Persona-driven Simulation Using Large Language Models for Multidimensional Diabetes Distress Assessment

Mark Szigeti[1,2], Barbara Simon[3,4], Marianna Matányi[5,6,7], Lehel Dénes-Fazakas[3,8], Gyula Péter Szigeti[7,8], Levente Kovács[3,8], and György Eigner[3,7,8]

[1]Applied Informatics and Applied Mathematics Doctoral School, Obuda University, Bécsi út 96/B, Budapest, H-1034, szigeti.mark@stud.uni-obuda.hu

[2]Institute of Cyberphysical Systems, John von Neumann Faculty of Informatics, Obuda University, Bécsi út 96/B, Budapest, H-1034, szigeti.mark.bence@uni-obuda.hu

[3]Physiological Controls Research Center, University Research and Innovation Center, Obuda University, Bécsi út 96/B, Budapest, H-1034, simon.barbara@uni-obuda.hu, denes-fazakas.lehel@nik.uni-obuda.hu, eigner.gyorgy@nik.uni-obuda.hu, kovacs@uni-obuda.hu

[4]John von Neumann Faculty of Informatics, Obuda University, Bécsi út 96/B, Budapest, H-1034

[5]Janos Szentágothai Neurosciences Division, Semmelweis University, Üllői út 26, Budapest, H-1085, matanyi.marianna@stud.semmelweis.hu

[6]Cortex Research Group, HUN-REN Institute of Experimental Medicine, Szigony utca 43, Budapest, H-1083, matanyi.marianna@koki.hun-ren.hu

[7]MedTech Innovation and Education Center, University Research and Innovation Center, Obuda University, Bécsi út 96/B, Budapest, H-1034, matanyi.marianna@uni-obuda.hu, szigeti.gyula@uni-obuda.hu

[8]Biomatics and Applied Artificial Intelligence Institute, John von Neumann Faculty of Informatics, Obuda University, Bécsi út 96/B, Budapest, H-1034

Abstract: Diabetes distress, a psychological response to the burdens of managing diabetes, significantly impacts patient well-being and treatment adherence. However, its routine assessment in clinical settings remains limited due to time and resource constraints. This study explores the potential of large language models (LLMs) to assess diabetes distress through persona simulation. Using various large language models we simulated responses to the validated Diabetes Distress Scale (DDS17) across multiple patient personas representing a diverse range of diabetes experiences. A structured pipeline implemented via

Python facilitated the scoring and analysis of model responses across emotional, regimen-related, physician-related, and interpersonal distress subdomains. Results revealed clinically meaningful distinctions in distress profiles across patient types and highlighted key differences in model sensitivity, range, and domain-specific strengths. Our findings suggest that LLMs can augment traditional psychological assessments in diabetes care by offering scalable, interpretable tools for early distress detection and monitoring.

Keywords: diabetes distress, large language models, artificial intelligence, patient personas, health informatics, digital health

1 Introduction

Diabetes mellitus represents one of the most significant global health challenges of the 21st century, affecting hundreds of millions of adults worldwide, with projections suggesting this figure will continue to rise substantially in the coming decades [1]. Beyond its physiological complications, diabetes imposes substantial psychological burdens on those living with the condition. This psychological dimension, termed ”diabetes distress,” encompasses the emotional, cognitive, and behavioral challenges associated with managing a complex, demanding chronic illness [2].

Diabetes distress manifests across multiple domains: emotional burdens (feelings of anger, fear, or being overwhelmed), physician-related concerns (communication issues or lack of confidence in healthcare providers), regimen-related distress (challenges with self-care behaviors), and interpersonal concerns (inadequate support from family and friends) [3]. Unlike clinical depression, diabetes distress is specifically tied to the demands of managing this condition and affects a substantial portion of those living with the disease, with rates varying by population and illness severity [4].

Assessment of diabetes distress typically relies on validated questionnaires such as the Diabetes Distress Scale (DDS), which helps clinicians identify patients experiencing significant distress who may benefit from targeted interventions [5]. However, routine assessment remains inconsistent in clinical practice due to time constraints, competing priorities, and limited resources [6]. This gap highlights the need for innovative approaches to diabetes distress screening and evaluation.

Recent advances in artificial intelligence, extensive language models (LLMs), present promising opportunities for healthcare applications [7]. These models have demonstrated remarkable capabilities in understanding and generating human language, reasoning about complex concepts, and even simulating human-like assessments across various domains [8]. Their application in healthcare contexts ranges from diagnostic support to patient communication analysis [9].

A common technique in LLM research and application development is persona creation—the development of simulated patient profiles with specific characteristics, medical histories, and psychological states. These personas enable systematic evaluation of LLM performance across varied scenarios while maintaining patient privacy [10]. In diabetes care research, such personas can represent the diverse spectrum of diabetes experiences, from newly diagnosed patients to those managing longstanding complications.

This study explores the potential of three distinct large language models such as gpt-4.1-mini, gpt-4.1-nano, and gpt-4o-mini to evaluate diabetes distress across different patient personas [11]. By simulating how these models would complete the DDS17 questionnaire for six varied patient profiles, we aim to assess their ability to recognize and differentiate patterns of diabetes distress consistent with clinical expectations [12]. This approach offers insights into both the capabilities of current AI systems in understanding psychosocial aspects of chronic disease and potential applications for enhancing diabetes distress assessment in clinical settings.

Our investigation addresses several key questions: Can AI models accurately differentiate levels of distress across varied patient profiles? Do these models demonstrate sensitivity to the multidimensional nature of diabetes distress? What unique characteristics do different model architectures bring to psychological assessment tasks? The answers to these questions may inform future development of AI-assisted screening tools that could supplement traditional clinical approaches to diabetes distress assessment [13] [14].

2 Related Works

2.1 Large language models for diabetes training: a prospective study

Li and colleagues conducted a comprehensive prospective study evaluating ten large language models' diabetes-specific knowledge performance: ChatGPT-3.5, ChatGPT-4.0, Google Bard, LlaMA variants, and several Chinese-specific models using standardized diabetes examinations [15]. The evaluation was based on the Chinese National Certificate Examination for Primary Diabetes Care and the UK Royal College of Physicians' Specialty Certificate Examination in Endocrinology and Diabetes. Results showed ChatGPT-4.0's superiority, reaching 62.50% accuracy on the English examination and significantly outperforming other models. When assisting primary care physicians, ChatGPT-4.0 achieved 84.82% performance, surpassing all physicians and improving their results by 1-6.13% when functioning as an assistant.

This research provides a theoretical foundation for our current study by validating LLMs' applicability in diabetes-specific contexts and demonstrating their capability to perform medical-level assessments. While Li's study examined general diabetes knowledge using standardized examinations, our research advances toward the psychological dimension, specifically focusing on diabetes distress assessment. The performance differences between models identified by Li are particularly relevant, as we can expect similar variability in different LLMs' diabetes distress evaluation capabilities, justifying our multi-model comparative approach.

2.2 Advice for Diabetes Self-Management by ChatGPT Models: Challenges and Recommendations

Hussain and Grundy systematically evaluated ChatGPT-3.5 and ChatGPT-4 performance in diabetes self-management education, examining 20 diabetes-related queries across diet, exercise, hypoglycemia/hyperglycemia education, and insulin management [16]. Healthcare professionals assessed the models on consistency, reliability, and accuracy. While ChatGPT-4 showed improvements in organization and communication over ChatGPT-3.5, both models demonstrated significant limitations in providing personalized diabetes advice.

Critical deficiencies included assumptions about blood glucose units without clarification, generic meal planning lacking individual customization, inability to differentiate insulin regimens, and misclassification of pseudo-hypoglycemia. Compared to other large language models like Claude 3.5 Sonnet and Mistral Large 2, ChatGPT models performed poorly, with dangerous medical misinterpretations such as incorrectly classifying critically high blood sugar as low.

This research establishes important performance benchmarks for LLMs in specialized medical contexts, showing both capabilities and limitations of general-purpose models in healthcare. Their systematic evaluation methodology and documentation of performance variations across model iterations provide valuable insights for medical AI assessment, validating our multi-model comparative approach for diabetes-related psychological evaluation and highlighting the need to identify which models perform most effectively in specific psychological domains.

2.3 PATIENTSIM: A Persona-Driven Simulator for Realistic Doctor-Patient Interactions

Kyung and colleagues developed PATIENTSIM, a persona-driven simulator that generates realistic doctor-patient interactions for clinical training and evaluation [17]. The system uses 170 clinical profiles derived from real-world data in the MIMIC-IV and MIMIC-ED datasets, creating diverse patient personas across four

key dimensions: personality type, language proficiency, medical history recall level, and cognitive confusion level, resulting in 37 unique combinations.

The researchers evaluated eight large language models as the simulator backbone, ultimately selecting Llama 3.3 based on its superior performance. The evaluation addressed three main research questions: (1) whether LLMs naturally reflect diverse persona traits in their responses, (2) whether they accurately derive responses based on given profiles, and (3) whether they can reasonably fill in information gaps.

Results demonstrated that ChatGPT-4.0 achieved the highest accuracy across both languages (62.5% on English endocrinology examinations, 90.98% on Chinese primary care examinations), while the Llama series excelled in persona simulation tasks. Four clinicians validated the system, providing an average quality score of 3.89 out of 4 across six evaluation criteria. The study revealed that PATIENTSIM shows significant potential as a cost-effective and scalable alternative to standardized patients for medical education, though continued development and clinical oversight remain essential for safe and effective implementation.

This research establishes important benchmarks for LLM-based patient simulation, demonstrating both the capabilities and current limitations of AI systems in healthcare education while highlighting the need for specialized training and validation in medical contexts.

3 Methodology and Implementation

3.1 Assessment of Diabetes-Related Distress: The DDS-17 questionnaire

Psychosocial burden is a critical factor in diabetes management, influencing self-care behaviors, glycemic control, and quality of life. One of the most widely used tools to assess the emotional impact of living with diabetes is the Diabetes Distress Scale (DDS-17), developed by Polonsky et al. [18]. This 17-item self-report questionnaire is designed to measure diabetes-related emotional distress experienced by individuals over the preceding month, offering a nuanced understanding of patient well-being beyond traditional clinical metrics such as HbA1c.

The DDS-17 is grounded in a multidimensional framework that captures four distinct domains of diabetes-related distress: Emotional Burden, Regimen-Related Distress, Interpersonal Distress, and Physician-Related Distress. Each item is rated on a 6-point Likert scale (1 = ”not a problem” to 6 = ”a very serious problem”), and subscale scores are calculated as the mean of relevant items. A mean item score of

larger or equal than 3.0 on any domain is typically interpreted as clinically significant distress warranting intervention [19].

3.2 Persona Simulation Setup

To simulate realistic human-like responses from large language models (LLMs), we created structured personas based on individualized patient data. Each *persona* (a fictionalized diabetic patient) was defined through a spreadsheet uploaded by the user. This spreadsheet contained detailed biographic and behavioral metadata, including age, gender, diabetes type, years since diagnosis, BMI, exercise habits, diet, smoking and alcohol use, medications, comorbidities, and a brief narrative of medical history.

The persona definition was passed to the generate system prompt() function, which dynamically constructed a contextual system message for the LLM. This prompt tasked the model with role-playing the patient in a psychologically consistent, empathetic tone. It included both structured metadata and an embedded instruction (default task) directing the model to respond in the form of a numeric score (1–6) with an accompanying rationale.

3.3 Model Configuration and Prompting Pipeline

The simulation tested the capacity of five LLMs to emulate these personas. Three OpenAI models were used: gpt-4.1-mini, gpt-4.1-nano, and gpt-4o-mini. These models span a range of architectural sizes and inference modalities.

The experiment was orchestrated through a loop over all defined personas. For each model-persona pair, the system prompt was concatenated with a fixed set of 17 psychological assessment items (the Diabetes Distress Scale). Each question was posed individually via the run llm() function, which abstracts API interactions for OpenAI based backends. The first character of each model's response was parsed as an integer score between 1 and 6. If parsing failed (due to malformed output or non-numeric tokens), a fallback score of 0 was assigned.

Example instruction enforced by the prompt:

> Answer only with numbers, and give a small explanation in this format:
> (number); Explanation: (your explanation)

This enforced constrained generation to facilitate robust numeric scoring.

3.4 Evaluation Metrics and Analysis

Responses were passed into a scoring function, examine_results(), which computed five distinct psychological subscales, corresponding to domains from the Diabetes Distress Scale (DDS):

- Total DDS Score (mean of all 17 items):

$$\text{total_dds} = \frac{1}{17}\sum_{i=1}^{17} a_i$$

- Emotional Burden (EB):

$$\text{emotional_burden} = \frac{1}{5}\sum_{i\in\{0,2,7,10,13\}} a_i$$

- Physician-Related Distress (PD):

$$\text{physician_related} = \frac{1}{4}\sum_{i\in\{1,3,8,14\}} a_i$$

- Regimen-Related Distress (RRD):

$$\text{regimen_related} = \frac{1}{5}\sum_{i\in\{4,5,9,11,15\}} a_i$$

- Interpersonal Distress (ID):

$$\text{interpersonal} = \frac{1}{3}\sum_{i\in\{6,12,16\}} a_i$$

Scores were evaluated against a threshold of clinical relevance (score > 3) and flagged with warnings if distress was detected. This enabled nuanced cross-model comparisons for fidelity in emotional realism and psychological consistency.

3.5 Application Interface and Execution Workflow

To make the pipeline accessible and interactive, the experiment was deployed using the Gradio framework. A web-based interface allowed users to:

- Select one or more models from a dropdown list
- Upload an Excel file containing one or more personas
- Trigger the evaluation process via a simple UI

Internally, the frontend executed the core evaluation pipeline by invoking the generate_results() function in main.py. This function orchestrated persona loading (load personas from excel), model interactions (run llm), results aggregation, and scoring. Outputs were streamed in real time to the interface, providing two logs:

1. Answers Log – raw LLM responses to each question
2. Scoring Summary – computed metric scores with threshold warnings

All results were stored in an Excel file by export results to excel(), which structured the file with per-persona sheets, model-level scores, and the original prompts used for reproducibility and auditing.

3.6 Stability Analysis and Variable Importance

All simulations were repeated ten times for each model–persona combination, yielding a total of 180 independent assessment runs (3 models × 6 personas × 10 repetitions). For each run, the complete DDS-17 questionnaire was administered independently, and subscale scores were computed identically to the primary analysis. Descriptive statistics (mean and sample standard deviation) were calculated across the ten repetitions for each model, persona, and subscale combination. Stability was quantified using the coefficient of variation, where values below 10% were considered indicative of acceptable run-to-run consistency.

To complement the quantitative stability assessment, a qualitative variable importance analysis was conducted by systematically examining the free-text explanations generated by the models alongside their numeric scores. Across all 10 runs, 3 models, and 17 questions, each persona yielded 510 explanation segments. These were analysed for the frequency with which specific patient attributes, including dietary habits, physical activity, stress and workload, complications, medication regimen, comorbidities, and blood glucose control, were explicitly referenced as justification for the assigned score. Additionally, pairwise patient comparisons were used to isolate the approximate contribution of individual variables to total DDS score differences, holding other characteristics as constant as the persona set permitted. Results reported in Tables 1-3 correspond to a single representative simulation run selected prior to the stability analysis. Aggregate statistics derived from all ten repetitions are presented separately in Table 5 to facilitate direct comparison.

4 Results

4.1 Model Performance in Diabetes Distress Assessment

Our analysis examined how three different AI models (gpt-4.1-mini, gpt-4.1-nano, and gpt-4o-mini) evaluated diabetes distress across six patients with varying profiles. The results demonstrate distinct patterns in how each model interprets and quantifies distress levels. The models showed consistent ability to differentiate between clinical categories, with total distress scores ranging from minimal to high levels across the patient spectrum.

As shown in Table 1, all three models appropriately identified patients with poorly controlled Type 2 diabetes (Patients 1 and 2) as having the highest distress levels, while correctly assessing the patient without diabetes (Patient 6) with minimal distress. The models demonstrated clinically meaningful distinctions across the spectrum of patients.

Table 1
Overall Diabetes Distress Scores by Model and Patient

Patient	Condition	gpt-4.1-mini	gpt-4.1-nano	gpt-4o-mini	Clinical Category
1	T2DM, poor control	4.00	3.29	3.76	High distress
2	T2DM, poor control	3.88	3.06	3.82	High distress
3	Prediabetes	2.94	2.76	3.29	Moderate-high distress
4	Gestational diabetes	2.53	2.53	2.82	Moderate distress
5	T2DM, good control	1.94	2.53	2.65	Low-moderate distress
6	No diabetes	1.00	2.00	1.59	Minimal distress

4.2 Assessment Patterns Across Patient Profiles

Analysis revealed clear correlations between patient characteristics and distress evaluations across all tested models. Patients with poorly controlled Type 2 diabetes consistently showed high distress scores across all models, with particularly elevated regimen-related distress scores ranging from 3.4 to 4.6. Well-controlled Type 2 diabetes patients demonstrated significantly lower distress levels, reflecting successful disease management strategies and effective therapeutic adherence.

The prediabetes patient exhibited moderate-high distress levels, with significant regimen-related concerns emerging across all models, suggesting that early-stage diabetes management creates a substantial psychological burden even before full disease manifestation. This finding indicates that the uncertainty and lifestyle

adjustments required during the prediabetic phase may generate considerable anxiety about future health outcomes. Gestational diabetes patients showed moderate distress levels, particularly around treatment regimen management. However, scores remained lower than those of chronic diabetes cases, likely reflecting the temporary nature of the condition and different coping mechanisms during pregnancy.

The non-diabetic control case was correctly identified with minimal distress by two models. However, one model showed some overestimation tendencies, highlighting the importance of model selection for accurate psychological assessment. The differential performance across models suggests that some AI systems may have inherent biases toward detecting distress even in non-relevant clinical populations, which has important implications for clinical screening applications.

4.3 Model-Specific Characteristics

The three AI models demonstrated distinct evaluation patterns that reflect their underlying architectures and training approaches, with each showing strengths and limitations in diabetes distress assessment. As shown in Table 2, the gpt-4.1-mini model showed the widest evaluation range and strongest differentiation between clinical categories, with the most polarized assessments and excellent identification of non-relevant cases, though occasionally prone to extreme evaluations that may not reflect nuanced clinical realities.

The model's tendency toward polarization manifested particularly in its assessment of severe cases, where it consistently assigned maximum or near-maximum scores to patients with poorly controlled diabetes. This sensitivity to extreme cases makes it potentially valuable for urgent case identification, but may limit its utility for tracking subtle changes in patient status over time. The model demonstrated exceptional ability to distinguish between relevant and non-relevant cases, suggesting sophisticated pattern recognition capabilities that could improve efficiency.

The gpt-4.1-nano model provided the most conservative assessments with minimal variance, showing consistent moderate evaluations across cases, but demonstrating poor discrimination of non-relevant cases. This conservative approach resulted in compressed scoring ranges that may obscure important clinical distinctions between patient groups. However, the model's consistency suggests potential value for longitudinal monitoring where stability of assessment is prioritized over sensitivity to acute changes.

The gpt-4o-mini model showed particular sensitivity to regimen-related distress, consistently assigning the highest scores to treatment routine questions while occasionally showing anomalous responses in other domains. This model appeared to weight treatment adherence challenges more heavily than other psychological factors, which aligns with clinical literature emphasizing the central role of self-care behaviors in diabetes management. The focused sensitivity to treatment issues suggests this model might be particularly useful for identifying patients struggling with medication adherence or lifestyle modifications. However, its occasional anomalies warrant careful interpretation of results.

Table 2
Model Evaluation Characteristics

Model	Evaluation Range	Differentiation	Key Strengths	Key Weaknesses
gpt-4.1-mini	1.00-4.00 (widest)	Strongest	Extreme case identification, clear differentiation	Occasional over-reaction, variable strategy
gpt-4.1-nano	2.00-3.29 (narrowest)	Most conservative	Consistent assessments, longitudinal reliability	Poor extreme case discrimination, non-relevant case handling
gpt-4o-mini	1.59-3.82 (moderate)	Regimen-focused	Treatment challenge sensitivity, detailed evaluation	Occasional anomalies, treatment-focused bias

Table 3
Key Subscale Patterns Across Patient Groups

Patient Group	Regimen-related	Emotional Burden	Physician-related	Interpersonal
Poorly controlled T2DM	3.4-4.6 (consistently high)	3.0-4.2 (high)	2.75-3.25 (moderate-high)	3.0-4.0 (high)
Well-controlled T2DM	1.8-3.2 (moderate)	2.2-2.4 (low)	1.5-2.75 (low-moderate)	2.0-2.33 (low)
Prediabetes	3.2-4.2 (high)	2.8-3.2 (moderate-high)	2.0-2.75 (moderate)	2.33-3.0 (moderate)
Gestational diabetes	2.8-3.6 (moderate-high)	2.6-2.8 (moderate)	1.75-2.5 (low-moderate)	2.0-2.67 (moderate)
No diabetes	1.0-2.0 (minimal)	1.0-2.0 (minimal)	1.0-2.25 (minimal)	1.0-2.0 (minimal)

4.4 Subscale Analysis

As detailed in Table 3, regimen-related distress emerged as the most problematic area across all diabetic patients, with all models consistently scoring this subscale highest across patient groups. This finding aligns with clinical literature suggesting treatment adherence challenges as a primary source of diabetes distress. The analysis revealed that Question 6, which addresses feelings of failure with diabetes routine, received exceptionally high scores across models, particularly from gpt-4o-mini, which consistently rated this item at maximum levels for patients with diabetes management challenges.

The emotional burden subscale showed strong differentiation between patient groups, with poorly controlled Type 2 diabetes patients demonstrating clinically significant distress levels across all models. Interestingly, the prediabetes patient showed higher emotional burden scores than the well-controlled diabetic patient, suggesting that uncertainty and early adaptation to diabetes management may create significant psychological stress. Physician-related distress showed the most variability across models, with some patients showing clinically significant scores

in only one model, indicating potential inconsistency in how AI systems interpret healthcare provider relationships.

Interpersonal distress patterns revealed important insights about social support systems and diabetes management. The patients with poorly controlled diabetes showed consistently high interpersonal distress, suggesting that treatment struggles may strain relationships with family and friends. The gestational diabetes patient showed moderate interpersonal distress, likely reflecting the temporary nature of the condition and different social support dynamics during pregnancy. The well-controlled diabetic patient maintained low interpersonal distress scores, indicating that successful disease management may preserve social relationships and reduce diabetes-related interpersonal conflicts.

4.5 Clinical Relevance

The models demonstrated good alignment with expected clinical patterns, showing strong correlations between established risk factors and distress levels that validate the clinical utility of AI-based diabetes distress assessment. Higher BMI values (greater than 30) correlated with increased distress scores across all models and subscales, reflecting the well-documented relationship between obesity and psychological burden in diabetes management. This correlation was particularly pronounced in the regimen-related distress domain, where patients with elevated BMI showed consistently higher scores, likely reflecting the additional complexity of weight management alongside diabetes care.

Regular physical activity was consistently associated with lower distress levels across all assessment models, supporting the established benefits of exercise on both physical and psychological well-being in diabetes populations. The models successfully captured this relationship across multiple domains, with physically active patients showing reduced emotional burden, lower interpersonal distress, and

Table 4
Clinical Factor Correlations with Distress Scores

Clinical Factor	Pattern Observed Distress	Cases
BMI ⟩ 30	Higher overall	P. 1, 2
Regular physical activity	Lower levels	P. 5, 6
Diabetes complications	Higher emotional burden	P. 1, 2, 5
Treatment complexity	Higher regimen-related	All diabetic P.
Good glycemic control	Lower overall	P. 5 vs. P. 1/2

improved physician-related satisfaction scores. This pattern validation suggests that AI assessment tools can effectively recognize the multifaceted benefits of lifestyle interventions.

The presence of diabetes complications was accurately reflected in higher emotional burden scores across all assessment models, demonstrating the models' ability to detect the psychological impact of disease progression. Patients with established complications showed elevated fear-based responses and increased anxiety about future health outcomes, patterns that were consistently identified across all three AI models. This correlation provides evidence for the clinical validity of AI-based psychological assessment in identifying patients who may benefit from additional psychological support.

Treatment complexity corresponded with elevated regimen-related distress across all models, supporting the clinical understanding that complex medication regimens create a substantial psychological burden for patients. The models successfully identified that patients requiring multiple medications, insulin therapy, or frequent monitoring showed increased distress specifically in treatment-related domains, while maintaining appropriate discrimination in other psychological areas. This specificity suggests that AI assessment tools can provide targeted insights for clinical intervention planning.

The comprehensive analysis of clinical factor correlations presented in Table 4 provides robust evidence for the clinical validity of AI-based diabetes distress assessment, demonstrating that the models can recognize and appropriately weigh established risk factors in their evaluations. The consistency of these patterns across different AI architectures suggests that AI system approaches can reliably capture clinically meaningful relationships in psychological assessment data, supporting their potential integration into routine diabetes care protocols.

4.6 Stability of Repeated Simulations

To evaluate the consistency of model outputs, all six patient personas were assessed ten times independently by each model. The results demonstrate that LLM-generated DDS scores are highly reproducible across repeated runs. As shown in Table 5, the coefficient of variation remained low for the vast majority of model–persona–subscale combinations, confirming that the primary findings are not artefacts of stochastic sampling variability.

The most stable assessments were observed for patients with clearly defined clinical profiles. For poorly controlled Type 2 diabetes patients (Patients 1 and 2), all three models produced CV% values consistently below 2.7% on the Total DDS score, indicating near-deterministic scoring behaviour in high-distress scenarios. Similarly, the non-diabetic control case (Patient 6) was assessed with low variability by gpt-4.1-mini (CV% = 2.5%), confirming the robustness of its discrimination between relevant and non-relevant populations.

Elevated variability was observed specifically in the physician-related distress subscale, particularly for the non-diabetic patient assessed by gpt-4o-mini (CV% =

24.3%) and for the prediabetes patient's interpersonal subscale assessed by gpt-4.1-nano (CV% = 15.1%). These cases share a common characteristic: the underlying true score is low (approximately 1.5–2.8), meaning that a single-point variation in any constituent item produces proportionally large relative changes. This finding suggests that physician-related distress and interpersonal distress in borderline or non-clinical populations represent the dimensions most susceptible to run-to-run fluctuation, and should be interpreted with greater caution in screening applications.

Across all 180 runs, the rank ordering of patients by Total DDS score was largely stable across repeated runs. In gpt-4.1-mini and gpt-4o-mini, the only observed instability was an occasional swap between the two poorly-controlled T2DM patients, who consistently occupied the top two positions, the ordering of all remaining patients was fully consistent across all ten runs. As both patients belong to the same clinical category, no clinically meaningful rank change occurred in either model. In gpt-4.1-nano, greater instability was observed: the Júlia Varga–Kenji Yamamoto transposition appeared in six of ten runs, and one run produced a more substantial reordering in which Kenji Yamamoto ranked above Ákos Takács. The positions of Anna Szabó (lowest) and the two poorly-controlled T2DM patients (highest) remained consistent across all ten runs in all three models.

4.7 Qualitative Variable Importance Analysis

To gain insight into which patient characteristics most strongly influence model scoring behaviour, the free-text explanations accompanying each numeric response were analysed across all 510 explanation segments per patient. Results reveal a consistent hierarchy of referenced variables that is largely shared across all three models.

Dietary habits and meal regularity emerged as the most frequently cited variable for patients with active diabetes management challenges, appearing in over 50% of explanations for poorly controlled Type 2 diabetes and gestational diabetes patients, while citation rates were notably lower for well-controlled (∼25%) and prediabetes (∼37%) cases. Models consistently referenced irregular meal patterns, carbohydrate-heavy diets, or adherence to a structured diet plan as primary

Table 5
Run-to-run stability of Total DDS scores across 10 independent repetitions per model-persona pair.

Model	Patient	Condition	Total DDS Score		
			Mean	SD	CV%
gpt-4.1-mini	István Kovács	T2DM, poor control	3.876	0.105	2.7
	Zita Kovács	T2DM, poor control	3.935	0.081	2.0
	Ákos Takács	Prediabetes	3.012	0.061	2.0
	Júlia Varga	Gestational diabetes	2.500	0.064	2.5

	Kenji Yamamoto	T2DM, good control	1.988	0.037	1.9
	Anna Szabó	No diabetes	1.012	0.025	2.5
gpt-4.1-nano	István Kovács	T2DM, poor control	3.235	0.083	2.6
	Zita Kovács	T2DM, poor control	3.247	0.054	1.7
	Ákos Takács	Prediabetes	2.835	0.129	4.6
	Júlia Varga	Gestational diabetes	2.506	0.063	2.5
	Kenji Yamamoto	T2DM, good control	2.594	0.076	2.9
	Anna Szabó	No diabetes	2.053	0.065	3.2
gpt-4o-mini	István Kovács	T2DM, poor control	3.776	0.067	1.8
	Zita Kovács	T2DM, poor control	3.912	0.064	1.6
	Ákos Takács	Prediabetes	3.294	0.062	1.9
	Júlia Varga	Gestational diabetes	2.835	0.072	2.6
	Kenji Yamamoto	T2DM, good control	2.600	0.087	3.3
	Anna Szabó	No diabetes	1.647	0.107	6.5

justifications for regimen-related distress scores. Occupational stress and workload appeared as the second most prominent factor exclusively for the patient whose narrative explicitly included work-related burden (Patient 1), but were rarely cited for all other patients, indicating that models respond to explicit textual cues rather than implying psychosocial context from indirect evidence.

BMI was notable for its near-absence in explicit model explanations (cited in under 3% of responses for any patient), despite its substantial implicit influence. A pairwise comparison of patients with equivalent diabetes duration and type but differing BMI, activity level, and diet quality (Patients 1 and 5) revealed a mean Total DDS difference of 1.24 points, the largest isolated variable effect observed in the dataset. This suggests that BMI exerts its influence indirectly, mediated through its co-occurrence with poor dietary habits and low physical activity, rather than being processed as an independent numeric input.

The presence of diabetes complications was cited in 22-26% of explanations for patients with documented complications (Patients 1 and 5), contributing primarily to emotional burden scores. However, when complications co-occurred with good glycaemic control (as in Patient 5), their distress-amplifying effect was partially attenuated, consistent with clinical evidence that perceived management efficacy moderates the psychological impact of disease progression.

The diagnosis itself accounted for an estimated 0.82-point elevation in Total DDS relative to the non-diabetic control, as evidenced by the Anna Szabó–Kenji Yamamoto comparison. Pregnancy context dominated the explanation profile of the gestational diabetes patient (65% citation rate), functioning as an overriding contextual frame that systematically suppressed physician-related distress scores relative to chronic diabetes cases. These findings collectively suggest that narrative-level variables carry the greatest explicit weight in model decision making, while numerically encoded variables such as BMI and age operate primarily through their covariance with these narrative elements.

4.8 Comparison with Large Language Models for Diabetes Training

To contextualize our findings, we compared our results with Li et al. (2025), who evaluated ten LLMs on standardized diabetes examinations [15]. While their study measured percentage accuracy on medical examinations, our investigation employed continuous scoring for diabetes distress assessment, enabling more nuanced evaluation of model behavior.

Both studies identified distinct performance characteristics among AI models. Li et al. found ChatGPT-4.0 achieved 84.82% accuracy, significantly outperforming specialized medical LLMs like MedGPT and HuatuoGPT. Similarly, our analysis revealed that gpt-4.1-mini demonstrated the widest evaluation range (1.00-4.00) and strongest differentiation between clinical categories. Both investigations observed that specialized medical models did not necessarily outperform general-purpose models.

The studies showed convergent findings regarding clinical validation. Li et al. demonstrated that LLMs improved physicians' examination performance by 1-6.13%, while our study found strong correlations between LLM assessments and established clinical risk factors. Both studies identified treatment-related challenges as critical areas, with our detailed subscale analysis revealing regimen-related distress as consistently the most problematic domain across all diabetic patients.

The complementary nature of these findings highlights that while examination-based assessments demonstrate knowledge retention, patient-centered distress evaluation provides insights into the psychological burden of diabetes management that may not be captured through traditional medical assessments, allowing for identification of specific distress patterns that could inform targeted interventions and personalized care strategies.

4.9 Comparison with ChatGPT Models for Diabetes Self Management

Our diabetes distress assessment findings provide a complementary perspective to the recent evaluation by Hussain and Grundy (2025), who examined ChatGPT models' capabilities in providing diabetes self-management advice [16]. While their study focused on the accuracy and safety of AI-generated diabetes advice through direct patient query evaluation, our investigation addresses the psychological burden assessment aspect of diabetes care.

Hussain and Grundy found that ChatGPT models, including GPT-4, GPT-4o, and GPT-4o mini, frequently provided generalized advice insufficient for individualized diabetes management needs, particularly struggling with blood glucose unit assumptions and insulin regimen differentiation. Similarly, our analysis revealed that all three models showed distinct evaluation patterns, with gpt-4.1-nano demonstrating conservative assessments that showed poor discrimination of non-relevant cases, suggesting similar challenges in contextual understanding.

The studies converged on the critical importance of avoiding assumptions in diabetes care scenarios. Hussain and Grundy highlighted dangerous misinterpretations when ChatGPT models assumed blood glucose readings were in mg/dL rather than mmol/L, potentially leading to life-threatening advice. Our findings complement this concern through the observation that gpt-4o-mini occasionally showed anomalous responses, despite its sensitivity to treatment-related distress, indicating similar risks when models make unwarranted assumptions about patient contexts.

The cultural and contextual limitations observed by Hussain and Grundy, particularly regarding Western-centric dietary recommendations and economic insensitivity, parallel our findings about the importance of patient-specific factors in distress assessment. While their study demonstrated how meal plans reflected affluent Western preferences unsuitable for diverse populations, our patient-centered approach revealed that individual characteristics such as BMI, physical activity, and complication status significantly influenced distress patterns, emphasizing the need for culturally and individually adapted AI applications in diabetes care that extend beyond standardized medical knowledge to encompass the full spectrum of patient experiences.

4.10 Comparison with PATIENTSIM

Our diabetes distress assessment framework demonstrates distinct evaluation patterns compared to the PATIENTSIM study's patient simulation approach [17]. While PATIENTSIM evaluated LLM performance across persona fidelity, factual accuracy, and clinical plausibility using standardized metrics, our study focused specifically on diabetes-related distress evaluation capabilities across different

patient profiles. The PATIENTSIM study found that Llama 3.3 consistently outperformed other models across multiple evaluation criteria, achieving an average score of 3.68 out of 4 in persona fidelity evaluation. Similarly, our analysis revealed that certain models demonstrated superior performance in diabetes distress assessment, though the evaluation dimensions differed significantly.

PATIENTSIM's comprehensive evaluation included 37 distinct persona combinations across four axes, evaluated through both automated LLM assessment and human clinician validation. In contrast, our study employed a more focused approach, examining how AI models interpret and quantify distress levels across six carefully selected patient profiles representing different diabetes management scenarios. The factual accuracy evaluation in PATIENTSIM showed models achieving high entailment rates (94.4-98.1%) for supported statements, with larger models demonstrating superior performance. Our findings complement this by showing that model performance in diabetes distress assessment correlates with clinical complexity rather than model size alone.

Our focused approach to diabetes distress assessment provides deeper clinical insights into a specific healthcare domain, while PATIENTSIM offers broader general medical conversation capabilities. Our analysis of correlations between patient characteristics (BMI, physical activity, complications) and distress scores provides evidence-based insights for clinical practice. The differential performance patterns we observed across patient profiles suggest that specialized clinical assessment tools may offer advantages over general-purpose medical conversation systems, with our finding that poorly controlled T2DM patients consistently showed high distress scores (3.06-4.00), providing specific clinical benchmarks that complement PATIENTSIM's broader simulation capabilities.

5 Conclusions

This study presents a validated proof-of-concept demonstrating the feasibility of applying AI models to the assessment of diabetes-related distress. The work should be interpreted as a methodological demonstration rather than a completed clinical investigation. Our comparative analysis of three AI models in evaluating diabetes distress reveals promising applications for these technologies in clinical settings.

While gpt-4.1-mini excelled in identifying extreme cases, gpt-4.1-nano provided consistent baseline assessments, and gpt-4o-mini demonstrated particular sensitivity to treatment adherence issues—the area consistently identified as most problematic for diabetic patients.

The stability analysis further supports the reliability of these findings: across 180 independent simulation runs, the coefficient of variation remained below 10% for the vast majority of model–persona combinations. The most pronounced

variability was observed in low-scoring subscales, particularly physician-related and interpersonal distress in borderline or non-clinical populations, where single-item score variations produce proportionally large relative changes. Rank ordering of patients by Total DDS score was largely preserved across repetitions, with instability primarily observed among patients occupying similar or proximate positions in the distress spectrum, and more pronounced in models with narrower scoring ranges. These results suggest that LLM-generated distress profiles are sufficiently reproducible for longitudinal monitoring applications, though borderline subscale scores and intermediate rank positions warrant cautious interpretation in clinical screening contexts.

This study has several important implications. First, it suggests that AI-based assessment tools could serve as valuable supplements to traditional diabetes care, potentially enabling more frequent monitoring of patient distress between clinical visits. Second, the models' consistent identification of regimen-related distress as a primary concern across patient groups underscores the need for targeted interventions addressing treatment adherence challenges.

The next phase of our research focuses on enhancing the realism and applicability of the personas by grounding them in real-world clinical data. Specifically, we aim to construct digital twins based on real-life diabetic patients, allowing for a more accurate and individualized representation of diabetes experiences.

To evaluate the fidelity of these personas, we will implement a dual-assessment approach using the DDS-17 questionnaire. Both the real patients and their corresponding digital personas will complete the questionnaire independently. By comparing the responses, we aim to assess how closely the emotional and psychological states modeled by the digital twins align with those reported by actual patients.

This comparison will serve as a critical validation step for the digital persona simulation framework, providing insights into its capacity to replicate complex psychosocial factors. Additionally, the results may reveal opportunities for further refinement in the modeling process, especially in capturing aspects of diabetes related distress.

Ultimately, this work will contribute toward the development of intelligent, patient centered simulation tools that can support both clinical decision making and personalized diabetes management strategies.

Acknowledgment

SUPPORTED BY THE 2024-2.1.1 UNIVERSITY RESEARCH SCHOLARSHIP

PROGRAM OF THE MINISTRY FOR CULTURE AND INNOVATION FROM THE SOURCE OF THE NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION FUND.

References

[1] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda. IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract*, 138:271–281, February 2018.

[2] L. Fisher, J. S. Gonzalez, and W. H. Polonsky. The confusing tale of depression and distress in patients with diabetes: a call for greater clarity and precision. *Diabet Med*, 31(7):764–772, July 2014.

[3] N. E. Perrin, M. J. Davies, N. Robertson, F. J. Snoek, and K. Khunti. The prevalence of diabetes-specific emotional distress in people with type 2 diabetes: a systematic review and meta-analysis. *Diabet Med*, 34(11):1508–1520, August 2017.

[4] K. Dennick, J. Sturt, and J. Speight. What is diabetes distress and how can we measure it? a narrative review and conceptual model. *J Diabetes Complications*, 31(5):898–911, February 2017.

[5] L. Fisher, D. Hessler, W. Polonsky, and J. Mullan. When is diabetes distress clinically meaningful?: Establishing cut points for the diabetes distress scale. *Diabetes care*, 35:259–64, 02 2012.

[6] D. Hessler, L. Fisher, R. E. Glasgow, L. A. Strycker, L. M. Dickinson, P. A. Arean, and U. Masharani. Reductions in regimen distress are associated with improved management and glycemic control over time. *Diabetes Care*, 37(3):617–624, October 2013.

[7] R. Bommasani, D. Hudson, E. Adeli, R. Altman, S. Arora, S. Arx, M. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Davis, D. Demszky, and P. Liang. On the opportunities and risks of foundation models, 08 2021.

[8] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. *Nat Med*, 29(8):1930–1940, July 2023.

[9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[10] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021.

[11] T. Wu, M. Terry, and C. J. Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts, 2022.

[12] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

[13] P. Szalay, D. A. Drexler, and L. Kovács. Exploring robustness in blood glucose control with unannounced meal intake for type-1 diabetes patient. *Acta Polytechnica Hungarica*, 20(8):123–142, 2023. Retrieved from Acta Polytechnica Hungarica Vol. 20, No. 8 (2023).

[14] G. Eigner, J. K. Tar, I. Rudas, and L. Kovács. Lpv-based quality interpretations on modeling and control of diabetes. *Acta Polytechnica Hungarica*, 13(1):171–186, 2016. Acta Polytechnica Hungarica Vol. 13, No. 1 (2016).

[15] H. Li, Z. Jiang, Z. Guan, Y. Bao, Y. Liu, T. Hu, J. Li, R. Liu, L. Wu, D. Cheng, et al. Large language models for diabetes training: a prospective study. *Science Bulletin*, 2025.

[16] W. Hussain and J. Grundy. Advice for diabetes self-management by chatgpt models: Challenges and recommendations, 2025.

[17] D. Kyung, H. Chung, S. Bae, J. Kim, J. Sohn, T. Kim, S. Kim, and E. Choi. Patientsim: A persona-driven simulator for realistic doctor-patient interactions, 05 2025.

[18] W. H. Polonsky, L. Fisher, J. Earles, R. J. Dudl, J. Lees, J. Mullan, and R. A. Jackson. Assessing psychosocial distress in diabetes: development of the diabetes distress scale. *Diabetes Care*, 28(3):626–631, 2005.

[19] L. Fisher, R. E. Glasgow, J. T. Mullan, M. M. Skaff, and W. H. Polonsky. Development of a brief diabetes distress screening instrument. *Annals of Family Medicine*, 6(3):246–252, 2008.