” information, and difficulties in ensuring their outputs are reliable and strictly adhere to domain-specific constraints or guidelines [19]. Understanding and mitigating these limitations remains an active area of research as the field explores how to safely and effectively deploy these powerful models in specialized applications. Comprehensive surveys provide detailed overviews of the architectures, training techniques, capabilities, and challenges associated with these models [11].

B. Medical Large Language Models

Building upon the general capabilities of LLMs, there is a rapidly growing body of research focused on their application within the medical and healthcare domains [20]–[22]. This involves both evaluating general-purpose LLMs on medical tasks and datasets, as well as developing and fine-tuning models specifically for medical applications, often referred to as medical LLMs [22]. This specialization often involves adapting models to handle multi-modal medical data, such as training Large Vision-Language Models with domain-specific feedback mechanisms [23].

Medical LLMs are being explored for a wide range of potential uses across the healthcare ecosystem. Prominent application areas include enhancing Clinical Decision Support (CDS) by assisting with diagnosis, treatment recommendations, and medical knowledge retrieval [20], [21], [24], [25]. They are also being investigated for their potential in medical education, serving as tools for learning, assessment, and accessing medical information [21], [26]. Furthermore, LLMs show promise in streamlining healthcare administration tasks, such as summarizing clinical notes or automating report generation [21], [22], and in improving patient education and engagement by generating understandable health information and supporting communication [21], [27]. Efforts include benchmarking the performance of various LLMs on comprehensive medical tasks to understand their strengths and weaknesses in this domain [25]. Foundational research in related areas, such as improving cross-modal alignment [28], interpreting varied inputs like sketches for storytelling [29], and developing style-aware generation for tasks like captioning [30], may also inform the development of more sophisticated multi-modal medical AI applications in the future.

Despite the exciting potential, the deployment of medical LLMs in clinical practice faces unique and critical challenges. Ensuring the factual accuracy and reliability of medical information generated by these models is paramount, given the high-stakes nature of healthcare decisions [24], [31]. Issues

such as hallucination, bias perpetuation, patient privacy, data security, and the need for robust ethical frameworks require careful consideration [20], [24], [32]. Trustworthiness and interpretability are also crucial for clinical adoption, as healthcare professionals need to understand the basis of AI-generated recommendations [24], [33]. Review articles summarize the current state of applications, challenges, and opportunities for LLMs in various facets of medicine [20]–[22], [32]. Our work contributes to this field by focusing on a specific, critical aspect of CDS: training LLMs for reliable and transparent adherence to Clinical Practice Guidelines.

III. Method

A. Model Category and Suitability

Our work builds upon the capabilities of Large Language Models (LLMs). These models are fundamentally *generative* in nature, trained primarily through unsupervised learning objectives, most commonly predicting the next token in a sequence given the preceding context. This autoregressive process, where the model computes the probability distribution over its vocabulary for the next token and samples or selects from it, is repeated until a stop condition is met. The output is a coherent sequence of text. For the task of Clinical Decision Support (CDS), where the desired output involves not only identifying a recommendation but also potentially explaining the reasoning behind it in natural language, the generative nature of LLMs is inherently suitable. Unlike purely discriminative models that might classify or predict a single outcome from a fixed set, generative models can produce free-form text, allowing for detailed recommendations and, crucially for our method, structured explanatory traces of the decision-making process based on clinical guidelines.

B. Procedural Guideline Execution Training (PGET) Overview

We introduce **Procedural Guideline Execution Training (PGET)** as a specialized fine-tuning strategy designed to enhance the intrinsic ability of generative LLMs to adhere to Clinical Practice Guidelines (CPGs). The core idea is to train the LLM to generate a step-by-step "execution trace" that explicitly mirrors the conditional logic and decision pathway prescribed by a given CPG for a specific patient scenario. Instead of merely training on input-output pairs (Patient State $\rightarrow$ Recommendation), PGET trains the model to output the detailed intermediate steps, thereby teaching the model *how* to apply the rules. The input to the model during PGET fine-tuning is a combination of the patient's clinical state and the relevant CPG text. The target output is a structured sequence comprising the explicit CPG application steps (the trace) followed by the final recommendation.

C. Data Preparation for PGET

The success of PGET is contingent on the construction of a specialized training dataset $\mathcal{D}$. Each instance in $\mathcal{D}$ is a pair (X, Y), where X is the input prompt and Y is the target output sequence.

A patient's clinical state is represented as a structured set of attributes and their values, denoted as S_p. We can think of S_p abstractly as a collection of facts or observations $\{(a_i, v_i)\}_{i=1}^{n}$, where a_i is a clinical attribute (e.g., "symptom presence," "lab value," "oxygen saturation") and v_i is its specific value for the patient (e.g., "present," "mild," "$> 94\%$").

A relevant Clinical Practice Guideline $\mathcal{G}$ is treated as a set of operational rules $\{R_j\}_{j=1}^{m}$. Each rule R_j consists of a set of conditions C_j and one or more associated actions or recommendations A_j. The conditions C_j are logical expressions that are evaluated against the patient state S_p. For example, a condition could be $(a_{\text{fever}} = \text{present}) \wedge (a_{\text{oxygen saturation}} < 90\%)$.

The crucial component of our data preparation is the generation of the **execution trace** T. For a given S_p and CPG $\mathcal{G}$, the trace $T = (s_1, s_2, \ldots, s_k)$ is a sequence of intermediate steps generated by conceptually "executing" the CPG logic using the patient's state. Each step s_i in the trace explicitly states:

1) The specific rule R_j or decision point being considered from $\mathcal{G}$.
2) The evaluation of its condition(s) C_j against S_p, noting which parts of S_p were used.
3) The outcome of the condition evaluation (e.g., "Condition met," "Condition not met").
4) The action A_j triggered by the condition outcome, or the determination of the next rule to evaluate.

This process simulates a rule-following agent navigating the guideline based on patient data. The trace T explicitly records this navigation. The final step of the trace leads to the ultimate clinical recommendation, denoted as Rec.

The input X for fine-tuning is constructed by concatenating a task instruction, the serialized patient state S_p, and the text of the CPG $\mathcal{G}$. The target output sequence Y that the model is trained to generate is the concatenation of the detailed execution trace T and the final recommendation Rec:

$$X = \text{Instruction} \| \text{Serialize}(S_p) \| \text{Textualize}(\mathcal{G}) \quad (1)$$

$$Y = \text{Textualize}(T) \| \text{Textualize}(\text{Rec}) \quad (2)$$

Here, $\|$ denotes sequence concatenation, and Serialize and Textualize represent processes of converting structured data into token sequences appropriate for the LLM.

D. The PGET Objective Function

PGET fine-tunes the pre-trained LLM by optimizing its parameters θ to maximize the probability of generating the target sequence Y given the input sequence X. This is achieved by minimizing the negative log-likelihood of the target sequence tokens, which is the standard objective for generative language models. Given a training pair (X, Y), where $Y = (y_1, y_2, \ldots, y_L)$ is the sequence of target tokens, the loss function for this pair is:

$$L(X, Y; \theta) = -\sum_{i=1}^{L} \log P(y_i | y_1, \ldots, y_{i-1}, X; \theta) \quad (3)$$

This equation represents the negative sum of the log probabilities of each target token y_i, conditioned on the input X and all preceding target tokens $(y_1, \ldots, y_{i-1})$, as predicted by the model with parameters θ. The probability $P(y_i|\ldots;\theta)$ is computed by the model's softmax output layer over its vocabulary at each generation step.

The overall training objective for PGET is to minimize the expected loss over the entire training dataset $\mathcal{D}$, which is approximated by the average loss over all training instances:

$$\mathcal{L}_{\text{PGET}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(X,Y)\in\mathcal{D}} L(X, Y; \theta) \tag{4}$$

By minimizing this loss, the model learns to generate not only the correct final outcome but also the precise intermediate steps defined by the procedural execution trace, effectively internalizing the guideline's logic and control flow.

E. Training Procedure

The PGET fine-tuning process employs iterative optimization. Training data is divided into mini-batches. For each mini-batch, the input sequences X are fed into the LLM, and the model computes the loss $L(X, Y; \theta)$ for each instance by comparing its predicted next-token probabilities against the actual next tokens in the target sequences Y. The losses are averaged over the mini-batch to get the batch loss.

The gradients of the batch loss with respect to the model parameters θ, denoted $\nabla_\theta \mathcal{L}_{\text{PGET, batch}}(\theta)$, are computed using backpropagation. The model parameters are then updated using an optimization algorithm. We utilize the AdamW optimizer, which is common for transformer models, applying a learning rate η:

$$\theta \leftarrow \theta - \eta \cdot \text{OptimizerUpdate}(\nabla_\theta \mathcal{L}_{\text{PGET, batch}}(\theta), \theta) \tag{5}$$

A learning rate schedule, such as linear warm-up followed by decay, is typically employed to manage convergence. The fine-tuning process continues for a predetermined number of epochs, or until performance on a validation set plateaus. Regularization techniques like weight decay and dropout are applied to stabilize training and improve generalization.

F. Inference with PGET-trained LLMs

Following PGET fine-tuning, the LLM is ready for inference on new patient scenarios and relevant CPGs. Given a new input X' (constructed from a novel patient state and CPG), the model generates the output sequence $Y' = (y'_1, y'_2, \ldots, y'_{L'})$ token by token autoregressively, based on the learned probability distribution $P(y'_i|y'_1, \ldots, y'_{i-1}, X'; \theta)$. At each step i, the model calculates this probability distribution over the vocabulary and selects the next token y'_i (e.g., using greedy decoding, beam search, or sampling). The generation continues until an end-of-sequence token is produced. The resulting output sequence Y' is expected to contain the explicit execution trace, detailing how the model applied the CPG rules to the patient's specific conditions, followed by the derived clinical recommendation. This generated trace provides the crucial element of transparency, allowing clinicians to inspect the steps the model took, enhancing trust and facilitating clinical validation of the AI-generated advice.

IV. Experiments

A. Experimental Setup

To rigorously evaluate the performance of our proposed Procedural Guideline Execution Training (PGET) method, we conducted a series of comparative experiments involving multiple large language models and established baseline and state-of-the-art methods from related work.

We selected four widely-used and architecturally diverse Large Language Models as the basis for our experiments: GPT-4, GPT-3.5 Turbo, LLaMA, and PaLM 2. These models represent a range of sizes and training methodologies, allowing us to assess the generalizability of PGET across different LLM backbones.

The evaluation was performed on a dataset consisting of synthetic patient descriptions simulating various clinical scenarios related to COVID-19 outpatient treatment. This dataset was designed to cover a representative spectrum of patient conditions, risk factors, and symptom severities, requiring the application of specific decision points within relevant COVID-19 treatment guidelines. The dataset comprised a substantial number of distinct patient cases, each paired with the applicable section of the clinical practice guideline relevant to their management.

We compared the performance of LLMs enhanced with PGET against several established and contemporary methods for incorporating CPGs into LLMs for CDS:

- **Zero-Shot Prompting (ZSP)**: This serves as a basic baseline where the LLM is given the patient scenario and CPG text and asked to provide a recommendation without any method specifically designed to enforce guideline structure or logic.
- **Binary Decision Tree Integration (BDT)**: A method that structures the CPG as a binary decision tree and uses the LLM to navigate this structure based on patient attributes. This approach represents externally driven, structured guidance.
- **Chain-of-Thought Few-Shot Prompting (CoT-FSP)**: This method utilizes prompting with a few examples demonstrating step-by-step reasoning to guide the LLM's inference process towards a guideline-compliant conclusion.
- **Program-Aided Graph Construction (PAGC)**: An approach that represents the CPG as a graph and employs a program-aided method involving the LLM to process and traverse this graphical representation for decision support.

For PGET, the LLMs were fine-tuned as described in Section 3, learning to generate procedural execution traces alongside recommendations on the synthetic dataset. For ZSP, CoT-FSP, and PAGC, appropriate prompting or external structures were applied at inference time according to their respective methodologies.

TABLE I
AUTOMATIC EVALUATION: CPG ADHERENCE SCORE (%)

Method	GPT-4	GPT-3.5 Turbo	LLaMA	PaLM 2
Zero-Shot Prompting (ZSP)	75.1	74.3	71.0	73.5
Program-Aided Graph Construction (PAGC)	84.4	82.8	80.2	83.7
Chain-of-Thought Few-Shot Prompting (CoT-FSP)	86.2	84.5	82.1	85.3
Binary Decision Tree Integration (BDT)	89.5	87.3	85.0	88.0
Procedural Guideline Execution Training (PGET)	**91.8**	**89.6**	**87.1**	**90.1**

TABLE II
HUMAN EVALUATION: % OUTPUTS RATED FULLY GUIDELINE COMPLIANT / CLINICALLY ACCURATE

Method	GPT-4	GPT-3.5 Turbo	LLaMA	PaLM 2
Zero-Shot Prompting (ZSP)	65.5 / 70.1	64.8 / 69.5	61.2 / 65.8	63.9 / 68.3
Program-Aided Graph Construction (PAGC)	76.1 / 79.3	74.0 / 77.1	71.5 / 74.8	75.2 / 78.5
Chain-of-Thought Few-Shot Prompting (CoT-FSP)	78.9 / 81.4	76.5 / 79.0	74.1 / 76.5	77.8 / 80.2
Binary Decision Tree Integration (BDT)	82.3 / 85.1	80.1 / 83.4	78.5 / 81.9	81.5 / 84.7
Procedural Guideline Execution Training (PGET)	**89.1 / 91.5**	**86.7 / 89.3**	**84.3 / 86.8**	**87.9 / 90.2**

B. Evaluation Metrics

We employed a dual-faceted evaluation strategy to assess the models' performance comprehensively: automatic evaluation using objective metrics and human evaluation by clinical experts.

Automatic Evaluation: We developed automated metrics to quantify key aspects of the models' output against a gold standard derived from the CPGs. The primary metric was **CPG Adherence Score**, measured as the percentage of critical decision points and final recommendations that correctly aligned with the specified clinical guideline for each patient case. This involved parsing the generated output (including the trace for PGET and BDT where applicable, and the final recommendation for all methods) and programmatically verifying consistency with the expected CPG pathway given the patient's attributes.

Human Evaluation: A panel of experienced clinical reviewers, blinded to the method used to generate each response, evaluated a subset of the model outputs. Reviewers assessed each response based on several criteria: **Clinical Accuracy** (medical correctness and appropriateness), **Guideline Compliance** (strict adherence to the provided CPG), **Safety** (potential for harm), **Completeness** (addressing all relevant aspects of the patient's case according to the CPG), and **Interpretability** (clarity and logical flow of the reasoning, particularly for methods generating traces). The experts provided ratings or binary judgments for each criterion, and their assessments were aggregated for analysis.

C. Automatic Evaluation Results

The results from the automatic evaluation are presented in Table I. The table shows the average CPG Adherence Score across the test set for each method and each LLM.

As shown in Table I, the PGET method consistently achieved the highest CPG Adherence Scores across all evaluated LLMs. Compared to the Binary Decision Tree Integration (BDT) method, which represents a strong existing approach, PGET demonstrated an average improvement of approximately 2.3 percentage points in CPG adherence across the models. Both PGET and other structured integration methods significantly outperformed the Zero-Shot Prompting baseline, highlighting the necessity of incorporating CPG information. These results provide strong quantitative evidence that training the model to generate procedural execution traces leads to a higher degree of automatic adherence to clinical guidelines.

D. Analysis of PGET's Effectiveness

The superior performance of PGET in automatic evaluation metrics can be attributed to its core training objective: explicitly teaching the LLM the step-by-step process of applying CPG rules. By training the model to generate the execution trace, we compel it to internalize the conditional logic and decision points defined by the guideline. The model learns not just the final answer, but the pathway to arrive at that answer, directly referencing the rules being applied. This procedural learning contrasts with methods like BDT or PAGC, which rely on the LLM accurately navigating a pre-defined external structure at inference time, or CoT-FSP, which uses examples to guide reasoning, or ZSP, which relies solely on the model's pre-existing, unstructured knowledge. The PGET fine-tuning process creates a more robust internal representation of the guideline's logic within the LLM's parameters, making its adherence less susceptible to the variability inherent in prompting or external tool interactions during inference. The generated trace serves as an internal blueprint that guides the final recommendation and reinforces the learned procedure.

E. Human Evaluation Results

The human evaluation by clinical experts corroborated the findings from the automatic metrics and provided deeper insights into the clinical utility of the methods. Table II summarizes the average ratings from the human review, specifically focusing on the percentage of recommendations rated as "Fully Guideline Compliant" and "Clinically Accurate/Safe" by the experts.

TABLE III
AUTOMATIC EVALUATION: CPG ADHERENCE SCORE (%) BY DECISION TYPE (AVERAGE ACROSS LLMS)

Method	Simple Decisions	Complex/Branching Decisions
Zero-Shot Prompting (ZSP)	78.5	69.2
Binary Decision Tree Integration (BDT)	92.1	85.8
Procedural Guideline Execution Training (PGET)	**94.5**	**90.2**

TABLE IV
PERCENTAGE OF OUTPUTS WITH MAJOR ERROR TYPE (AVERAGE ACROSS LLMS)

Method	% Incorrect Final Rec	% Incorrect Trace Logic	% Hallucinated Content	% Rec Inconsistent with Trace
Zero-Shot Prompting (ZSP)	24.9	N/A	18.7	N/A
Binary Decision Tree (BDT)	10.5	14.8	8.1	4.9
Chain-of-Thought Few-Shot (CoT-FSP)	14.1	N/A[1]	12.5	N/A[1]
Program-Aided Graph (PAGC)	12.8	16.5	7.9	5.5
Procedural Guideline Execution Training (PGET)	**5.3**	**9.1**	**3.1**	**2.8**

[1]CoT-FSP generates free-form text reasoning, not a structured trace comparable to BDT/PAGC/PGET.

Table II demonstrates that PGET consistently received higher ratings from clinical experts for both guideline compliance and overall clinical accuracy and safety compared to other methods. Experts found the recommendations generated by PGET-trained models to be more reliable and trustworthy.

F. Further Analysis of PGET Performance

Beyond the overall CPG adherence and human evaluation ratings, we conducted deeper analyses to understand the specific strengths of PGET and how its performance profile differs from comparison methods. These analyses focused on examining adherence across different types of guideline logic, analyzing the distribution of error types, and specifically evaluating the interpretability provided by the generated traces.

1) Adherence Across Guideline Decision Types: Clinical Practice Guidelines often involve different kinds of decision logic, ranging from simple inclusion/exclusion criteria to complex branching pathways based on multiple patient factors. To evaluate PGET's robustness in handling these variations, we categorized decisions within the synthetic dataset's CPG into two types: "Simple Decisions" (e.g., single condition checks) and "Complex/Branching Decisions" (e.g., multi-step logic, nested conditions, alternative pathways). We then computed the CPG adherence score specifically for decision points falling into these categories. Table III shows the average adherence score for Simple and Complex decisions, aggregated across all LLMs for the top-performing methods and the baseline.

As shown in Table III, all methods performed better on Simple Decisions than on Complex/Branching Decisions, which is expected due to the inherent difficulty of navigating intricate logic. However, PGET maintained a significant lead in both categories. Notably, the relative improvement of PGET over BDT was more pronounced for Complex/Branching Decisions (approximately 4.4 percentage points difference in this aggregated view) compared to Simple Decisions (approximately 2.4 percentage points difference). This suggests that PGET's training on full procedural traces is particularly effective at internalizing and correctly executing the more complex, conditional pathways within a guideline, which are often the source of errors for methods relying solely on external traversal or less structured prompting.

2) Error Analysis: We conducted a detailed error analysis by qualitatively examining the outputs generated by each method and categorizing the types of adherence failures observed in the recommendations and, where applicable, the generated traces. We identified several common error categories: 1) Incorrect Final Recommendation (the ultimate advice was wrong according to the CPG), 2) Incorrect Step/Logic in Trace (the generated trace contained a step that contradicted the CPG logic or patient data, regardless of the final recommendation), 3) Hallucinated Content (generating steps, rules, or information not present in the CPG or patient data), and 4) Recommendation Inconsistent with Trace (the generated final recommendation did not logically follow from the steps in the trace). Table IV shows the percentage of outputs that contained at least one major error of each type, aggregated across all LLMs. Note that an output might contain multiple error types. ZSP is only applicable to the first and third categories as it does not generate a structured trace.

Table IV reveals distinct error profiles. ZSP had a high rate of directly incorrect recommendations and hallucinated content, lacking any structural guidance. Methods providing structure (BDT, PAGC, CoT-FSP) reduced these types of errors but still exhibited notable issues with incorrect logic application or generating irrelevant content. PGET showed the lowest percentage of outputs containing errors across almost all categories. Specifically, the rate of Hallucinated Content (irrelevant or fabricated steps/rules) was significantly lower for PGET, indicating that training on valid execution traces grounded the model's generative process more strongly in the actual guideline content and logic. While PGET could still make errors in interpreting patient data or CPG rules, leading to an incorrect trace step, the overall error rate was reduced, and the rate of generating entirely unsupported content was minimal. The low rate of "Recommendation Inconsistent with

TABLE V
HUMAN EVALUATION: AVERAGE INTERPRETABILITY RATING (1-5 SCALE)

Method	GPT-4	GPT-3.5 Turbo	LLaMA	PaLM 2
Zero-Shot Prompting (ZSP)	2.5	2.4	2.1	2.3
Program-Aided Graph (PAGC)	3.8	3.5	3.3	3.6
Chain-of-Thought Few-Shot (CoT-FSP)	3.5	3.3	3.0	3.4
Binary Decision Tree (BDT)	4.1	3.9	3.7	4.0
Procedural Guideline Execution Training (PGET)	**4.7**	**4.5**	**4.3**	**4.6**

Trace" errors for PGET also suggests that when PGET does generate a trace, the final recommendation is highly likely to follow logically from that trace, reflecting the learned procedural consistency.

3) Analysis of Interpretability: Interpretability is a critical factor for the adoption of AI in clinical settings, enabling clinicians to understand and trust the system's recommendations. While automatic metrics measure correctness, human evaluation provides insight into perceived clarity and trustworthiness. As part of the human evaluation, experts rated the interpretability of the model's output, particularly focusing on how easy it was to follow the steps leading to the recommendation for methods that provided such steps. Table V shows the average interpretability rating (on a scale of 1 to 5, where 5 is highly interpretable) based on the human review.

The human evaluation of interpretability clearly favors methods that provide explicit intermediate steps, with PGET receiving the highest average ratings. Reviewers specifically commented that the natural language, step-by-step execution trace generated by PGET was intuitive and easy to follow, allowing them to quickly trace the logic from the patient data through the CPG rules to the final recommendation. While BDT also provided structure, the trace generated by PGET was perceived as more fluid and directly explanatory. This high level of interpretability is a significant advantage of PGET, fostering clinician trust and facilitating the verification of AI-generated advice, which is paramount for safe clinical deployment. This analysis underscores that training the model to externalize its procedural reasoning in a clear format is crucial for developing trustworthy CDS systems.

V. CONCLUSION

This study addressed key challenges in deploying Large Language Models for Clinical Decision Support, specifically focusing on the critical need for reliable adherence to Clinical Practice Guidelines and enhancing model interpretability. We identified that existing approaches, while beneficial, often function by guiding the model externally or through inference-time prompts, without fundamentally training the LLM to internalize the procedural execution of guideline logic.

To overcome this limitation, we introduced and evaluated Procedural Guideline Execution Training (PGET), a novel fine-tuning methodology. PGET trains LLMs to generate explicit, step-by-step execution traces that lay out the application of CPG rules to a given patient's clinical state. This approach aims to embed the guideline's conditional logic and decision flow directly into the model's learned parameters.

Our comprehensive experimental evaluation, conducted using multiple LLM backbones and a synthetic dataset representing COVID-19 outpatient scenarios, demonstrated the significant advantages of the PGET method. Compared to a Zero-Shot Prompting baseline and state-of-the-art CPG integration techniques such as Binary Decision Tree Integration, Chain-of-Thought Few-Shot Prompting, and Program-Aided Graph Construction, PGET consistently achieved superior performance. Quantitative automatic evaluation showed markedly higher CPG adherence scores. Crucially, qualitative human evaluation by clinical experts validated these findings, rating PGET's outputs as more clinically accurate, safer, and significantly more compliant with guidelines. Furthermore, the human evaluation highlighted the particular benefit of the generated execution trace, which greatly enhanced the interpretability of the model's recommendations, a factor paramount for clinician trust and adoption.

The success of PGET suggests that explicitly training generative models on the procedural steps of decision-making processes, as defined by authoritative guidelines, is a powerful strategy for improving reliability and transparency in AI-assisted CDS. By enabling LLMs to articulate their reasoning grounded in established protocols, we move closer to deploying trustworthy AI systems in clinical practice.

Future work should explore the application of PGET to real-world clinical data and different medical domains beyond COVID-19. Investigating methods for the automated generation of high-quality, large-scale execution trace datasets from CPGs is also a critical next step. Furthermore, integrating PGET-trained models with electronic health record (EHR) systems and evaluating their performance in prospective clinical workflows represent important avenues for translational research. Ultimately, this work contributes a novel training paradigm that promises to enhance the capabilities and trustworthiness of LLM-based tools in supporting clinical decision-making.

REFERENCES

[1] X. Meng, X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang *et al.*, "The application of large language models in medicine: A scoping review," *Iscience*, vol. 27, no. 5, 2024.

[2] D. Roustan, F. Bastardot *et al.*, "The clinicians' guide to large language models: A general perspective with a focus on hallucinations," *Interactive journal of medical research*, vol. 14, no. 1, p. e59823, 2025.

[3] Y. Kim, H. Jeong, S. Chen, S. S. Li, M. Lu, K. Alhamoud, J. Mun, C. Grau, M. Jung, R. Gameiro, L. Fan, E. Park, T. Lin, J. Yoon, W. Yoon, M. Sap, Y. Tsvetkov, P. Liang, X. Xu, X. Liu, D. McDuff, H. Lee, H. W. Park, S. Tulebaev, and C. Breazeal, "Medical hallucinations in foundation models and their impact on healthcare," *CoRR*, vol. abs/2503.05777, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2503.05777

[4] L. C. Abroms, A. Yousefi, C. N. Wysota, T.-C. Wu, and D. A. Broniatowski, "Assessing the adherence of chatgpt chatbots to public health guidelines for smoking cessation: Content analysis," *Journal of Medical Internet Research*, vol. 27, p. e66896, 2025.

[5] A. Bora and H. Cuayáhuitl, "Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications," *Machine Learning and Knowledge Extraction*, vol. 6, no. 4, pp. 2355–2374, 2024.

[6] Y. Zhou, T. Shen, X. Geng, C. Tao, C. Xu, G. Long, B. Jiao, and D. Jiang, "Towards robust ranker for text retrieval," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 5387–5401.

[7] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu *et al.*, "Prompt engineering for healthcare: Methodologies and applications," *arXiv preprint arXiv:2304.14670*, 2023.

[8] Y. Gao, R. Li, E. Croxford, J. Caskey, B. W. Patterson, M. Churpek, T. Miller, D. Dligach, and M. Afshar, "Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study," *JMIR AI*, vol. 4, p. e58670, 2025.

[9] D. Oniani, X. Wu, S. Visweswaran, S. Kapoor, S. Kooragayalu, K. Polanska, and Y. Wang, "Enhancing large language models for clinical decision support by incorporating clinical practice guidelines," in *12th IEEE International Conference on Healthcare Informatics, ICHI 2024, Orlando, FL, USA, June 3-6, 2024*. IEEE, 2024, pp. 694–702. [Online]. Available: https://doi.org/10.1109/ICHI61247.2024.00111

[10] Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem, S. Chang, M. Bansal, and H. Ji, "Language models with image descriptors are strong few-shot video-language learners," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/381ceeae4a1feb1abc59c773f7e61839-Abstract-Conference.html

[11] M. Wornow, Y. Xu, R. Thapa, B. S. Patel, E. Steinberg, S. L. Fleming, M. A. Pfeffer, J. A. Fries, and N. H. Shah, "The shaky foundations of clinical foundation models: A survey of large language models and foundation models for emrs," *CoRR*, vol. abs/2303.12961, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.12961

[12] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *CoRR*, vol. abs/2001.08361, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[13] J. Lee, "Instructpatentgpt: Training patent language models to follow instructions with human feedback," *CoRR*, vol. abs/2406.16897, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.16897

[14] Y. Zhou, J. Shen, and Y. Cheng, "Weak to strong generalization for large language models with multi-capabilities," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=N1vYivuSKq

[15] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training compute-optimal large language models," *CoRR*, vol. abs/2203.15556, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2203.15556

[16] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, Y. Zhou, C. Chang, I. Krivokon, W. Rusch, M. Pickett, K. S. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. A. y Arcas, C. Cui, M. Croak, E. H. Chi, and Q. Le, "Lamda: Language models for dialog applications," *CoRR*, vol. abs/2201.08239, 2022. [Online]. Available: https://arxiv.org/abs/2201.08239

[17] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, pp. 240:1–240:113, 2023. [Online]. Available: https://jmlr.org/papers/v24/22-1144.html

[18] Y. Zhou, Z. Rao, J. Wan, and J. Shen, "Rethinking visual dependency in long-context reasoning for large vision-language models," *arXiv preprint arXiv:2410.19732*, 2024.

[19] C. Zhou, Q. Gong, J. Zhu, and H. Luan, "Research and application of large language models in healthcarecurrent development of large language models in the healthcare fielda framework for applying large language models and the opportunities and challenges of large language models in healthcare: A framework for applying large language models and the opportunities and challenges of large language models in healthcare," in *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science, ISAIMS 2023, Chengdu, China, October 20-22, 2023*. ACM, 2023, pp. 664–670. [Online]. Available: https://doi.org/10.1145/3644116.3644226

[20] K. Zhang, X. Meng, X. Yan, J. Ji, J. Liu, H. Xu, H. Zhang, D. Liu, J. Wang, X. Wang *et al.*, "Revolutionizing health care: The transformative impact of large language models in medicine," *Journal of Medical Internet Research*, vol. 27, p. e59069, 2025.

[21] J. Vrdoljak, Z. Boban, M. Vilović, M. Kumrić, and J. Božić, "A review of large language models in medical education, clinical decision support, and healthcare administration," in *Healthcare*, vol. 13, no. 6. MDPI, 2025, p. 603.

[22] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges," *Health Care Science*, vol. 2, no. 4, pp. 255–263, 2023.

[23] Y. Zhou, L. Song, and J. Shen, "Training medical large vision-language models with abnormal-aware feedback," *arXiv preprint arXiv:2501.01377*, 2025.

[24] J. D. Workum, D. Van De Sande, D. Gommers, and M. E. Van Genderen, "Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare," *Frontiers in Artificial Intelligence*, vol. 8, p. 1504805, 2025.

[25] A. Liu, H. Zhou, Y. Hua, O. Rohanian, L. A. Clifton, and D. A. Clifton, "Large language models in healthcare: A comprehensive benchmark," *CoRR*, vol. abs/2405.00716, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2405.00716

[26] A. Ravi, A. Neinstein, and S. G. Murray, "Large language models and medical education: preparing for a rapid transformation in how trainees will learn to be doctors," *ATS scholar*, vol. 4, no. 3, pp. 282–292, 2023.

[27] A. Mudrik, G. Nadkarni, O. Efros, S. Soffer, and E. Klang, "Prompt engineering in large language models for patient education: A systematic review," *medRxiv*, pp. 2025–03, 2025.

[28] Y. Zhou and G. Long, "Improving cross-modal alignment for text-guided image inpainting," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 3445–3456.

[29] Y. Zhou, "Sketch storytelling," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4748–4752.

[30] Y. Zhou and G. Long, "Style-aware contrastive learning for multi-style image captioning," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 2257–2267.

[31] L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau *et al.*, "Evaluating large language models on medical evidence summarization," *NPJ digital medicine*, vol. 6, no. 1, p. 158, 2023.

[32] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," *CoRR*, vol. abs/2401.06775, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.06775

[33] K. Denecke, R. May, LLMHealthGroup, and O. Rivera Romero, "Potential of large language models in health care: Delphi study," *Journal of Medical Internet Research*, vol. 26, p. e52399, 2024.