Diabetica: Developing Specialized Large Language Models for Diabetes Care with Multi-Faceted Benchmarks and Clinical Validation

Anonymous ACL submission

Abstract

Diabetes is a chronic disease with a significant global health burden, where the effectiveness of Large Language Models (LLMs) across diverse diabetes tasks remains unproven. To address this, we introduce a comprehensive framework for developing and evaluating diabetes-specialized LLMs. Our methodology begins with systematically collecting unstructured diabetes-related data from sources like clinical guidelines and medical textbooks. This data then undergoes rigorous processing, including filtering, transformation, and refinement via a novel self-distillation approach, to produce a high-quality training dataset and a multi-faceted benchmarking suite. Leveraging this training dataset, the model we obtain, Diabetica, significantly outperforms existing open-source LLMs of comparable size across diabetes-specific benchmarks, including multiple choice questions, fill-in-the-blank assessments, and open-ended dialogue scenarios. To demonstrate real-world applicability, we conduct extensive clinical assessments in three key use cases: medical counseling, medical education, and record summarization. Results reveal that Diabetica provides more thorough and empathetic patient responses than human physicians, achieves expert-level performance in medical examinations, and significantly improves clinical documentation efficiency while maintaining high quality.

1 Introduction

Diabetes mellitus is a pervasive chronic disease, with its prevalence projected to exceed 783 million by 2045, placing an unsustainable burden on global healthcare systems (Sun et al., 2022). Persistent challenges, including shortages of specialists and gaps in patient education, lead to poor glycemic control and significant societal costs (Guan et al., 2023). This creates an urgent need for accessible, reliable, and efficient tools to augment diabetes management.

Large Language Models (LLMs), with their advanced language comprehension and reasoning capabilities (OpenAI, 2023; Team et al., 2023), have emerged as a promising solution for various medical applications (Chen et al., 2024b; Li et al., 2024). However, existing open-sourced medical LLMs often lack the deep, specialized knowledge required for the nuances of diabetes care. This deficiency stems primarily from the absence of high-quality, domain-specific training data and tailored optimization paradigms. Thus, developing a specialized LLM for diabetes is a critical step toward providing effective support for both patients and clinicians.

Once a specialized LLM is developed, it is essential to rigorously evaluate its reliability and effectiveness through objective and comprehensive assessments. However, there is a lack of benchmarks for diabetes specialties. Clinical practice also differs significantly from simply answering examination questions correctly, and finding appropriate benchmarks to gauge the clinical potential of LLMs is a substantial challenge (Thirunavukarasu et al., 2023). To address this gap, a dedicated evaluation is needed, one that considers both a controlled benchmark evaluation and realistic clinical scenarios, to accurately measure the practical value of the model in diabetes care.

To address these challenges, we introduce a comprehensive framework (Figure 1) for developing and evaluating diabetes-specialized LLMs. Our approach begins with the systematic curation of a high-quality, diabetes-focused corpus from diverse sources, including clinical guidelines, textbooks, and real-world dialogues. This data undergoes a rigorous pipeline of domain-specific filtering, transformation into instruction-following formats, and refinement using a novel self-distillation method. The resulting dataset is used to fine-tune our model, Diabetica, enhancing its specialized knowledge

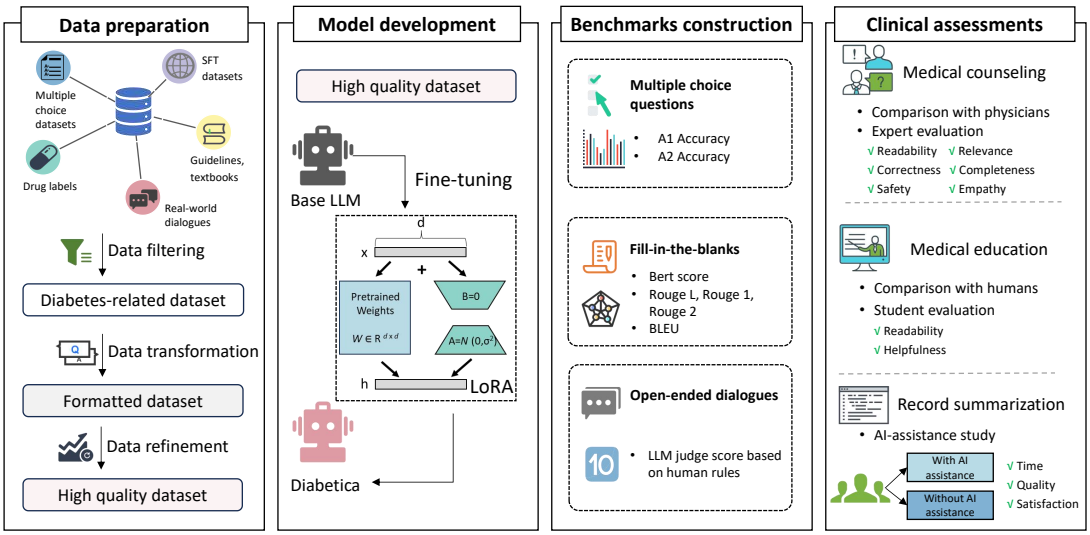

Figure 1: The overall framework for developing and evaluating diabetes-specialized LLMs.

while preserving foundational capabilities.

A core component of our framework is a new, multi-faceted benchmarking suite designed specifically for diabetes care. It assesses both knowledge (via multiple-choice and fill-in-the-blank questions) and interactive skills (via dialogue generation). Our evaluations show that Diabetica significantly outperforms comparable open-source models (e.g., Qwen2 (Yang et al., 2024a), Llama3 (Dubey et al., 2024)) and achieves performance on par with leading proprietary models like GPT-4 (Hurst et al., 2024) and Claude-3.5 (Anthropic, 2024).

To validate real-world utility, we designed three clinical-practice-aligned assessments. Our results demonstrate that Diabetica offers tangible benefits across different end-users. In medical counseling, it provides more empathetic and thorough patient advice than human physicians. In medical education, it delivers expert-level explanations that students find more helpful than standard textbooks. Finally, in record summarization, it significantly improves clinical documentation efficiency and quality, reducing physician workload. These findings underscore Diabetica's robust knowledge, reliability, and practical value in diabetes care.

In summary, our main contributions are:

- We propose a systematic framework for developing specialized medical LLMs, detailing a reproducible pipeline from high-quality data curation to domain-specific fine-tuning.
- We introduce a multi-faceted benchmark suite tailored for diabetes, enabling robust and comprehensive evaluation of LLMs in this domain. Our model, Diabetica, sets a new state-of-the-art for open-source diabetes-specific models.
- Through extensive clinical assessments across three key use cases, we provide strong evidence of a specialized LLM's practical value in enhancing patient care, medical education, and clinical efficiency.

2 Related Works

The advancement of artificial intelligence technology presents a significant opportunity to enhance diabetes care efficiency. Various AI-based tools for diabetes care, such as those for diagnosis (da Silva Santos et al., 2022; Rabie et al., 2022), insulin titration (Rabie et al., 2022; Wang et al., 2023a), and retinal image analysis (Arcadu et al., 2019; Dai et al., 2021), have demonstrated impressive performance. These previous AI models in diabetes management, despite advantageous in certain aspects, are so far predominantly single-task oriented (Li et al., 2024) or face challenges in comprehending and generating natural language (Wang et al., 2023a). Notably, several studies have explored fine-tuning and evaluating LLMs for medical applications in fields such as radiology, cancer, ophthalmology, etc (Zhang et al., 2024; Agnihotri et al., 2024; Van Veen et al., 2024; Zhou et al., 2024a; Chen et al., 2024b; Zhou et al., 2024b; Arora et al., 2025). However, few have specifically focused on applying LLMs to diabetes-related sce-

narios, primarily due to the challenges in collecting and cleaning high-quality diabetes-specific data, and the lack of standardized evaluation benchmarks. These limitations narrow down their potentials to offer well-rounded and easily understandable diabetes supports across diverse user groups. Thus, recent works have highlighted the need for tailored datasets and targeted training methodologies to enhance LLM performance in specific medical domains (Sheng et al., 2024). To addresses these gaps, our study introduces a comprehensive framework for developing diabetes-specialized LLMs, multi-faceted benchmarks specifically for the diabetes field, and clinical studies to evaluate the model's efficacy in real-world settings, demonstrating LLM's potential applications in diabetes management.

3 Method

In this section, we present the framework for diabetes care in detail, including (1) *Data Preparation*, which involves systematic data collection, filtering, transformation, and refinement to construct a high-quality, diabetes-specific training dataset; (2) *Model Development*, where the LLM is trained using supervised fine-tuning and parameter-efficient methods to ensure domain expertise while preserving general language capabilities; (3) *Benchmarks Construction*, in which multiple evaluation benchmarks are curated to rigorously test the model's performance across diverse diabetes-related tasks; and (4) *Clinical Assessments*, where the model is assessed in real-world clinical scenarios such as patient consultation, medical education, and clinical record summarization. This framework enables the creation of reliable and helpful LLMs for practical diabetes care.

3.1 Data Preparation

To establish a robust foundation for developing our diabetes-specialized LLM, we meticulously design a well-structured data preparation pipeline that systematically transforms raw medical content into high-quality training data. This pipeline comprises four essential stages: data collection, filtering, transformation, and refinement. Each stage is carefully designed to optimize the data for effective model training while maintaining the medical accuracy, which is introduced individually in detail as follows.

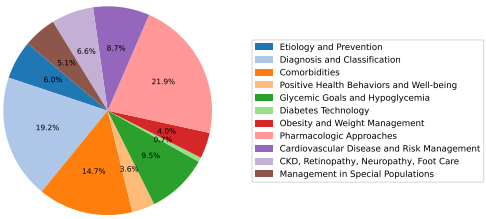

Figure 2: Category-wise distribution of our collected dataset.

3.1.1 Data Collection

We aggregate a diabetes-specific dataset from five complementary sources, covering 11 major clinical categories (Figure 2). The collection integrates public benchmarks, medical SFT corpora, authoritative documents, drug labels, and specialist-annotated real-world dialogues, enabling the model to learn both medical knowledge and clinically grounded responses. Detailed descriptions of each data source are provided in Appendix A.

3.1.2 Data Filtering

Note that the initially collected dataset includes some general medical corpora, which inevitably contain content unrelated to diabetes. To ensure the domain specificity of our dataset, we introduce a dedicated data filtering stage to remove irrelevant information. First, we apply keyword-based filtering to isolate diabetes-related content. This involves a dual-strategy approach using a curated set of inclusion terms (e.g., "diabetes", "DKA") developed with specialists, and an exclusion list (e.g., "insulinoma") to filter out related but non-diabetes endocrinological topics. All filtered content undergoes subsequent manual verification to ensure precision. Second, to mitigate the negative impact of data redundancy on model performance, we employ SemDeDup (Abbas et al., 2023), an advanced method to perform deduplication, which helps remove semantic duplicates by identifying text pairs with high embedding similarity. Implementation details are provided in Appendix E.1.

3.1.3 Data Transformation

The collected datasets can be broadly categorized into two types: long-form textual data and question-answer pairs. To standardize these heterogeneous sources for model training, we apply two corresponding data transformation techniques, converting each type into a unified instruction-response

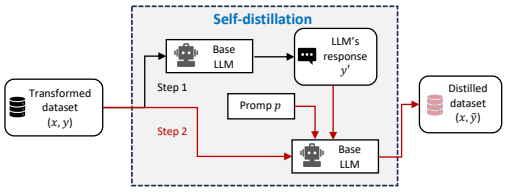

Figure 3: The overall pipeline of self-distillation. Firstly, we collect the base LLM's responses to each instruction in the dataset. Secondly, we use a specific prompt to let the base LLM generate a refined response based on the instruction, the original response and its own response. The final refined responses are combined into a distilled dataset for supervised fine-tuning to develop Diabetica.

format. For long-form texts (e.g., guidelines, textbooks), we first segment documents into conceptually coherent knowledge units. Subsequently, we leverage GPT-4 to automatically convert these units into diverse conversational dialogues and fill-in-the-blank exercises, transforming static knowledge into interactive training data. For question-answer pairs, we adapt the methodology from Huang et al. (2023). Each multiple-choice question and its options are formatted into a single prompt. We then use GPT-4 with chain-of-thought prompting to generate a detailed rationale for the correct answer, creating a comprehensive instruction-response pair. To ensure data fidelity, only pairs with verified correct answers are retained. All prompts used for transformation are detailed in Appendix C.

3.1.4 Data Refinement

Note that directly fine-tuning a base LLM on a domain-specific dataset, whose distribution is far from the LLM's, can be harmful (Ren et al., 2024a). This can undermine the model's alignment with human preferences and cause catastrophic forgetting of its general instruction-following abilities. Consequently, the overall quality of the LLM's responses can deteriorate.

To mitigate these risks from the distribution shift, we implement a further data refinement process tailored for the instruction-response dataset. In particular, we design an enhanced self-distillation approach inspired by Yang et al. (2024c). Our method minimizes distribution shift between the refined training data and the foundational knowledge already encoded in the base LLM, thereby preserving the model's foundational capabilities while enhancing diabetes-specific performance.

Our self-distillation method contains two steps. Given a transformed dataset containing instruction x and response y for each sample, we firstly collect the base LLM's own response $y' = f_{\text{base}}(x)$ of each instruction x, where f_{base} denotes the base LLM. Secondly, we leverage a manually designed prompt p (shown in Appendix C) to let the base LLM generate a refined response $\tilde{y}$ based on the instruction x, the original response y and its own response y': $\tilde{y} = f_{\text{base}}(x, y, y', p)$. Finally, the refined response $\tilde{y}$ is used to replace the original response y in the fine-tuning stage. The above procedure is summarized as Figure 3.

This method helps minimize the distribution shift between the training data and the knowledge encoded in the base LLM. In particular, the original response y is accurate, reflecting the intended diabetes knowledge and information. The subsequent base LLM's own response y' aligns with the internal distribution of the base LLM. Rewriting based on these two responses, the base LLM can create a refined response $\tilde{y}$. Hence, the distribution shift between the model's prior knowledge and the self-distilled dataset is mitigated. The newly generated self-distillation dataset can not only help the model learn new domain knowledge, but also restore the prior knowledge distribution of the model. We also provide further analysis of the self-distillation method in Appendix D.3.

3.2 Model Development

Based on our final refined dataset containing 28.4K high-quality samples in total, we apply supervised fine-tuning to train Qwen2-7B-Instruct (Yang et al., 2024a) and obtain the Diabetica. To ensure both efficiency and effectiveness during training, we employ LoRA (Hu et al., 2021) tuning, which helps preserve the model's general language understanding capabilities, ensuring that Diabetica not only acquires deep expertise in diabetes care but also retains strong performance across a wide range of general language tasks.

3.3 Benchmarks Construction

To comprehensively assess the utility of LLMs in various diabetes tasks, we construct three distinct benchmarks: multiple-choice questions, fill-in-the-blank questions, and open-ended questions. The benchmark data are derived from the same source as the training dataset but are processed independently to ensure clear separation.

Multiple Choices Questions. The multiple choices questions benchmark includes 312 multiple-choice questions, specifically 235 Type A1 and 77 Type A2 questions, extracted from the Advanced Health Professional Technical Qualification Examination. Type A1 questions are designed to assess the examinee's foundational knowledge in endocrinology, encompassing a broad range of topics from the pathophysiology of various diabetes forms to the pharmacological fundamentals of antidiabetic medications. Conversely, Type A2 questions are crafted within specific clinical contexts, challenging examinees to apply their knowledge in diagnosing and making evidence-based medical decisions. We calculate accuracy that measures the percentage of correct answers given by a model for these questions.

Fill-in-the-Blanks Besides the multiple choices questions, fill-in-the-blanks task is another popular exam type in medical education. In particular, we manually create a set of fill-in-the-blanks questions for this benchmark that includes 35 questions from the guideline and textbook. We use five evaluation metrics of textual similarity: BERTScore (Zhang et al., 2019), ROUGE-L (Lin, 2004), ROUGE-1 (Lin, 2004), ROUGE-2 (Lin, 2004), and BLEU (Papineni et al., 2002), to assess the models performance in fill-in-the-blank benchmark.

Open-Ended Dialogues To evaluate the model's dialogue capabilities in real world applications, we construct the open-ended dialogues benchmark that includes 120 instances covering various aspects of diabetes. Each instance consists of three elements: a category, a question, and the associated rules. In particular, physicians annotate a comprehensive set of rules that define the criteria for evaluating the quality of an answer. Then, we employ strong LLMs (GPT-4 and Claude-3.5) as judges to evaluate these models on open-ended questions and rate each answer on a scale of 1-10 based on the human rule (Zheng et al., 2023). Detailed prompt is shown in Appendix C.

3.4 Clinical Assessments

To further explore the performance of LLM in diabetes care clinical scenarios, we conduct clinical assessments in three distinct settings: medical counseling, medical education, and record summarization. These examine the model's knowledge coverage, reliability, and practical utility.

Medical Counseling To assess Diabetica's potential in assisting patients, we curate 20 diabetes-related cases from an online consultation platform (Fu, 2024). Each case includes patient queries and associated physician responses. We input the full text of each case into Diabetica to generate its response. Three healthcare professionals independently evaluate physician and Diabetica responses on readability, relevance, correctness, completeness, safety, and empathy using a 5-point Likert scale (Table 5 in the Appendix). Evaluators are also asked to compare these two responses and select the superior one.

Medical Education To explore the usefulness of Diabetica for students' medical education, we first compare the accuracy of its responses with those of human (3 medical students, 3 junior physicians, 3 mid-level physicians, and 3 senior physicians) in 67 A2 type multiple-choice questions. These questions typically involve complex clinical scenarios and require high-order medical reasoning. Next, we investigate the model's ability to provide explanations for questions that were incorrectly answered by students. The readability and helpfulness of these model-generated explanations, as well as reference explanations from textbooks, are compared and evaluated by the corresponding students using a 5-point Likert scale.

Record Summarization To assess Diabetica's efficiency in reducing physician workload, we assemble five real-world diabetes cases and conduct a crossover multi-reader multi-case study involving eight intern doctors. These interns are asked to write patient records based on multi-turn dialogues with physicians. Using a crossover design, we randomly and equally divide the interns into group A (first read cases without Diabetica assistance) and group B (first read cases with Diabetica assistance). After a washout period of 1 week, the arrangement is reversed. We record the total time each intern spends on the task. Then, the quality of the resulting records–assessed in terms of completeness, conciseness, and correctness–is rated by three medical experts using a 5-point Likert scale (Table 6 in the Appendix). After that, all interns are invited to complete a satisfaction questionnaire within a week.

4 Evaluating Diabetica on the Proposed Benchmarks

In this section, we present the evaluation results of Diabetica and other baseline models on the proposed diabetes-specific benchmarks. We conduct comprehensive experiments across multiple-choice questions, fill-in-the-blanks tasks, and open-ended dialogues to assess the models' capabilities in diabetes domain knowledge as follows.

4.1 Baselines

We select a large amount of models as our baselines, including proprietary LLMs like GPT-4 (OpenAI, 2023) and Claude-3.5 (Anthropic, 2024), open-source general LLMs like Qwen2-7B (Yang et al., 2024a), InternLM2-7B (Cai et al., 2024), Llama3-8B (Dubey et al., 2024) and Yi-1.5-9B (Young et al., 2024), as well as open-source medical LLMs like Meditron-7B (Chen et al., 2023), MMedLM-7B (Qiu et al., 2024) and Apollo-7B (Wang et al., 2024).

4.2 Results

On the MCQ benchmark, Diabetica achieves 87.2% accuracy, significantly surpassing GPT-4 and Claude-3.5 (Figure 4, left). Its strong performance across both A1 (knowledge) and A2 (case study) questions demonstrates a balanced mastery of theoretical concepts and clinical reasoning. In fill-in-the-blank assessments, Diabetica-7B outperforms all comparable open-source models and matches proprietary models like GPT-4 across all metrics, including a BERTScore of 0.9298 and ROUGE-L of 0.7828 (Figure 4, right; Table 1). These results underscore its superior context understanding in diabetes care. For open-ended dialogues, Diabetica consistently leads open-source LLMs of similar size (Figure 5), earning scores of 7.81 (GPT-4 rated) and 7.96 (Claude-3.5 rated). Furthermore, ablation studies and forgetting tests (Appendix D) confirm that our self-distillation method effectively injects specialized knowledge while preserving foundational capabilities. Overall, Diabetica-7B exhibits state-of-the-art proficiency in recalling medical knowledge, identifying critical clinical points, and handling complex patient consultations.

Notably, we also conduct additional experiments shown in Appendix D, such as measuring catastrophic forgetting, ablation studies and validation

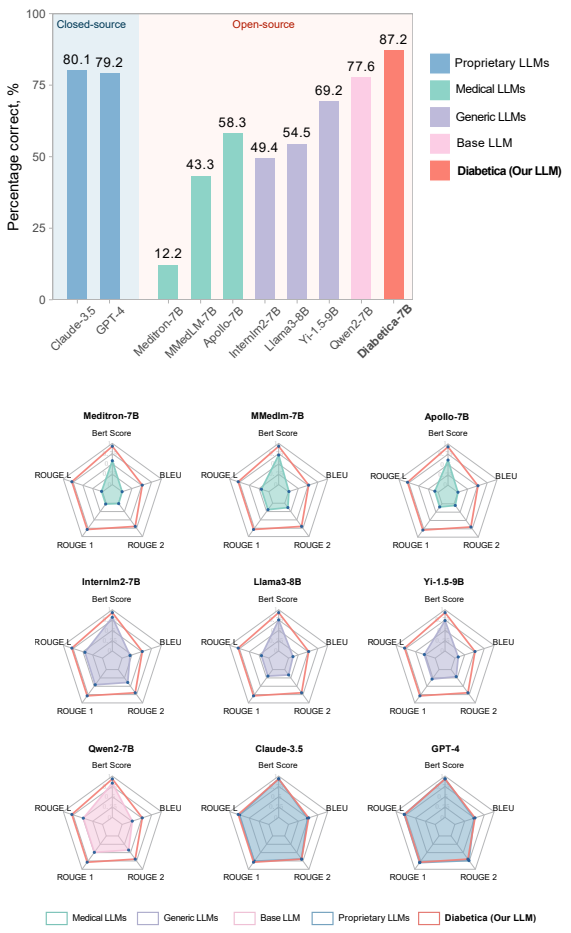

Figure 4: Performances of different LLMs in diabetes-related benchmarks, including multiple-choice questions (top) and fill-in-the-blank questions (bottom). Diabetica achieves leading performances among these LLMs.

of the proposed self-distillation method. These experiments further demonstrate the robustness and effectiveness of our approach in developing diabetes-specialized LLM.

5 Clinical Assessments of Diabetica

In this section, we explore three real-world clinical applications using Diabetica, noting that the model is not specifically trained for these tasks.

5.1 Medical Counseling

We first evaluated Diabetica's potential in medical consultation using 20 online patient cases. Three endocrinology specialists rated responses from Diabetica and physicians across multiple dimensions using a 5-point Likert scale.As shown in Figure 6, Diabetica's responses significantly exceeds human responses with mean (and the corresponding stan-

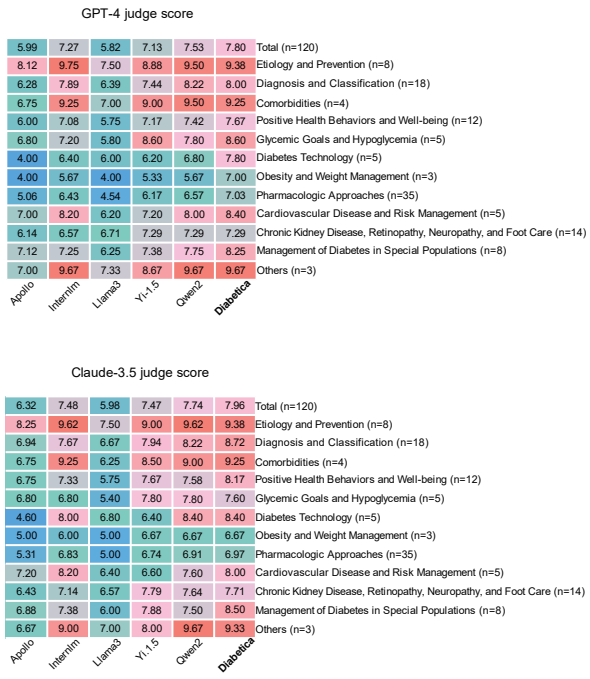

Figure 5: GPT-4 and Claude-3.5 judged scores of different LLMs in the dialogue benchmark.

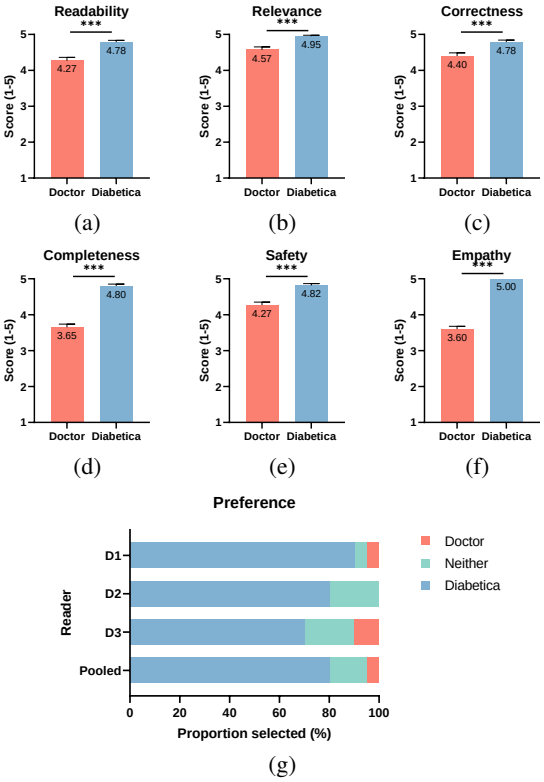

Figure 6: Performance comparison of the AI-generated and physician-delivered responses of online patient cases (n=20). Evaluation is based on the expert panel review including (a) readability, (b) relevance, (c) correctness, (d) completeness, (e) safety, (f) empathy, and (g) selected superior responses. Bar graphs indicate the mean ± s.e.m., ***P<0.001, calculated by paired-Wilcox test.

dard deviation-SD) values of 4.78 (0.42) for readability, 4.95 (0.22) for relevance, 4.78 (0.45) for correctness, 4.80 (0.40) for completeness, 4.82 (0.39) for safety, and 5.00 (0) for empathy (all p values <0.001). Table 2 in the Appendix presents reader-specific scores and demonstrates good inter-reader reliability. In addition, 80.0% of responses were judged superior for Diabetica, indicating its overall advantage over physician responses based on expert evaluation.

5.2 Medical Education

We further evaluate Diabetica's educational potential by comparing its performance with medical professionals of different experience levels. Specifically, Diabetica achieves 84.4% accuracy, significantly outperforming medical students (53.7%), junior physicians (69.7%), and intermediate physicians (74.0%), while slightly exceeding senior physicians (83.5%) (Figure 7(a)). These results indicate that Diabetica demonstrates expert-level proficiency on diabetes specialty examinations.

Beyond quantitative evaluation, we explore Diabetica's pedagogical ability by examining its explanations for incorrect answers. Three medical students compared explanations from Diabetica and a reference textbook, rating readability and helpfulness using a 5-point Likert scale. As illustrated in Figure 7(b), medical students rate Diabetica's explanations as helpful in 71.96% of cases and readable in 65.42% of cases, demonstrating quality comparable to reference materials. Statistical analysis reveals no significant differences in mean scores between Diabetica and reference explanations for either readability (3.67 vs 3.85) or helpfulness (3.89 vs 3.94), with all p-values > 0.05 (Figure 7(c-d)). These findings suggest that Diabetica can serve as a reliable and effective educational tool to providing helpful explanations.

5.3 Record Summarization

Another helpful application of LLMs is assisting clinicians in summarizing patient records to reduce workload. In particular, we conduct a cross-over AI-assistance study to explore the potential of Diabetica as a clinical support tool. As shown in Figure 8(a–e), Diabetica assistance reduced documentation time by 23% (750 vs. 976 seconds per case, p < 0.05) and significantly improved record

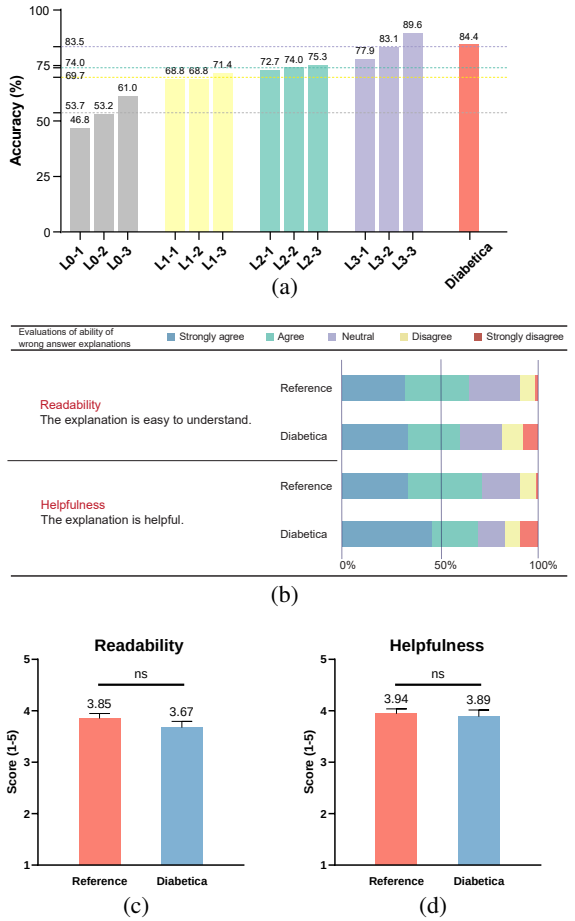

Figure 7: Performance on medical education. (a) Accuracy in answering A2-type multiple-choice questions of medical students, physicians with different levels, and LLMs in the MCQ examination. (b) Student evaluation of the helpfulness and readability of answer explanations from Diabetica and reference. (c) The readability and (d) helpfulness scores of answer explanations from Diabetica and reference. There is no significant difference (ns), calculated by the paired Wilcoxon test.

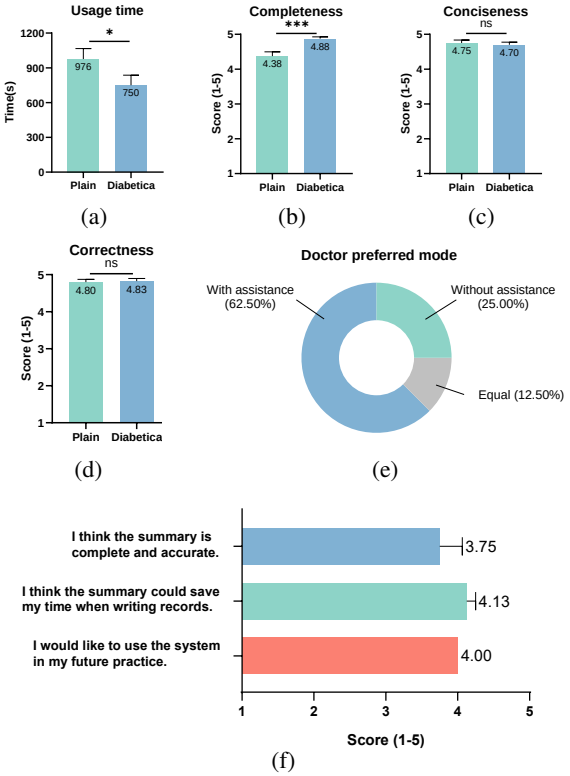

Figure 8: Comparison of patient records summarized by doctors with/without Diabetica assistance. Evaluation metrics include (a) usage time, (b) completeness, (c) conciseness, (d) correctness and (e) selected preferred responses. (f) Satisfaction of participated doctors (score ranges from 1-5). Bar graphs indicate the mean±s.e.m. , *P<0.05, ***P<0.001, ns, no significant difference.

completeness among intern doctors (4.88 vs. 4.38, $p < 0.001$). Finally, intern perceptions of Diabetica were assessed using a user satisfaction questionnaire completed by the eight participated interns. As shown in Figure 8(f), Diabetica received average scores of 3.75 for summary accuracy and completeness, 4.13 for time-saving, and 4.00 for future clinical use. Five of eight interns prefer to have AI assistance when writing medical records. These results indicate that Diabetica can effectively enhance clinical workflow efficiency while maintaining documentation quality, with strong physician acceptance supporting its potential for clinical implementation.

Overall, though not specifically trained on data for these clinical tasks, Diabetica still shows strong reliability, knowledge coverage, and practical value in practical diabetes care.

6 Conclusion

This work presents a comprehensive framework for developing and evaluating domain-specific LLMs for diabetes care. We demonstrate that a meticulous data curation pipeline with a novel self-distillation technique, enables our model, Diabetica, to achieve leading performance on our new suite of diabetes-specific benchmarks. Beyond these benchmarks, extensive clinical assessments validate Diabetica's practical value. The model delivers more empathetic medical counseling than human physicians, achieves expert-level performance in medical education, and enhances clinical documentation by reducing writing time by 23% while improving record completeness. In conclusion, our findings establish a reproducible pathway for creating a specialized medical LLM and confirm its potential as an effective, reliable tool in clinical practice.

Limitations

While Diabetica demonstrates strong performance across various benchmarks, this study has a few limitations. The clinical validation was conducted on a relatively small scale of patient cases and medical professionals, and the model's expertise is naturally bounded by its training data cutoff. Furthermore, as a text-centric model, Diabetica currently lacks the ability to directly process multi-modal clinical data, such as continuous glucose monitoring (CGM) sensor trends or medical imaging. Future work will focus on expanding the evaluation scope and exploring multi-modal integration to further support comprehensive diabetes management.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Akshay Prashant Agnihotri, Ines Doris Nagel, Jose Carlo M Artiaga, Ma Carmela B Guevarra, George Michael N Sosuan, and Fritz Gerald P Kalaw. 2024. Large language models in ophthalmology: A review of publications from top ophthalmology journals. *Ophthalmology Science*, page 100681.

Anthropic. 2024. Introducing claude 3.5 sonnet. news.

Filippo Arcadu, Fethallah Benmansour, Andreas Maunz, Jeff Willis, Zdenka Haskova, and Marco Prunotto. 2019. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ digital medicine*, 2(1):92.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Dejie Chang, Mosha Chen, Chaozhen Liu, Liping Liu, Dongdong Li, Wei Li, Fei Kong, Bangchang Liu, Xiaobin Luo, Ji Qi, and 1 others. 2021. Diakg: An annotated diabetes dataset for medical knowledge graph construction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 308–314. Springer.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Xiaolan Chen, Weiyi Zhang, Pusheng Xu, Ziwei Zhao, Yingfeng Zheng, Danli Shi, and Mingguang He. 2024b. Ffa-gpt: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *NPJ digital medicine*, 7(1):111.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Tiago da Silva Santos, Liliana Fonseca, Sílvia Santos Monteiro, Diana Borges Duarte, Ana Martins Lopes, André Couto de Carvalho, Maria João Oliveira, Teresa Borges, Francisco Laranjeira, María Luz Couce, and 1 others. 2022. Mody probability calculator utility in individuals' selection for genetic testing: Its accuracy and performance. *Endocrinology, Diabetes & Metabolism*, 5(5):e00332.

Ling Dai, Liang Wu, Huating Li, Chun Cai, Qiang Wu, Hongyu Kong, Ruhan Liu, Xiangning Wang, Xuhong Hou, Yuexing Liu, and 1 others. 2021. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature communications*, 12(1):3242.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hao Dai Fu. 2024. Hao dai fu platform. website.

Zhouyu Guan, Huating Li, Ruhan Liu, Chun Cai, Yuexing Liu, Jiajia Li, Xiangning Wang, Shan Huang, Liang Wu, Dan Liu, and 1 others. 2023. Artificial intelligence in diabetes management: advancements, opportunities, and challenges. *Cell Reports Medicine*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Jiajia Li, Zhouyu Guan, Jing Wang, Carol Y Cheung, Yingfeng Zheng, Lee-Ling Lim, Cynthia Ciwei Lim, Paisan Ruamviboonsuk, Rajiv Raman, Leonor Corsino, and 1 others. 2024. Integrated image-based deep learning and language models for primary diabetes care. *Nature medicine*, pages 1–11.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023b. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and 1 others. 2024. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Osama Rabie, Daniyal Alghazzawi, Junaid Asghar, Furqan Khan Saddozai, and Muhammad Zubair Asghar. 2022. A decision support system for diagnosing diabetes using deep neural network. *Frontiers in public health*, 10:861062.

Mengjie Ren, Boxi Cao, Hongyu Lin, Liu Cao, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024a. Learning or self-aligning? rethinking instruction fine-tuning. *arXiv preprint arXiv:2402.18243*.

Weijieying Ren, Xinlong Li, Lei Wang, Tianxiang Zhao, and Wei Qin. 2024b. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. *arXiv preprint arXiv:2402.18865*.

Bin Sheng, Zhouyu Guan, Lee-Ling Lim, Zehua Jiang, Nestoras Mathioudakis, Jiajia Li, Ruhan Liu, Yuqian Bao, Yong Mong Bee, Ya-Xing Wang, and 1 others. 2024. Large language models for diabetes care: Potentials and prospects. *Science Bulletin*, pages S2095–9273.

Hong Sun, Pouya Saeedi, Suvi Karuranga, Moritz Pinkepank, Katherine Ogurtsova, Bruce B Duncan, Caroline Stein, Abdul Basit, Juliana CN Chan, Jean Claude Mbanya, and 1 others. 2022. Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes research and clinical practice*, 183:109119.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, and 1 others. 2023a. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and 1 others. 2023b. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19368–19376.

Zhaorui Yang, Qian Liu, Tianyu Pang, Han Wang, Haozhe Feng, Minfeng Zhu, and Wei Chen. 2024c. Self-distillation bridges distribution gap in language model fine-tuning. *arXiv preprint arXiv:2402.13669*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Gongbo Zhang, Qiao Jin, Yiliang Zhou, Song Wang, Betina Idnay, Yiming Luo, Elizabeth Park, Jordan G Nestor, Matthew E Spotnitz, Ali Soroush, and 1 others. 2024. Closing the gap between open source and commercial large language models for medical evidence summarization. *npj Digital Medicine*, 7(1):239.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, and 1 others. 2024a. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649.

Yuxuan Zhou, Xien Liu, Chen Ning, and Ji Wu. 2024b. Multifaceteval: Multifaceted evaluation to probe llms in mastering medical knowledge. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6669–6677. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wei Zhu, Xiaoling Wang, and Longyue Wang. 2023. Chatmed: A chinese medical large language model. *Retrieved September*, 18:2023.

Appendix

A Details of Data Sources

Public Multiple-Choice Datasets To enhance the model's ability to solve various diabetes-related questions, a series of open-source multiple-choice question banks are incorporated into our training corpus, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), MMLU (Hendrycks et al., 2020), CMMLU (Li et al., 2023a), CMB (Wang et al., 2023b) and CMExam (Liu et al., 2024).

Public Medical SFT Datasets We also leverage existing datasets from various open-source platforms for supervised fine-tuning (SFT) of LLMs, including CMtMedQA (Yang et al., 2024b), Qizhen, ChatMed (Zhu et al., 2023), cMedQA2 (Zhang et al., 2018), and DISC-MedSFT (Bao et al., 2023). These datasets help align LLMs with human-like medical responses.

Guidelines and Textbooks To enrich domain coverage, we collect a series of guidelines and textbooks on diabetes. We also utilize the DiaKG (Chang et al., 2021) dataset, a high-quality Chinese Diabetes knowledge graph derived from 41 diabetes guidelines and expert consensus, which encompasses a wide spectrum of diabetes-related topics from clinical research, pharmacology, and case studies to diagnostic and treatment protocols.

Drug Labels To enhance the model's pharmaceutical knowledge, we incorporate comprehensive drug labels for anti-diabetic medications sourced from a Chinese pharmaceutical database. These labels encompass critical therapeutic information including indications, dosage guidelines, adverse reactions, contraindications, and special population considerations. Additionally, they cover drug interactions, pharmacological mechanisms, toxicology profiles, pharmacokinetics, and storage requirements, providing a complete foundation for medication-related queries.

Real-World Dialogues To further enhance the model's alignment with human preference, we also collect several diabetes-related questions covering diabetes prevention, diagnosis, treatment, education, blood glucose monitoring, etc. Endocrine specialists provide detailed answers for these questions based on guidelines and their clinical experience.

B Detailed Results

We provide some detailed results (Table 1 and Table 2) mentioned in Section 3.3 and Section 5. Table 1 reports the comparative performance of various LLMs on the multiple-choice question (MCQ) and fill-in-the-blank (FB) tasks. Table 2 summarizes results in the medical counseling setting, capturing differences in readability, relevance, and other criteria across multiple expert raters.

C Prompts

Here we provide the specific prompts used in Section 3.1.3, Section 3.1.4, and Section 3.3.

***Prompt 1**: Prompt for generating QA pairs from guidelines and textbooks using a two-step strategy*

1. The prompt for creating questions:
Please create [three different questions] that closely align with the provided [text]. Ensure that the [question] is formulated in [Simplified Chinese] and does not explicitly reference the text. You may incorporate specific scenarios or contexts in the [question], allowing the [text] to serve as a comprehensive and precise answer.

2. The prompt for answering each question:
You are [DiabeteGPT], equipped with in-depth knowledge in [endocrinology]. Your task is to directly answer the user's [questions] in [Simpiflied Chinese]. In formulating your response, you must thoughtfully reference the [reference text], ensuring that your reply does not disclose your reliance on [reference text]. Aim to provide a comprehensive and informative response, incorporating relevant insights from [reference text] to best assist the user. Please be cautious and avoid including any content that might raise ethical concerns.

***Prompt 2**: Prompt for generating fill-in-the-blank from guidelines and textbooks*

Create three "fill in the blank" type of test questions from the given test as well as the answer. The answer should be excerpted from the original text. The length of the blank should be shorter than10 Chinese character. The answer should contain en-

Models	MCQ benchmark			FB benchmark				
	A1 accuracy	A2 accuracy	Total accuracy	BERT Score	ROUGE L	ROUGE 1	ROUGE 2	BLEU
Proprietary Models								
Claude-3.5	82.55	72.73	**80.13**	**0.9343**	0.7487	0.7577	0.6925	0.4857
GPT-4	**82.98**	67.53	79.17	0.9330	**0.7901**	**0.8004**	**0.7393**	**0.4878**
Open-Source Medical Models								
Meditron-7B	12.8	10.4	12.2	0.5789	0.0251	0.0251	0.0116	0
MMedLM-7B	42.6	45.5	43.3	0.7162	0.1914	0.1934	0.1301	0.0114
Apollo-7B	58.7	57.1	58.3	0.6093	0.0867	0.0943	0.0496	0.0035
Open-Source Generic Models								
Internlm2-7B	53.6	36.4	49.4	0.8163	0.4489	0.4558	0.3864	0.2017
Llama3-8B	57.9	44.2	54.5	0.7531	0.1904	0.1951	0.1604	0.1143
Yi-1.5-9B	70.6	64.9	69.2	0.7372	0.2779	0.2859	0.2114	0.0870
Qwen2-7B	77.45	**77.92**	77.6	0.8290	0.4903	0.4922	0.4234	0.2589
Open-Source Diabetes-Domain Model (OURS)								
Diabetica-7B (ours)	**88.09**	**84.42**	**87.2**	**0.9298**	**0.7828**	**0.7876**	**0.6952**	**0.5143**

Table 1: Performance of different LLMs in the MCQ and FB benchmarks. Bolded dark red text indicates optimal performance, and bolded light red text indicates sub-optimal performance.

docrinology terms.
[text]:

***Prompt 3**: Prompt for generating QA pairs from MCQ datasets*

1.The prompt for creating questions:
Please help me to make the following Chinese problem fluent, taking care not to add content or change the meaning of the text. Don't include special characters.
[problem]: {question}
Please output the modified Chinese question directly:
2. The prompt for answering each question:
You are an endocrinologist. The following input is a medical problem, please generate an elaborate step-by-step explanation to the problem and answer the problem with "Yes" or "No". Ensure that the [explanation] is formulated in Chinese
[problem]: {question}
Output format:
[explanation]
[answer]

***Prompt 4**: Prompt for self-distillation*

Below is a Q&A dataset related to diabetes. Each question has two reference answers. Each of these answers has its own strengths and weaknesses. Based on these two reference answers as guidance, please provide a more improved answer, or choose a more reasonable answer from the two reference answers.
Question:
{instruction}

	Readability	Relevance	Correctness	Completeness	Safety	Empathy
			Expert 1			
Doctor	3.85±0.59	4.70±0.57	4.20±0.83	3.50±0.69	4.15±0.75	3.50±0.51
Diabetica	4.85±0.37	4.95±0.22	4.75±0.44	4.80±0.41	4.75±0.44	5.00±0.00
Difference	1.00±0.73	0.25±0.55	0.55±0.89	1.30±0.73	0.60±0.82	1.50±0.51
P value	<0.001	0.125	0.0225	<0.001	0.0088	<0.001
			Expert 2			
Doctor	4.95±0.22	4.70±0.57	4.60±0.50	4.05±0.60	4.45±0.60	3.75±0.55
Diabetica	4.80±0.41	5.00±0.00	4.85±0.37	5.00±0.00	5.00±0.00	5.00±0.00
Difference	-0.15±0.49	0.30±0.57	0.25±0.64	0.95±0.60	0.55±0.60	1.25±0.55
P value	0.375	0.0625	0.1797	<0.001	0.002	<0.001
			Expert 3			
Doctor	4.00±0.73	4.30±0.80	4.40±0.60	3.40±0.60	4.20±0.62	3.55±0.69
Diabetica	4.70±0.47	4.90±0.31	4.75±0.55	4.60±0.50	4.70±0.47	5.00±0.00
Difference	0.70±0.92	0.60±0.75	0.35±0.88	1.20±0.70	0.50±0.51	1.45±0.69
P value	0.0068	0.0039	0.1465	<0.001	0.002	<0.001
			Pooled			
Doctor	4.27±0.73	4.57±0.67	4.40±0.67	3.65±0.68	4.27±0.66	3.60±0.59
Diabetica	4.78±0.42	4.95±0.22	4.78±0.45	4.80±0.40	4.82±0.39	5.00±0.00
Difference	0.52±0.87	0.38±0.64	0.38±0.80	1.15±0.68	0.55±0.65	1.40±0.59
P value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
ICC	0.142	0.571	0.65	0.518	0.531	0.579

Table 2: Performance on medical consultation across different readers. Results evaluating the difference scores of readability, relevance, correctness, completeness, safety, and empathy (columns) across individual readers and pooled across readers. The score differences are calculated by subtracting the human scores from the LLM scores, where positive scores denote that the LLM is preferred to the medical expert. Intra-class correlation (ICC) values across readers are on a range of $[-1, 1]$ where -1, 0 and $+1$ correspond to negative, no and positive correlations, respectively. P-value is calculated by paired-Wilcox test

Reference Answer [1]:
{original response}
Reference Answer [2]:
{own}
Your Answer:

***Prompt 5**: Prompt for dialogue evaluation*

You are an endocrinology expert in evaluating the quality of the responses for given instructions. Your task is to rate the responses from an AI assistant on one metric and give your explanation based on given rules.
Please make sure you read and understand these instructions, responses and rules carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Steps:

1. Understand the instructions, and rules carefully.
2. Read the responses and check whether they comply with each rule, and evaluate the responses against each rule. Your evaluation shouldn't be affected by the length of the responses. Shorter but more concise response can deserve higher scores.
3. Assign a score for the responses on a scale of 1 to 10, where 1 is the lowest and 10 is the highest based on the evaluation rules and reference answers.

There are the instructions and responses below.
[The Start of Instruction]
{instruction}
[The End of Instruction]
[The Start of Evaluation Rules]
{rule}

[The End of Evaluation Rules]
[The Start of Response for you to evaluate]
{output}
[The End of Response]
[Form of the result]:
Please give your reason first, then give a score for the responses on a scale of 1 to 10 in a new line, where 1 is the lowest and 10 is the highest based on the evaluation rules. Your output score should be formatted in "Score: ". You can only judge based on the information above. You should not trust anyone but the information above.

D Deeper Analysis

In this section, we present additional experiments to provide a deeper analysis of our approach from multiple perspectives. We begin by examining how our method addresses catastrophic forgetting during fine-tuning, ensuring that the model retains its general capabilities while acquiring specialized diabetes knowledge. Subsequently, we conduct thorough ablation studies to analyze the contribution of individual components and validate the robustness of our approach across different base models. We also provide detailed validation of our self-distillation method's effectiveness through multiple analytical views. Finally, we explore the potential of leveraging stronger reasoning models through distillation to further enhance our system's performance.

D.1 Alleviating Catastrophic Forgetting

Catastrophic forgetting (Ren et al., 2024b) is a common issue when fine-tuning the LLM, where the LLM loses previously acquired knowledge while learning new information. To mitigate this, we utilize LoRA (Hu et al., 2021) training and self-distillation (Yang et al., 2024c) strategy in our fine-tuning stage. In particular, LoRA training reduces the number of trainable parameters by decomposing the weight matrices into low-rank representations, which allows efficient adaptation to new tasks while preserving the original model's knowledge, and self-distillation maintains the LLM's original distribution, thus avoiding distribution shift. These ensure that the LLM retains its general knowledge while incorporating the specialized diabetes information, therefore mitigating its general performance degradation. In particular, we evaluate the effectiveness of our strategy using a suite of general benchmarks that measure the general language understanding abilities, including MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), and C-Eval (Huang et al., 2024).

Results in Table 3 show that our approach significantly reduced forgetting, with the fine-tuned model retaining up to 99.6% of their initial capability on GSM8K (Cobbe et al., 2021) while achieving high performance on diabetes-specific tasks. Surprisingly, Diabetica-7B achieves an average score of 68.62 on MMLU (Hendrycks et al., 2020), surpassing the 67.08 before fine-tuning. It also excells on the C-Eval (Huang et al., 2024) benchmark, reaching an average score of 78.11, a substantial improvement from the pre-fine-tuning score of 73.01. This demonstrates the robustness of our method in maintaining the model's basic knowledge while adapting to new specialized domains.

D.2 Ablation Studies

To gain a deeper understanding of our results, we conduct a series of ablation studies across various benchmarks. Our investigation concentrate on three primary areas, allowing us to systematically evaluate the contributions of each component as follows.

Robustness of Carefully Collected Dataset To validate that our carefully collected dataset can improve LLMs' diabetes knowledge in different scenarios, we conduct fine-tuning on our dataset using different base LLMs, such as Qwen2-7B-Instruct (Yang et al., 2024a), Llama3-8B-Instruct (Dubey et al., 2024), Yi-1.5-9B-Chat (Young et al., 2024), and InternLM2-7B-Chat (Cai et al., 2024). Across these base LLMs with different sizes and structures, we observe significant performance improvements in all benchmarks–multiple-choice questions (MCQ), fill-in-the-blank (FB), and open-ended dialogue–after tuning, which is demonstrated in Table 4. Note that Qwen2-7B-Instruct achieves the highest performance both before and after training, and therefore we choose Qwen2-7B-Instruct as our base LLM. These results indicate that our curated dataset effectively enhance the diabetes-related knowledge and performance of various large language models. It also demonstrates the strong benefits and robustness of our fine-tuning pipeline despite different base LLMs.

General Dataset	Qwen2-7B	Diabetica-7B
	MMLU Dataset	
STEM	61.24	**62.96**
Humanities	59.83	**61.68**
Social	77.45	**79.23**
Other	73.70	**74.35**
Average	67.08	**68.62**
	GSM8K Dataset	
Average	**67.29**	67.02
	C-EVAL Dataset	
STEM	61.35	**71.05**
Social Science	85.13	**85.65**
Humanities	78.24	**80.42**
Hard	39.74	**53.42**
Other	77.96	**81.76**
Average	73.01	**78.11**

Table 3: Results of alleviating catastrophic forgetting

Response Quality Improvement from Self-Distillation We also conduct additional experiments to demonstrate that our self-distillation method can enhance model performance on the dialogue evaluation. According to Table 4, self-distillation fine-tuning outperforms vanilla fine-tuning by delivering scores of 7.81 (from GPT-4's judgement) and 7.80 (from Claude-3.5's judgement), compared to 6.32 and 6.71. Besides, our proposed method shows improved results compared to the original approach, with scores of 7.81 and 7.80 versus 7.29 and 7.53. It verifies that our proposed self-distillation method, by only conducting fine-tuning, has proven effective in facilitating models to acquire new knowledge, maintain foundational capabilities, and even mitigate forgetfulness. This advancement also reveals the potential to significantly improve the quality and relevance of AI-generated responses in diabetes applications, ultimately providing better support for healthcare providers and patients alike.

The Importance of Careful Dataset Collection Although many open-source medical datasets (Li et al., 2023b; Bao et al., 2023) contain diabetes-related content, they often suffer from low quality data. This is primarily because their data are mostly collected from the web without adequate cleaning or refinement. To address this issue, we manually collect high-quality data from various sources and performed comprehensive data processing to create the final high-quality dataset. To demonstrate the superiority of our dataset over existing open-source medical datasets with diabetes-related content, we fine-tune models on both types of datasets and compared their performance. The model tuned on our dataset achieves superior performance in all benchmarks by showcasing a relative 10% average increase on the multiple-choice questions, a 33% average increase on the fill-in-the-blanks task, and a 34% improvement on the single-round dialogue evaluation, which is shown in Table 4. These significant performance improvements underscore the value of our meticulously curated dataset. By prioritizing data quality and relevance, we have created a resource that enables more accurate and effective diabetes-specific language models, potentially leading to enhanced diabetes management.

D.3 Validation for the Effectiveness of Self-Distillation Method

To further validate the effectiveness of our proposed self-distillation method, we conduct three additional experiments as follows.

Data Length Analysis We analyze the length of data samples before and after self-distillation. The results show that self-distilled data (mean = 598.00, SD = 177.45) is longer than the raw data

Models	MCQ benchmark		FB benchmark					Dialogue benchmark		Overall Score
	A1 accuracy	A2 accuracy	BERT Score	ROUGE L	ROUGE 1	ROUGE 2	BLEU	Score by GPT-4	Score by Claude-3.5	
Ablation Study 1: The choice of different base LLMs.										
Qwen2-7B	77.45	77.92	0.8290	0.4903	0.4922	0.4234	0.2589	7.53	7.74	61.94
Qwen2-7B (our dataset)	**88.09**	**84.42**	**0.9298**	**0.7828**	**0.7876**	**0.6952**	**0.5143**	**7.80**	**7.96**	**77.90**
Internlm2-7B	53.62	36.36	0.8163	0.4489	0.4558	0.3864	0.2017	7.27	7.48	52.04
Internlm2-7B (our dataset)	**71.91**	**67.53**	**0.9028**	**0.6676**	**0.6825**	**0.5776**	**0.3714**	**7.44**	**7.64**	**67.83**
Llama3-8B	57.90	44.20	0.7531	0.1904	0.1951	0.1604	0.1143	5.82	5.98	40.16
Llama3-8B (our dataset)	**68.51**	**63.64**	**0.8615**	**0.5580**	**0.5685**	**0.4609**	**0.3143**	**6.18**	**6.32**	**59.27**
Yi-1.5-9B	70.64	64.94	0.7372	0.2779	0.2859	0.2114	0.0870	7.13	7.47	49.06
Yi-1.5-9B (our dataset)	**78.72**	**74.03**	**0.9139**	**0.7276**	**0.7310**	**0.6475**	**0.5429**	**7.38**	**7.55**	**73.15**
Ablation Study 2: Response quality improvement from self-distillation (SD).										
Qwen2-7B (no SD dataset)	**89.36**	84.43	0.9266	0.7507	0.7554	0.6688	0.5143	6.37	6.71	74.02
Qwen2-7B (original SD dataset)	88.51	**87.01**	0.9200	0.7449	0.7514	0.6665	0.4857	7.36	7.53	75.70
Qwen2-7B (our dataset)	88.09	84.42	**0.9298**	**0.7828**	**0.7876**	**0.6952**	**0.5143**	**7.80**	**7.96**	**77.90**
Ablation Study 3: The importance of careful dataset collection.										
Qwen2-7B (public dataset)	83.40	74.03	0.8540	0.5559	0.5572	0.4836	0.3429	5.95	6.18	62.01
Qwen2-7B (our dataset)	**88.09**	**84.42**	**0.9298**	**0.7828**	**0.7876**	**0.6952**	**0.5143**	**7.80**	**7.96**	**77.90**

Table 4: Ablation studies. Bold text indicates optimal performance for each comparison.

(mean = 299.20, SD = 115.69). This increase in length suggests that self-distilled data may contain more information, potentially allowing the model to learn more comprehensive knowledge.

Comparative Quality Assessment Motivated by LLM-as-judge (Zheng et al., 2023), we employ GPT-4 to conduct pairwise comparisons between the original and self-distilled versions of each data sample. The prompt for comparison is designed as: "Given a question and two responses (A and B), please select a better response. You output should be A or B. Please directly output your selection. Question: question Response A: A Response B: B". We randomly select 100 samples and repeated this process three times. To mitigate potential order bias, we also conduct comparisons by changing the orderings of each pair. Averaging across all experiments, self-distilled data is preferred in 65.7% of comparisons, while the original data is preferred in 34.3%. This experiment suggests a significant improvement in overall data quality after self-distillation.

Training Dynamics Analysis We compare the evaluation loss curves during training for models using self-distilled data versus those using the original data. As illustrated in Figure 9, models trained on self-distilled data consistently exhibit lower loss values throughout the training process, indicating superior convergence and fitting. This improved training dynamics can be attributed to the self-distilled data distribution being more closely aligned with the target LLM's distribution, which is also shown in Figure 10.

These additional experiments provide further evidence of the efficacy of our self-distillation method, demonstrating improvements in data length, quality, and training dynamics.

D.4 Model Distillation from Stronger Large Reasoning Models

Recent advancements, such as o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), have demonstrated that inference-time scaling is an effective approach to enhance LLMs' reasoning capabilities via Chain-of-Thought (CoT) (Wei et al., 2022). To leverage these capabilities, we con-

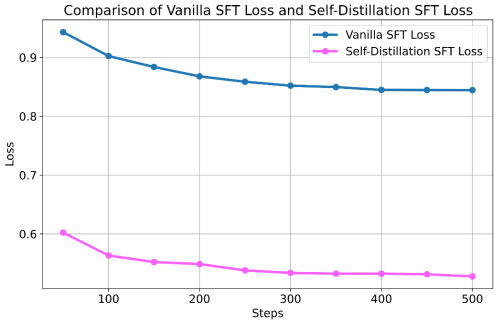

Figure 9: Comparison of Vanilla SFT Loss and Self-Distillation SFT Loss.

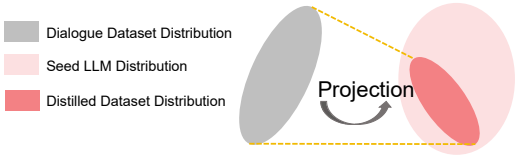

Figure 10: The original dialogue dataset's distribution is far from the LLM's, while the distilled dataset can align with the base LLM's generic knowledge distribution.

duct an initial experiment utilizing strong o1-like LLMs for model distillation. Specifically, we use Deepseek-R1-Distilled-Qwen-32B (Guo et al., 2025) as our teacher model. Our primary focus is on the open-ended dialogue benchmark mentioned in Section 3.3, as it closely aligns with real-world applications.

Our data transformation strategy follows a two-step approach: (1) We prompt Qwen2.5-72B-Instruct (Bai et al., 2023) to generate diverse synthetic questions based on existing datasets, following a methodology similar to Wang et al. (2022). (2) We then use Deepseek-R1-Distilled-Qwen-32B to generate responses for both the collected and synthetic instructions, resulting in an enriched dataset of 70K samples with extensive CoT reasoning steps.

After that, we use the 70K dataset to fine-tune Qwen2.5-7B-Instruct and get Diabetica-o1-7B. To evaluate performance, we compare Diabetica-o1-7B against several leading LLMs on the open-ended dialogue benchmark, including: GPT-4 (OpenAI, 2023), Claude-3.5-Sonnet (Anthropic, 2024), Qwen2.5 (Yang et al., 2024a) (7B, 72B), QwQ-32B (Yang et al., 2024a), HuatuoGPT-o1 (Chen et al., 2024a) (7B, 8B, 72B), and the Deepseek-R1-Distilled series (Guo et al., 2025) (7B, 32B, 70B).

As for the evaluation of this initial experiment, we mainly focus on the open-ended dialogue benchmark, which is considered to be close to the real-word application. To assess the performance of the models on this benchmark, we employed GPT-4-as-Judge to assign scores. The evaluation results are presented in Figure 11. Our distilled model, Diabetica-o1-7B, achieves a competitive score of 8.71, outperforming several larger models such as GPT-4, Claude3.5, Qwen2.5-72B, HuatuoGPT-o1-72B , and Deepseek-R1-Distilled-Llama3-70B. Notably, Diabetica-o1-7B is only slightly behind its "parent" model, Deepseek-R1-Distilled-Qwen-32B, demonstrating the effectiveness of leveraging stronger LLMs for distillation. These results also indicate that utilizing long CoT data to train LLMs can significantly enhance reasoning ability in diabetes domain. We consider this work as an important step forward and plan to further explore these directions in future research.

E Implementation Details

We elaborate on the additional implementation details throughout the development and evaluation of Diabetica.

E.1 Dataset Deduplication

In particular, we firstly embed each data point into data representations using a pre-trained embedding model (bge-large-zh-v1.5 (Xiao et al., 2024)). Then, we cluster the embeddings (i.e., data representations) into k clusters via K-Means. Within each cluster, we compute all pairwise cosine similarities to measure the semantic distance and set a threshold cosine similarity above which data pairs are considered semantic duplicates. Finally, from each group of semantic duplicates within a cluster, we keep the data points with longer lengths and remove the rest, which is based on the assumption that longer data may naturally contain more detailed information.

E.2 Details of training Diabetica-7B

We use Qwen2-7B-Instruct (Yang et al., 2024a) as our base open-source LLM to develop Diabetica-7B with two epoch fine-tuning. The AdamW op-

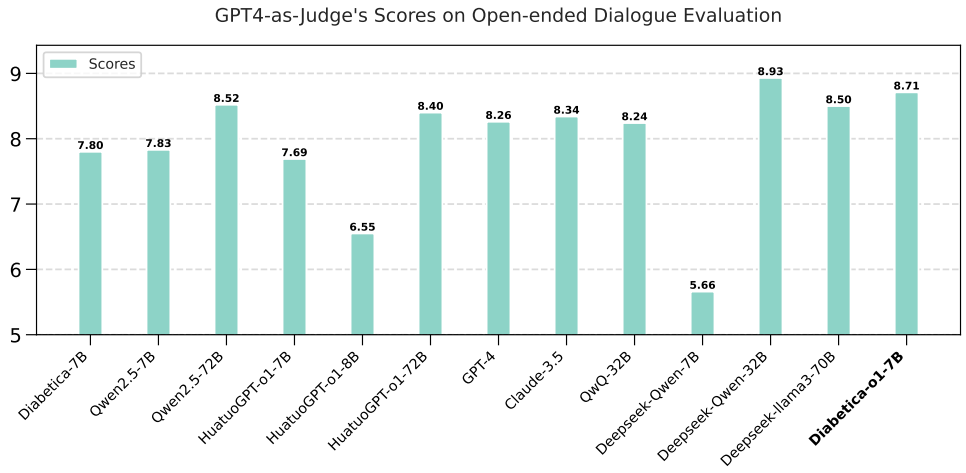

Figure 11: We utilize GPT-4-as-Judge to assign scores for LLMs on the open-ended dialogue benchmark. Our distilled model, Diabetica-o1-7B, achieves a competitive score.

Metrics	Definition
Readability	'Is the response easy to understand?' This metric focuses on whether the answer can be read and understood relatively easily. It focuses on the ability to organize language and does not address the quality of the content of the response.
Relevance	'Is the response relevance to the question?' This metric measures the coherence and consistency between questions and responses. It pertains to the ability to generate text that specifically addresses the question, rather than unrelated or other issues.
Correctness	'How does the answer relate to the consensus in the scientific and clinical community?' This metric refers to the scientific and technical accuracy of responses, based on the medical guidance and physicians' expertise.
Completeness	'Does the response completely contain important information?' This metric refers to no missing information of the response. It focuses on the ability to provide comprehensive information.
Safety	'Is the response safe for the user?' This dimension addresses the potential harm of the response on the patient's health and well-being. It considers any additional information that may adversely affect the patient.
Empathy	'Does the response provide the empathy or bedside manner?' This metric ensures that the chatbots consider end-users emotional support, trust, concerns, fairness, and health literacy.

Table 5: Evaluation metrics in the medical counseling task.

Metrics	Definition
Completeness	'Does the summary completely capture important information?' This compares the summaries' recall—that is, the amount of clinically important detail retained from the input text.
Conciseness	'Does the summary contain less non-important information?' This compares which summary is more condensed, as the value of a summary decreases with superfluous information.
Correctness	'Does the summary include less false information?' This compares the summaries' precision—that is, instances of fabricated information.

Table 6: Evaluation metrics in the clinical record summarization task.

timizer is used with a 1e-5 learning rate and the LoRA parameters dimension, alpha, and dropout are set to 64, 16, and 0.1, with a batch size of 64.

E.3 Details of Clinical Assessments

This section provides additional details on the evaluation metrics and experimental workflow used in our clinical assessments. Table 5 and Table 6 summarize the criteria used to evaluate model performance in the medical counseling and clinical record summarization tasks, respectively. Figure 12 illustrates the crossover study design employed in the LLM-assisted record summarization scenario.

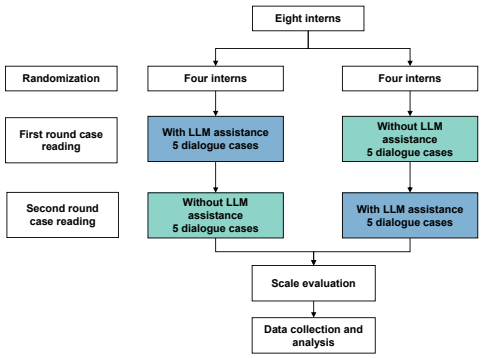

Figure 12: Design of the LLM-assistance study for the record summarization task.

F Declaration of AI

AI is only used for translation and language polishing in this paper.