

Original Paper

Reflection- and Curriculum-Based Instruction Tuning for Diabetes-Specialized Large Language Models: Model Development and Evaluation Study

Jaesung Hwang¹, MS; Deniise Liz Namayanja¹, BS; Donghyeon Park², PhD

¹Department of Software, Sejong University, Seoul, Republic of Korea

²Department of Artificial Intelligence & Data Science, Sejong University, Seoul, Republic of Korea

Corresponding Author:

Donghyeon Park, PhD
Department of Artificial Intelligence & Data Science
Sejong University
Daeyang AI Center 707, 209 Neungdong-ro, Gwangjin-gu
Seoul 05006
Republic of Korea
Phone: 82 10-4860-7491
Email: parkdh@sejong.ac.kr

Abstract

Background: Effective diabetes management requires continuous interpretation of glycemic trends, personalized dietary guidance, and sustained patient education. Although large language models (LLMs) are increasingly being explored for health-related applications, existing general-purpose and biomedical models often struggle with diabetes-specific reasoning and instruction-following, limiting their reliability for domain-focused tasks such as clinical question answering and dietary recommendation tasks.

Objective: This study aimed to develop and evaluate a diabetes-specialized LLM optimized for diabetes-specific reasoning, instruction-following, and dietary recommendation tasks.

Methods: This was a model development and benchmark evaluation study. We developed a model-centric instruction refinement framework using instruction-following difficulty and reversed instruction-following difficulty to identify and replace suboptimal instruction-response pairs during instruction tuning. Curriculum-based instruction tuning was applied by sequencing instructions from lower to higher difficulty. The resulting diabetes-specialized model was evaluated across diabetes-related question answering, natural language inference, information extraction, summarization, clinical answer generation, and dietary recommendation tasks. Performance was compared with biomedical LLMs and general-purpose baselines using diabetes-specific benchmark subsets and simulation-based glycemic evaluation.

Results: Across diabetes-related benchmark tasks, the proposed model demonstrated improved performance in question answering, information extraction, and generative tasks. Ablation experiments showed that the full reflection- and curriculum-based instruction tuning strategy improved overall performance by 59.03% relative to the zero-shot LLaMA3.1 8B baseline. For dietary recommendation tasks, the proposed model achieved a 0.66% improvement in the Diet Quality Index—International score compared with ChatGPT (GPT-4). Simulation-based evaluation using the SimGlucose simulator further showed that meal plans generated by the proposed model resulted in a reduced postprandial glycemic burden, as measured by a lower incremental area under the curve, compared with GPT-4-generated meal plans.

Conclusions: This study demonstrates that domain-specific instruction tuning can effectively adapt general-purpose LLMs for diabetes management. By combining reflection-based instruction replacement with curriculum-based instruction tuning, the proposed approach enhances instruction-following, reasoning capability, and dietary guidance for diabetes. The results highlight the potential of specialized LLMs to provide more reliable and clinically aligned support for diabetes-related decision-making and self-management, offering a promising direction for safe and effective AI deployment in chronic disease care.

JMIR AI 2026;5:e92843; doi: [10.2196/92843](https://doi.org/10.2196/92843)

Keywords: diabetes mellitus; large language models; instruction tuning; instruction-following difficulty; curriculum learning; dietary recommendation; glycemic simulation; AI in health care

Introduction

Background

Diabetes mellitus remains one of the most significant global health challenges, with prevalence rates nearly doubling over the past 3 decades [1]. Managing diabetes effectively requires not only pharmacological treatment but also ongoing interpretation of blood glucose fluctuations, dietary adaptations, personalized lifestyle interventions, and continuous patient education. As the demand for scalable, patient-centered care rises, the health care community increasingly seeks robust AI tools, such as large language models (LLMs), that can support clinical decision-making and self-management.

LLMs have gained significant traction in the biomedical domain due to demonstrated advances in natural language understanding [2], large-scale incorporation of biomedical literature [3], and emerging capabilities in complex clinical reasoning tasks [4]. In the context of diabetes management, LLMs have been leveraged across a diverse set of tasks tailored to patient needs. For instance, integrated systems, such as DeepDR-LLM, combine image-based data with LLM-generated responses [5]. Retrieval-augmented generation architectures have been used to support diabetes education by grounding LLM outputs in trusted reference sources, thereby improving factual reliability [6]. Other efforts have used LLMs to analyze continuous glucose monitoring (CGM) data directly and summarize trends over time for patients and clinicians [7].

Despite these advances, existing LLM-based approaches for diabetes management still show gaps in consistency, clinical alignment, and personalization. General biomedical language models such as BioBERT (biomedical bidirectional encoder representations from transformers) [2] and BioGPT (biomedical generative pretrained transformer) [3] achieve strong performance on broad biomedical benchmarks, but they are not explicitly optimized for diabetes-specific reasoning or patient-facing counseling. Similarly, evaluations of general-purpose chatbots such as ChatGPT have identified deficiencies in diabetes-related responses, including incomplete reasoning and insufficient dietary sensitivity, thereby raising concerns about their reliability for unsupervised diabetes management support [8,9]. Collectively, these shortcomings underscore a critical gap: the absence of LLMs explicitly optimized for diabetes-specific reasoning, instruction-following, and patient-centered guidance, including accurate responses to diabetes-related queries and diabetes-appropriate dietary recommendations.

To address this gap, we propose a diabetes-specialized LLM trained on a curated, diabetes-specific instruction dataset designed to capture the clinical and practical nuances of diabetes management. Importantly, our approach moves beyond simple fine-tuning on domain-specific text. Instead, we adopt a model-centric perspective that treats the LLM as a student whose learning must be aligned with its capacity and

receptiveness. Inspired by the notion of student-guided data selection introduced in selective reflection-tuning [10], we adopt a reflection-based replacement strategy using instruction-following difficulty (IFD) and reversed IFD (r-IFD) metrics to systematically filter and refine instruction-response pairs, ensuring that the training data are both cognitively challenging and instructionally coherent.

In addition, we use a curriculum-based instruction tuning framework [11] to promote stable knowledge acquisition and reduce catastrophic forgetting. By ordering training samples from lower- to higher-difficulty instructions, the model progressively acquires diabetes-specific reasoning skills while preserving foundational biomedical knowledge.

We evaluate our model across a diverse range of diabetes-specific tasks, including question answering (QA), natural language inference (NLI), and personalized dietary recommendation. Across these evaluations, the model consistently outperforms strong biomedical and general-purpose baselines. Notably, in dietary recommendation tasks measured by the Diet Quality Index–International (DQI-I) score [12], our model achieves a 0.66% improvement over ChatGPT (GPT-4), indicating superior nutritional alignment in generated meal plans.

The primary contributions of our work are as follows:

1. We adapt and apply a reflection-based filtering strategy (IFD or r-IFD) to selectively refine a diabetes-specific instruction dataset.
2. We integrate a task complexity-aware curriculum-based instruction tuning strategy that sequences diabetes-related instructions from simpler to more challenging cases, improving training stability and domain adaptation.
3. This work demonstrates strong performance across multiple diabetes-relevant tasks, including QA, NLI, IE, and dietary recommendation, and is supported by ablation experiments that evaluate the contributions of instruction optimization (IFD or r-IFD) and curriculum-based instruction tuning.
4. Finally, we assess the physiological relevance of the model's dietary recommendations using the SimGlucose simulator [13], which uses the UVA/Padova physiological model [14] to predict glucose kinetics. Our results show reduced postprandial glycemic responses compared with those from GPT-4-generated meal plans.

Overall, this study shows that domain-aware instruction tuning, when combined with reflection-guided data refinement and curriculum learning, can effectively adapt general-purpose LLMs into clinically meaningful assistants for diabetes management, narrowing the gap between general linguistic capability and specialized clinical reasoning.

Related Work

LLMs in Biomedical and Clinical Natural Language Processing

LLMs have demonstrated strong performance across a wide range of biomedical natural language processing (NLP) tasks, including medical QA, clinical summarization, and IE. Domain-adapted models such as BioBERT [2] and BioGPT [3] have shown improved performance on biomedical benchmarks by incorporating large-scale biomedical corpora during pretraining. More recent instruction-tuned biomedical models, including MedAlpaca [15], PMC-LLaMA [16], ChatDoctor [17], and BioMistral [18], have further expanded the scope of clinical language understanding by aligning model behavior with natural language instructions.

Despite these advances, most biomedical LLMs are trained to optimize general medical competence rather than task consistency within a specific disease domain. As a result, their performance often varies substantially across task types, particularly when moving from factual QA to generative or reasoning-intensive tasks. This limitation is especially pronounced in chronic disease management, where models must integrate clinical context, lifestyle constraints, and patient-specific considerations.

Instruction Tuning

Instruction tuning [19,20] has emerged as a widely adopted posttraining paradigm for enhancing the ability of LLMs to follow natural language instructions across a wide range of downstream tasks. Early research efforts in this area relied primarily on manually creating task-specific instruction tuning datasets [21,22]. Subsequent work has shifted toward scalable data construction strategies that leverage more powerful LLMs to generate synthetic data [23-27]. These studies have laid the foundation for instruction tuning, demonstrating that large task-specific or multitask instruction tuning datasets can substantially boost the instruction-following behavior and generalization in language models.

LLM-Based Approaches for Diabetes Management

Recent studies have explored the use of LLMs in diabetes-related applications, primarily focusing on patient education, lifestyle, and self-management [28-30]. Prior work has examined conversational agents that provide nutritional guidance, carbohydrate-related explanations, and guideline-informed recommendations to assist daily decision-making in diabetes care [31,32].

Other efforts have focused on using LLMs to translate CGM data into natural language summaries that improve interpretability for patients and clinicians [7,33].

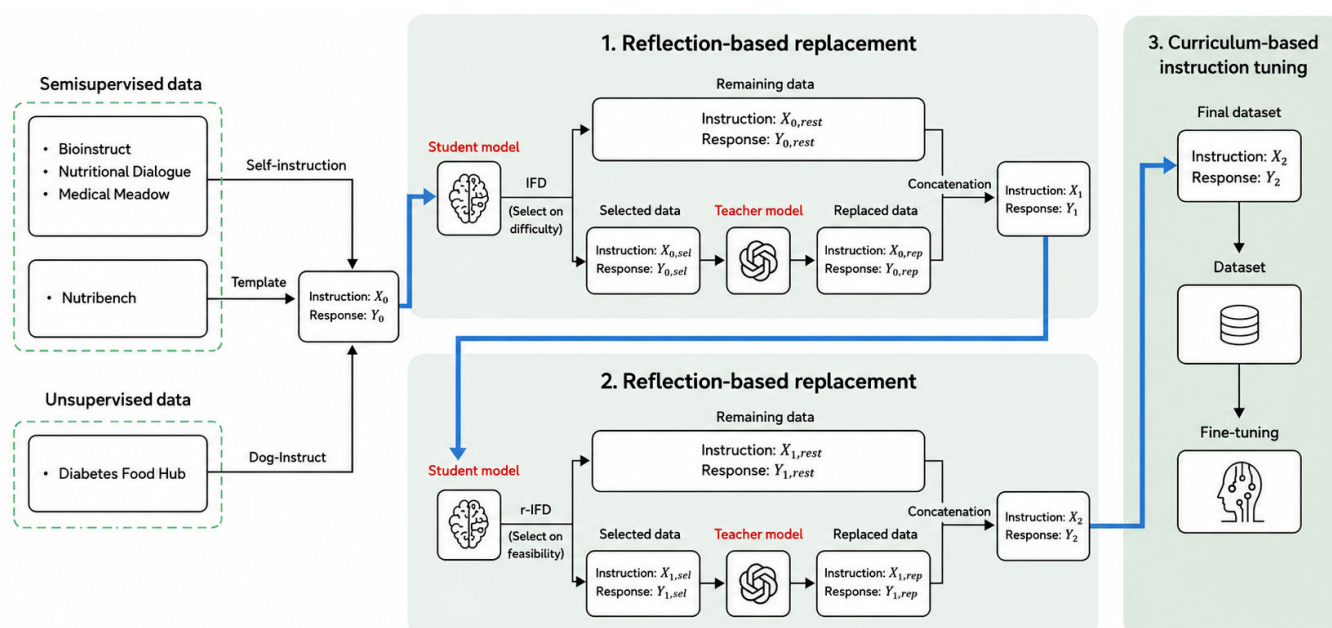
Despite these advances, existing LLM-based systems for diabetes management are typically evaluated in narrow task settings or rely on general-purpose models without explicit optimization for diabetes-specific instruction-following and reasoning. To date, there has been limited work on systematically adapting LLMs to perform consistently across a diverse set of diabetes-relevant tasks, such as clinical QA, reasoning over patient scenarios, and generating diabetes-appropriate dietary recommendations, within a unified, disease-specialized framework.

Methods

Study Design

To address the limitations of general biomedical LLMs in performing diabetes-specific reasoning, we developed and evaluated a diabetes-specialized LLM. The model was fine-tuned using reflection-based replacement and curriculum-based instruction tuning. Performance was assessed across multiple diabetes-related tasks, including QA, NLI, IE, summarization, generation, and dietary recommendation. An overview of the proposed instruction refinement and curriculum-based instruction tuning pipeline is shown in [Figure 1](#).

Figure 1. Overall architecture of the proposed diabetes-specialized large language model training pipeline, including instruction dataset construction, reflection-based replacement, and curriculum-based instruction tuning. IFD: instruction-following difficulty; r-IFD: reversed instruction-following difficulty.



Instruction Dataset Construction

To construct a domain-specific instruction dataset that captures the clinical and practical complexity of diabetes management, we aggregated content from 4 complementary sources. These included BioInstruct [34], a biomedical NLP dataset comprising 25,005 instruction-response pairs generated using GPT-4; Medical Meadow [15], a collection of clinical NLP prompts covering disease definitions, treatment recommendations, and patient counseling scenarios; Diabetes Food Hub [35], a curated repository of diabetes-friendly recipes and nutritional information; and Nutribench [36], a structured nutrition dataset providing recipe-level guidance for glycemic control.

Each dataset entry was structured as a triplet consisting of (1) a natural language instruction describing the task, (2) an input argument that instantiated the instruction, and (3) an output reflecting the correct execution of the instruction. To generate new samples, we adopted the Self-Instruct [25] methodology and FLAN (Fine-Tuned Language Net)-style prompting [37,38], where GPT-4 [39] was guided by 3 randomly selected seed tasks to produce diverse triplets. For long-form or informal content, such as recipes or patient narratives, we applied Dog-Instruct [40] techniques, which use back-translation and structural rewriting to transform text into high-quality instruction-response pairs. For datasets such as Nutribench, customized templates were applied to standardize nutritional tasks.

To ensure topical focus, we applied a diabetes-specific keyword filtering procedure to retain samples relevant to diabetes management. The filtering included terms related to diabetes and glycemic control, such as diabetes, diabetic, glucose, insulin, glycemic, hyperglycemia, hypoglycemia, glycosylated hemoglobin (HbA_{1c}), and blood sugar. Diversity was promoted by discarding instructions

with a ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence) similarity score greater than 0.7 relative to existing samples. Invalid generations were identified and removed based on heuristics, including instructions that were too short (<3 words) or too long (>150 words), outputs that repeated inputs, or tasks requiring unsupported modalities, such as images or graphs. The resulting dataset encompassed a wide range of diabetes-related tasks, including disease definition and explanation, diabetes-friendly dietary recommendations, glycemic trend analysis, and diet optimization.

Dataset Optimization Strategy

Overview

After constructing the diabetes-specific dataset, we implemented a multistage optimization strategy to improve both the quality and use of the instruction-response pairs. This strategy was inspired by recent advances in reflection-tuning [41] and selective reflection-tuning [10], which show that low-quality or poorly aligned samples can undermine instruction tuning by inducing superficial reasoning, ambiguity, or inconsistent model behavior. By adapting these approaches to the diabetes domain, we ensured that our dataset provided clear, challenging, and semantically aligned samples for instruction tuning.

Instruction Optimization via IFD

Instruction optimization focused on refining the clarity and difficulty of prompts so that they provided meaningful guidance to the model. We adopted a reflection-based replacement mechanism following [41], in which a teacher model (eg, GPT-4) critically evaluated each instruction and generated improved versions when necessary.

To quantify instructional difficulty from the student's perspective, we used IFD, a metric that measures how

much the presence of an instruction reduces uncertainty in generating the correct response. Intuitively, IFD captures how cognitively demanding it is for the model to follow a given instruction.

Formally, let f_θ denote the student model with parameters θ , x the instruction, and y the corresponding response. The IFD score is defined as:

$$\text{IFD}_\theta(y | x) = \frac{\text{ppl}(y | x)}{\text{ppl}(y)} - \exp(L_\theta(y | x) - L_\theta(y)) \quad (1)$$

where $L_\theta(y|x)$ denotes the negative log-likelihood loss of generating response y conditioned on instruction x , and $\text{ppl}(\cdot)$ denotes perplexity. A higher IFD score indicates that the instruction substantially influences the generation of the response, suggesting greater instructional difficulty.

For each original instruction-response pair and its reflected counterpart, the student model retained the version with the higher IFD score:

$$(x_1, y_1) = \arg \max_{(x, y)} \text{IFD}_\theta(y | x) \quad (2)$$

This selection process filtered out overly trivial or redundant instructions, resulting in a dataset that emphasized cognitively meaningful tasks capable of promoting deeper reasoning and robust instruction-following behavior.

Response Optimization via r-IFD

While IFD assesses the difficulty of instructions, it does not directly evaluate the clarity or semantic grounding of responses. To address this limitation, we further optimized responses using r-IFD, which measures how easily the original instruction can be inferred from the response alone.

The motivation behind r-IFD is that a high-quality response should clearly and unambiguously reflect the intent of the instruction. Responses that are vague, underspecified, or weakly grounded provide poor learning signals and may hinder generalization.

Formally, r-IFD is defined as:

$$\text{r-IFD}_\theta(x | y) = \frac{\text{ppl}(x | y')}{\text{ppl}(x)} = \exp(L_\theta(x | y') - L_\theta(x)) \quad (3)$$

where y' denotes a transformed version of the response used to prompt the model to recover the instruction. A lower r-IFD score indicates that the model can more easily infer the instruction from the response, reflecting stronger semantic alignment and instructional clarity.

For each instruction-response pair after instruction optimization, we selected the response with the lowest r-IFD score:

$$(x_2, y_2) = \arg \min_{(x_1, y_1)} \text{r-IFD}_\theta(x_1 | y_1) \quad (4)$$

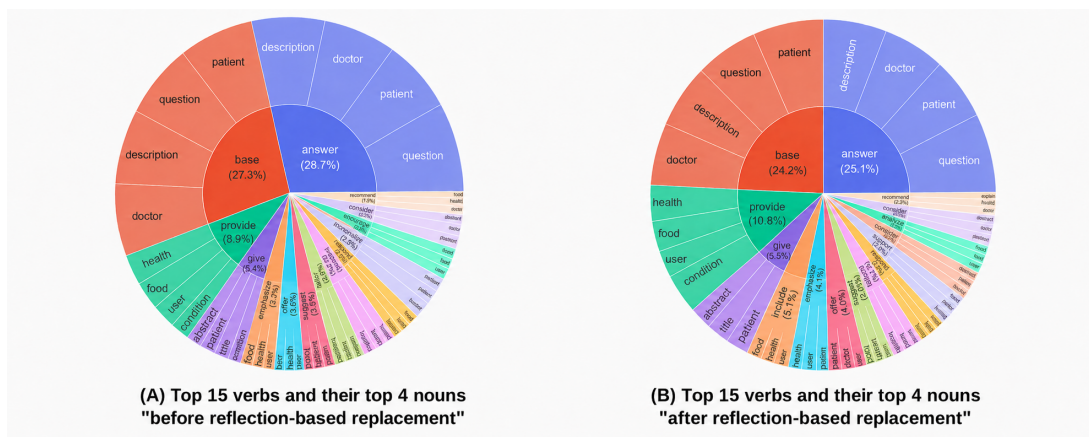
Based on IFD and r-IFD, we filter the dataset to retain instruction-response pairs characterized by high IFD scores and low r-IFD scores, respectively. Samples falling outside these optimal thresholds are flagged as suboptimal. The prompt instructs the model to revise instructions by improving topic complexity, required level of detail, domain knowledge requirements, ambiguity, and logical reasoning or problem-solving components. Corresponding responses are refined to improve helpfulness, relevance, accuracy, and level of detail while maintaining consistency with the intended instructional objective.

After refinement, IFD and r-IFD scores were recalculated to determine whether the revised instruction-response pairs satisfied the predefined filtering thresholds. Samples that continued to fall outside the acceptable range after refinement were excluded from the final instruction tuning dataset. The selection logic is formalized as follows:

$$(x', y') = \begin{cases} \text{Replace}(x, y), & \text{if } \text{IFD}_\theta(y | x) \leq 0.45 \text{ and } \text{r-IFD}_\theta(x | y) \geq 0.98 \\ (x, y), & \text{otherwise} \end{cases} \quad (5)$$

Instruction characteristics before and after reflection-based replacement are illustrated in Figure 2, showing a shift toward more diverse and reasoning-oriented instructions following optimization with IFD and r-IFD.

Figure 2. Instruction characteristics before (A) and after (B) reflection-based replacement using instruction-following difficulty and reversed instruction-following difficulty.

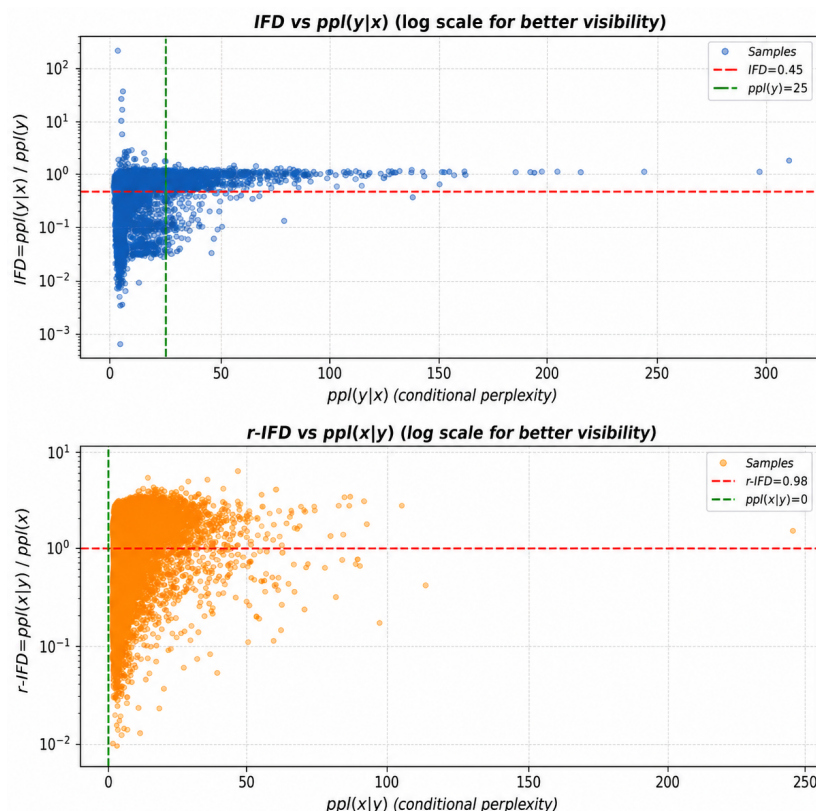


Reflection-Based Evaluation Criteria

Figure 3 shows the distribution of IFD and r-IFD scores in the instruction dataset. Based on the inspection of the log-scale distributions, we empirically selected threshold values to retain instruction-response pairs consistent with the selected

filtering criteria while filtering out ambiguous or weakly instructive samples. This dual-filtering strategy reduced noise in the instruction dataset while preserving task diversity prior to curriculum-based instruction tuning.

Figure 3. Reflection-based evaluation criteria for instruction dataset optimization. IFD: instruction-following difficulty; r-IFD: revised instruction-following difficulty; ppl: perplexity.



Curriculum-Based Instruction Tuning

Given the heterogeneity of tasks and response lengths, including QA, IE, NLI, summarization, generation, and diabetes-friendly dietary recommendation—the resulting instruction-response pairs vary widely in both structure and response length. Fine-tuning on such heterogeneous data without any ordering can destabilize training and hinder the model’s ability to internalize complex tasks.

To address this, we apply a staged curriculum-based instruction tuning strategy that sequences instructions from easier to more challenging, helping the model gradually adapt to increasing task complexity and response diversity. We first rank all instruction-response pairs based on their IFD scores, which reflect the relative difficulty of each sample from the perspective of the student model. An overview of the curriculum-based instruction tuning procedure is shown in Figure 4.

The dataset was subsequently partitioned into minibatches arranged in increasing order of difficulty, where difficulty

was quantified using the IFD score. A higher IFD score indicated that the sample was more difficult for the student model to follow. All instruction-response pairs were first sorted in ascending order of the IFD score, from easier to more challenging samples, and then divided into minibatches with similar difficulty levels to minimize intrabatch variance.

Training proceeded sequentially across these ordered minibatches. The model progressed from easier to more difficult samples according to the predetermined IFD-based ordering. This progressive training approach stabilized learning by mitigating task-level heterogeneity and reducing divergence in difficulty across batches.

This facilitated robust convergence and prevented catastrophic forgetting by reinforcing previously acquired general medical competencies before introducing more specialized domain-specific instructions. As a result, the model can effectively acquire expert-level reasoning for diabetes-related tasks without compromising its foundational biomedical knowledge.

Figure 4. Overview of the reflection-based instruction replacement and curriculum-based instruction tuning procedure used to optimize the diabetes-specific instruction dataset prior to fine-tuning. IFD: instruction-following difficulty.

Algorithm 1 IFD-Based Curriculum Instruction Tuning

Require: Instruction dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, model θ , number of stages B

Ensure: Fine-tuned model via staged curriculum instruction tuning

```

1: Initialize empty list  $\mathcal{D}'$ 
2: for each  $(x, y) \in \mathcal{D}$  do
3:   Compute  $\text{IFD}_{\theta}(y | x)$ 
4:   Append  $(x, y, \text{IFD}_{\theta}(y | x))$  to  $\mathcal{D}'$ 
5: end for
6: Sort  $\mathcal{D}'$  in ascending order of IFD scores
7: Partition  $\mathcal{D}'$  into  $B$  mini-batches:  $\{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_B\}$ 
8: for  $b = 1$  to  $B$  do
9:   Fine-tune model  $\theta$  on batch  $\mathcal{B}_b$ 
10: end for
  
```

Experimental Setup

In this section, we detail our experimental setup, the benchmark datasets, the baseline models, and the evaluation strategy used to assess the performance and effectiveness of our diabetes-specialized instruction-tuned LLM across multiple diabetes-relevant biomedical NLP tasks and dietary recommendation settings.

Implementation Details

We instruction-tuned our model using the LLaMA3 architecture [42], experimenting with both 3B- and 8B-parameter variants, and implementing using the *Transformers* library from Hugging Face [43]. To ensure computational efficiency, we used low-rank adaptation [44] via the parameter-efficient fine-tuning framework [45], setting the low-rank adaptation rank to 16, the scaling factor α to 16, and the dropout rate to 0.05.

Training was conducted with an effective batch size of 64, achieved using a per-device batch size of 16 and gradient accumulation over 4 steps. We optimized the model using the AdamW optimizer with a learning rate of 2×10^{-5} and trained for 2 epochs. The maximum sequence length was set to 2048 tokens.

All experiments were performed on a single NVIDIA A100 GPU with 40 GB of memory. Instruction filtering and reflection-based replacement were applied before tokenization to ensure alignment between the refined instruction set and the final training batches, including the ordering required for curriculum-based instruction tuning. This preprocessing step ensured consistent batching and preserved the intended difficulty progression during training.

Benchmark Tasks and Datasets

To align the evaluation benchmarks with the targeted diabetes domain, we constructed diabetes-specific subsets rather than directly using the original open-source benchmark datasets in full. For MedQA-USMLE, MedMCQA, PubMedQA, MedNLI, and the ADE Corpus, candidate samples were first identified using diabetes-related keyword filtering, including terms such as diabetes, diabetic, glucose, insulin, glycemic, hyperglycemia, hypoglycemia, HbA_{1c}, and blood sugar. The

filtered samples were then manually reviewed to confirm their relevance to diabetes-related clinical knowledge, reasoning, or adverse event extraction. Samples in which diabetes was mentioned only incidentally or that were not directly relevant to diabetes management were excluded from the final evaluation subsets.

The resulting evaluation framework encompasses the following benchmark tasks and datasets:

- Multiple-choice question answering (MCQA) includes the following:
 - MedQA-USMLE [46]: this dataset contains questions based on the US Medical Licensing Examination and is used to evaluate professional-level medical knowledge.
 - MedMCQA [47]: this large-scale dataset contains medical entrance examination questions drawn from various textbooks, with a particular focus on clinical scenarios.
 - PubMedQA [48]: this benchmark evaluates a model's ability to reason over biomedical abstracts and provide "yes," "no," or "maybe" answers.
- NLI includes the following:
 - MedNLI [49]: this dataset uses clinical narratives from MIMIC-III to evaluate a model's ability to perform medical natural language inference.
- Clinical information extraction (IE) includes the following:
 - ADE Corpus [50]: this corpus is used to evaluate the extraction of adverse drug events and related entities from medical text.
- Medical summarization includes the following:
 - NoteChat [51]: this dataset focuses on summarizing patient-doctor dialogues into structured, professional clinical notes.
- Clinical answer generation includes the following:
 - iCliniq [17]: this real-world doctor-patient question-and-answer corpus is used to assess a model's ability to provide accurate and compassionate clinical advice.
- Diabetes-specific specialized task includes the following:

- Nutritional plan generation: this novel task evaluates a model's ability to generate nutritionally balanced and diabetes-friendly meal plans based on patient-specific constraints.

Evaluation Metrics

For the classification-based tasks—QA, NLI, and IE—we report accuracy as the primary evaluation metric. Accuracy provides a direct and interpretable measure of task performance across these diabetes-related benchmarks and is commonly used in prior biomedical language model evaluations [34].

For generative tasks, including medical dialog summarization and clinical answer generation, we used a combination of automated semantic similarity metrics and model-based qualitative evaluation. Specifically, we computed BERTScore (Bidirectional Encoder Representations from Transformers Score) [52], which measures semantic similarity between generated outputs and reference texts using contextualized embeddings, and BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) [53], a learned evaluation metric designed to assess both syntactic and semantic alignment with human-written references. These metrics jointly capture content fidelity and linguistic adequacy in long-form medical text generation.

In addition to automated metrics, we evaluated generation quality using GPT-4 as an external evaluator. Each generated output was assessed using 3 criteria: coherence, completeness, and naturalness. Coherence was defined as the logical consistency, contextual relevance, and intersentence connectivity of the response. Completeness was defined as the extent to which the response addressed all clinically relevant aspects of the question without omitting key diabetes-related information. Naturalness was defined as the fluency, readability, and clinical appropriateness of the response, including whether the answer maintained an appropriate balance between patient-friendly language and medical professionalism.

For each evaluation instance, GPT-4 was provided with the input question, the generated response, and the reference answer, where applicable. The prompt included explicit definitions of coherence, completeness, and naturalness, together with a predefined scoring rubric, and GPT-4 was instructed to assign separate scores from 1 to 5 for each criterion. The full GPT-4 evaluation prompts and scoring rubrics used for answer correctness and generation-quality assessment are provided in [Multimedia Appendix 1](#) to improve transparency and reproducibility.

The use of GPT-4 for generation evaluation was motivated by prior studies demonstrating strong alignment between GPT-4-based assessments and human judgments in natural language generation tasks, including clinical dialog and medical text generation. However, no separate human evaluation or human-GPT-4 calibration was conducted in this study.

All evaluations were conducted on held-out test data that were not used during training.

Baseline Models

To benchmark the performance of our model, we conducted a comparative analysis against several prominent biomedical LLMs, including the following:

- MedAlpaca 7B is an instruction-tuned biomedical language model derived from LLaMA and fine-tuned on medical QA and dialog-style datasets.
- PMC-LLaMA is a domain-adapted LLaMA model pretrained on a large-scale corpus of biomedical literature, including PubMed Central articles and medical textbooks.
- ChatDoctor is a conversational medical language model fine-tuned on synthetic doctor-patient interactions.
- AlpaCare LLaMA2 7B [54] is an instruction-tuned medical assistant model trained on curated health care dialog data.
- BioMistral is a biomedical LLM based on the Mistral-7B architecture, pretrained on PubMed Open Access content.
- ME-LLaMA [55] is a model focused on evidence-based medicine and designed to prioritize grounding in verifiable clinical evidence.

Results

Overview

We evaluated the proposed diabetes-specialized instruction-tuned LLM across multiple diabetes-related task categories, including QA, NLI, IE, summarization, generation, and diabetes-friendly dietary recommendation. Performance was compared with several state-of-the-art biomedical and medical instruction-tuned baselines. Unless otherwise stated, all results are reported on held-out test sets not used during training.

Performance on QA and NLI Tasks

We first evaluated our model on MCQA (MedQA-USMLE, MedMCQA, and PubMedQA) and NLI (MedNLI). Results are summarized in [Table 1](#).

Across QA tasks, the proposed model demonstrated strong performance relative to existing biomedical baselines. It achieved accuracy scores of 0.44 on MedQA-USMLE, 0.53 on MedMCQA, and 0.78 on PubMedQA, yielding the highest average QA performance among all evaluated models. The improvement was particularly pronounced on PubMedQA, where the proposed model exceeded the next-best baseline by a substantial margin, highlighting its enhanced ability to retrieve and reason over diabetes-related biomedical facts.

Table 1. Zero-shot and instruction-tuned performance of diabetes-specialized large language models compared to baseline models on MCQA^a and NLI^b tasks.

Model	MedQA-USMLE ^c	MedMCQA	PubMedQA	MedNLI	ADE ^d Corpus
MedAlpaca 7B, accuracy	0.34	0.28	0.05	0.55	0.31
PMC ^e -LLaMA, accuracy	0.20	0.14	0.40	0.13	0.02
ChatDoctor, accuracy	0.18	0.17	0.37	0.21	0.21
AlpacaCare LLaMA2 7B, accuracy	0.09	0.35	0.60	0.13	0.22
BioMistral, accuracy	0.36	0.42	0.78	0.48	0.35
Me-LLaMA, accuracy	0.36	0.33	0.26	0.23	0.31
LLaMA3.2 3B (zero-shot), accuracy	0.24	0.30	0.20	0.07	0.29
LLaMA3.2 3B (instruction-tuned), accuracy	0.29	0.31	0.78	0.25	0.34
LLaMA3.1 8B (zero-shot), accuracy	0.34	0.30	0.55	0.21	0.30
LLaMA3.1 8B (ours), accuracy	0.44	0.53	0.78	0.39	0.39

^aMCQA: multiple-choice question answering.^bNLI: natural language inference.^cUSMLE: US Medical Licensing Examination.^dADE: adverse drug event.^ePMC: PubMed Central.

For NLI, when evaluated on the MedNLI dataset, the proposed model achieved an accuracy of 0.39. While this performance was slightly lower than that of MedAlpaca 7B, which attained an accuracy of 0.55, this difference aligns with the composition of the training data. Specifically, the diabetes-specific instruction dataset contained a limited number of entailment-focused examples, many of which were filtered during reflection-based replacement due to low instructional value. Despite this, the model maintained competitive NLI performance through broader exposure to reasoning-oriented instructions.

Performance on IE Tasks

We assessed IE performance using the ADE Corpus, focusing on accuracy as the primary metric. As shown in Table 1, the proposed model achieved an accuracy of 0.39, matching the best-performing baseline, BioMistral.

These results indicate that the diabetes-specialized instruction tuning process preserved the model's ability to extract structured biomedical information while prioritizing diabetes-specific reasoning and instruction-following. Importantly, the reflection-based replacement strategy did not degrade extraction performance despite selectively refining the instruction dataset.

We carried out additional experiments to evaluate GPT-4o-mini on the same diabetes-specific benchmark subsets. As shown in Table 2, GPT-4o-mini achieved higher performance on QA and NLI tasks, whereas the proposed model achieved substantially higher performance on the ADE IE task.

Table 2. Comparison of GPT-4o-mini and the proposed model on diabetes-specific benchmark subsets.

Task	Proposed model (8B), accuracy	GPT-4o-mini, accuracy
MedQA ^a -USMLE ^b	0.44	0.75
MedMCQA ^c	0.53	0.75
PubMedQA	0.78	1.00
MedNLI ^d	0.39	0.75
ADE ^e (IE ^f)	0.39	0.05

^aQA: question answering.^bUSMLE: US Medical Licensing Examination.^cMCQA: multiple-choice question answering.^dNLI: natural language inference.^eADE: adverse drug event.^fIE: information extraction.

Performance on Medical Dialog Summarization and Answer Generation

Furthermore, we evaluated generative performance on medical dialog summarization (NoteChat) and medical

answer generation (iCliniq), and the results are reported in Tables 3 and 4, respectively.

Table 3. Performance on medical dialog summarization using the NoteChat dataset.

Model	Coherence	Completeness	Naturalness	BLEURT ^a	BERTScore ^b	Average (SD)
MedAlpaca 7B	0.772	0.470	0.810	0.72	0.61	0.676 (0.123)
PMC ^c -LLaMA	0.013	0	0.130	0.55	0.50	0.239 (0.239)
ChatDoctor	0.405	0.315	0.485	0.41	0.43	0.409 (0.055)
AlpacaCare LLaMA2 7B	0.802	0.675	0.847	0.67	0.62	0.723 (0.086)
BioMistral	0.693	0.682	0.820	0.52	0.73	0.689 (0.097)
Me-LLaMA	0.755	0.645	0.835	0.55	0.64	0.685 (0.099)
LLaMA3.2 3B (zero-shot)	0.320	0.235	0.335	0.41	0.34	0.328 (0.056)
LLaMA3.2 3B (instruction-tuned)	0.817	0.667	0.863	0.61	0.54	0.699 (0.123)
LLaMA3.1 8B (zero-shot)	0.355	0.287	0.382	0.35	0.33	0.341 (0.032)
LLaMA3.1 8B (ours)	0.823	0.693	0.863	0.67	0.65	0.740 (0.086)

^aBLEURT: Bilingual Evaluation Understudy with Representations from Transformers.

^bBERTScore: Bidirectional Encoder Representations from Transformers score.

^cPMC: PubMed Central.

Table 4. Performance on medical answer generation using the iCliniq dataset.

Model	Coherence	Completeness	Naturalness	BLEURT ^a	BERTScore ^b	Average (SD)
MedAlpaca 7B	0.63	0.45	0.70	0.56	0.61	0.590 (0.083)
PMC ^c -LLaMA	0.245	0.162	0.38	0.60	0.62	0.401 (0.184)
ChatDoctor	0.62	0.453	0.698	0.57	0.62	0.592 (0.081)
AlpacaCare LLaMA2 7B	0.827	0.665	0.935	0.56	0.44	0.685 (0.178)
BioMistral	0.502	0.312	0.613	0.55	0.52	0.499 (0.101)
Me-LLaMA	0.68	0.483	0.798	0.60	0.67	0.646 (0.103)
LLaMA3.2 3B (zero-shot)	0.745	0.573	0.90	0.56	0.47	0.650 (0.154)
LLaMA3.2 3B (instruction-tuned)	0.775	0.603	0.907	0.69	0.50	0.695 (0.140)
LLaMA3.1 8B (zero-shot)	0.588	0.44	0.645	0.56	0.55	0.557 (0.067)
LLaMA3.1 8B (ours)	0.83	0.655	0.925	0.57	0.63	0.722 (0.133)

^aBLEURT: bilingual evaluation understudy with representations from transformers.

^bBERTScore: Bidirectional Encoder Representations from Transformers Score.

^cPMC: PubMed Central.

On the NoteChat summarization task, the proposed model achieved the highest average score (0.74), outperforming all baseline models across coherence, completeness, and naturalness. Similarly, on the iCliniq clinical answer generation task, the model achieved an average score of 0.722, demonstrating a strong capability in producing clinically appropriate, context-aware, and user-aligned responses.

Automatic semantic similarity metrics (BLEURT and BERTScore) further supported these findings, indicating that the generated outputs closely aligned with reference responses both syntactically and semantically. GPT-4-based evaluations showed that the proposed model maintained fluent and natural language generation while delivering responses that were complete and tailored to diabetes-related clinical contexts.

Evaluation of Diabetes-Friendly Dietary Recommendations

To assess domain specialization beyond standard NLP tasks, we evaluated the model's ability to generate diabetes-friendly

dietary recommendations. In this setting, we compared the proposed model with GPT-4, a strong general-purpose baseline.

Each model generated 10 full-day meal plans for individuals with diabetes. The generated plans were evaluated using two complementary criteria: (1) language quality, assessed via GPT-4-based coherence, completeness, and naturalness scores; and (2) nutritional quality, assessed using the DQI-I score.

As shown in [Figure 5](#), the proposed model achieved higher average DQI-I scores than those of GPT-4 across multiple dietary components, indicating superior nutritional balance and closer alignment with dietary guidelines for diabetes management.

Language quality scores are summarized in [Table 5](#).

Figure 5. Comparison of Diet Quality Index–International (DQI-I) component scores for diabetes-friendly meal plans generated by GPT-4 and the proposed diabetes-specialized model: (A) DQI-I scores for GPT-4, (B) DQI-I scores for the proposed model, and (C) comparison of DQI-I component scores between GPT-4 and the proposed model.

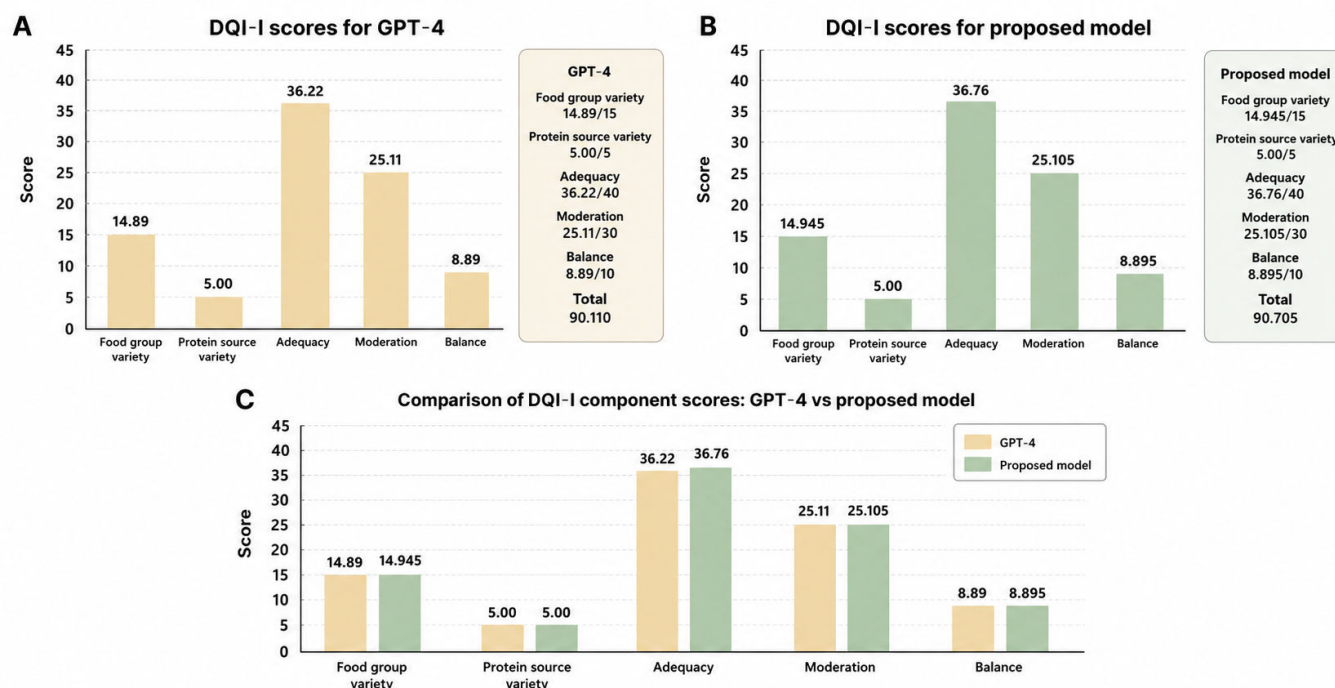


Table 5. Language quality scores using GPT-4–based ratings of coherence, completeness, and naturalness.

Model	Coherence	Completeness	Naturalness
GPT-4	0.847	0.930	0.972
Proposed model	0.825	0.915	0.915

GPT-4 achieved slightly higher scores for coherence, completeness, and naturalness; however, the differences were small, and the proposed model maintained competitive language quality while producing more nutritionally appropriate meal plans for diabetes-specific needs.

Although the improvement in DQI-I was modest, the proposed model generated meal plans with higher nutritional quality and lower simulated postprandial glycemic burden than those generated by GPT-4. This finding suggests that domain-specific instruction tuning may provide benefits for diabetes-focused dietary recommendations despite relatively small differences in language-quality metrics.

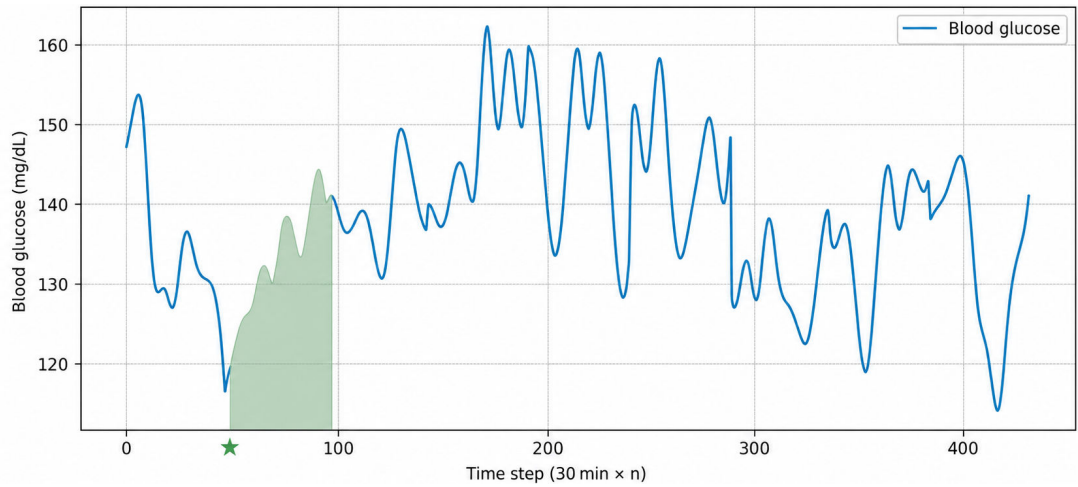
Simulation-Based Evaluation of Glycemic Response

To further validate the physiological relevance of the generated dietary recommendations, we conducted simulation-based evaluations using the SimGlucose simulator, an open-source simulator based on the US Food and Drug Administration–validated UVA/Padova type 1 diabetes model.

We simulated postprandial glycemic responses under 2 conditions: carbohydrate-only intake and carbohydrate intake combined with insulin dosing. In both scenarios, meal plans generated by the proposed model resulted in lower incremental area under the curve values compared with those generated by GPT-4. Specifically, the proposed model achieved a lower postmeal incremental area under the curve (161.53 mg/dL/min) than GPT-4 (163.97 mg/dL/min).

The corresponding blood glucose trajectories over a 9-day simulation period without insulin are illustrated in Figure 6, showing reduced glycemic excursions when following meal plans generated by the proposed model. These findings indicate that the model's dietary recommendations are not only nutritionally aligned but also physiologically favorable for glycemic control. Representative qualitative examples of dietary recommendations, failure cases, carbohydrate distributions, and corresponding glycemic response simulations are provided in Multimedia Appendix 1.

Figure 6. Simulated blood glucose trajectory over a 9-day period without insulin. The shaded star indicates the start of the highlighted interval, and the shaded area represents the incremental area under the curve used to assess the postprandial glycemic response, corresponding to the highlighted window.



Ablation Study of Instruction Optimization and Curriculum-Based Instruction Tuning

To directly evaluate the contribution of the proposed instruction optimization and curriculum-based instruction tuning strategies, we conducted an ablation study examining the impact of each major training component. Specifically,

we compared the base model with progressively enhanced variants, including instruction tuning with seed diabetes-specific data, IFD-based instruction optimization, combined IFD and r-IFD optimization, and the final curriculum-based instruction tuning strategy. Results are reported across diabetes-related benchmark tasks in Table 6, medical dialog summarization in Table 7, and clinical answer generation in Table 8.

Table 6. Ablation study results across diabetes-related benchmark tasks.

Model	MedQA-USMLE ^a	MedMCQA ^b	PubMedQA ^c	MedNLI ^d	ADE ^e Corpus
LLaMA3.1 8B (zero-shot), accuracy	0.34	0.30	0.55	0.21	0.30
+ Seed, accuracy	0.41	0.37	0.69	0.41	0.32
+ IFD ^f , accuracy	0.47	0.52	0.72	0.40	0.34
+ r-IFD ^g , accuracy	0.43	0.52	0.82	0.37	0.36
+ Curriculum-based instruction tuning, accuracy	0.44	0.53	0.78	0.39	0.39

^aUSMLE: US Medical Licensing Examination.

^bMCQA: multiple-choice question answering.

^cQA: question answering.

^dNLI: natural language inference.

^eADE: adverse drug event.

^fIFD: instruction-following difficulty.

^gr-IFD: revised instruction-following difficulty.

Table 7. Ablation study results for medical dialog summarization using the NoteChat dataset.

Model	Coherence	Completeness	Naturalness	BLEURT ^a	BERTScore ^b
LLaMA3.1 8B (zero-shot)	0.355	0.287	0.382	0.35	0.33
+ Seed	0.837	0.670	0.853	0.41	0.45
+ IFD ^c	0.780	0.710	0.875	0.52	0.48
+ r-IFD ^d	0.823	0.718	0.875	0.65	0.60
+ Curriculum-based instruction tuning	0.823	0.693	0.863	0.67	0.65

^aBLEURT: Bilingual Evaluation Understudy with Representations from Transformers.

^bBERTScore: Bidirectional Encoder Representations from Transformers Score.

^cIFD: instruction-following difficulty.

^dr-IFD: revised instruction-following difficulty.

Table 8. Ablation study results for clinical answer generation using the iCliniq dataset.

Model	Coherence	Completeness	Naturalness	BLEURT ^a	BERTScore ^b
LLaMA3.1 8B (zero-shot)	0.588	0.440	0.645	0.56	0.55
+ Seed	0.810	0.608	0.913	0.61	0.67
+ IFD ^c	0.835	0.650	0.927	0.56	0.63
+ r-IFD ^d	0.808	0.635	0.883	0.63	0.57
+ Curriculum-based instruction tuning	0.830	0.655	0.925	0.57	0.63

^aBLEURT: Bilingual Evaluation Understudy with Representations from Transformers.

^bBERTScore: Bidirectional Encoder Representations from Transformers Score.

^cIFD: instruction-following difficulty.

^dr-IFD: revised instruction-following difficulty.

Overall, the ablation results show that each component contributed to model performance. IFD-based optimization improved instruction selection by prioritizing samples that were more informative for the student model, while r-IFD-based optimization improved semantic alignment between instructions and responses. Curriculum-based instruction tuning further supported performance by progressively exposing the model to increasingly difficult instruction-response pairs.

Discussion

Principal Findings

In this study, we developed and evaluated a diabetes-specialized, instruction-tuned LLM designed to support diabetes-related reasoning, instruction-following, and dietary recommendation tasks. Across a diverse set of diabetes-relevant benchmarks, including QA, NLI, IE, medical dialog summarization, answer generation, and nutrition-focused planning, the proposed model consistently demonstrated strong performance relative to existing biomedical and general-purpose LLMs.

The model achieved strong performance on diabetes-focused QA tasks and generated competitive medical summaries and responses, as measured by both automated metrics and GPT-4-based evaluations of coherence, completeness, and naturalness. In dietary recommendation tasks, the proposed model produced meal plans with higher DQI-I scores than those generated by GPT-4, indicating improved nutritional balance and suitability for diabetes management. Simulation-based evaluation using the SimGlucose simulator further suggested that these meal plans were associated with more favorable postprandial glycemic responses, supporting the physiological plausibility of the generated recommendations.

Importantly, the model maintained competitive performance on general biomedical tasks such as NLI and adverse event IE, suggesting that domain specialization did not substantially degrade broader biomedical capabilities. Together, these results indicate that targeted instruction tuning may improve diabetes-specific performance while preserving broader biomedical capabilities.

Limitations

This study has several limitations. First, although we evaluated the model across multiple diabetes-related tasks, the benchmarks remain limited to publicly available datasets and simulated scenarios. Real-world clinical deployment would require rigorous prospective evaluation, including clinician oversight and patient safety assessments.

Second, dietary recommendation quality was assessed using established nutritional indices and glycemic simulations rather than by direct clinical outcomes. While these measures provide useful proxies, they cannot fully capture individual variability, comorbidities, or behavioral factors that influence diabetes management.

Third, GPT-4 was used as an evaluator for several generative metrics. Although prior studies have shown strong alignment between GPT-4-based evaluation and human judgment, automated evaluation may still introduce biases or overlook subtle clinical errors. In addition, no separate human evaluation or human-GPT-4 calibration was performed in this study. Future work should include expert clinical evaluation to further validate the reliability of GPT-4-based assessment, particularly for detecting subtle medical inaccuracies or clinically inappropriate recommendations.

Finally, the model was fine-tuned for a specific architecture and parameter scale. The generalizability of the proposed instruction optimization strategy to other model families and larger-scale systems warrants further investigation.

Comparison With Prior Work

Prior work has demonstrated that general-purpose and biomedical LLMs can support a range of health care-related tasks, including medical QA, dialog summarization, and patient education. However, multiple studies have also reported that these models often struggle with chronic disease management scenarios that require sustained reasoning, contextual interpretation, and domain-specific judgment, particularly in diabetes care.

Unlike earlier approaches that rely primarily on large-scale biomedical pretraining or generic instruction tuning, this study emphasizes model-centric instruction optimization, in which training data are refined based on the student model's learning behavior. By incorporating reflection-based replacement and curriculum-based instruction tuning, the

proposed approach extends prior instruction tuning frameworks by explicitly accounting for instruction difficulty, ambiguity, and learning progression. This contrasts with existing diabetes-related LLM efforts that either focus on narrow tasks (eg, CGM summarization or education chatbots) or depend heavily on retrieval-based augmentation without optimizing instruction-level learning dynamics.

The observed improvements in diabetes-focused reasoning and dietary planning suggest that aligning instruction data with the model's receptiveness and sequencing tasks by difficulty may provide complementary benefits for domain-specific adaptation.

Implications for Diabetes Management and Clinical Informatics

The findings of this study have several implications for the development of AI-assisted tools in diabetes care. First, these findings highlight the potential of domain-specialized LLMs to support complex, multifaceted diabetes management tasks that extend beyond factual QA. The ability to generate nutritionally appropriate meal plans, explain glycemic trends, and provide context-aware responses may be particularly valuable for patient education and decision support.

Second, the results suggest that instruction-level data curation and training strategies play a critical role in determining model behavior in clinical domains. Rather than treating instruction datasets as static resources, adaptive refinement based on model feedback may help mitigate common issues, such as vague responses, superficial reasoning, and poor alignment with clinical intent.

Finally, the use of simulation-based evaluation provides a complementary perspective on text-based metrics by assessing downstream physiological plausibility. Although not a substitute for clinical validation, such simulations may help identify potentially harmful or suboptimal recommendations during model development.

Conclusions

This study presents a diabetes-specialized instruction-tuned LLM optimized through reflection-based replacement and curriculum-based instruction tuning. By aligning instruction data with the model's learning capacity and sequencing tasks by difficulty, the proposed approach enables robust diabetes-specific reasoning and instruction-following across diverse tasks, including QA, NLI, medical dialog processing, and dietary recommendations.

The model consistently outperformed existing biomedical and general-purpose LLMs on diabetes-focused benchmarks while maintaining competitive performance on general biomedical tasks. Simulation-based evaluation further supported the physiological plausibility of the generated dietary recommendations.

These findings suggest that carefully designed domain-specific instruction tuning can improve the clinical relevance and diabetes-specific capabilities of general-purpose language models for chronic disease management. Future work should focus on prospective clinical evaluation, integration with real-world decision-support systems, and extension of the proposed framework to other chronic conditions.

Acknowledgments

The authors used generative AI technology (ChatGPT; OpenAI) during the final editing stage for minor image revisions and quality enhancement.

Funding

This research was supported by the National Research Foundation of Korea grant funded by the Ministry of Science and ICT (RS-2025-25460384) and the Regional Innovation System & Education (RISE) through the Seoul RISE Center, funded by the Ministry of Education and the Seoul Metropolitan Government (2025-RISE-01-019-04), and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00442006).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

JH developed the research idea, designed the methodology, conducted the experiments, and drafted the initial manuscript. DLN collected and preprocessed the datasets. DP supervised the research, contributed to the interpretation of the results, and critically revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Evaluation details, including GPT-4 evaluation prompts and scoring rubrics, and case studies of dietary recommendation and glycemic response simulation.

[\[DOCX File \(Microsoft Word File\), 42 KB-Multimedia Appendix 1\]](#)

References

1. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. *Lancet*. Nov 23, 2024;404(10467):2077-2093. [doi: [10.1016/S0140-6736\(24\)02317-1](https://doi.org/10.1016/S0140-6736(24)02317-1)] [Medline: [39549716](https://pubmed.ncbi.nlm.nih.gov/39549716/)]
2. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Feb 15, 2020;36(4):1234-1240. [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
3. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. Nov 19, 2022;23(6):bbac409. [doi: [10.1093/bib/bbac409](https://doi.org/10.1093/bib/bbac409)] [Medline: [36156661](https://pubmed.ncbi.nlm.nih.gov/36156661/)]
4. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Aug 2023;620(7972):172-180. [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
5. Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat Med*. Oct 2024;30(10):2886-2896. [doi: [10.1038/s41591-024-03139-8](https://doi.org/10.1038/s41591-024-03139-8)] [Medline: [39030266](https://pubmed.ncbi.nlm.nih.gov/39030266/)]
6. Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. *J Med Internet Res*. Nov 8, 2024;26(1):e58041. [doi: [10.2196/58041](https://doi.org/10.2196/58041)] [Medline: [39046096](https://pubmed.ncbi.nlm.nih.gov/39046096/)]
7. Healey E, Kohane IS. LLM-CGM: a benchmark for large language model-enabled querying of continuous glucose monitoring data for conversational diabetes management. *Pac Symp Biocomput*. 2025;30:82-93. [doi: [10.1142/9789819807024_0007](https://doi.org/10.1142/9789819807024_0007)] [Medline: [39670363](https://pubmed.ncbi.nlm.nih.gov/39670363/)]
8. Hussain W, Grundy J. Advice for diabetes self-management by ChatGPT models: challenges and recommendations. *arXiv*. Preprint posted online on Jan 14, 2025. [doi: [10.48550/arXiv.2501.07931](https://doi.org/10.48550/arXiv.2501.07931)]
9. Li H, Jiang Z, Guan Z, et al. Large language models for diabetes training: a prospective study. *Sci Bull (Beijing)*. Mar 30, 2025;70(6):934-942. [doi: [10.1016/j.scib.2025.01.034](https://doi.org/10.1016/j.scib.2025.01.034)] [Medline: [39947986](https://pubmed.ncbi.nlm.nih.gov/39947986/)]
10. Li M, Chen L, Chen J, He S, Gu J, Zhou T. Selective reflection-tuning: student-selected data recycling for LLM instruction-tuning. In: Ku LW, Martins A, Srikumar V, editors. *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics; 2024:16189-16211. [doi: [10.18653/v1/2024.findings-acl.958](https://doi.org/10.18653/v1/2024.findings-acl.958)]
11. Min Y, Zhou K, Gao D, Zhao WX, Hu H, Li Y. DATA-CUBE: DATA curriculum for instruction-based sentence representation learning. In: Ku LW, Martins A, Srikumar V, editors. *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics; 2024:13748-13761. [doi: [10.18653/v1/2024.findings-acl.816](https://doi.org/10.18653/v1/2024.findings-acl.816)]
12. Kim S, Haines PS, Siega-Riz AM, Popkin BM. The Diet Quality Index-International (DQI-I) provides an effective tool for cross-national comparison of diet quality as illustrated by China and the United States. *J Nutr*. Nov 2003;133(11):3476-3484. [doi: [10.1093/jn/133.11.3476](https://doi.org/10.1093/jn/133.11.3476)] [Medline: [14608061](https://pubmed.ncbi.nlm.nih.gov/14608061/)]
13. Xie J. jxx123/simglucose. GitHub. 2026. URL: <https://github.com/jxx123/simglucose> [Accessed 2026-01-25]
14. Man CD, Micheletto F, Lv D, Breton M, Kovatchev B, Cobelli C. The UVA/Padova type 1 diabetes simulator: new features. *J Diabetes Sci Technol*. Jan 2014;8(1):26-34. [doi: [10.1177/1932296813514502](https://doi.org/10.1177/1932296813514502)] [Medline: [24876534](https://pubmed.ncbi.nlm.nih.gov/24876534/)]
15. Han T, Adams LC, Papaioannou JM, et al. MedAlpaca: an open-source collection of medical conversational AI models and training data. *arXiv*. Preprint posted online on Apr 14, 2023. [doi: [10.48550/arXiv.2304.08247](https://doi.org/10.48550/arXiv.2304.08247)]
16. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inform Assoc*. Sep 1, 2024;31(9):1833-1843. [doi: [10.1093/jamia/ocae045](https://doi.org/10.1093/jamia/ocae045)] [Medline: [38613821](https://pubmed.ncbi.nlm.nih.gov/38613821/)]
17. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus*. Jun 2023;15(6):e40895. [doi: [10.7759/cureus.40895](https://doi.org/10.7759/cureus.40895)] [Medline: [37492832](https://pubmed.ncbi.nlm.nih.gov/37492832/)]
18. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains. In: Ku LW, Martins A, Srikumar V, editors. *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics; 2024:5848-5864. [doi: [10.18653/v1/2024.findings-acl.348](https://doi.org/10.18653/v1/2024.findings-acl.348)]
19. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc; 2020. URL: <https://dl.acm.org/doi/abs/10.5555/3495724.3495883> [Accessed 2026-01-30]
20. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. *arXiv*. Preprint posted online on Jul 18, 2023. [doi: [10.48550/arXiv.2307.09288](https://doi.org/10.48550/arXiv.2307.09288)]
21. Ye Q, Lin BY, Ren X. CrossFit: a few-shot learning challenge for cross-task generalization in NLP. In: Marie-Francine M, Xuanjing H, Lucia S, Scott Wen-tau Y, editors. *Proceedings of the 2021 Conference on Empirical Methods in Natural*

- Language Processing. Association for Computational Linguistics; 2021:7163-7189. [doi: [10.18653/v1/2021.emnlp-main.572](https://doi.org/10.18653/v1/2021.emnlp-main.572)]
22. Khashabi D, Min S, Khot T, et al. UNIFIEDQA: crossing format boundaries with a single QA system. In: Cohn T, He Y, Liu Y, editors. Findings of the Association for Computational Linguistics. Association for Computational Linguistics; 2020:1896-1907. [doi: [10.18653/v1/2020.findings-emnlp.171](https://doi.org/10.18653/v1/2020.findings-emnlp.171)]
 23. Honovich O, Scialom T, Levy O, Schick T. Unnatural instructions: tuning language models with (almost) no human labor. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2023:14409-14428. [doi: [10.18653/v1/2023.acl-long.806](https://doi.org/10.18653/v1/2023.acl-long.806)]
 24. Taori R, Gulrajani I, Zhang T, et al. Tatsu-lab/stanford_alpaca. GitHub. 2023. URL: https://github.com/tatsu-lab/stanford_alpaca [Accessed 2026-01-20]
 25. Wang Y, Kordi Y, Mishra S, et al. Self-instruct: aligning language models with self-generated instructions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2023:13484-13508. [doi: [10.18653/v1/2023.acl-long.754](https://doi.org/10.18653/v1/2023.acl-long.754)]
 26. Xu C, Sun Q, Zheng K, et al. WizardLM: empowering large pre-trained language models to follow complex instructions. Presented at: International Conference on Learning Representations 2024 (ICLR 2024); May 7-11, 2024; Vienna, Austria. URL: <https://openreview.net/forum?id=CfXh93NDgH> [Accessed 2026-07-16]
 27. Wang Y, Mishra S, Alipoormolabashi P, et al. Super-naturalinstructions: generalization via declarative instructions on 1600+ NLP tasks. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2022:5085-5109. [doi: [10.18653/v1/2022.emnlp-main.340](https://doi.org/10.18653/v1/2022.emnlp-main.340)]
 28. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. Diabetes Care. May 1, 2023;46(5):e103-e105. [doi: [10.2337/dc23-0197](https://doi.org/10.2337/dc23-0197)] [Medline: [36920843](https://pubmed.ncbi.nlm.nih.gov/36920843/)]
 29. Nayak A, Vakili S, Nayak K, et al. Use of voice-based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: a randomized clinical trial. JAMA Netw Open. Dec 1, 2023;6(12):e2340232. [doi: [10.1001/jamanetworkopen.2023.40232](https://doi.org/10.1001/jamanetworkopen.2023.40232)] [Medline: [38039007](https://pubmed.ncbi.nlm.nih.gov/38039007/)]
 30. Sheng B, Guan Z, Lim LL, et al. Large language models for diabetes care: potentials and prospects. Sci Bull. Mar 2024;69(5):583-588. [doi: [10.1016/j.scib.2024.01.004](https://doi.org/10.1016/j.scib.2024.01.004)]
 31. Abbasian M, Yang Z, Khatibi E, et al. Knowledge-infused LLM-powered conversational health agent: a case study for diabetes patients. In: 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2024. [doi: [10.1109/EMBC53108.2024.10781547](https://doi.org/10.1109/EMBC53108.2024.10781547)]
 32. Yang Z, Khatibi E, Nagesh N, et al. ChatDiet: empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. Smart Health. Jun 2024;32:100465. [doi: [10.1016/j.smhl.2024.100465](https://doi.org/10.1016/j.smhl.2024.100465)]
 33. Martinez-Cruz C, Guerrero JFG, Ruiz JLL, Rueda AJ, Espinilla M. A first approach to the generation of linguistic summaries from glucose sensors using GPT-4. In: Bravo J, Urzáiz G, editors. Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023). Springer; 2023:33-43. [doi: [10.1007/978-3-031-48642-5_4](https://doi.org/10.1007/978-3-031-48642-5_4)]
 34. Tran H, Yang Z, Yao Z, Yu H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. J Am Med Inform Assoc. Sep 1, 2024;31(9):1821-1832. [doi: [10.1093/jamia/ocae122](https://doi.org/10.1093/jamia/ocae122)] [Medline: [38833265](https://pubmed.ncbi.nlm.nih.gov/38833265/)]
 35. Diabetes Food Hub. URL: <https://diabetesfoodhub.org/> [Accessed 2026-02-01]
 36. Dhaliwal M, Hua H, Pullela L, Burke R, Qin Y. NutriBench: a dataset for evaluating large language models on nutrition estimation from meal descriptions. Presented at: International Conference on Learning Representations 2025 (ICLR 2025); Apr 24-28, 2025; Singapore. URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/ef3a57e4f26b640e6f90d78cbb011feb-Abstract-Conference.html [Accessed 2026-07-16]
 37. Longpre S, Hou L, Vu T, et al. The Flan collection: designing data and methods for effective instruction tuning. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research; 2023:22631-22648. URL: <https://proceedings.mlr.press/v202/longpre23a.html> [Accessed 2026-07-16]
 38. Wei J, Bosma M, Zhao V, et al. Finetuned language models are zero-shot learners. Presented at: The Tenth International Conference on Learning Representations (ICLR 2022); Apr 25-29, 2022; Virtual (online). URL: <https://openreview.net/pdf?id=gEZrGCozdqR> [Accessed 2026-07-16]
 39. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: [10.48550/arXiv.2303.08774](https://doi.org/10.48550/arXiv.2303.08774)]
 40. Chen Y, Jiang H, Huang X, Shi S, Qi G. DoG-instruct: towards premium instruction-tuning data via text-grounded instruction wrapping. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics; 2024:4125-4135. [doi: [10.18653/v1/2024.naacl-long.230](https://doi.org/10.18653/v1/2024.naacl-long.230)]

41. Li M, Chen L, Chen J, et al. Reflection-tuning: data recycling improves LLM instruction-tuning. Presented at: Instruction Tuning and Instruction Following Workshop at the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023); Dec 15, 2023; New Orleans, Louisiana, USA. URL: <https://openreview.net/pdf?id=xqoZZqkPU> [Accessed 2026-07-16]
42. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: [10.48550/arXiv.2407.21783](https://doi.org/10.48550/arXiv.2407.21783)]
43. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics; 2020:38-45. [doi: [10.18653/v1/2020.emnlp-demos.6](https://doi.org/10.18653/v1/2020.emnlp-demos.6)]
44. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. Presented at: The Tenth International Conference on Learning Representations (ICLR 2022); Apr 25-29, 2022. URL: <https://openreview.net/forum?id=nZeVKeeFYf9> [Accessed 2026-07-16]
45. Mangrulkar S, Gugger S, Debut L, et al. PEFT: state-of-the-art parameter-efficient fine-tuning methods. GitHub. 2022. URL: <https://github.com/huggingface/peft> [Accessed 2026-07-16]
46. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11(14):6421. [doi: [10.3390/app11146421](https://doi.org/10.3390/app11146421)]
47. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Flores G, Chen GH, Pollard T, Ho JC, Naumann T, editors. Proceedings of the Conference on Health, Inference, and Learning. Proceedings of Machine Learning Research; 2022:248-260. URL: <https://proceedings.mlr.press/v174/pal22a.html> [Accessed 2026-07-16]
48. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019:2567-2577. [doi: [10.18653/v1/D19-1259](https://doi.org/10.18653/v1/D19-1259)]
49. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2018:1586-1596. [doi: [10.18653/v1/D18-1187](https://doi.org/10.18653/v1/D18-1187)]
50. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform. Oct 2012;45(5):885-892. [doi: [10.1016/j.jbi.2012.04.008](https://doi.org/10.1016/j.jbi.2012.04.008)] [Medline: [22554702](https://pubmed.ncbi.nlm.nih.gov/22554702/)]
51. Wang J, Yao Z, Yang Z, et al. NoteChat: a dataset of synthetic patient-physician conversations conditioned on clinical notes. Find ACL ACL. Aug 2024;2024:15183-15201. [doi: [10.18653/v1/2024.findings-acl.901](https://doi.org/10.18653/v1/2024.findings-acl.901)] [Medline: [42273713](https://pubmed.ncbi.nlm.nih.gov/42273713/)]
52. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. Presented at: The Eighth International Conference on Learning Representations (ICLR 2020); Apr 26-30, 2020; Addis Ababa, Ethiopia. URL: <https://openreview.net/forum?id=SkeHuCVFDr> [Accessed 2026-07-16]
53. Sellam T, Das D, Parikh A. BLEURT: learning robust metrics for text generation. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020:7881-7892. [doi: [10.18653/v1/2020.acl-main.704](https://doi.org/10.18653/v1/2020.acl-main.704)]
54. Zhang X, Tian C, Yang X, Chen L, Li Z, Petzold LR. AlpaCare: instruction-tuned large language models for medical application. arXiv. Preprint posted online on Oct 23, 2023. [doi: [10.48550/arXiv.2310.14558](https://doi.org/10.48550/arXiv.2310.14558)]
55. Xie Q, Chen Q, Chen A, et al. Me-LLaMA: foundation large language models for medical applications. Research Square. Preprint posted online on May 22, 2024. [doi: [10.21203/rs.3.rs-4240043/v1](https://doi.org/10.21203/rs.3.rs-4240043/v1)]

Abbreviations

BERTScore: Bidirectional Encoder Representations from Transformers Score

BioGPT: biomedical generative pretrained transformer

BLEURT: Bilingual Evaluation Understudy with Representations from Transformers

CGM: continuous glucose monitoring

DQI-I: Diet Quality Index–International

FLAN: Fine-Tuned Language Net

HbA_{1c}: glycosylated hemoglobin

IE: information extraction

IFD: instruction-following difficulty

LLM: large language model

MCQA: multiple-choice question answering

NLI: natural language inference

NLP: natural language processing

QA: question answering

r-IFD: revised instruction-following difficulty

ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence

Edited by Hongfang Liu; peer-reviewed by Gilbert Lim, Irene Li; submitted 04.Feb.2026; final revised version received 01.Jun.2026; accepted 12.Jun.2026; published 25.Aug.2026

Please cite as:

Hwang J, Namayanja DL, Park D

Reflection- and Curriculum-Based Instruction Tuning for Diabetes-Specialized Large Language Models: Model Development and Evaluation Study

JMIR AI 2026;5:e92843

URL: <https://ai.jmir.org/2026/1/e92843>

doi: [10.2196/92843](https://doi.org/10.2196/92843)

© Jaesung Hwang, Deniise Liz Namayanja, Donghyeon Park. Originally published in JMIR AI (<https://ai.jmir.org>), 25.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.