SWEET: Large Language Model Benchmark for Scalable Diabetes Patient Education

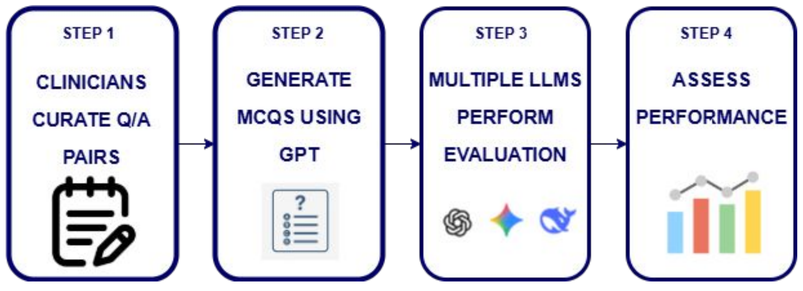
Syna[1], Bhavana Kunisetty[2], Chuyi Zhang[2], Yang Li[1], Elizabeth Healey[3], Agatha F. Scheideman[4], Mandy M. Shao[4], Anna Simos[2], Helge Ræder[5], Yanfu Zhang[1], David C. Klonoff[6], Marina Basina[2], Michael Snyder[2], Haipeng Chen[1], Tao Wang[2]

[1]College of William & Mary, USA · [2]Stanford University, USA · [3]Boston Children's Hospital, USA · [4]Diabetes Technology Society, USA · [5]University of Bergen, Norway · [6]Mills-Peninsula Medical Center (Sutter Health), USA

Introduction

- **Diabetes** affects >529M people worldwide (GBD 2021).
- **Patient education** is critical but limited by clinician time.
- LLMs such as GPT-5 and Gemini may help bridge this gap, but reliability remains uncertain.
- We develop **SWEET (Scalable diabetes patient Education Workbench for Evaluating Emerging Transformers),** the first benchmark for patient diabetes education.
- SWEET evaluates state-of-the-art LLM models on accuracy and test–retest reliability (self-consistency).
- **Evaluated models selection criteria:**
  1) top performance, 2) balance of open vs closed source, 3) clinician preference, and 4) API availability.

Methods



Unified Prompt Template (with Difficulty Annotations)

Task
Transform a clinician-curated Q&A into a multiple-choice question for diabetes education based on the following instructions.

Requirements
- Reading level: High school (Easy) / Undergraduate (Medium) / Graduate (Difficult)
- Options: 4 / 7 / 10
- Correct answers: 1 / 1 / 2
- Distractors: plausible but wrong
- Difficult: may include multi-step reasoning, lab values, contextual clues

Format
- Input: Original Question + Answer
- Output: JSON with transformed question, options (A…J), and correct answer(s)

Results

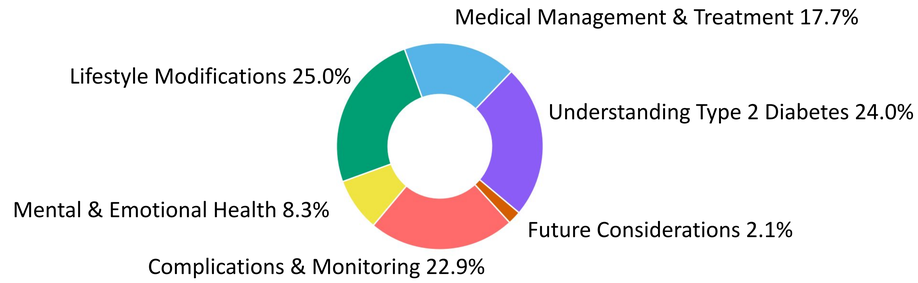


Figure 1. Distribution of 96 benchmark questions across clinical domains. *Lifestyle modifications and understanding type 2 diabetes dominate the dataset, with smaller contributions from other areas.*

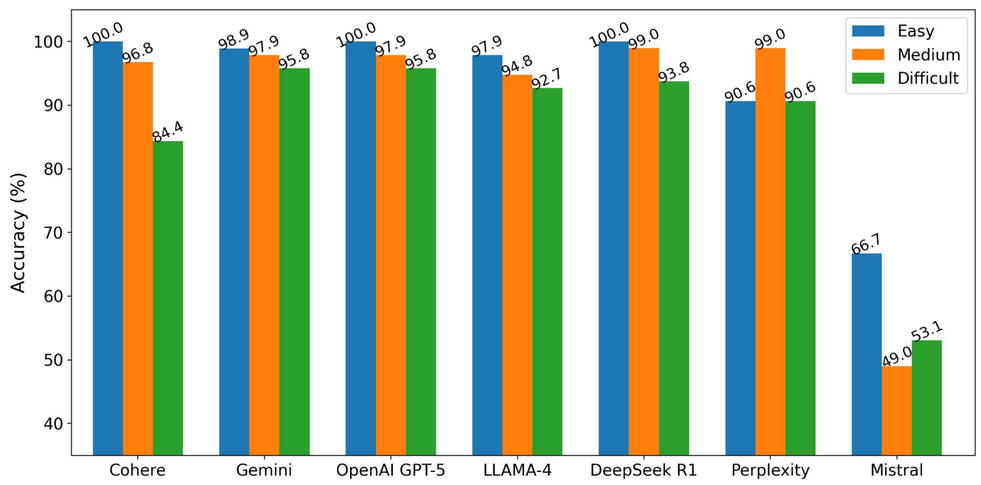


Figure 2. Accuracy of LLMs across difficulty levels, *via majority vote of 3 runs per question.* ***OBSERVATION-*** *GPT-5, Gemini, and DeepSeek R1 consistently achieved >95%, while Mistral underperformed. Cohere and Perplexity showed reduced robustness at higher difficulty.*

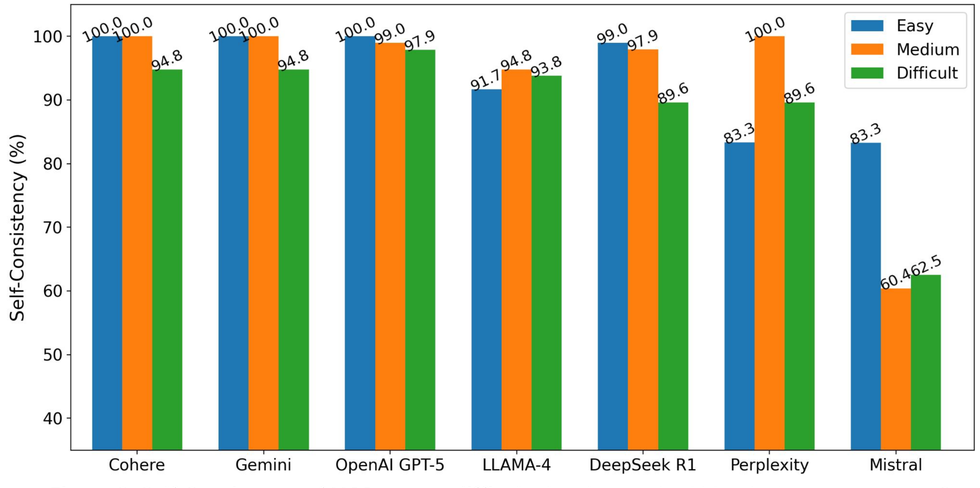
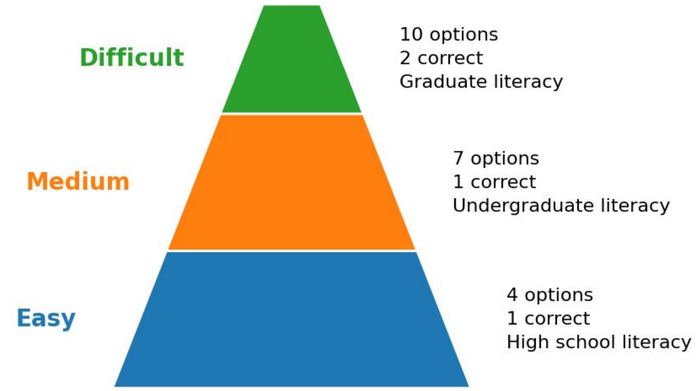


Figure 3. Self-Consistency of LLMs across difficulty levels, *required identical answers across 3 runs.* ***OBSERVATION-*** *GPT-5, Gemini, and Cohere were highly stable, DeepSeek R1 reliable, while Perplexity and Mistral showed inconsistencies.*

Dataset Difficulty Structure



96 clinician-curated Q/As → converted to MCQs

Conclusions

- **SWEET is the first benchmark for diabetes education.** Co-designed by clinicians and computer scientists, tailored to patient queries.
- Provides a **scalable and reproducible pipeline** for quantifying LLM accuracy and reliability.
- **Model choice impacts reliability.** GPT-5, Gemini, and DeepSeek R1 performed strongly, while Mistral and Cohere showed variability.
- **Human oversight remains essential.** Even high-performing LLMs struggle with multi-step reasoning and analogy-heavy questions.

References

1. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet. 2023
2. Gwira JA, Fryar CD, Gu Q. Prevalence of total, diagnosed, and undiagnosed diabetes in adults: United States, August 2021–August 2023. NCHS Data Brief, no 516. Hyattsville, MD: National Center for Health Statistics. 2024.
3. Lucas, M. M., Yang, J., Pomeroy, J. K., & Yang, C. C. (2024). Reasoning with large language models for medical question answering. Journal of the American Medical Informatics Association, 31(9), 1964-1975.