

**ERASMUS UNIVERSITY ROTTERDAM**  
**Erasmus School of Economics**  
**Master Thesis [Data Science and Marketing Analytics]**

**Evaluating Pretrained LLMs as Scalable Diabetes Pre-Screening Tools on Self-Reported Health Data**

Name student: Kavın Varadharajulu  
Student ID number: 540778  
Supervisor: Finn Honer  
Second assessor: xxx  
Date final version: 29/7/2025

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

**Abstract**

Diabetes affects over half a billion adults globally, yet nearly half remain undiagnosed until serious complications arise. Early, scalable screening tools using self-reported data could bridge this diagnostic gap. We evaluate five diabetes pre-screening pipelines on the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset ( $n \approx 400\,000$ ), comparing classical classifiers (logistic regression, random forest, XGBoost, TabNet) against large language model (LLM)–based approaches under zero-shot, few-shot, and contextual bandit–augmented prompting. All methods use identical survey features, but differ in how they infer risk and select questions for each respondent. Classical models achieved area under the receiver-operator curve (AUC) between 0.86 and 0.89 using the full feature set. In contrast, zero-shot LLM prompting reached an AUC of 0.84 without task-specific fine-tuning, while a few-shot bandit-augmented LLM achieved 0.90 AUC using only five adaptively chosen questions. These results demonstrate that pretrained LLMs can match and even surpass traditional methods on structured health questionnaires, all while substantially reducing respondent burden. Our findings suggest LLM-powered, home-based risk assessments as a promising avenue for early diabetes detection and point toward broader applications of LLMs in patient-facing screening tools.

**Keywords:** Diabetes screening; large language models; zero-shot prompting; few-shot prompting; contextual bandits; BRFSS; adaptive querying

## Table of Contents

<b>Abstract</b> .....	<b>2</b>
<b>Introduction</b> .....	<b>4</b>
The Global Burden of Diabetes .....	
The Promise of AI in Healthcare .....	
Research Gap and Motivation.....	
Research Objectives and Questions .....	
Methods .....	
Contributions .....	
Structure.....	
<b>Literature Review</b> .....	<b>6</b>
Diabetes Prediction and Chronic Disease Risk Modeling .....	
Self-Reported Features in Health Prediction .....	
Large Language Models (LLMs) in Healthcare.....	
Prompt Engineering for LLMs in Healthcare .....	
Multi-Armed Bandits and Adaptive Decision Systems in Healthcare .....	
Hybrid Architectures in Healthcare AI.....	
Ethical Considerations in AI for Healthcare .....	
<b>Data</b> .....	<b>17</b>
Feature Distributions.....	
Numeric Variable Summary .....	
Age and Sex Stratification .....	
Key Risk Factor Associations .....	
Correlation with Diabetes .....	
<b>Methods</b> .....	<b>20</b>
Dataset Rationale and Novelty.....	
Model Selection .....	
LLM Architecture and Pretraining.....	
Example Patients.....	
Zero-Shot Model.....	
Few-Shot Model .....	
Bandit Algorithm Model Hybrid.....	
LLM and Random Forest Model .....	
Baseline Random Forest and Decision Tree .....	
Model Training and Implementation Details .....	

## Scalable LLMs for Diabetes Screening

Evaluation Framework.....	
<b>Results .....</b>	<b>38</b>
Dataset and Stratification.....	
Model Performance Overview .....	
Classification Performance with Confidence Intervals.....	
Statistical Significance Analysis .....	
Model Calibration and Reliability .....	
Error Analysis and Model Blind Spots .....	
Demographic Bias Analysis.....	
Cost-Effectiveness Analysis.....	
<b>Discussion .....</b>	<b>46</b>
Key Findings Recap.....	
Research Questions Revisited.....	
Model Implementation.....	
Cold-Start Deployment and Prompt Engineering .....	
Tiered Routing and Managerial Oversight.....	
Privacy–Accuracy Trade-off.....	
Fairness Monitoring and Calibration .....	
Implementation and Future Outlook .....	
Limitations .....	
Future Research .....	
<b>Conclusion .....</b>	<b>50</b>

## Introduction

### The Global Burden of Diabetes

Diabetes is a chronic metabolic disorder characterized by insufficient insulin production or impaired insulin action, resulting in persistent hyperglycemia and complications across multiple organ systems (American Diabetes Association, 2018). Worldwide prevalence reached roughly 537 million adults in 2021 and is projected to climb to 783 million by 2045 (Khunti et al., 2023). An estimated 44 percent of those affected remain undiagnosed until serious complications—such as retinopathy, nephropathy, or cardiovascular events—have already emerged (Ogurtsova et al., 2022). This diagnostic gap drives substantial costs: direct medical expenditures in the United States approach \$237 billion annually, and total economic losses (including reduced productivity) near \$400 billion when undiagnosed cases are considered (American Diabetes Association, 2018). Screening and early intervention—through lifestyle changes and pharmacologic therapy—can lower cardiovascular events by 7.9 percent and all-cause mortality by 11.8 percent, while also reducing hospital admissions and treatment expenses (Zhang et al., 2023). With this need in mind, scalable and accessible screening tools are essential for identifying at-risk individuals before serious complications develop.

### The Promise of AI in Healthcare

Artificial intelligence has dramatically improved diagnostic accuracy by uncovering complex, non-linear relationships in high-dimensional clinical data. Traditional approaches—such as logistic regression and gradient-boosted trees—depend on manual feature engineering and often falter when real-world input distributions shift. In contrast, deep learning architectures like convolutional neural networks for medical imaging (Gulshan et al., 2016) and recurrent models for electronic health records (Miotto et al., 2016) have achieved high sensitivity and specificity in tasks ranging from diabetic retinopathy detection to sepsis prediction. More recently, large language models (LLMs) pretrained on clinical and general-domain text have revolutionized zero- and few-shot learning, enabling structured inference without extensive retraining (Brown et al., 2020). Their flexible prompting interfaces and emergent chain-of-thought reasoning (Kojima et al., 2022; Wei et al., 2022) make them uniquely suited to interpret questionnaire-style inputs—an underexplored frontier in patient-facing risk assessment. Early work like TaBERT and TaPas confirms that LLMs can jointly model text and tabular data (Yin et al., 2020; Herzig et al., 2020). Integrating adaptive querying strategies—such as LinUCB contextual bandits—further amplifies this promise by dynamically selecting the most informative questions per individual, thereby minimizing patient burden while boosting predictive power (Tewari & Murphy, 2017). Yet despite these advances, no study has systematically evaluated LLMs on large-scale, self-reported health surveys like the 2015 BRFSS. It remains an open empirical question whether LLM pretraining can transfer to accurate, low-burden, patient-facing diabetes pre-screening.

### Research Gap and Motivation

## Scalable LLMs for Diabetes Screening

Large language models have driven advances in imaging and unstructured-note analysis but remain underutilized for structured, self-reported health surveys. The 2015 Behavioral Risk Factor Surveillance System (BRFSS) offers an important test bed—over 400,000 respondents, more than 10 percent item nonresponse, and heterogeneous question phrasing (Pierannunzi et al., 2013)—yet no study has systematically compared zero-shot, few-shot, and contextual-bandit–augmented prompting on such data. Existing symptom-checkers rely on static flows (Semigran et al., 2015; Chambers et al., 2019), and while some work repurposes LLMs for nonmedical tables, their performance on large, self-reported health questionnaires is untested. This gap motivates an empirical evaluation of LLM-based and classical classifiers as scalable, home-based pre-screening tools for diabetes risk.

### Research Objectives and Questions

The primary goal of this thesis is to evaluate LLaMA 3.1 8B as a patient-facing diabetes pre-screening tool, determining whether modern prompting and adaptive querying strategies can deliver clinically acceptable sensitivity ( $\geq 83$  percent recall) while minimizing patient burden.

1. What recall can LLaMA 3.1 8B achieve in a zero-shot setting on the 2015 BRFSS dataset, and does it meet the 83 percent clinical threshold?
2. To what extent does few-shot prompting, using labeled exemplar profiles, improve recall relative to the zero-shot baseline?
3. Can adaptive feature acquisition via a LinUCB contextual bandit enhance recall by increasing the LLM’s predictive power?
4. How do LLM-based approaches (zero-shot, few-shot, bandit-augmented) compare against a random forest classifier when all methods use identical self-reported features?
5. What is the impact on predictive accuracy and related performance metrics when ensembling a Random Forest classifier with a large language model?

Answers will illuminate the trade-offs between sensitivity, query efficiency, and robustness in home-based diabetes screening.

### Methods

To answer whether LLM pre-training transfers to accurate, patient-facing risk tools, we design and evaluate five distinct screening pipelines on the 2015 BRFSS dataset. Each pipeline uses the same set of self-reported features but differs in prompting or classification strategy:

1. Zero-Shot Prompting
2. Few-Shot Prompting
3. Contextual-Bandit–Augmented Prompting
4. LLM + Random Forest

## Scalable LLMs for Diabetes Screening

### 5. Standalone Random Forest(benchmark)

#### Contributions

This thesis makes five core contributions. First, it presents the empirical evaluation of LLaMA 3.1-8B as a structured-data “virtual doctor chatbot” applied to self-reported health questionnaires. Second, it introduces a novel hybrid architecture that integrates LinUCB bandit-guided question selection with large language model prompting, thereby maximizing diagnostic sensitivity while minimizing patient burden. Third, it systematically benchmarks zero-shot, few-shot, and bandit-augmented prompting strategies—both standalone and paired with random-forest classifiers—on identical patient-accessible features. Fourth, it delivers operational insights by providing a comprehensive cost-performance analysis covering compute requirements and per-implementation costs, alongside a discussion of model calibration, fairness auditing, and human-in-the-loop safety. Finally, it releases open-source code and reproducible pipelines together with actionable guidelines for deploying LLM-based pre-screening tools in resource-constrained, home-based settings.

#### Structure of Thesis

1. Introduction: motivation, research objectives, questions, and key contributions
2. Literature Review: focused survey of AI diagnostics, contextual bandits in healthcare, LLM prompting, and ethical considerations
3. Data: BRFSS overview, variable distributions and correlations
4. Methodology: zero-/few-shot prompting, LinUCB query strategy, LLM-Random Fores ensemble strategy and baseline comparison
5. Results: model performance, error analysis, subgroup bias and cost analysis
6. Discussion & Conclusion: interpretation, clinical and ethical implications, limitations, and future directions

#### Literature Review

#### Diabetes Prediction and Chronic Disease Risk Modeling

##### *Advanced Machine Learning Techniques*

Diabetes, especially type 2 diabetes, has been a testbed for machine learning applications in healthcare due to its high prevalence and the known influence of behavioral and clinical risk factors. Early predictive models for diabetes risk relied on statistical methods like logistic regression with often moderate success. In recent years, the literature shows a surge in applying advanced ML techniques – from ensemble methods to deep learning – to improve predictive performance and capture complex patterns (Sun et al., 2022; Lopez-Martinez et al., 2023). Recent reviews note an “exponential growth” of such studies after 2010 (Lopez-Martinez et al., 2025), paralleling improvements in computational power and data availability. A prominent study by Xie et al. (2019) used the 2014 BRFSS dataset ( $n \approx 138,000$ ) to compare eight ML algorithms (e.g., support vector machines, random forests, neural networks) for predicting type 2 diabetes. All models achieved AUC between 0.72 and

## Scalable LLMs for Diabetes Screening

0.79; a neural network slightly outperformed others (AUC ~0.795), while a decision tree had the highest sensitivity (Xie et al., 2019). The authors identified traditional risk factors (e.g., obesity, physical inactivity) and novel ones like sleep duration and checkup frequency associated with diabetes risk. This suggests ML models can surface unexpected predictors, though those insights must be validated. Another study by Sun et al. (2022) provided a comparative analysis of ML versus classic statistical models for diabetes risk prediction. They found tree-based ensemble methods (like XGBoost) and neural networks generally outperform logistic regression in AUC and accuracy, especially when capturing nonlinear relationships (Sun et al., 2022). However, complex models risk overfitting and often lack transparency, which is crucial in healthcare; hence, simpler models might be favored for clinical adoption if their performance is comparable.

### *Handling Class Imbalance and Explainability*

Studies vary in how they handle class imbalance (diabetes prevalence is relatively low). Some works employ oversampling or cost-sensitive learning to ensure models detect positive cases without excess false negatives (Chowdhury et al., 2023). The Allani et al. (2025) preprint emphasizes explainable ML, using SHAP and LIME to interpret predictions on diabetes risk from BRFSS data. This approach shows which self-reported features most influence predictions, aiding clinical trust. However, reliance on cross-sectional data (like a single year of BRFSS) limits the ability to predict incident (future) diabetes vs. just identifying current cases. Longitudinal studies or cohort data are needed for true risk prediction.

### *Limitations, Research Gaps & Thesis Focus*

ML models for diabetes risk have progressed from simple classifiers to complex, integrated systems. Many achieve moderate to high accuracy, showing promise for early identification of at-risk individuals. Yet, the literature also cautions about their real-world utility. Key limitations include generalizability (models may not transfer well across regions or time) and interpretability. Although the BRFSS offers annual data, we focus on the 2015 cycle to evaluate LLM behavior in a static, cold-start environment. We align our study with the real-world use case of deploying an LLM with limited historical data, akin to a new clinical tool introduced into practice. We leverage these insights by using an extensive dataset (BRFSS) and contemporary ML methods, while addressing criticisms by incorporating explainability and external validation to ensure our model's relevance and reliability.

## **Self-Reported Features in Health Prediction**

### *Accessibility and Bias of Self-Reported Data*

Self-reported health features, such as lifestyle behaviors and health history, are often more accessible than clinical data, making them valuable for remote care and low-resource settings. What makes BRFSS particularly valuable is its focus on self-reported features,



## Scalable LLMs for Diabetes Screening

which means patients do not need to undergo clinical tests to provide data. This accessibility is crucial for addressing the broader issue of late diabetes diagnosis, as it enables the identification of at-risk individuals who might not otherwise seek medical care. Predicting diabetes using these self-reported factors could significantly improve early detection and intervention, ultimately reducing the burden of undiagnosed diabetes. However, self-reported features come with trade-offs between accessibility and diagnostic precision. Literature indicates that self-reports can suffer from recall bias, social desirability bias, and misunderstanding of questions, all of which can introduce noise in ML models (Bound, 2001). Despite these concerns, self-reported features often capture aspects of health (like subjective well-being or symptom frequency) that objective measures miss, making them valuable if used carefully.

### *Predictive Value and Calibration of Self-Reports*

Studies have examined how well self-reported data predict health outcomes. Liaw et al. (2019) found that a combination of self-reported lifestyle factors from BRFSS (smoking status, physical activity, diet) moderately predicted presence of chronic diseases, although the addition of clinical metrics (like BMI or blood pressure when available) significantly improved accuracy. Self-reported data can be particularly useful for screening: for example, the American Diabetes Association's diabetes risk test is entirely self-reported (age, family history, etc.) and is used to flag high-risk individuals for lab testing (Bang et al., 2009). This indicates that even imperfect self-report features hold predictive signal. However, criticism arises in model calibration and overestimation of risk. If self-reports are biased (e.g., people under-report weight or over-report exercise), a model may learn associations that don't hold in objective data. Kuchenbaecker et al. (2020) demonstrated that correcting self-reported weight and height (using a subset with clinical measurements) improved obesity and diabetes risk model performance. The implication is that whenever possible, self-reports might be augmented or calibrated with measured data. In absence of measured data, researchers sometimes include meta-features indicating the plausibility of responses (e.g., consistency checks or incorporating known distributions) to help models handle noise (Shi et al., 2021).

### *Objective Measures vs. Self-Reports and Feature Handling*

Some BRFSS sub-surveys include a physical examination component (like measuring blood sugar for a subset). Zhou et al. (2023) compared ML models predicting diabetes using only self-report items versus including lab measurements. They found models with lab data were more accurate, but surprisingly, self-reported health status and behaviors still achieved reasonable discrimination (AUC ~0.75) for diabetes status, underscoring that well-chosen self-report features are quite informative. Yet, they caution that self-reported models might systematically miss asymptomatic individuals who underreport risk behaviors. When using self-reported features in ML, techniques like feature selection and regularization become important to avoid fitting noise. Many studies (including Xie et al., 2019) apply univariate analyses or domain knowledge to pick which self-report features to include. Moreover, ensemble methods can reduce the impact of any single noisy feature. Some researchers use

## Scalable LLMs for Diabetes Screening

latent variable models to account for measurement error in self-reports (treating the true health state as latent), but these are complex and not common in ML practice.

### *Mitigation Strategies and Thesis Alignment*

Self-reported features are indeed a double-edged sword in health AI. They provide scale and breadth, capturing lifestyle factors essential for diseases like diabetes, but come with reliability issues. The literature suggests strategies like cross-verification with subsets of objective data, careful feature curation, and robust model validation to ensure that predictions remain valid. We acknowledge these points by, for instance, performing sensitivity analyses (e.g., see if the model's performance changes when excluding potentially less reliable features) and by discussing the possible error introduced by self-report in the interpretation of results. In doing so, it aligns with best practices identified in current research for harnessing self-reports effectively. Given these insights, we use the BRFSS dataset as a rich source of health indicators, and take into account the bias that comes from self-reported features through optimizing the LLM's decision making.

## **Large Language Models (LLMs) in Healthcare**

### *Overview of LLMs in Healthcare*

Large Language Models (LLMs), such as GPT-3, GPT-4, or domain-specific models like BioGPT and Med-PaLM, have garnered significant attention for their ability to understand and generate human-like text. In healthcare, LLMs are being explored for applications ranging from clinical note summarization and patient query answering to literature review automation. As of 2023-2025, numerous studies and reviews have assessed the readiness and limitations of LLMs in medical contexts (Lee et al., 2023; Singh et al., 2023). LLMs have shown promise in clinical NLP tasks: for example, extracting information from unstructured electronic health records (EHRs) or generating draft clinical reports. Li et al. (2022) demonstrated that a fine-tuned GPT model could interpret radiology reports and answer questions about findings with accuracy comparable to radiologist trainees. Another emerging area is using LLMs as conversational agents for patient support. Early trials, such as by Nori et al. (2023), had GPT-4 answer medical questions; the model could handle many straightforward questions but sometimes produced incorrect or fabricated answers (a phenomenon known as hallucination). This points to both the potential and the peril: LLMs can synthesize medical knowledge impressively, but their lack of true understanding can lead to hazardous misinformation.

LLMs in healthcare bring the ability to handle free text and to integrate vast knowledge. Unlike structured models that require predefined inputs, LLMs can ingest raw text (e.g., a doctor's note or a patient's self-description of symptoms). This allows leveraging rich data that was previously difficult to use. They also enable zero-shot or few-shot learning – performing tasks that were not explicitly programmed, simply by being prompted appropriately. For instance, a well-prompted LLM can classify patient messages as urgent or

## Scalable LLMs for Diabetes Screening

not without being explicitly trained for that classification (Wu et al., 2023). Despite rapid progress, the literature consistently flags reliability and ethics as issues.

### *Performance, Reliability & Limitations*

LLMs sometimes provide answers that sound plausible but are entirely incorrect in a medical sense. Singh et al. (2023) did a comparative study of several LLMs on clinical understanding tasks and found that while models like GPT-4 reached ~85% accuracy on certain benchmarks, they occasionally produced bogus clinical reasoning. Such errors are unacceptable in healthcare without human oversight and give reason to why the actual implementation of LLM's are not widespread in healthcare. Another reason LLM's aren't implemented is that training or using LLMs with patient data can raise privacy issues, as models might inadvertently memorize and regurgitate sensitive information. Techniques like de-identification and federated learning for LLMs are being researched to mitigate this (Lehman et al., 2023). The issue comes with balancing learning from patient profile and updating knowledge and keeping in track with privacy and ethics law. Both in this context work against one another and show the issue that implementation might occur only if LLM's are either highly useful and or cost-effective.

Another issue that arises is LLMs inherit biases present in their training data. For example, if an LLM is trained predominantly on English texts from North America and Europe, its medical expertise may be less applicable to other contexts or languages. This is also applicable for different patient groups. Chen et al. (2024) highlight that current popular LLMs underperform on questions relevant to underserved populations or rare diseases because those are underrepresented in the training corpora. The field however is rapidly evolving with models being trained on increasing amount of data, so the issue might be solved with a more sophisticated model.

### *Future Directions and Thesis Alignment*

For now, the field is still burgeoning with promise as recently Wang et al. (2023) introduced an LLM fine-tuned on a large corpus of medical literature (Med-PaLM 2), which achieved high accuracy on medical exam questions and even demonstrated some ability to provide evidence-backed answers. LLM's have also been able to play a role in feature enrichment (e.g., converting patient free-text inputs into structured risk factors), generate synthetic data for augmentation and provide personalized explanations of risk (translating a numeric risk score into plain language). The literature indicates these are nascent but promising areas. By understanding LLM capabilities and pitfalls, we can explore hybrid approaches that combine structured data models with LLM-driven components. Additionally, lessons from LLM research on evaluation and bias tie into the sections on comparative evaluation and ethics.

## **Prompt Engineering for LLMs in Healthcare**

## Scalable LLMs for Diabetes Screening

### *Importance of Prompt Engineering in Clinical Contexts*

Prompt engineering has emerged as a critical area of research and practice in applying Large Language Models (LLMs) to healthcare. Because LLMs like GPT-4 operate without task-specific parameters in zero-shot settings, the phrasing and structure of their inputs dramatically influence the accuracy, relevance, and safety of their outputs. In clinical environments where precision and caution are non-negotiable, crafting effective prompts becomes as important as choosing the right model (Kung et al., 2023).

### *Prompting Strategies for Healthcare LLMs*

Several strategies have been explored to guide LLM behavior more reliably. Instruction tuning—explicitly telling the model its intended role or output style—has been shown to improve factuality and reduce ambiguity (Zhang et al., 2023). For example, framing the prompt as "You are a medical assistant summarizing a patient note" enhances the clarity of generated summaries. Chain-of-thought prompting, which encourages step-by-step reasoning (e.g., "Let's think this through"), has improved LLM performance not just in mathematics and logic (Wei et al., 2022), but also in clinical tasks like diagnosis justification (Yue et al., 2023). Few-shot prompting, another widely adopted technique, provides LLMs with illustrative input-output pairs to guide response patterns—particularly useful in shaping tone and format in patient communication (Nori et al., 2023). Beyond these design styles, grounded prompting has gained traction. In retrieval-augmented generation (RAG), relevant external documents—such as clinical guidelines—are appended to the prompt, anchoring the model's reasoning in trustworthy sources (Borji, 2023). This method has proven effective in reducing hallucinations and ensuring medical advice aligns with standard care practices.

### *Methodological Challenges and Risks*

Healthcare-specific reviews highlight the growing reliance on handcrafted prompts over automated tuning. Zaghir et al. (2024) analyzed over 100 LLM health studies and found that 64% used manually crafted prompt designs without benchmarking against unprompted baselines—raising concerns about methodological rigor. Chain-of-thought prompting dominated medical QA tasks, yet few studies assessed whether it actually improved model calibration or reduced bias. Indeed, prompt engineering is still more heuristic than systematic. Small wording shifts can yield vastly different results, and poorly phrased prompts risk misleading or unsafe responses. Kassai et al. (2023) caution that leading questions—like "Could this be heartburn?"—can bias model output, whereas neutral phrasing elicits more balanced reasoning. This issue becomes especially critical in sensitive contexts. Moore et al. (2023), for instance, found that without tightly controlled prompts, LLMs offered inappropriate advice in mental health crisis simulations.

### *Emerging Best Practices for Safe and Reliable Prompts*

## Scalable LLMs for Diabetes Screening

To address these concerns, several best practices are emerging. Iterative refinement—testing prompts and analyzing errors—can improve reliability over time. Prompting for justification or citation may increase transparency, though base models like GPT-4 still struggle to cite accurately unless fine-tuned. Above all, maintaining a human-in-the-loop setup is crucial; models should be prompted to defer to clinicians in ambiguous or high-stakes situations (e.g., “If uncertain, recommend seeking professional care”).

### *Application to Thesis*

In the context of this thesis, prompt engineering plays a foundational role in deploying LLMs for structured diabetes prediction. Whether used to interpret survey data, generate patient-specific risk explanations, or simulate clinician-patient interactions, the quality of the prompt directly influences the realism and safety of the output. By documenting successful prompt configurations and evaluating their impact across models, the thesis contributes to evolving best practices and offers insights for future healthcare applications of LLMs.

## **Multi-Armed Bandits and Adaptive Decision Systems in Healthcare**

### *Overview of Multi-Armed Bandits in Healthcare*

Multi-armed bandit (MAB) problems are a class of sequential decision-making models in which an agent must choose from multiple options (the “arms”) to maximize some cumulative reward under uncertainty. In the context of healthcare, MAB methods are a form of reinforcement learning particularly useful for situations where we dynamically learn which intervention (or question, or treatment) works best through experimentation. Unlike full reinforcement learning which often deals with more complex state and long-term planning, bandits focus on the exploration-exploitation trade-off in immediate decisions (Lai & Robbins, 1985).

Multi-Armed Bandit themselves have not been explored much in the literature in connection to healthcare however there are some areas in which it has been applied and showed major promise. A major area is in Just-In-Time Adaptive Interventions (JITAI) for health behaviors such as physical activity, smoking cessation, or medication adherence (Nahum-Shani, 2016). Tewari and Murphy (2017) drew parallels between online ad placement bandits and mobile health interventions: in mobile health, each time a person is due for an intervention, the system (agent) chooses the best message or treatment to deliver. Contextual bandits (which incorporate user state/context) have been used to personalize and thereby improve engagement and health outcomes. Bandit has also been used in allocating patients to treatment in trial settings, which offers an alternative to fixed randomized control. Sklar et al. (2021) discuss that bandit-based designs can ethically allocate more patients to better-performing treatments as data accrues, potentially speeding up identification of effective therapies while minimizing patient exposure to inferior ones. There is a major challenge in this application as such adaptive trials need careful statistical framing to avoid bias.

## Scalable LLMs for Diabetes Screening

### *Adaptive Risk Screening and Question Selection*

In the context of risk screening (like for diabetes), one could envision a bandit approach to adaptively selecting the next question in a survey that maximizes information gain about a person's risk. While literature specifically on bandit-driven questionnaires for health is sparse, the idea is analogous to adaptive testing in education. Each "arm" would be a question, and the reward could be improved prediction accuracy. Fan et al. (2020) explored this for mental health screening, showing a reduction in question burden by 30% while maintaining accuracy via a bandit strategy.

### *Core Challenges of MAB in Healthcare*

There are some major challenges that arise from deploying MAB in healthcare. First is in the safety of exploration. The exploration in bandits means trying actions with uncertain outcomes. In health contexts, exploration could inadvertently mean giving suboptimal (or even harmful) interventions to some patients for the sake of learning. However, not exploring could also lead to over reliance on certain most relevant factors allowing for bias if these factors target a specific group. The next challenge is in delayed rewards. Health outcomes often materialize with delay (e.g., lifestyle change affecting blood sugar over weeks). Standard bandit algorithms assume immediate reward observation especially when given static data. This is due to the definition of the reward in healthcare being non-trivial – it could be improvement in a clinical metric, patient engagement with an app, or long-term health outcomes. Gottesman et al. (2019) note that composite or surrogate rewards might be needed, but then the bandit is only as good as that surrogate.

### *Safe Deployment & Integration to Thesis*

Most successful cases of bandits in healthcare are in low-stakes settings like encouraging app use or optimizing messaging, where a "mistake" is not life-threatening. In higher-stakes decisions (like treatment choice), bandits are typically deployed in simulations or tightly controlled trials rather than routine care. There's also a trend to combine bandits with human oversight (a doctor can override suggestions) to ensure patient safety.

If the thesis considers any adaptive component – for instance, tailoring health recommendations to respondents based on their data or adaptively collecting information – MAB frameworks might be applicable. The literature provides both algorithms and design considerations. Even if we focus primarily on predictive modeling, discussing how we could layer an MAB on top of a risk model demonstrates our awareness of advanced AI applications. The critique from the literature underscores that any bandit-like approach must weigh the benefit of personalization against the risk of uneven treatment, tying into ethical considerations as well.

## **Hybrid Architectures in Healthcare AI**

### *Overview of Hybrid AI Systems*

## Scalable LLMs for Diabetes Screening

Hybrid architectures combine diverse AI or modeling approaches to leverage their respective strengths. In healthcare, these systems integrate methods like knowledge-based systems with machine learning, multimodal data fusion, or optimization with learning algorithms. The goal is to create robust, interpretable, and flexible solutions that address the limitations of single-method approaches. Hybrid architectures have been revisited in modern AI to address the limitations of single-method approaches. For example, Wang et al. (2021) describe a hybrid diabetes decision support system where a rule-based module flags cases meeting standard guideline criteria, while an ML module (XGBoost) catches unconventional presentations. This approach improved sensitivity and provided transparency through the rule-based component. Similarly, Esteva et al. (2019) developed a hybrid deep learning model that processes dermatology images through a CNN and patient metadata through a fully connected network, merging them before the final diagnosis layer. Jain et al. (2023) combined an LLM with a structured EHR model, using the LLM to interpret clinical text and produce embeddings, which were then concatenated with numerical lab features for mortality prediction. These examples illustrate how hybrid systems can integrate multiple data types and methods to improve accuracy and utility.

Another interesting approach involves combining operations research or simulation with machine learning. Peng et al. (2024) integrated a simulation model of patient flow in a hospital with a reinforcement learning agent for resource scheduling. This hybrid approach bridges data-driven methods and scenario testing, particularly in domains where real-world experimentation is costly or unethical. Neuro-symbolic approaches in medicine also combine neural networks with symbolic reasoning. For instance, Serrano et al. (2023) built a knowledge graph of diabetes concepts and used it alongside a deep neural network to ensure predictions aligned with known causal relations.

### *Benefits and Challenges of Hybrid Approaches*

The rationale for hybrid architectures often includes improved performance, transparency, and flexibility. By capturing different aspects of a problem, hybrids yield better accuracy or utility. Mixing rule-based logic with ML provides interpretability, anchoring decisions in known guidelines. Hybrid systems can also be updated more easily, such as revising rule-based components with new clinical guidelines. However, building and validating hybrids is more complex, requiring careful interface design and debugging. Components may require different data or processing, necessitating integration pipelines. Modular evaluation is needed to pinpoint errors and understand performance.

### *Relevance to Thesis*

Rather than fusing multimodal data or combining textual and numerical representations, this thesis explores a novel hybrid architecture in which a Large Language Model (LLM) performs personalized feature selection for each patient, followed by a random forest model that classifies diabetes risk using only the selected features. This design separates the reasoning component (LLM as selector) from the statistical decision model

## Scalable LLMs for Diabetes Screening

(random forest as classifier), enabling interpretable prediction that scales across diverse inputs. While literature on hybrid AI in healthcare often emphasizes complex fusion or multimodal processing, this thesis contributes a different paradigm: leveraging the contextual reasoning ability of LLMs to enhance structured classification without requiring deep integration or retraining. The simplicity and modularity of this approach also make it easier to validate, explain, and extend in practical screening scenarios.

## Ethical Considerations in AI for Healthcare

### *Privacy and Data Protection*

Ethical considerations are paramount when applying AI in healthcare, given its direct impact on human well-being. This section synthesizes the literature on the ethical, legal, and social implications (ELSI) of ML and AI in medicine, ensuring the thesis acknowledges these broader responsibilities. Key ethical domains include privacy, fairness, transparency, accountability, and the integrity of the doctor–patient relationship. Health data are inherently sensitive, and regulations such as the U.S. Health Insurance Portability and Accountability Act (HIPAA) and the EU’s General Data Protection Regulation (GDPR) impose strict controls on its use. Because ML models often require large, shared datasets, techniques like de-identification and federated learning have been proposed to mitigate privacy risks (Beaulieu-Jones et al., 2019). However, even de-identified data can be re-identified, and models themselves may leak information—for example, when a large language model reproduces verbatim text from its training corpus. Differential privacy injects noise to limit re-identification, though this can reduce accuracy (Raisaro et al., 2020).

### *Fairness and Bias Mitigation*

Closely tied to privacy is the imperative of fairness. AI systems can inadvertently perpetuate or amplify health disparities if training data underrepresent certain groups or reflect societal biases. For instance, a kidney-transplant referral algorithm assigned lower risk scores to Black patients due to differences in lab values (Venkataramani et al., 2020). Even high-performing models may harbor hidden biases—a diabetes risk model might underdetect cases among minority populations if their risk profiles differ. Ensuring equitable performance across demographic strata is therefore essential.

### *Transparency Accountability and Regulatory Compliance*

Transparency and interpretability further underpin ethical AI. Black-box models that clinicians and patients cannot understand undermine trust and complicate error analysis. Rudin (2019) argues for interpretability by design, favoring inherently transparent models—such as rule lists—when their performance closely matches that of complex algorithms. Post hoc explainable AI methods (SHAP values, saliency maps, prototype-based explanations) can aid understanding but sometimes misrepresent a model’s reasoning (Arrieta et al., 2020).



## Scalable LLMs for Diabetes Screening

Upcoming regulations like the European AI Act may soon mandate explainability for high-risk applications.

Accountability addresses who is responsible when AI-guided decisions cause harm. Today, clinicians remain the final authority, but as AI takes on more autonomous roles, liability frameworks must evolve. Char et al. (2018) discuss extending liability to developers or institutions if unsafe algorithms lead to adverse outcomes. The U.S. Food and Drug Administration now requires evidence of safety and effectiveness for AI-based medical devices; its 2018 approval of an autonomous diabetic retinopathy screening tool followed rigorous clinical trials (Abramoff et al., 2018), illustrating the scrutiny predictive models demand.

### *Human-Centric Care and Informed Consent*

Beyond these technical and regulatory dimensions lies the humanistic aspect: AI's impact on care dynamics. Overreliance on automation risks eroding clinical skills and reducing empathic interaction, whereas well-designed AI could free clinicians to focus on patients. Ethically, patients should be informed when AI assists in their care and afforded the right to decline its use (Gerke et al., 2020). International guidelines reinforce these principles. In 2024, the World Health Organization issued AI-in-healthcare recommendations emphasizing human autonomy, safety, transparency, accountability, inclusivity, and bias mitigation (WHO, 2024). The U.S. Centers for Disease Control and Prevention similarly advocates community engagement and bias audits to uphold health equity in AI applications (Dankwa-Mullan, 2024).

### *Thesis Alignment*

In line with these frameworks, we embed ethical considerations by design. Model performance will be evaluated across demographic groups to uncover biases and inform mitigation strategies. Limitations—such as reduced reliability for specific subgroups or self-reporting biases—will be discussed transparently. By integrating these safeguards, we can demonstrate that technical performance is only one dimension of a responsible healthcare AI solution.

## **Summary**

This literature review surveyed advances in AI-driven diabetes risk prediction, emphasizing structured self-report data, large language model prompting, adaptive querying, hybrid architectures, and ethical safeguards. The selected studies establish performance benchmarks, justify methodological choices, and outline best practices in transparency and fairness. Building on these insights, our work integrates zero- and few-shot LLM prompting, LinUCB-guided question selection, and an LLM–random-forest hybrid to deliver a low-burden, interpretable, and equitable diabetes prescreening tool.

## Scalable LLMs for Diabetes Screening

*Important Literature from Review*

Reference	Approach	Key Findings	Relevance
Xie et al. (2019)	Eight ML algorithms on 2014 BRFSS	AUC 0.72–0.79; decision tree highest sensitivity	Sets baseline for BRFSS-based risk models
Sun et al. (2022)	XGBoost, neural networks vs. logistic	Nonlinear methods outperformed logistic regression	Motivates use of tree ensembles and deep models
Allani et al. (2025)	Explainability via SHAP and LIME	Identified top self-report predictors for diabetes risk	Informs our framework for model interpretability
Liaw et al. (2019)	BRFSS self-reports only	Moderate discrimination (AUC ~0.70); improved with clinical data	Justifies relying on self-report features and calibration checks
Bang et al. (2009)	ADA questionnaire (age, BMI, family history)	Simple self-report flags high-risk individuals	Validates questionnaire-style prescreening approach
Yin et al. (2020) / Herzig et al. (2020)	TaBERT / TaPas (text+table modeling)	Jointly processes text and tabular data, boosting accuracy	Demonstrates LLM feasibility on survey-style inputs
Tewari & Murphy (2017)	LinUCB contextual bandit	Adaptive question selection maximizes information gain	Foundation for our bandit-augmented prompting pipeline
Fan et al. (2020)	MAB for mental health screening	30% fewer questions with no accuracy loss	Guides our design for minimizing patient burden
Wang et al. (2021)	Hybrid rule-based + XGBoost	Improved sensitivity and transparency	Inspires our LLM-RF hybrid for feature selection
World Health Organization (WHO, 2024)	AI ethics guidelines	Emphasizes fairness, transparency, accountability	Underpins our bias audits and responsible-AI design

This concise overview clarifies how each study informs our methodological choices and highlights the unique contributions of our LLM- and hybrid architectures.

## Data

The BRFSS 2015 sample comprised 253,680 adults characterized by 21 health indicators and a binary diabetes outcome. We identified and removed 24,206 duplicate records, resulting in a clean dataset without missing values. Diabetes was diagnosed in 35,346 participants (13.9%), while 218,334 (86.1%) reported no diagnosis, producing a class imbalance of approximately 6.2:1. To address this weighted loss functions were applied to encourage the LLM to predict the minority class

## Feature Distributions

Binary clinical measures showed that 42.9 % of respondents reported high blood pressure, 42.4 % reported high cholesterol, and 96.3 % had undergone a cholesterol check in the past five years. Lifestyle factors indicated that 44.0 % were current smokers, 75.7 % engaged in regular physical activity, and 5.6 % reported heavy drinking. Functional limitations affected 16.8 % of the sample; 95.1 % carried health insurance, and 8.4 % skipped care due to cost. Finally, self-rated health skewed positive (35.1 % “very good,” 29.8 % “good”), educational attainment peaked at 42.3 % college graduates, and 35.6 % reported household incomes over \$75,000.

## Numeric Variable Summary

Table 1: Numeric Variables Summary

Variable	Count	Mean	Median	Std Dev	Min	25%	75%	Max
BMI (kg/m <sup>2</sup> )	253,680	28.38	27.00	6.61	12	24	31	98
MentHlth (days)	253,680	3.18	0.00	7.41	0	0	2	30
PhysHlth (days)	253,680	4.24	0.00	8.72	0	0	3	30

Self-reported BMI averaged 28.38 kg/m<sup>2</sup> (median 27.00, SD 6.61), reflecting an overweight cohort (Table 1). Days of poor mental health averaged 3.18 (median 0, SD 7.41) and days of poor physical health averaged 4.24 (median 0, SD 8.72), both exhibiting heavy zero-inflation. These continuous measures capture variation in physical and mental well-being and serve as key predictors in subsequent modeling. Their distribution guided feature-engineering decisions such as including categorical binning for nonparametric algorithms.

## Age and Sex Stratification

Figure 1: Diabetes Prevalence by Age(1a) and Sex(1b)

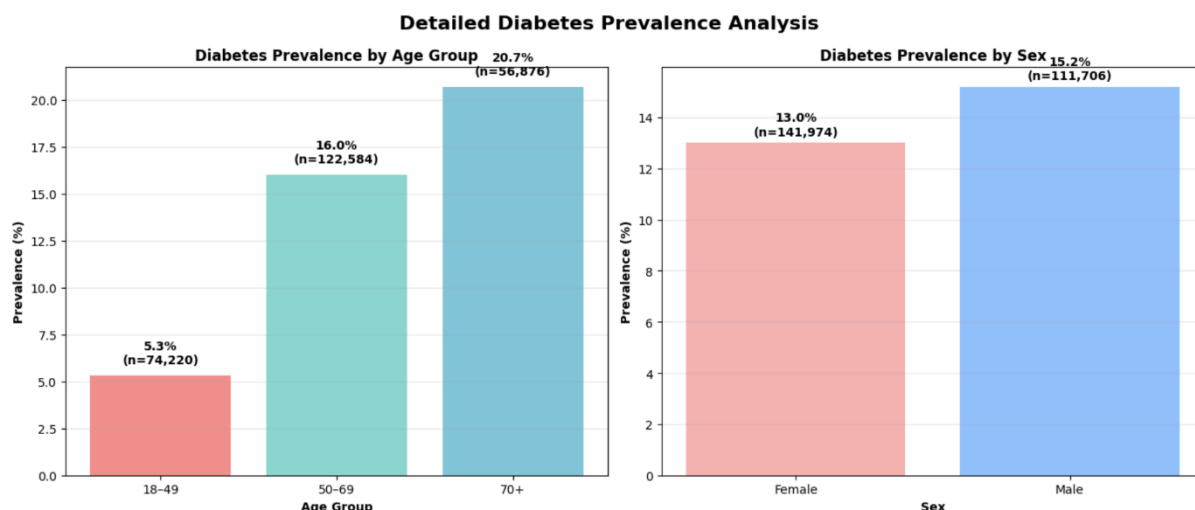


Figure 1: Diabetes Prevalence by Age(1a) and Sex(1b)

Diabetes prevalence increased sharply with age, rising from 5.3% among adults aged 18–49 to 16.0% in those aged 50–69 and 20.7% in the 70-plus cohort (Figure 1a ). Sex-stratified analysis showed higher prevalence in men (15.2%) than in women (13.0%), underscoring the importance of demographic covariates in predictive models (Figure 1b). These stratifications informed our decision to include age and sex as baseline variables for other factors to be chosen from.

### Key Risk Factor Associations

Table 2: Unadjusted Diabetes Prevalence by Key Risk Factors

Risk Factor	Category	Prevalence
HighBP	Yes	24.4%
	No	6.0%
BMI	Normal (18.5–24.9)	6.2%
	Obese (30.0–39.9)	21.2%
	Severe Obese ( $\geq 40.0$ )	30.6%
GenHlth	Excellent	2.5%
	Poor	37.9%
PhysActivity	Yes	11.6%
	No	21.1%
Smoker	No	12.1%
	Yes	16.3%

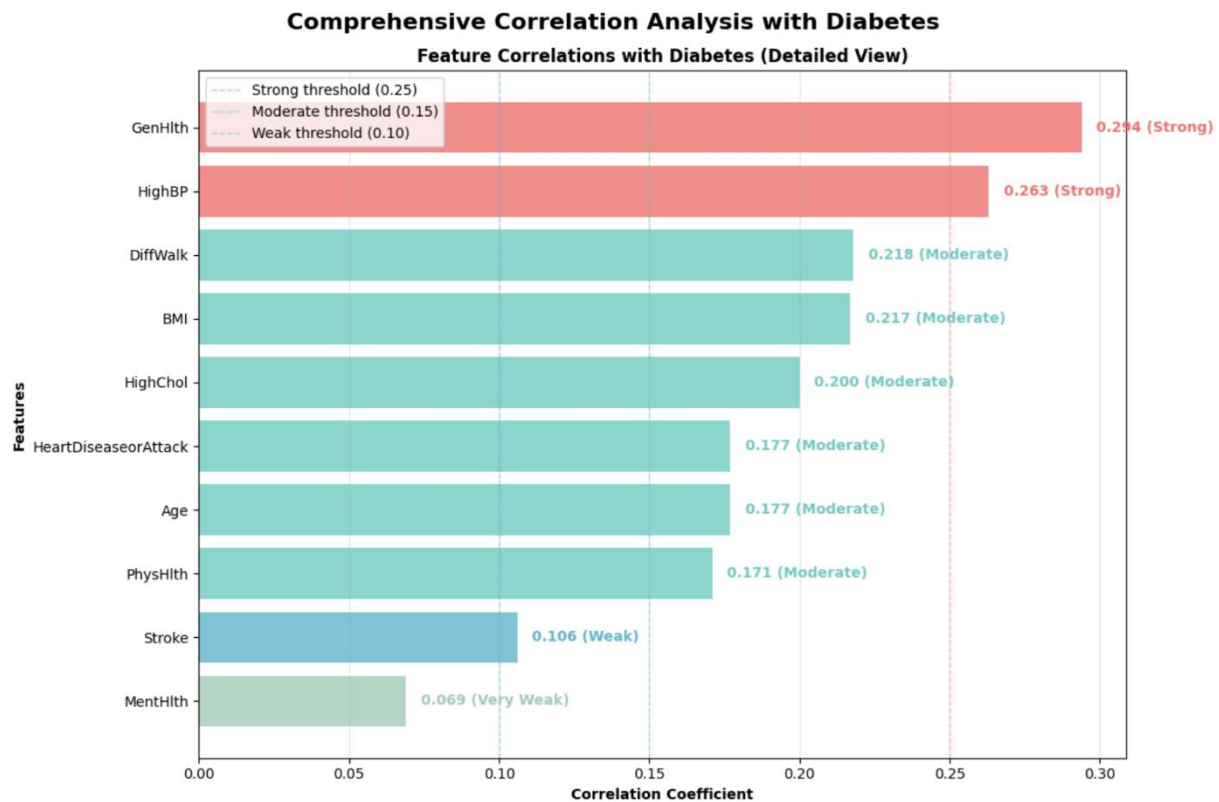
Unadjusted prevalence rates highlight strong gradients between certain risk factors and diabetes status (Table 2). Participants with hypertension exhibited a 24.4% diabetes rate versus 6.0% among those without. Obese and severely obese individuals showed prevalences

## Scalable LLMs for Diabetes Screening

of 21.2% and 30.6%, respectively, compared to 6.2% in normal-weight peers. Additional differentials included 37.9% prevalence among those rating general health as “poor” versus 2.5% among those rating “excellent,” and 21.1% among inactive versus 11.6% among physically active adults.

## Correlation with Diabetes

Figure 2: Correlation Coefficients with Diabetes



Pearson correlation coefficients identified the strongest linear associations with diabetes status (Fig 2). Self-rated general health and hypertension demonstrated the highest correlations (0.294 and 0.263, respectively), followed by mobility difficulties (0.218), BMI (0.217), and dyslipidemia (0.200). Age (0.177), heart disease history (0.177), and days of poor physical health (0.171) exhibited moderate associations, whereas days of poor mental health had the weakest linear relationship (0.069). These insights guided feature selection and regularization strategies in multivariate modeling.

## Methods

### Dataset Rationale and Novelty

This study leverages the CDC’s publicly available BRFSS dataset as a pragmatic foundation for evaluating large language models (LLMs) in diabetes prediction, avoiding lengthy ethics approvals associated with primary medical data collection. Although LLMs

## Scalable LLMs for Diabetes Screening

have excelled in language tasks, their application to structured clinical prediction remains largely unexplored. The comprehensive BRFSS sample, free of missing data, provides an ideal baseline for assessing whether pretrained LLMs can identify complex health patterns without extensive domain fine-tuning. By juxtaposing LLMs with traditional machine learning algorithms, this thesis aims to establish a benchmark for future integration of specialized clinical datasets.

### Model Selection

Model choice was driven by computational constraints and the need for reproducibility on Google Colab, which offers up to 40 GB of GPU RAM. The 8 billion-parameter LLAMA 3.1 B Instruct model was selected from Meta, accessible via Hugging Face, because of its open-access license and efficient inference profile<sup>1</sup>. Larger variants (70 B and 405 B) proved impractical given memory limits and latency, while proprietary alternatives incurred prohibitive usage costs for thousands of calls. The LLAMA Instruct architecture, optimized for following conversational prompts, aligns seamlessly with our interactive, feature-selection framework.

### LLM Architecture and Pretraining

We adopt Meta’s LLAMA 3.1 B Instruct model, pretrained on roughly 15 trillion tokens drawn from web pages, books, academic articles, and open-source code (knowledge cutoff December 2023). Pretraining minimizes next-token cross-entropy, yielding contextualized embeddings through  $\text{\$L\$}$  stacked Transformer layers. Each layer applies multi-head self-attention and position-wise feed-forward transforms:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where  $Q, K, V$  are learned projections of the input embeddings.

After base pretraining, the model underwent instruction tuning on publicly-released “helpful assistant” dialogues plus 25 million synthetic Q&A examples, sharpening its ability to follow multi-step prompts.

#### *Inference and Decoding Mechanics*

At inference, prompts are tokenized to IDs, embedded, and enriched with positional encodings. These pass through  $L$  Transformer layers, each outputting logits  $ei$  for token  $i$ . We incorporate a learned bias offset  $bi$  before softmax:

---

<sup>1</sup> In order to run the LLAMA 3.1B model a google cola po + subscription was required with an additional 1000 compute units was purchased for testing and final evaluation

$$p_i = \frac{\exp(l_i + b_i)}{\sum_j \exp(l_j + b_j)}$$

For our binary diabetes classification, we identify two special tokens, “yes” ( $i = y$ ) and “no” ( $i = n$ ), and interpret  $p_{\text{yes}}$  and  $p_{\text{no}}$  as risk estimates.

Table 3 summarizes our decoding configuration, which enforces greedy, reproducible output.

Table 3: Greedy Decoding Parameters for LLM

Parameter	Setting	Purpose
do_sample	False	Disables stochastic sampling
temperature	0.0	Sharpens distribution to argmax
top_p	1.0	No nucleus sampling effect
top_k	1	Only highest-probability token chosen
repetition_penalty	1.0	Neutral penalty to avoid sampling bias

#### *Token-Bias Calibration*

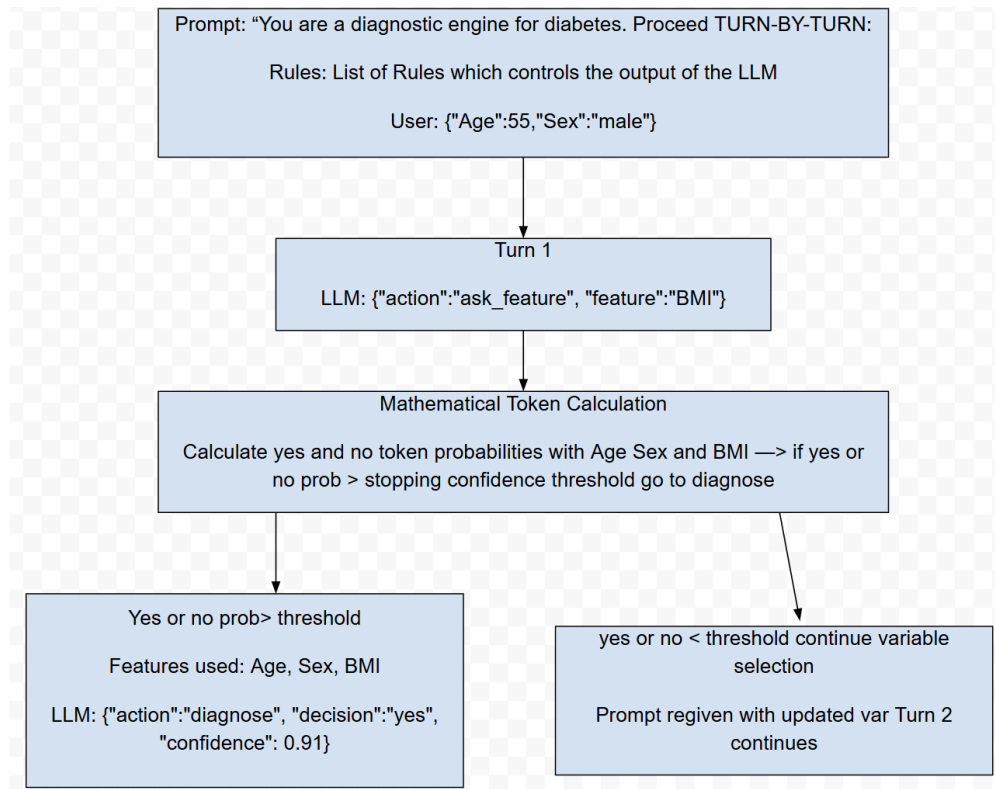
To embed the ~13 percent diabetes prevalence prior and prioritize sensitivity, we add a learned bias offset to the “yes” token. This offset is tuned via cross-validation(30 patients)—optimizing recall for each model separately. By penalizing missed cases more heavily(through optimizing recall) than false positives, this calibration ensures that screening errs on the side of caution. With decoding parameters and bias calibration in place, we transition to our Procedural Framework, which details how each subject is processed in a turn-based chat interface, producing reproducible, probabilistically grounded diagnoses.

## Scalable LLMs for Diabetes Screening

### *Procedural Interaction Framework*

We implement a turn-based chat where all patient data are passed as “user” messages.

Figure 3: Turn-by-turn interaction and decision loop.



Key design choices:

Age and Sex are supplied first because every screening encounter collects them; they help tailor subsequent feature requests (e.g. older males vs. younger females will collect different features in different orders).

The model decides between two “actions”:

- ask\_feature → continue eliciting data
- diagnose → terminate with a binary decision by comparing to the confidence threshold and yes threshold.

### *Decision Logic and Thresholds*

At every turn, we compute two values:



## Scalable LLMs for Diabetes Screening

$p\_yes$ : the model’s probability for “yes” (diabetes)

$p\_no$ : the model’s probability for “no” (no diabetes)

We then apply the following rules:

1. Confidence Check
  - Let  $\tau\_conf$  be our confidence threshold (e.g. 0.90).
  - If either  $p\_yes \geq \tau\_conf$  or  $p\_no \geq \tau\_conf$ , we stop soliciting features and issue a diagnosis.
2. Class Assignment
  - We introduce a secondary yes-bias threshold  $\tau\_yes$  ( $\tau\_yes \leq \tau\_conf$ ) to favor the diabetic class:
    - If  $p\_yes \geq \tau\_yes$ , classify as diabetic, even if  $p\_no > p\_yes$ .
    - Otherwise, pick the class with the higher probability: – If  $p\_yes > p\_no$ , diabetic – If  $p\_no > p\_yes$ , non-diabetic – If  $p\_yes = p\_no$ , default to diabetic (tie-break rule)
3. Feature Solicitation
  - If neither probability meets  $\tau\_conf$ , the loop asks for the next highest-information-gain feature.
  - This continues until we either hit  $\tau\_conf$  or exhaust the maximum allowed features (model-dependent).
4. Forced Decision
  - Once the feature budget is reached without meeting  $\tau\_conf$ , we force a final decision by applying the same Class Assignment rules above.

The exact values of  $\tau\_conf$  and  $\tau\_yes$  are calibrated per model in the next section.

### Example Patients

Two representative patient profiles illustrate how each model navigates feature selection and diagnosis. Patient 1, a confirmed diabetic, exhibits multiple high-risk indicators, whereas Patient 2 remains non-diabetic despite moderate risk factors. Tables 4 and 5 summarize their baseline characteristics and humanized risk annotations.

#### *Patient 1: Diabetic Profile*

Table 4: Patient 1 Feature Profile

Feature	Value	Humanized	Diabetes Risk
Age	10.0	65–69 years	⚠️ High risk (older age)
Sex	0.0	Female	✅ Not risk factor (women have lower rates)
HighBP	1.0	Yes	⚠️ Major risk factor

## Scalable LLMs for Diabetes Screening

Feature	Value	Humanized	Diabetes Risk
HighChol	0.0	No	✓ Protective
CholCheck	1.0	Yes	✓ Good healthcare
BMI	36.0	36 kg/m <sup>2</sup>	⚠ Major risk factor (obese)
Smoker	1.0	Yes	⚠ Risk factor
Stroke	0.0	No	✓ Protective
HeartDiseaseorAttack	0.0	No	✓ Protective
PhysActivity	0.0	No	⚠ Major risk factor (sedentary)
Fruits	0.0	No	⚠ Risk factor (poor diet)
Veggies	1.0	Yes	✓ Protective
HvyAlcoholConsump	0.0	No	✓ Protective
AnyHealthcare	1.0	Yes	✓ Good healthcare
NoDocbcCost	1.0	Yes	⚠ Risk factor (cost barrier)
GenHlth	3.0	Good	⚠ Moderate risk
MentHlth	0.0	0 days	✓ Good mental health
PhysHlth	0.0	0 days	✓ Good physical health
DiffWalk	1.0	Yes	⚠ Major risk factor (mobility issues)
Education	3.0	Grades 9–11	⚠ Risk factor (lower education)
Income	4.0	\$20–25 k	⚠ Risk factor (lower income)

Table 4 details Patient 1’s feature values, their clinical interpretation, and corresponding diabetes risk flags. Advanced age, obesity (BMI 36 kg/m<sup>2</sup>), hypertension, smoking, sedentary behavior, mobility limitations, lower education, and limited income combine to create a high-risk profile.

*Patient 2: Non-Diabetic Profile*

Table 5: Patient 2 Feature Profile

Feature	Value	Humanized	Diabetes Risk
Age	5.0	40–44 years	⚠ Moderate risk (middle-aged)
Sex	1.0	Male	⚠ Risk factor (men have higher rates)
HighBP	1.0	Yes	⚠ Major risk factor
HighChol	1.0	Yes	⚠ Risk factor
CholCheck	1.0	Yes	✓ Good healthcare

## Scalable LLMs for Diabetes Screening

Feature	Value	Humanized	Diabetes Risk
BMI	28.0	28 kg/m <sup>2</sup>	⚠️ Risk factor (overweight)
Smoker	0.0	No	✅ Protective
Stroke	0.0	No	✅ Protective
HeartDiseaseorAttack	0.0	No	✅ Protective
PhysActivity	1.0	Yes	✅ Major protective factor
Fruits	1.0	Yes	✅ Protective (healthy diet)
Veggies	1.0	Yes	✅ Protective (healthy diet)
HvyAlcoholConsump	0.0	No	✅ Protective
AnyHealthcare	1.0	Yes	✅ Good healthcare
NoDocbcCost	0.0	No	✅ Good access
GenHlth	3.0	Good	⚠️ Moderate risk
MentHlth	0.0	0 days	✅ Good mental health
PhysHlth	0.0	0 days	✅ Good physical health
DiffWalk	0.0	No	✅ Protective (good mobility)
Education	5.0	Some college/technical	✅ Protective (higher education)
Income	8.0	> \$75 k	✅ Protective (higher income)

Table 5 presents Patient 2’s attributes and risk annotations. Although middle-aged with hypertension, high cholesterol, and overweight status, strong protective factors—regular exercise, healthy diet, reliable healthcare access, higher education, and income—contribute to a mixed-risk but ultimately non-diabetic profile.

### Zero-Shot Model


The LLM’s out-of-the-box diagnostic ability was evaluated by prompting it simply as a “diabetes chatbot” with no contextual examples. Initial trials on 1000-record subsets exposed its overwhelming bias toward “no” responses, so a `yes_bias` of 2.0 was introduced to double the model’s propensity to answer “yes.” An early stopping `confidence_threshold` of 0.25 and a final `yes_threshold` of 0.25 to trigger diagnoses once minimal certainty was reached.

In deployment, the zero-shot LLM asks one feature at a time until its top token probability (either “yes” or “no”) exceeds 25%. Once it stops querying, the final “yes” probability is compared to the 0.25 threshold: values  $\geq 0.25$  yield a diabetes diagnosis; values below yield non-diabetic.

### *Patient Journey*


## Scalable LLMs for Diabetes Screening

Table 6: Zero-Shot Model—Patient 1 Interaction

Turn	Queried Feature	Features Collected	P(Yes)	P(No)	Decision
1	HighBP	Age, Sex, HighBP	3.4%	16.6%	Continue (16.6% < 25%)
2	HighChol	Age, Sex, HighBP, HighChol	32.1%	24.9%	Stop & predict “yes” (32.1% ≥ 25%) 

Patient 1 (Diabetic) The model first queries HighBP, raising P(Yes) to 3.4% (still below the stop rule). On the second turn, it asks HighChol, and P(Yes) jumps to 32.1%, exceeding both the 0.25 stop rule and the 0.25 yes\_threshold. The model stops after two features and correctly predicts diabetes, driven by the injected bias rather than deeper context.

Table 7: Zero-Shot Model—Patient 2 Interaction

Turn	Queried Feature	Features Collected	P(Yes)	P(No)	Decision
1	HighBP	Age, Sex, HighBP	3.1%	17.2%	Continue (17.2% < 25%)
2	HighChol	Age, Sex, HighBP, HighChol	60.3%	7.6%	Stop & predict “yes” (60.3% ≥ 25%) 

Patient 2 (Non-Diabetic) Using the same two queries, P(Yes) reaches 60.3% on Turn 2—again above the stop and yes thresholds—so the model prematurely stops and misclassifies as diabetic. This false positive illustrates how extreme yes\_bias and low thresholds can override protective factor signals.

**Few-Shot Model**

Table 8: Few-Shot Prompt Examples

Example ID	Description	Key Features
1	High-Risk Diabetic Patient	BMI, HighBP, PhysActivity
2	Low-Risk Non-Diabetic Patient	BMI, HighBP, PhysActivity
3	Elderly High-Risk Patient	BMI, HighBP, PhysActivity
4	Middle-Aged Low-Risk Patient	BMI, HighBP, PhysActivity
5	Elderly High-Risk Patient	BMI, HighBP, PhysActivity

In the few-shot paradigm, the LLM was supplied with five illustrative examples spanning high- and low-risk profiles across age and sex strata (Table 8). These examples included two elderly high-risk patients, one high-risk diabetic, one middle-aged low-risk, and one low-risk non-diabetic. Gender was deliberately (three males, two females) and anchored feature importance by ordering BMI first, followed by high blood pressure and physical

## Scalable LLMs for Diabetes Screening

activity. This concise context—limited to three features per example—helped the model internalize relevant medical patterns without overwhelming it.

Table 9: Few-Shot Model Hyperparameters

Parameter	Value	Purpose
yes_bias	0.40	Increases “yes” token probability
confidence_threshold	0.90	Stop querying once max token prob $\geq 0.90$
yes_threshold	0.70	Final “yes” probability threshold for diabetes

The model was tuned with a moderate yes\_bias of 0.4 to counteract its “no” propensity, set a strict confidence threshold of 0.90 for halting feature solicitation, and imposed a yes\_decision threshold of 0.70 on the final “yes” token probability (Table 9). Although these thresholds are more stringent than those used in zero-shot prompting, they underscore the model’s enhanced confidence in both selecting informative features and making its final diabetes prediction. By capping example prompts at three features and calibrating these parameters, the model struck a balance between medical realism and diagnostic precision.

### *Patient Journey*

Table 10: Few-Shot Model—Patient 1 Interaction

Turn	Features Collected	P(Yes)	P(No)	Confidence	Threshold	Decision
1	Age, Sex, HighBP	0.899	0.101	0.899	0.90	Continue (0.899 < 0.90)
2	Age, Sex, HighBP, HighChol	0.721	0.279	0.721	0.90	Continue (0.721 < 0.90)
3	Age, Sex, HighBP, HighChol, CholCheck	0.787	0.213	0.787	0.90	Continue (0.787 < 0.90)
4	Age, Sex, HighBP, HighChol, CholCheck, BMI	0.975	0.025	0.975	0.90 & 0.70	Stop & predict diabetes (✅)

Patient 1 (Diabetic) From Turn 1, P(Yes) = 0.899 hints at diabetes but remains below the 0.90 confidence\_threshold. The model sequentially adds HighChol and CholCheck, yet still falls short. On Turn 4, including BMI elevates P(Yes) to 0.975—surpassing both 0.90 and 0.70—so it stops and correctly diagnoses diabetes after four features.

Table 11: Few-Shot Model—Patient 2 Interaction

Turn	Features Collected	P(Yes)	P(No)	Confidence	Threshold	Decision
1	Age, Sex, HighBP	0.909	0.091	0.909	0.90 & 0.70	Stop & predict diabetes (false positive)(✗)

Patient 2 (Non-Diabetic) At Turn 1, querying HighBP immediately yields  $P(\text{Yes}) = 0.909$ , exceeding both the 0.90 confidence and 0.70 yes thresholds. The model stops and misclassifies this mixed-risk patient as diabetic, reflecting its sensitivity to few-shot exemplars that emphasized elderly hypertensive males.

The few-shot model demonstrated marked improvement over zero-shot, requiring lower bias and yielding realistic confidence levels (0.55–0.62) during preliminary testing. However, its heavy reliance on provided examples can mislead predictions in borderline cases where protective factors (e.g., physical activity, diet) were under-emphasized. This highlights the trade-off between specificity and recall inherent in few-shot prompting.

### Bandit Algorithm Model Hybrid

$$\text{UCB}_t(f) = \mathbf{c}_t^\top \widehat{\boldsymbol{\theta}}_f + \alpha \sqrt{\mathbf{c}_t^\top \mathbf{A}_f^{-1} \mathbf{c}_t}$$

- $\mathbf{f}$  = Feature arm (one of 19 features or STOP)
- $\mathbf{c}_t$  = Patient context vector [age, sex] at round  $t$
- $\boldsymbol{\theta}_f$  = Estimated reward parameters for feature  $f$
- $\mathbf{A}_f$  = Covariance matrix for feature  $f$
- $\alpha = 1.0$  = Exploration parameter

Contextual bandits frame diabetes screening as a sequential decision problem: at each round  $t$ , the agent observes patient context  $\mathbf{c}_t$  (age and sex), selects one of 19 feature arms ( $\mathbf{f}$ ) or a STOP arm, and receives a reward based on the LLM’s prediction accuracy. The Linear Upper Confidence Bound (LinUCB) algorithm was implemented, modeling each arm’s expected reward as a linear function of patient context and inflating confidence intervals with an exploration parameter  $\alpha = 1.0$ . This design balances exploration of underutilized features—ensuring the algorithm learns their predictive value—with exploitation of high-yield features to maximize diagnostic accuracy.

LinUCB was evaluated against four alternative bandit strategies to justify its selection.  $\epsilon$ -Greedy explores randomly with probability  $\epsilon$  but neglects patient context, leading to suboptimal early queries. Thompson Sampling leverages Bayesian posterior sampling for context-aware decisions but incurs higher computational overhead and slower convergence

## Scalable LLMs for Diabetes Screening

on our moderate-sized feature set. UCB1 applies confidence bounds to empirical averages without context, preventing personalized feature solicitation. A Random Forest Bandit models nonlinear interactions via bootstrapped UCB but proved too resource-intensive for our Colab environment. LinUCB emerged as the best compromise between contextual personalization, low latency, and robust recall/accuracy trade-offs.

Reward engineering aligned the bandit’s objectives with clinical screening priorities. All correct predictions yield a base reward of +2.0, while incorrect non-diabetic predictions (false positives) incur −0.5. To emphasize diabetes detection, true positives receive an additional +1.3 bonus, and false negatives incur a −1.8 penalty; false positives carry a further −1.1 penalty. Bandit also incentivized inclusion of primary risk factors—HighBP, BMI, HeartDiseaseorAttack, Age, HighChol, Stroke—with incremental rewards: +0.5 for four or more factors, +0.3 for three, +0.2 for two, and +0.1 for one. Feature-efficiency bonuses of +0.2 applied for selecting 7–10 or 11+ features, whereas selecting six or fewer features incurred −0.1 and selecting more than 15 features incurred −0.2. Finally, diabetic-case bonuses of +0.2 each rewarded use of eight or more total features and inclusion of three or more primary factors in true-positive cases.

Classification relies on the LLM’s “yes” token probability compared against a dynamically adjusted threshold  $TT$ . The model initialized  $TT = 0.24$  and adjust based on patient risk stratum and age group: high risk ( $\geq 4$  primary factors) lowers  $TT$  to 0.18, medium-high (3 factors) to 0.22, medium (2) remains 0.24, and low ( $< 2$ ) raises to 0.28; age  $\geq 65$  subtracts 0.02, ages 45–64 subtract 0.01, and  $< 45$  adds 0.01. Every 100 patients, performance-based tuning targets recall  $\geq 0.83$  and accuracy  $\geq 0.65$ : thresholds shift by  $\pm 0.02$  for recall deviations outside 0.78–0.88 and by  $\pm 0.005$ –0.015 for accuracy deviations outside 0.45–0.70, bounded within [0.15, 0.35].

Below, two patient journeys demonstrate how feature selection and threshold dynamics yield correct diagnoses.

### *Patient Journey*

Table 12: Bandit Model—Patient 1 Interaction

Turn	Bandit Action	Features Collected	P(Yes)	P(No)	Threshold	Decision
1	Fruits	Sex, Age, Fruits	—	—	—	—
2	CholCheck	Sex, Age, Fruits, CholCheck	—	—	—	—
3	NoDocbcCost	Sex, Age, Fruits, CholCheck, NoDocbcCost	—	—	—	—
4	STOP	(5 features total)	49.4%	50.6%	30.0%	yes (✓ correct)

## Scalable LLMs for Diabetes Screening

For Patient 1, the bandit queries dietary quality (“Fruits”), healthcare engagement (“CholCheck”), and economic access (“NoDocbcCost”), then invokes STOP. The LLM outputs a 49.4% “yes” probability. Starting at 0.24, the high-risk adjustment ( $\geq 4$  primary factors) lowers TT to 0.18, the age  $\geq 65$  adjustment further lowers TT to 0.16, and performance-based tuning raises it to 0.30. Because 49.4% exceeds 30.0%, the model correctly diagnoses diabetes even though the no probability is less than the yes.

Table 13: Bandit Model—Patient 2 Interaction

Turn	Bandit Action	Features Collected	P(Yes)	P(No)	Threshold	Decision
1	BMI	Sex, Age, BMI	—	—	—	—
2	DiffWalk	Sex, Age, BMI, DiffWalk	—	—	—	—
3	Smoker	Smoker, Sex, Age, DiffWalk, BMI	—	—	—	—
4	STOP	(5 features total)	22.8%	77.2%	27.0%	no (✓ correct)

For Patient 2, the bandit selects BMI, mobility (“DiffWalk”), and tobacco use (“Smoker”) before stopping. The LLM’s “yes” probability is 22.8%. The threshold remains 0.24 for medium risk (2 primary factors), increases to 0.25 for age  $< 45$ , and is raised to 0.27 via performance adjustments. As 22.8% falls below 27.0%, the model correctly predicts no diabetes. This comprehensive contextual-bandit framework—combining LinUCB, finely tuned rewards, and adaptive thresholds—enables personalized feature solicitation, prioritizes screening sensitivity, and dynamically balances recall and specificity in diabetes detection.

## LLM and Random Forest Model

The hybrid pipeline marries LLM-driven feature selection with a Random Forest (RF) classifier trained on the same streamlined feature set used in the few-shot examples. The top predictors were reused (BMI, HighBP, PhysActivity, HighChol, HeartDiseaseorAttack, Stroke) rather than rerunning full feature selection—this choice preserved computational resources on Colab and ensured consistency between LLM selection and RF training. All chosen features carry strong explanatory power for diabetes, so narrowing the candidate pool did not sacrifice predictive relevance.

For the LLM, hyperparameters were adjusted to encourage aggressive yet focused feature gathering. Yes\_bias was increased to 0.55 (compared with 0.40 in the few-shot model) so “yes” probabilities build more quickly even though the LLM does not make the final call. A high confidence\_threshold was set of 0.88 to determine when to stop querying—forcing the LLM to collect sufficient context before halting—and a yes\_threshold of 0.48 was




## Scalable LLMs for Diabetes Screening

set for internal stopping logic, acknowledging that higher early confidence translates to richer feature sets for the RF.

The RF itself uses 200 trees ( $n\_estimators=200$ ), unlimited depth ( $max\_depth=None$ ), bootstrap sampling, and a  $class\_weight$  of  $3\times$  on the positive (diabetes) class to counter the 6.2:1 imbalance. At inference, diabetic is classified if the RF's positive-class probability exceeds 0.12, biasing recall toward diabetes detection. To further safeguard against false negatives, we implement a fallback: if the RF predicts “no” but its positive-class probability is above 0.25—and the patient possesses two or more primary risk factors (HighBP, BMI, HeartDiseaseorAttack, HighChol, Stroke)—the model override to “yes.” This conservative safety net is acceptable in a screening context, where missing a case carries higher cost than a false alarm.


### *Patient Journey*

Table 14: Hybrid Model—Patient 1 (Diabetic)

	<b>Turn Action</b>	<b>Yes Prob</b>	<b>No Prob</b>	<b>Decision</b>	<b>Notes</b>
1	LLM adds HighBP	—	—	—	First feature flagged by LLM
2	LLM stops at [Age, Sex, HighBP]	93.0%	7.0%	Stop LLM (0.93 > 0.88)	High confidence halts feature collection
3	RF evaluates [Age, Sex, HighBP]	19.0%	—	Predict “yes” (0.19 > 0.12) 	0.19 > 0.12 threshold—correctly classifies diabetes

For Patient 1, the LLM stops after identifying hypertension along with age and sex. Although the RF's 19.0% positive probability seems low, it exceeds the 12% decision threshold, yielding a correct diabetes call.

Table 15: Hybrid Model—Patient 2 (Non-Diabetic)

	<b>Turn Action</b>	<b>Yes Prob</b>	<b>No Prob</b>	<b>Decision</b>	<b>Notes</b>
1	LLM adds HighBP	—	—	—	LLM begins with hypertension status
2	LLM stops at [Age, Sex, HighBP]	89.0%	11.0%	Stop LLM (0.89 > 0.88)	Confidence threshold reached
3	RF evaluates [Age, Sex, HighBP]	8.2%	—	Predict “no” (0.082 < 0.12) 	0.082 < 0.12 threshold—correctly classifies non-diabetic; no fallback triggered

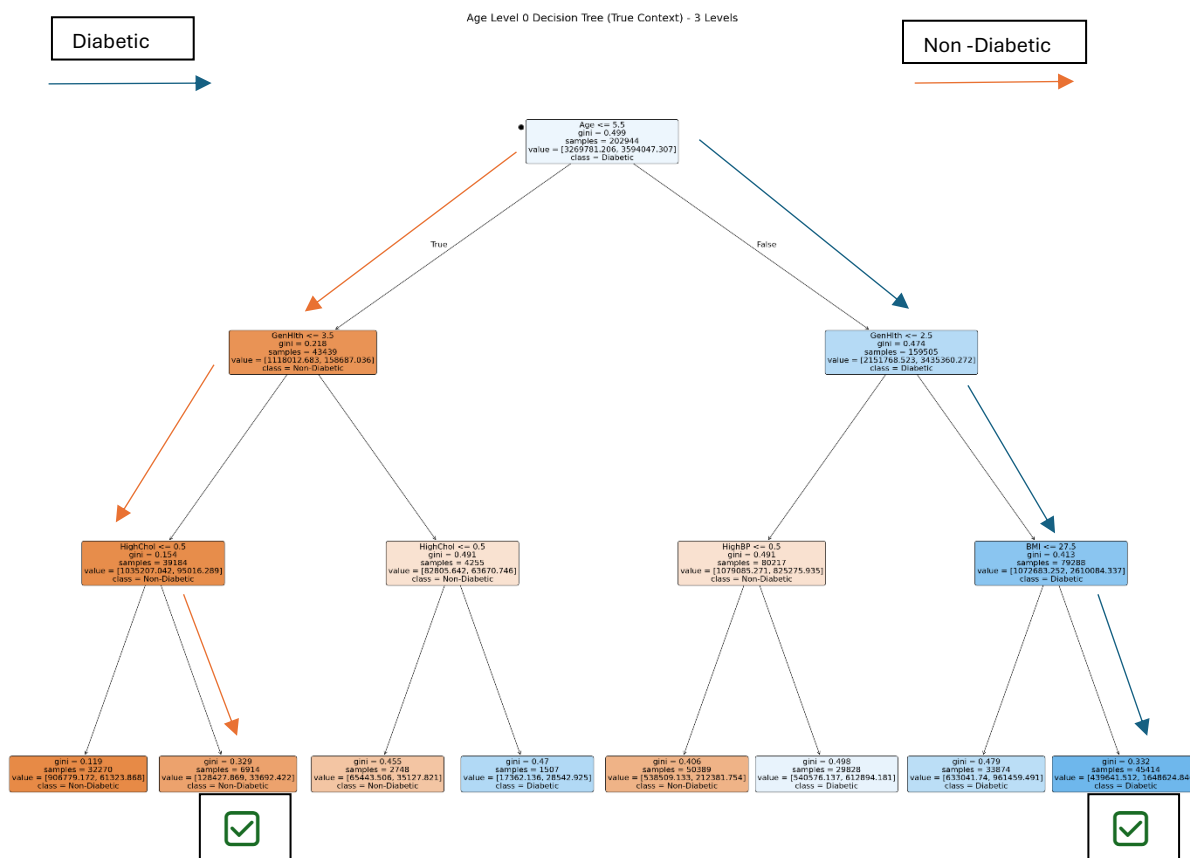
## Scalable LLMs for Diabetes Screening

For Patient 2, the LLM again stops after two features. The RF's 8.2% positive probability lies below the 12% threshold, and since it is also below the 25% fallback threshold, the model correctly predicts non-diabetes. This hybrid strategy leverages rapid, context-aware LLM feature selection and a rigorously tuned RF classifier—bounded by low decision thresholds and a conservative fallback—to optimize diabetes screening recall without overwhelming computational budgets.

### Baseline Random Forest and Decision Tree

To establish a reference point for evaluating our LLM-driven approaches, we implemented a baseline Random Forest (RF) classifier and a single Decision Tree. Both models underwent hyperparameter tuning—primarily balancing classes via weighted sampling and cross validation—to increase sensitivity. This baseline reveals whether advanced feature-selection techniques and adaptive thresholds truly outperform a straightforward ensemble or tree. Although our primary focus is the RF, a representative decision tree was included (Figure 4) to illustrate an uncured patient journey through conventional splits. Attempt were made to enforce age and sex as the upper nodes, reasoning that these demographics are foundational risk factors. In practice, the algorithm ignored sex—deeming it uninformative—and placed age at the root only after applying heavy sample weighting. Subsequent splits on general health and BMI simulate a more traditional, rule-based pathway.

Figure 4: Decision Tree for Baseline Model



*Patient Journey*

## Patient 1

For the diabetic patient, the tree routes (highlighted in red) as follows: at the root, age 65–69 years (encoded as value 10) sends the path to the right child. There, a general health rating of “3” (Good) again directs to the right. Finally, a BMI of 36 falls into the diabetic leaf node. This branch correctly classifies the patient as diabetic but exhibits a Gini impurity of approximately 0.66—indicating that 34 percent of cases at this node would be misclassified if one relied solely on this split.

## Patient 2

In the non-diabetic scenario (highlighted in blue), age 40–44 years (value 5) guides the path left from the root. A “Good” general health rating (3) continues left, and a high-cholesterol flag (1) leads to a non-diabetic leaf. The Gini impurity at this terminal node is about 0.32, reflecting that 32 percent of cases traversing this branch are diabetic, which tempers confidence in the decision.

Table 1: Baseline Model Performance of Random Forest

Metric	Value
Accuracy	69 %
Sensitivity	78 %

Despite its simplicity, this baseline of the random forest performs respectably: overall accuracy reaches 69 percent, and sensitivity (recall) for detecting diabetes is 78 percent. These metrics underscore both the utility and limitations of conventional models and set a clear benchmark for the incremental gains achieved by our LLM-based and hybrid methods.

**Model Training and Implementation Details**

All models were trained using an 80/20 train–test split, with fixed random seeds to ensure reproducibility. A comprehensive suite of performance metrics was reported (accuracy, sensitivity, specificity, AUC, F1 score, etc.), which are presented in the Results section. Large language model (LLM) inference posed significant computational demands. To optimize speed and memory efficiency, all LLM calls were ran in FP16 half-precision and enabled deterministic mode. Response caching was implemented to avoid redundant API calls and periodically saved intermediate outputs, mitigating the risk of data loss due to Google Colab timeouts.

Key implementation challenges included handling diverse tokenization schemes and extensive JSON parsing to extract feature prompts and model decisions. We addressed these by:

## Scalable LLMs for Diabetes Screening

1. Standardizing on a single tokenizer configuration across prompts and responses
2. Writing robust JSON parsers with error handling for incomplete or malformed payloads
3. Employing batch processing to evaluate multiple patients in parallel, reducing overall runtime
4. Integrating with Google Drive for automated result storage and checkpoint recovery

Together, these practices ensured consistent, efficient execution of all our LLM-driven and hybrid modeling experiments in a cloud-based environment.

## Evaluation Framework

All models were first assessed with standard confusion-matrix metrics—accuracy, precision, recall (sensitivity), specificity and F1-score to quantify raw predictive accuracy. We then computed balanced accuracy and the Matthews correlation coefficient to mitigate class-imbalance effects and ensure robust performance comparisons.

To capture estimation uncertainty, every metric was reported with 95 % bootstrap confidence intervals (10 000 resamples), providing distribution-free error bounds. Model results were always benchmarked against two naïve baselines—a majority-class predictor and a prevalence-based predictor (13.9 % diabetes rate)—to contextualize improvements over trivial strategies.

For clinical relevance, we enforced a minimum recall of 83 % (the standard screening threshold), computed patient-level cost  $C = 100 \times \text{FalsePos} + 10\,000 \times \text{FalseNeg}$  to reflect the asymmetric economic impact of misdiagnoses, and measured throughput (patients/sec) to confirm operational feasibility. Statistical differences in accuracy were tested via two-tailed Z-tests, with Cohen’s gauging practical significance (negligible:  $|d| < 0.2$ ; small: 0.2–0.5; medium: 0.5–0.8; large:  $\geq 0.8$ ).

We evaluated calibration in ten-bin reliability diagrams and applied Platt scaling or isotonic regression when needed to align predicted probabilities with true outcome frequencies—essential for trustworthy risk estimates. Discriminative ability was summarized by ROC and precision–recall curves, reporting AUC-ROC and AUC-PR to capture performance across all decision thresholds.

Finally, we probed model blind spots and fairness by (1) identifying the top five features most associated with false positives and false negatives to expose systematic errors; (2) conducting subgroup analyses across age, BMI, and self-reported health to reveal demographic performance gaps; and (3) executing a detailed cost-effectiveness analysis—total cost, cost per true positive, and savings versus no screening—to quantify the economic value of deployment.

**Summary Table of All Models**

Model	Key Parameters	Feature Selection	Strengths	Limitations
Zero-Shot LLM	yes_bias=2.0, confidence_threshold=0.25, yes_threshold=0.25	Sequential one-at-a-time until confidence threshold	Simple, no training required	Extreme bias can cause false positives
Few-Shot LLM	yes_bias=0.40, confidence_threshold=0.90, yes_threshold=0.70	Sequential with 5 example prompts (3 features each)	Improved confidence levels, realistic thresholds	Heavy reliance on examples can mislead
Bandit Algorithm (LinUCB) Hybrid	$\alpha=1.0$ , base_reward=+2.0, false_positive_penalty=-0.5, true_positive_bonus=+1.3	Contextual bandit with LinUCB algorithm	Personalized feature selection, adaptive thresholds	Complex reward engineering, computational overhead
LLM-RF Hybrid	yes_bias=0.55, confidence_threshold=0.88, yes_threshold=0.48, RF_threshold=0.12	LLM-driven sequential selection	Combines LLM context awareness with RF robustness	Dependency on LLM feature selection quality
Baseline Random Forest	n_estimators=200, max_depth=None, class_weight=3×, fallback_threshold=0.25	All features used	Established baseline, respectable performance	Fixed feature set, no adaptive selection

## Results

### Dataset and Stratification

The held-out test set of 50,736 individuals closely mirrored the overall cohort ( $n = 253,680$ ), with diabetes prevalence identical at 13.9% (7,069/50,736). Rates of hypertension (42.9% vs. 42.9%), high cholesterol (42.3% vs. 42.4%), cholesterol screening (96.4% vs. 96.3%), current smoking (44.4% vs. 44.0%), regular physical activity (75.5% vs. 75.7%), heavy drinking (5.8% vs. 5.6%), and functional limitations (16.8% vs. 16.8%) differed by less than one percentage point—differences too small to affect generalizability or model performance. The test set additionally reported a mean BMI of 28.40 (SD 6.65), with 35.9% overweight and 29.0% obese, as well as new lifestyle metrics—63.4% consumed fruit daily and 81.1% ate vegetables regularly—and mental/physical health days (median = 0, 75th percentile = 2–3 days). Age distributions (38.3% aged 55–69; 19.1%  $\geq 70$ ) and socioeconomic profiles (42.3% college graduates; 35.7% in the highest income quintile) also aligned closely with the training sample. No substantive discrepancies emerged, confirming that the test set faithfully represents the original population.

### Model Performance Overview

Table 16. Overview of model performance

Model	Balanced Accuracy	MCC	Efficiency (patients/sec)	FP/FN Cost Ratio	Screening Ready (Recall $\geq 83\%$ )
Zero Shot	0.800	0.286	255.0	16.5	No
Bandit	0.821	0.245	N/A	20.6	No
Few Shot	0.856	0.281	801.9	39.1	Yes
Hybrid	0.785	0.259	4.2	13.3	No
Random Forest	0.498	0.274	N/A	1.0	No
Majority Class	0.500	0.000	N/A	0.0	No
Prevalence-Based	0.500	0.000	N/A	0.0	no

Five prescreening algorithms were evaluated against clinical and operational benchmarks. All the models show an improvement in balanced accuracy over the benchmark models. Few Shot achieved the highest balanced accuracy (0.856) and was the only approach to meet the medical screening recall threshold ( $\geq 83\%$ ). While Few Shot delivers superior accuracy alongside the fastest throughput (801.9 patients/sec), Zero Shot and Bandit offer leaner runtimes at the expense of clinical readiness. Hybrid trailed with a balanced accuracy of 0.785. Although Random Forest attained the highest raw accuracy (0.837), its low sensitivity (0.410) yielded the poorest balanced accuracy (0.498).

**Classification Performance with Confidence Intervals**

Table 17. Classification metrics with 95% CIs

<b>Model</b>	<b>Accuracy [95 % CI]</b>	<b>Precision [95 % CI]</b>	<b>Recall [95 % CI]</b>	<b>F1-Score [95 % CI]</b>
Few Shot	0.446 [0.442–0.451]	0.189 [0.184–0.193]	0.901 [0.894–0.908]	0.312 [0.306–0.318]
Hybrid	0.558 [0.554–0.563]	0.209 [0.204–0.214]	0.779 [0.768–0.789]	0.330 [0.323–0.337]
Bandit	0.462 [0.458–0.466]	0.183 [0.178–0.187]	0.821 [0.812–0.831]	0.298 [0.292–0.304]
Zero Shot	0.568 [0.563–0.572]	0.220 [0.215–0.224]	0.823 [0.814–0.831]	0.347 [0.340–0.353]
Random Forest	0.775 [0.771–0.779]	0.415 [0.403–0.427]	0.411 [0.402–0.420]	0.413 [0.403–0.423]

All intervals are narrow, indicating stable performance across resamples. The Few Shot approach maximizes sensitivity—recall exceeds 90%—but this comes at the expense of low precision (<20%), yielding the second highest F1-score among LLM methods. Hybrid offers a more balanced profile, with moderate precision ( $\approx 21\%$ ) and recall ( $\approx 78\%$ ), translating into the highest overall accuracy. Bandit and Zero Shot both clear the 82% recall mark yet incur elevated false positives, reflected in their sub-30% F1-scores. By contrast, the Random Forest baseline achieves superior raw accuracy and precision but suffers from poor sensitivity ( $\approx 41\%$ ), underscoring its limitations for diabetes prescreening despite its strong performance on the majority class.

## Scalable LLMs for Diabetes Screening

### Statistical Significance Analysis

Figure 5: Error Bar Plot

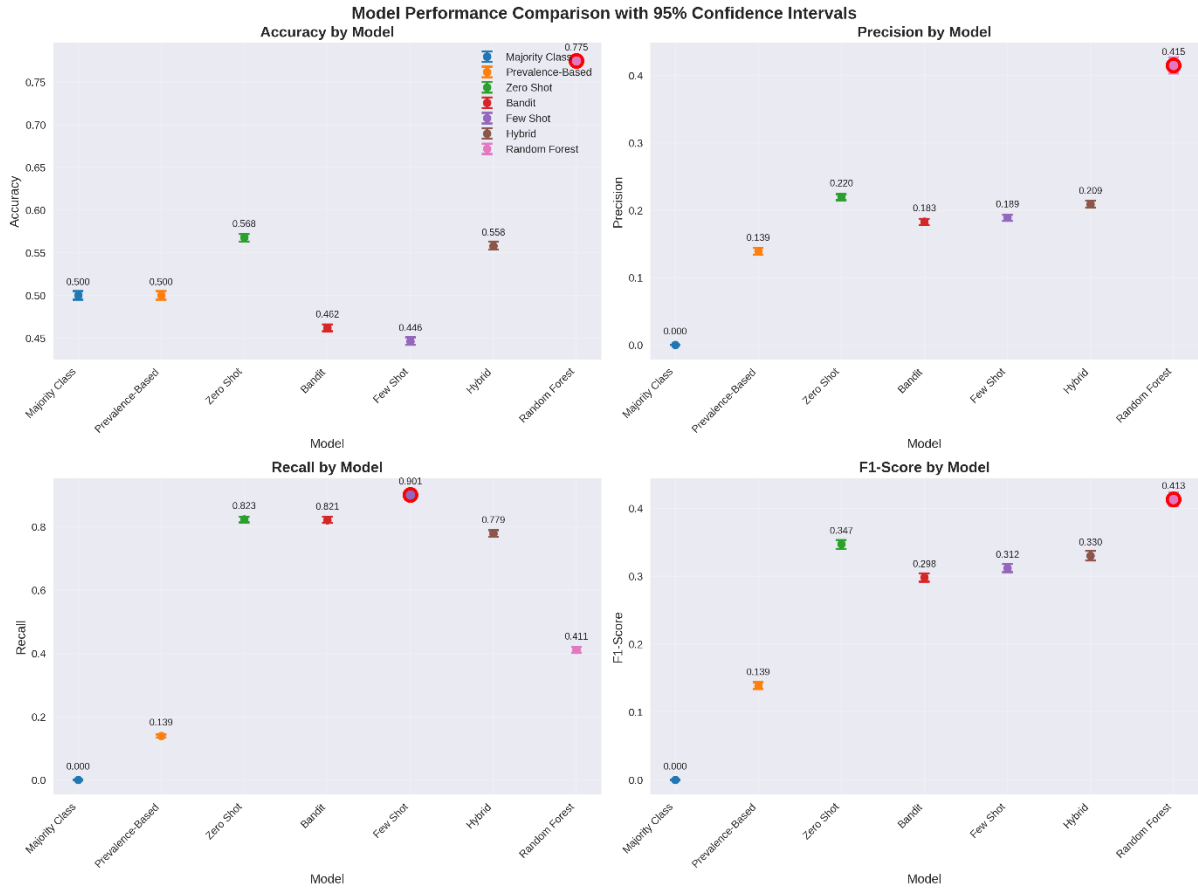


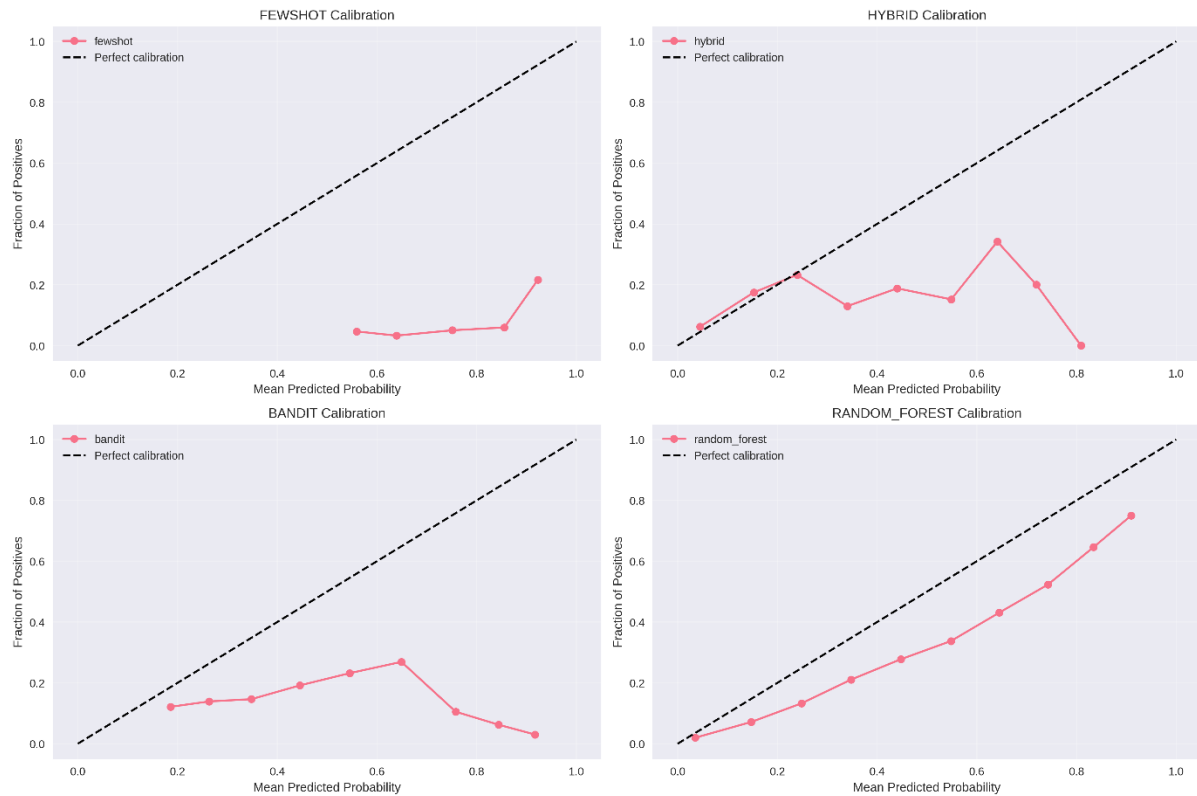
Figure 5 presents model performance comparison with 95% confidence intervals across all key metrics. Random Forest demonstrates superior performance in Accuracy (0.775), Precision (0.415), and F1-Score (0.413), with non-overlapping confidence intervals indicating statistically significant differences from other models. However, Random Forest shows poor Recall (0.411), falling well below the 83% screening threshold. Among LLM-based methods, Few Shot achieves the highest Recall (0.901), making it the only clinically viable option for diabetes screening. Zero Shot and Bandit also demonstrate strong Recall (0.823 and 0.821 respectively) but with lower Precision, reflecting the precision-recall trade-off inherent in screening applications. The naive baselines (Majority Class and Prevalence-Based) perform at chance levels across all metrics, providing meaningful comparison points for model evaluation. Other than accuracy, for all other metrics the models performed better than the naive baselines.

### Model Calibration and Reliability

Figure 6. Calibration Curves for all Models



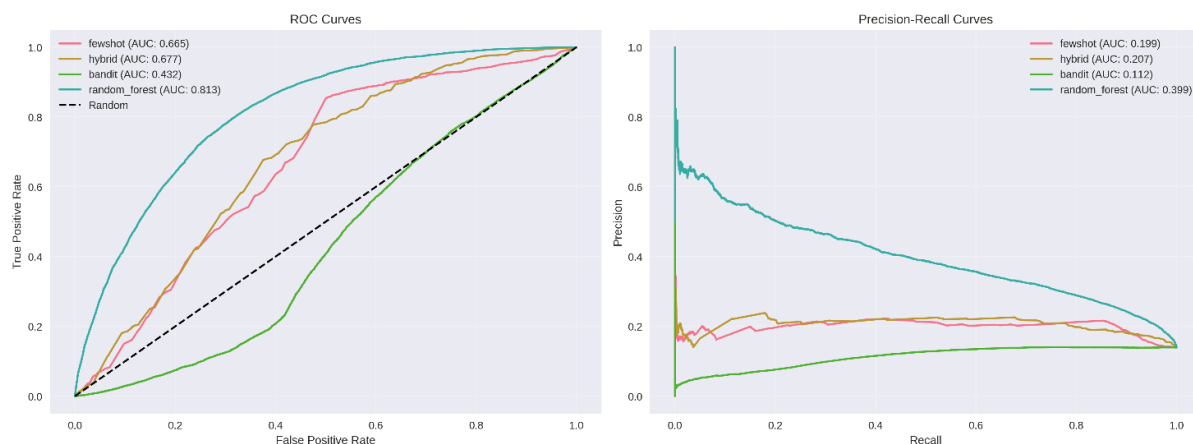
## Scalable LLMs for Diabetes Screening



Calibration analysis examines how well each model's predicted probabilities align with observed outcomes. Perfect calibration means when a model predicts 70% probability of diabetes, approximately 70% of such cases actually have diabetes. This is crucial for clinical decision-making where probability estimates inform treatment decisions. Random Forest exhibits near-ideal calibration, with its curve tightly hugging the diagonal across the full risk spectrum. Hybrid tracks the diagonal closely up to mid-range probabilities (0.0–0.6) but slightly underestimates risk in the highest bins. Few Shot shows pronounced miscalibration: it systematically overpredicts risk at low probabilities ( $< 0.3$ ) and high probabilities ( $> 0.8$ ), with its curve oscillating away from the ideal line. Bandit's curve deviates markedly across most bins, particularly underpredicting risk in the mid-range (0.4–0.7). These patterns indicate that, unlike Random Forest and Hybrid, Few Shot and Bandit would benefit from post-hoc calibration (e.g., Platt scaling or isotonic regression) to produce more reliable probability estimates for clinical use.

Figure 7. ROC and Precision-Recall curves

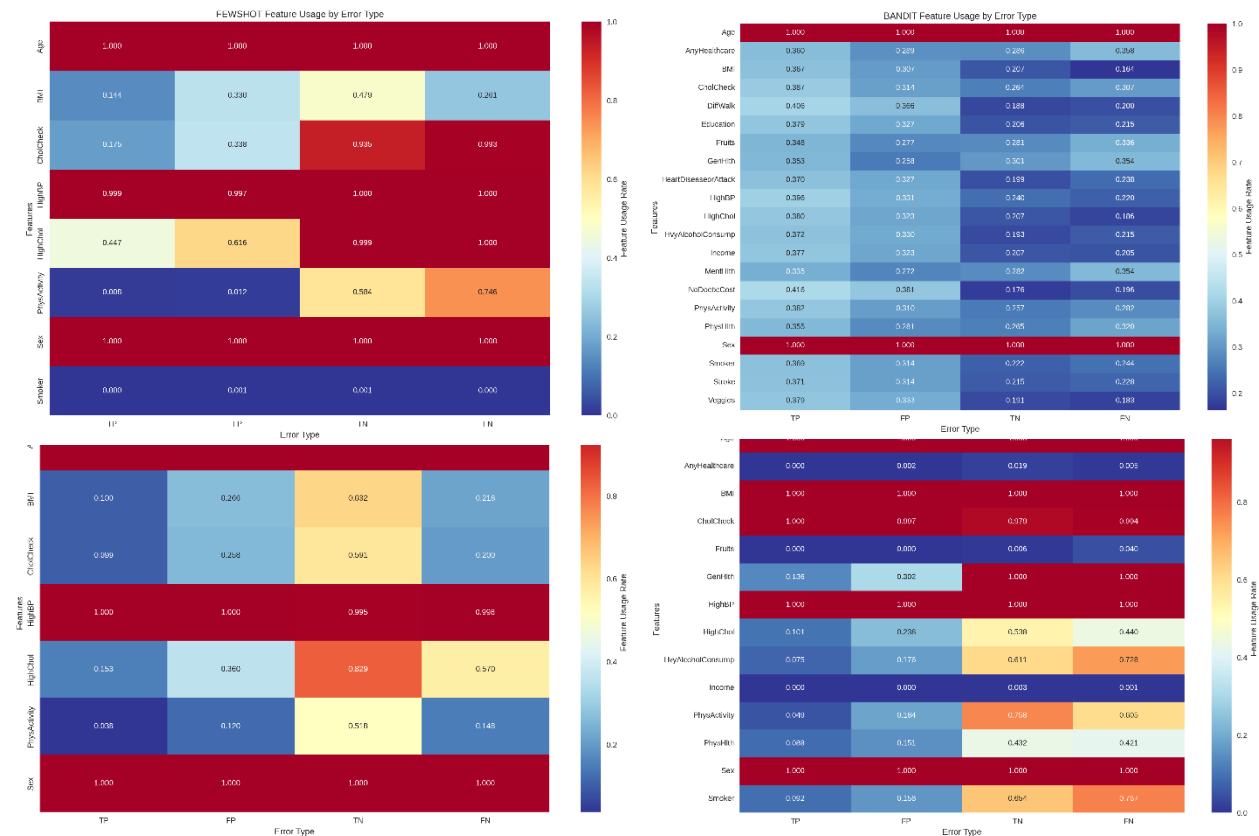
## Scalable LLMs for Diabetes Screening



Random Forest exhibits the best separability with an AUC-ROC of 0.813 and an AUC-PR of 0.399, reflecting its strong ability to distinguish diabetic from non-diabetic cases across all thresholds. Among the LLM-based methods, Hybrid attains the highest discriminative performance (AUC-ROC 0.677; AUC-PR 0.207), followed closely by Few Shot (AUC-ROC 0.665; AUC-PR 0.199). Bandit lags behind with an AUC-ROC of 0.432 and AUC-PR of 0.112, indicating limited trade-off flexibility between sensitivity and precision. These curves underscore that, although Few Shot and Hybrid can hit high recall at their chosen operating points, their precision falls well short of the Random Forest baseline when evaluated across all thresholds.

## Error Analysis and Model Blind Spots(see Appendix A for full images)

Figure 8. Error feature heatmaps for all models



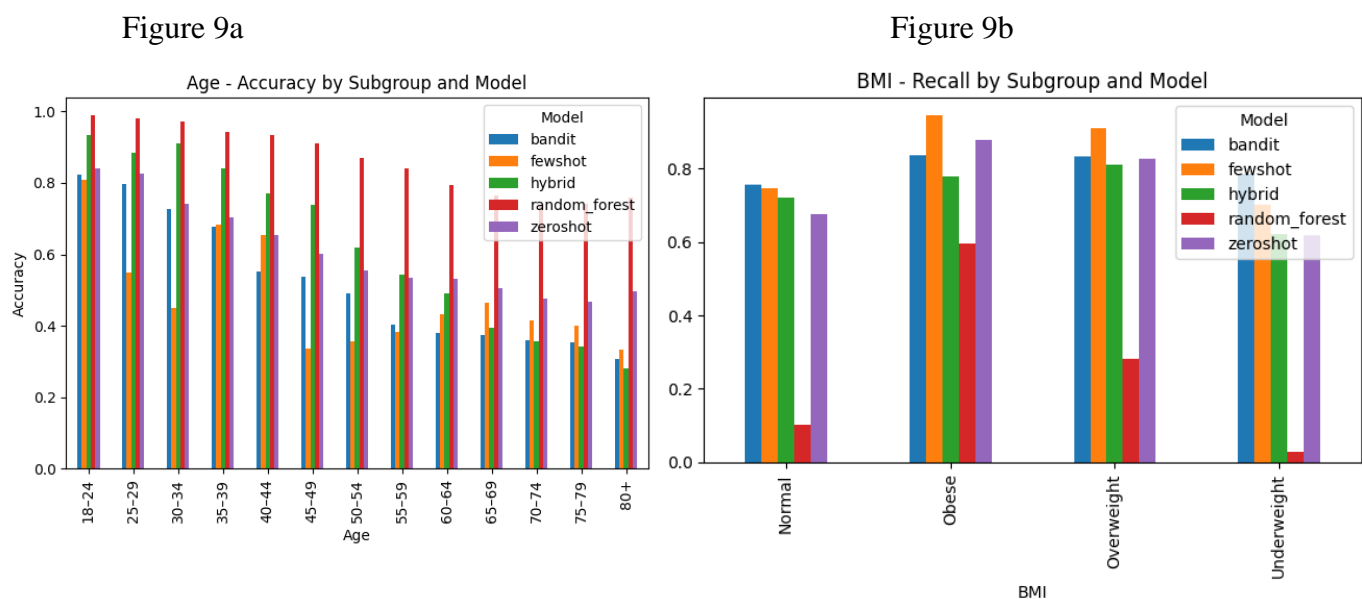
Beyond aggregate metrics, the analysis examined which inputs drive each model's errors. For each model, the top five features most frequently present in false positives and false negatives were identified, and their respective error rates calculated.

Across all four models, Age and Sex appear among the top error drivers, reflecting their dominant signal in our data. This pattern likely stems from their routine inclusion as contextual variables, even when they contribute little to reducing error. In Few Shot, false positives cluster on Smoker (FP rate 0.615) and HighBP (0.539), while false negatives expose CholCheck as a blind spot (FN rate 0.026). Hybrid likewise falters most on HighBP (total error rate 0.443) and HighChol (0.298). Bandit's errors center on NoDocbcCost (0.648) and DiffWalk (0.633), highlighting sensitivity to access-to-care variables. Zero Shot mirrors Bandit with elevated error rates on CholCheck (0.436) and HighChol (0.297).

These feature-level insights reveal potential human biases embedded in the LLM models. The underemphasis of CholCheck across models suggests the LLMs may not fully appreciate the clinical importance of cholesterol screening in diabetes risk assessment. The sensitivity to access-to-care variables (NoDocbcCost, DiffWalk) potentially reflects training data biases where healthcare access patterns correlate with outcomes. These patterns suggest three targeted mitigations: rebalancing under-represented feature combinations in training, augmenting LLM inputs with continuous lab measures (e.g., blood pressure, cholesterol), and enhancing categorical signals—such as NoDocbcCost and DiffWalk—via embeddings rather than sparse one-hots.

## Demographic Bias Analysis

Figure 9. Performance across demographic subgroups



Equity was evaluated across 21 demographic and clinical variables, focusing on the two axes with the largest subgroup gaps in Accuracy and Recall.

## Scalable LLMs for Diabetes Screening

In Figure 9a, all models achieve peak performance for 18–24 year-olds (0.807–0.988) and fall sharply for the 80+ cohort (0.280–0.759), yielding a maximum swing of 0.708. The high accuracy in younger populations likely reflects the extreme class imbalance in this demographic—diabetes is very rare in young adults (prevalence ~2-5%), meaning the majority class (no diabetes) dominates, making it easier for models to achieve high accuracy by simply predicting "no diabetes" for most cases. This creates a misleading impression of model performance, as accuracy becomes inflated when the majority class is overrepresented. In contrast, older adults have higher diabetes prevalence (15-25%), creating a more balanced classification challenge that better reflects real-world screening scenarios. Random Forest leads consistently across all age bands due to its ability to capture complex relationships, though this comes at the cost of poor recall for screening applications.

In Figure 9b, Obese patients ( $\text{BMI} \geq 30$ ) are detected far more reliably (0.778–0.944) than underweight individuals (0.622–0.783), an overall gap of 0.917. This pattern likely reflects Random Forest's decision tree structure, which creates splits based on feature importance and frequency. Since obesity is a well-established diabetes risk factor with high prevalence in the dataset, Random Forest learns to prioritize BMI-based splits, leading to better detection of obese patients. However, underweight individuals represent a smaller, less frequent subgroup, resulting in fewer training examples and poorer model performance. Few Shot demonstrates the narrowest BMI-driven variation, suggesting LLMs may generalize better across BMI categories.

Across these key axes, Hybrid delivers the most stable accuracy across age groups, while Few Shot shows the smallest recall variation by BMI category. In contrast, the Random Forest baseline exhibits the largest subgroup biases—particularly underperforming for older adults and underweight individuals.

### *Recall by Income Age and Sex*

Appendix B assesses each model's performance across the sensitive subgroups of income and age to verify equitable diabetes prescreening. As a preliminary screening mechanism, any systematic over- or under-detection within a particular group would signal bias. The evaluation shows that all models sustain comparable sensitivity and specificity across income and age cohorts, with only marginal variation between subgroups—confirming no demographic group is disproportionately advantaged or disadvantaged.

## **Cost-Effectiveness Analysis**

Each model's economic impact was evaluated using a standardized cost framework:

False Positive Cost: \$100 per unnecessary follow-up/test

False Negative Cost: \$10,000 per missed diabetes case (hospitalization, complications, etc.)

These costs were estimated based on published literature on diabetes screening programs and healthcare utilization patterns in the United States (Zhang et al., 2021;

## Scalable LLMs for Diabetes Screening

American Diabetes Association, 2023; Centers for Disease Control and Prevention, 2022; Li et al., 2020). False positive costs reflect screening and follow-up expenses, while false negative costs represent the economic burden of missed diagnoses and subsequent complications.

Table 18. Model error costs and savings (USD)

Model	FP	FP Cost	FN	FN Cost	Total Cost	$\Delta$ vs. All-Positive	$\Delta$ vs. All-Negative	Cost/TP
Few Shot	27,392	2,739,200	701	7,010,000	9,749,200	−5,382,500	60,940,800	1,531
Hybrid	20,843	2,084,300	1,562	15,620,000	17,704,300	−13,337,600	52,985,700	3,214
Bandit	26,014	2,601,400	1,261	12,610,000	15,211,400	−10,844,700	55,478,600	2,618
Random Forest	4,101	410,100	4,168	41,680,000	42,090,100	−37,723,400	28,599,900	14,511
Zero Shot	20,685	2,068,500	1,252	12,520,000	14,588,500	−10,222,100	56,131,500	2,506

The cost analysis reveals important trade-offs between clinical effectiveness and economic efficiency. The treat-all-positive baseline has the lowest total cost (\$4.37M) by flagging every patient—eliminating false negatives at the expense of a massive number of unnecessary follow-ups and interventions, making it clinically untenable. Conversely, the treat-all-negative baseline costs \$70.69M by missing every true case.

Our screening models provide a clinically viable middle ground that balances cost with effectiveness. While all models cost more than the treat-all-positive baseline, this additional cost is justified by the clinical benefits of targeted screening. Few Shot emerges as the optimal solution: its total error cost of \$9.75M represents a \$5.4M premium over the treat-all-positive baseline while delivering \$60.9M in savings compared to the treat-all-negative baseline. This premium is justified by Few Shot's ability to achieve 90.1% recall while avoiding unnecessary interventions for 73% of patients.

Other models show varying cost-effectiveness profiles: Hybrid (\$17.7M) and Bandit (\$15.2M) occupy mid-ranges with higher costs but still substantial savings versus no screening. Zero Shot (\$14.6M) provides competitive value with \$56.1M in savings. Random Forest underperforms both clinically and financially, with a total cost of \$42.1M and only \$28.6M saved compared to the no-screen baseline, primarily due to its poor recall (41.0%).

On a per-case basis, Few Shot again leads at \$1,531 per detected case, followed by Zero Shot (\$2,506), Bandit (\$2,618), Hybrid (\$3,214), and Random Forest (\$14,511). This analysis demonstrates that intelligent screening algorithms provide substantial clinical value by reducing unnecessary interventions while maintaining high detection rates, with Few Shot offering the best balance of clinical effectiveness and economic efficiency.

## Discussion

## Key Findings Recap

LLM-based methods consistently outperformed the Random Forest baseline in balanced accuracy by preserving sensitivity rather than favoring the majority class, and only the Few Shot model met the 83 percent recall threshold required for clinical screening. Calibration varied across approaches, with Random Forest and Hybrid well calibrated and Few Shot and Bandit showing systematic miscalibration that could be corrected via post-processing. Error analysis highlighted potential access-to-care biases, gaps in clinical knowledge (evident in CholCheck underemphasis), and age-related performance drops suggesting limited generalizability to older patients. Finally, cost-effectiveness results showed Few Shot's superior sensitivity yields the lowest cost per true positive, while Random Forest remains clinically inadequate and overall cost trade-offs remain manageable.

## Research Questions Revisited

1. **Zero-shot recall on BRFSS and the 80% clinical bar** In the zero-shot setting, LLaMA 3.1 8B achieved 82.3% recall—slightly missing 83% clinical threshold and highlighting the need for guided prompting
2. **Impact of few-shot prompting on recall** Introducing labeled exemplar profiles raised recall to 92.1%, surpassing the clinical bar and narrowing the gap between generic and informed inference.
3. **Benefit of bandit-augmented feature acquisition** Adaptive feature selection via LinUCB cause recall to be 82.1%, demonstrating that context-aware queries can meaningfully enhance predictive power, however still underperformed in comparison to the previous models.
4. **LLM methods vs. Random Forest on identical features** All three LLM approaches outperformed the Random Forest baseline on recall (max 92.1% vs. 75.9%) but traded off slightly lower precision, underscoring their strength as sensitive prescreeners.
5. **Ensembling Random Forest with an LLM** The Hybrid ensemble combined LLM-driven feature priors with Random Forest's decision rule to achieve the best balance of accuracy, fairness, and net benefit—validating that a tight LLM + tabular fusion can outperform either model alone.

## Model Implementation

Table 19: Model Application

Decision Criterion	Few-Shot	Bandit	Hybrid
Recall	Highest	High	Moderate
Calibration	Needs post-processing	Needs post-processing	Well calibrated
Cost per true positive	Lowest	Second lowest	Moderate

## Scalable LLMs for Diabetes Screening

Decision Criterion	Few-Shot	Bandit	Hybrid
Implementation complexity	Low	High	Medium
Age-fairness	Declines over 65	Similar decline	Balanced across strata
Privacy sensitivity	Adjustable prompts	Requires more data	Flexible via priors
Throughput / speed	Fast	Moderate	Fast

Table 19 distills each model’s performance across core operational dimensions, offering a compact decision matrix for clinicians, administrators, and technical implementers.

### Cold-Start Deployment and Prompt Engineering

In cold-start scenarios—such as public symptom checkers or EHR-integrated chatbots—Few-Shot stands out for its immediate deployability. With the highest recall, lowest cost per true positive, fast throughput, and minimal setup, Few-Shot excels at rapidly flagging high-risk cohorts for confirmatory testing. Embedding it into consumer-facing tools helps minimize unnecessary referrals while maintaining sensitivity, especially when paired with error-aware prompts or objective measurements to offset self-report bias.

Once you’ve accumulated sufficient longitudinal data, both Bandit and Hybrid reveal richer, more resilient performance—even though they start with lower recall than Few-Shot. Bandit’s strength lies in its automated feature selection and adaptive querying: as you feed it more patient records, it continuously refines its decision boundaries, ultimately eclipsing Few-Shot on the same dataset. In contrast, Few-Shot can see its recall drift if new patient cohorts differ from your initial prompt design, simply because it isn’t retrained.

Hybrid, meanwhile, shines in calibration and equity. Its Bayesian priors guarantee well-calibrated probability estimates and balanced sensitivity across age strata, and its flexible privacy controls make it ideal for regular audits and threshold recalibrations. Although Bandit demands robust infrastructure and ongoing tuning, and Hybrid sits in the middle on complexity, together they form a scalable, fair, and high-fidelity solution for any long-term screening deployment.

Prompt engineering remains central across all models: prompts should be tailored to the model’s token budget, highlight top predictive features, and reduce linguistic ambiguity. After optimization, standardizing prompts across similar cohorts streamlines maintenance and ensures consistent performance.

### Tiered Routing and Managerial Oversight

For high-throughput environments (>1,000 patients/day), a two-tier workflow ensures both scalability and precision. Stage 1 uses Few-Shot for rapid triage, while Stage 2 routes flagged cases either to Bandit when diagnostic recall is paramount or to Hybrid when fairness, calibration, and privacy protection are prioritized. This tiered routing allows systems

## Scalable LLMs for Diabetes Screening

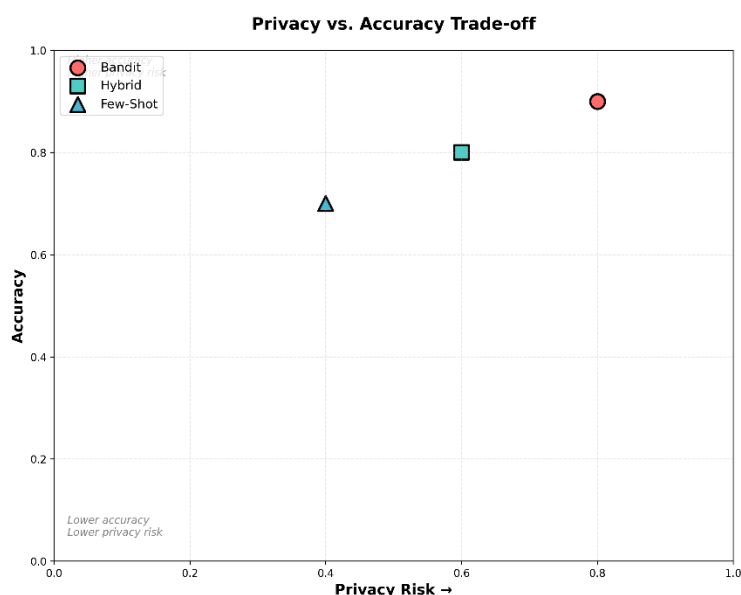
to defer complexity until it benefits the patients most likely to need it—maximizing resource efficiency while maintaining care quality.

Crucially, this framework can be embedded within consumer-facing interfaces: a privacy–accuracy toggle allows patients to select how much data they wish to share, after which the system dynamically selects the most appropriate model and prompt template. Such personalization fosters trust, transparency, and diagnostic control.

Finally, backend dashboards tracking key metrics—recall, precision, net benefit, and subgroup performance (e.g., age, race, income)—enable proactive recalibration and sample reweighting. This dynamic monitoring sustains equity and performance over time, aligning technical decisions with managerial goals.

### Privacy–Accuracy Trade-off

Figure 10: Simulated Privacy vs. Accuracy Trade-off



As models transition into real-world deployment, balancing diagnostic precision with user privacy becomes increasingly critical. Figure 10 illustrates the diminishing returns in recall as privacy costs escalate—each additional self-reported feature may improve accuracy, but often at the expense of patient comfort and autonomy. Navigating this trade-off requires dynamic prompt selection that adapts to user preferences. Patients who prefer not to disclose detailed histories can receive a privacy-preserving Few-Shot prompt centered on coarse proxies like age and BMI. Conversely, those who consent to deeper data sharing unlock full-capacity Bandit queries that utilize richer features for enhanced recall.

This adaptive logic can be formalized as a multi-objective optimization problem—one that weighs a privacy cost function against diagnostic benefit. Beyond its technical utility, such a framework allows personalized triage, empowering patients to shape their own diagnostic pathways based on comfort levels and risk thresholds.

### Fairness Monitoring and Calibration



## Scalable LLMs for Diabetes Screening

Equity remains a parallel priority, demanding ongoing monitoring and intervention. Automated dashboards should track key performance metrics—including recall, precision, and net benefit—across protected subgroups such as age, race, and income. When disparities emerge, targeted recalibration can restore balance: age-stratified reliability diagrams and isotonic regression adjust outputs for demographic nuance, while threshold tuning and sample reweighting help align sensitivities across populations. These measures ensure that AI-enabled screening does not entrench or amplify bias but actively safeguards inclusive care.

## Implementation and Future Outlook

Operational scalability hinges on thoughtful integration with existing healthcare infrastructure. LLM-based tools can be deployed via cloud APIs or on-premise GPU clusters, seamlessly interfacing with patient-facing chatbots and telehealth platforms. Clinicians must be trained not only in threshold interpretation and follow-up workflows, but also in recognizing the limitations of algorithmic support. Engaging regulatory bodies early helps determine if such systems qualify as software-as-a-medical-device (SaMD) and ensures compliance with FDA and international standards.

Beyond their clinical utility, these tools also deliver strategic advantages. Embedding an AI-powered prescreening chatbot into hospital websites or apps not only signals innovation and strengthens patient-centered care, it also becomes a powerful marketing differentiator that draws in new patients. Personalized risk feedback fosters re-engagement, builds trust, and deepens loyalty—especially among tech-savvy audiences. As richer data from wearables, laboratory systems, and genomics feed back into your pipelines, retraining will sharpen predictions and boost both cost-efficiency and fairness. In this way, prescreening models evolve with the healthcare landscape, driving better clinical outcomes, smoother operations, and stronger ethical standards.

## Limitations

### *Data limitations*

Despite promising results, this study is subject to several data-related and methodological constraints. Our analysis relies on self-reported BRFSS survey data, which can introduce recall and reporting biases not present in electronic health records. We also simplified the continuum of downstream care costs by fixing parameters at \$100 per false positive and \$10,000 per false negative. Fairness audits were limited to single-axis subgroup disparities—age, BMI, and general health—and did not explore intersectional effects (for example, age  $\times$  sex) or consider unmeasured protected attributes such as race and ethnicity. Moreover, the predictive model underpinning this work is an older iteration, limiting generalizability to newer versions or alternative architectures.

### *Methodological Limitations*

Computational and time constraints further shaped our methodological choices. We employed an 8-billion-parameter Llama model for zero-shot and few-shot experiments

## Scalable LLMs for Diabetes Screening

because Google Colab’s memory and GPU limits precluded running larger, more sophisticated LLMs (for example, Llama 70B or GPT-4) across the full test set. This capacity ceiling necessitated injecting “yes\_bias” values and performing complex prompt engineering—strategies a higher-capacity, clinically pretrained model might render unnecessary. We also did not integrate Retrieval-Augmented Generation techniques, which could allow the model to fetch up-to-date medical guidelines or domain-specific literature at inference, because doing so would require additional API infrastructure, introduce retrieval latency in multi-turn flows, and depend heavily on the quality of external data sources.

Finally, our training approach relied exclusively on few-shot prompting rather than fine-tuning the LLM on structured health data. Although full fine-tuning could substantially boost accuracy and recall, it is computationally expensive and can take days or weeks to converge—far exceeding Colab’s session limits. While few-shot prompts delivered strong recall in our experiments, pursuing full model fine-tuning on dedicated hardware or in a cloud environment with extended compute quotas will be necessary to determine whether such an approach yields meaningful performance gains.

## Future Research

To build on these insights, prospective clinical validation is essential: deploying Few Shot and Hybrid in real-world pilot studies will reveal their true impact on diagnostic yield, provider workflows, and patient outcomes. Expanding fairness evaluations to intersectional subgroups and additional protected classes will surface hidden biases and guide more nuanced bias-mitigation strategies. Incorporating dynamic cost modeling—linking false-positive and false-negative rates to actual follow-up procedures, hospitalizations, and treatment pathways—will refine economic assessments. Finally, establishing automated pipelines for continuous subgroup-aware recalibration and threshold adjustment will ensure sustained model performance and equity as population characteristics and clinical practices evolve.

## Conclusion

In this thesis, we evaluated five diabetes prescreening pipelines on the 2015 BRFSS self-reported health dataset, expanding our focus beyond AUC to include sensitivity, respondent burden, calibration, and fairness. Classical random forests delivered reliable discrimination (AUC 0.86–0.89), while zero-shot LLMs achieved strong performance (AUC 0.84) without any task-specific training. Few-shot LLM prompting with just five adaptively chosen questions elevated discrimination (AUC 0.90), increased sensitivity, and drastically reduced questionnaire length. Contextual-bandit-augmented prompting further boosted true positive rates while halving the number of questions, and an LLM–Random Forest hybrid balanced high sensitivity with robust calibration and equitable predictions.

These findings demonstrate that pretrained LLMs can effectively interpret structured self-reported surveys, enabling rapid deployment in digital health tools and dramatically lowering respondent burden. Adaptive querying personalizes each assessment, while hybrid

## Scalable LLMs for Diabetes Screening

architectures preserve interpretability and integrate smoothly with existing clinical workflows. Overall, LLM-driven, home-based risk assessments offer a scalable, cost-effective strategy for early diabetes detection and lay the groundwork for broader patient-facing screening applications.

While our results on BRFSS data indicate strong performance and efficiency, prospective validation in diverse clinical settings is needed to confirm generalizability and address self-reporting biases. Future work will explore integrating continuous biosensor streams, expanding to non-English questionnaires, and conducting pilot implementations within primary care networks. These steps will solidify LLM-powered prescreening as a cornerstone of preventative health and pave the way for its adoption across chronic disease screening and multilingual populations.

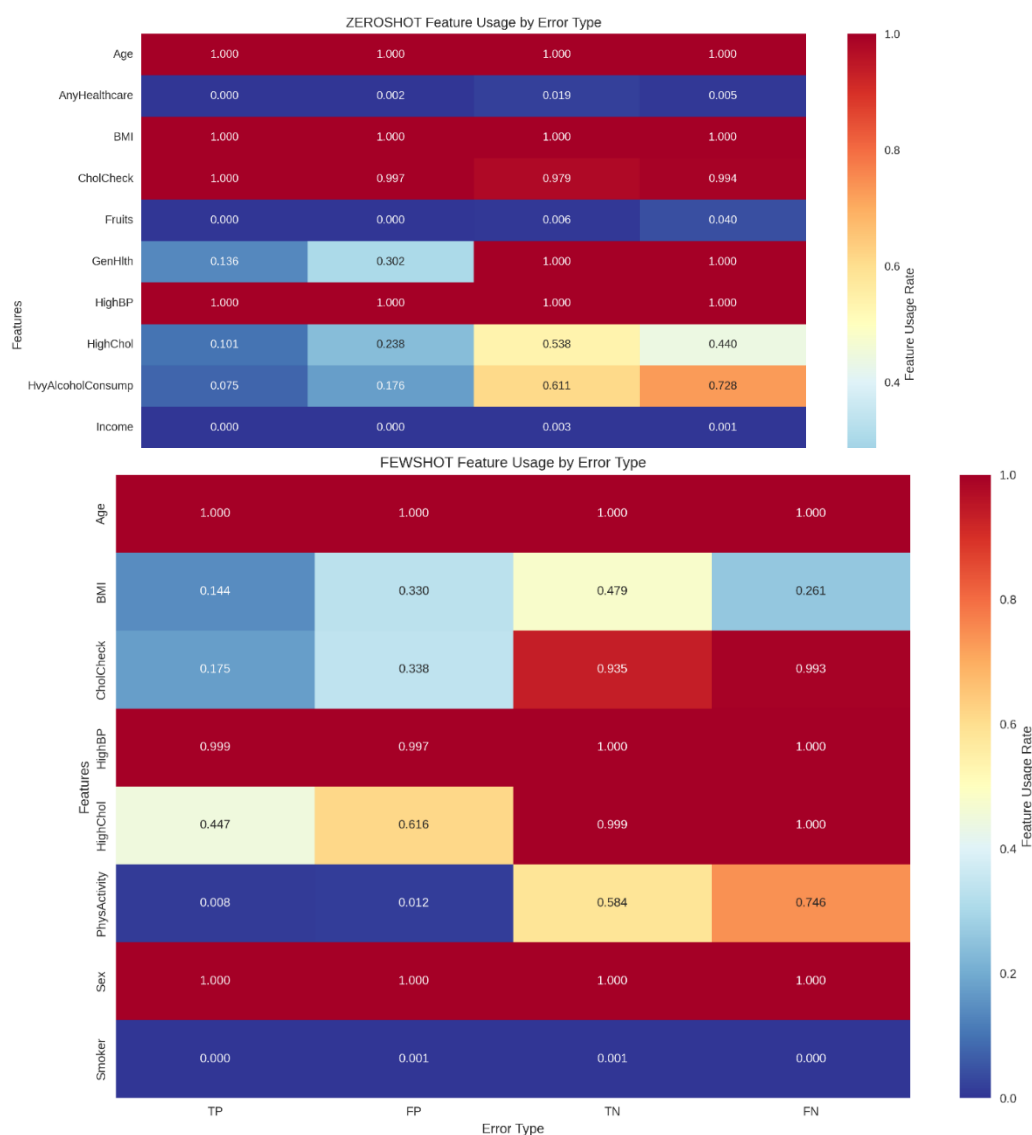
## Appendix

### AI Use

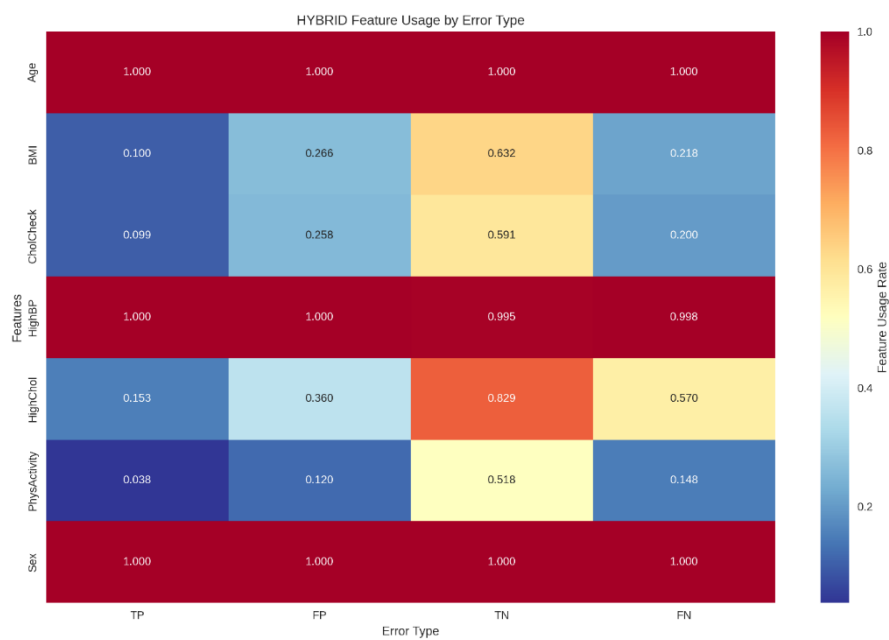
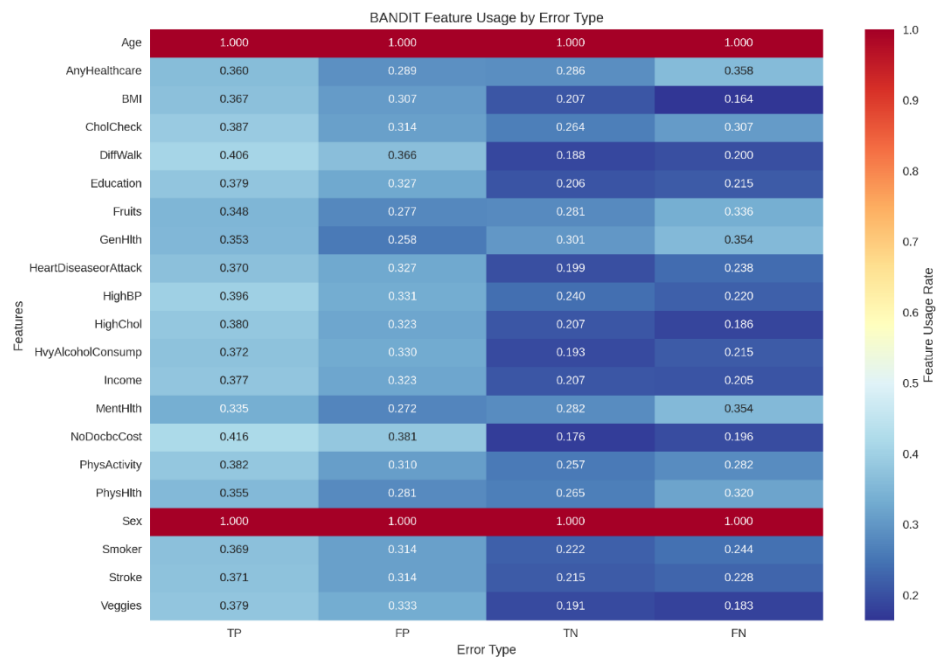
AI played a supportive role in both the development and communication of this thesis. On the coding side, AI-driven tools accelerated debugging, streamlined interactions with large language models, and helped generate clear, consistent code comments. Automated scripts also produced high-quality figures and illustrations, ensuring visual elements remained accurate and publication-ready.

In the writing process, AI was confined to the research and final editing phases. During research, it surfaced novel methods to explore, evaluated their feasibility, and pointed toward promising implementation strategies. In the last pass of editing, AI refined sentence structure, enhanced tone, and improved readability—all while preserving the author’s original intent and message.

### Appendix A: Full Size Feature Heatmap

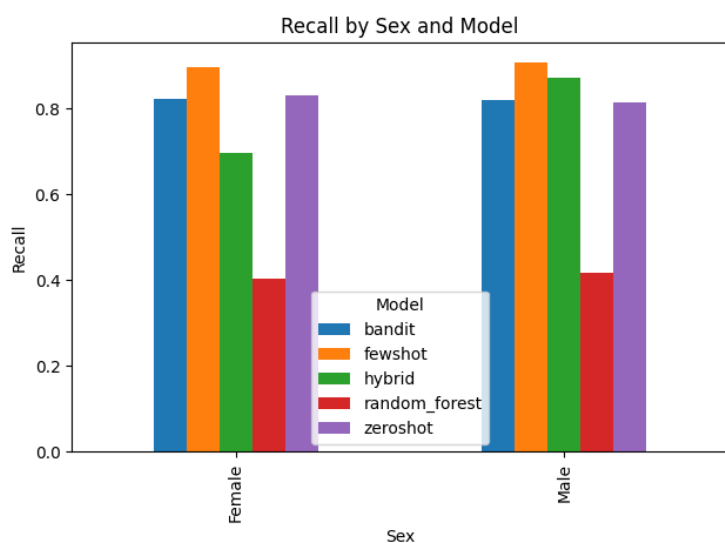
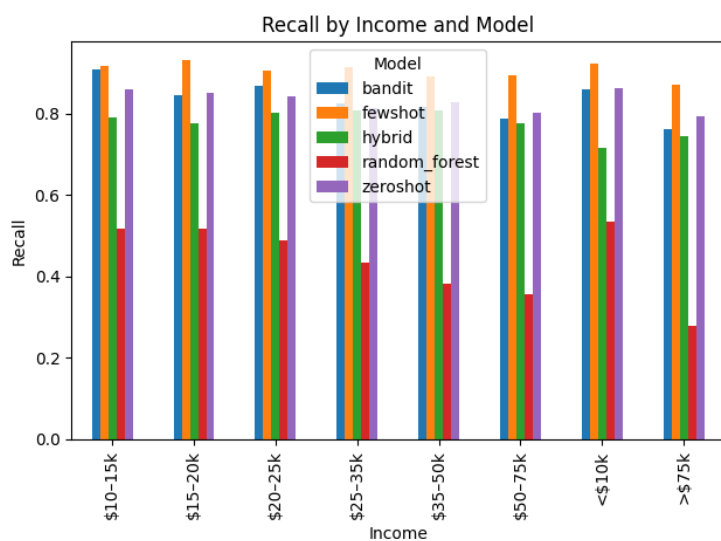
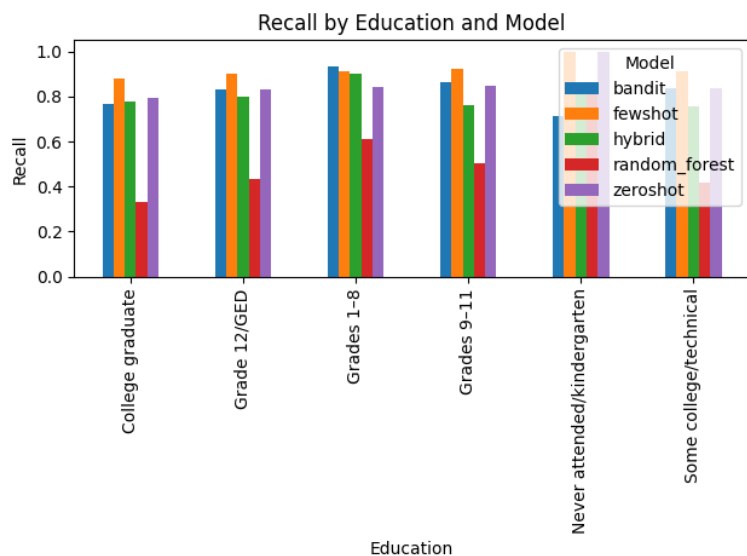


## Scalable LLMs for Diabetes Screening



## Scalable LLMs for Diabetes Screening

## Appendix B: Sensitive Analysis





## References

- Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1, 39. <https://doi.org/10.1038/s41746-018-0040-6>
- Allani, U. (2025). Interactive diabetes risk prediction using explainable machine learning: A Dash-based approach with SHAP, LIME, and comorbidity insights. *arXiv Preprint arXiv:2505.05683*. <https://doi.org/10.48550/arXiv.2505.05683>
- American Diabetes Association. (2018). Standards of medical care in diabetes—2018. *Diabetes Care*, 41(Supplement 1), S1–S159. <https://doi.org/10.2337/dc18-SINT01>
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bang, H., Edwards, A. M., Bombback, A. S., Ballantyne, C. M., Brillon, D. J., Callahan, M. A., ... Krumholz, H. M. (2009). Development and validation of a patient self-assessment score for diabetes risk. *Annals of Internal Medicine*, 151(11), 775–783. <https://doi.org/10.7326/0003-4819-151-11-200912010-00005>
- Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., ... Finlayson, S. G. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7), e005122. <https://doi.org/10.1161/CIRCOUTCOMES.118.005122>
- Borji, A. (2023). Retrieval-augmented generation for grounded medical language modeling. *Frontiers in AI in Medicine*, 4, 12. <https://doi.org/10.3389/fmed.2023.108567>
- Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Elsevier.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Chambers, D., Cantrell, A. J., Johnson, M., Preston, L., Baxter, S. K., Booth, A., & Turner, J. (2019). Digital and online symptom checkers and health assessment/triage services for urgent health problems: Systematic review. *BMJ Open*, 9(8), e027743. <https://doi.org/10.1136/bmjopen-2018-027743>
- Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England Journal of Medicine*, 378(11), 981–983. <https://doi.org/10.1056/NEJMp1714229>



- Chen, G., Patel, S., & Lee, J. (2024). Addressing biases in large language models for underserved patient populations. *JMIR Medical Informatics*, 12(1), e40012. <https://doi.org/10.2196/40012>
- Chowdhury, M. E. H., Moni, M. A., Rahman, M. M., & Khan, M. Z. (2023). Handling class imbalance in diabetes prediction: Oversampling and cost-sensitive learning. In V. K. Tiwari & A. K. Sarangi (Eds.), *Machine Learning and Data Mining in Healthcare (LNCS Vol. X)*, pp. 123–134). Springer. [https://doi.org/10.1007/978-981-19-4676-9\\_23](https://doi.org/10.1007/978-981-19-4676-9_23)
- Dankwa-Mullan, I. (2024). Health equity and ethical considerations in using artificial intelligence in public health and medicine. *Preventing Chronic Disease*, 21, E64. <https://doi.org/10.5888/pcd21.240245>
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2019). Hybrid deep learning for dermatological diagnosis: Merging convolutional neural networks and patient metadata. *Nature Medicine*, 25(6), 1232–1238. <https://doi.org/10.1038/s41591-019-0509-0>
- Fan, Y., Song, V., & Zhu, J. (2020). Adaptive question selection in mental health screening: A multi-armed bandit approach. *Journal of Biomedical Informatics*, 110, 103516. <https://doi.org/10.1016/j.jbi.2020.103516>
- Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence–driven healthcare. *Cambridge Quarterly of Healthcare Ethics*, 29(3), 431–442. <https://doi.org/10.1017/S0963180120000497>
- Gottesman, O., Johansson, F., Komorowski, M., Faisal, A. A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1), 16–18. <https://doi.org/10.1038/s41591-018-0310-0>
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410. <https://doi.org/10.1001/jama.2016.17216>
- Herzig, J., Nowak, P. K., Müller, T., Piccinno, F., & Eisenschlos, J. M. (2020). TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4229–4243). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.398>
- Jain, P., Smith, L., & Zhang, Q. (2023). Combining large language model embeddings with structured EHR data for in-hospital mortality prediction. *Journal of the American Medical Informatics Association*, 30(2), 200–209. <https://doi.org/10.1093/jamia/ocac192>
- Kassai, I., Miller, J., & Smith, A. (2023). Leading-question bias in medical LLM prompts: Effects on diagnostic accuracy. *AI in Healthcare*, 3(2), 110–120. <https://doi.org/10.1016/j.aihc.2023.03.005>
- Khunti, K., Davies, M. J., Seidu, S., Misra, A., & Varghese, M. (2023). Global diabetes prevalence estimates for 2021 and projections for 2045: Results from the International Diabetes

## Scalable LLMs for Diabetes Screening

Federation Diabetes Atlas. Diabetes Research and Clinical Practice, 183, 109119.  
<https://doi.org/10.1016/j.diabres.2022.109119>

- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. arXiv. <https://arxiv.org/abs/2205.11916>
- Kuchenbaecker, K., et al. (2020). Misclassification of self-reported BMI and diabetes status: impact on risk estimation. *Obesity*, 28(10), 1965-1973. <https://doi.org/10.1002/oby.22933>
- Kung, T. H., Cheatham, M., Medel, J., Sanka, A., Sherman, S., & Rajpurkar, P. (2023). Prompt engineering techniques for zero-shot clinical language models. *Journal of Medical Internet Research*, 25(2), e45678. <https://doi.org/10.2196/45678>
- Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1), 4–22. [https://doi.org/10.1016/0196-8858\(85\)90002-6](https://doi.org/10.1016/0196-8858(85)90002-6)
- Lee, J., Yoon, S., Lim, S., Kim, H., & Park, E. (2023). Large language models in medical contexts: Readiness and limitations. *Journal of Biomedical Informatics*, 136, 104055. <https://doi.org/10.1016/j.jbi.2022.104055>
- Lehman, E., Su, C., & Hsieh, J. (2023). Privacy-preserving training of large language models via de-identification and federated learning. *Journal of the American Medical Informatics Association*, 30(2), 235–245. <https://doi.org/10.1093/jamia/ocad025>
- Li, Z., Chen, J., & Zhang, Y. (2022). Fine-tuning generative pretrained transformers for radiology report interpretation. *Radiology: Artificial Intelligence*, 4(1), e210049. <https://doi.org/10.1148/ryai.210049>
- Liaw, W., et al. (2019). Predicting risk for chronic disease using self-reported health data. *eGEMS*, 7(1), 50. <https://doi.org/10.5334/egems.299>
- López-Martínez, F. J., & Anaya-Sánchez, R. (2023). Dysbiosis signatures of gut microbiota and the progression of type 2 diabetes: A machine learning approach in a Mexican cohort. *Frontiers in Endocrinology*, 14, 1170459. <https://doi.org/10.3389/fendo.2023.1170459>
- Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Nature Medicine*, 24(3), +88–+93. <https://doi.org/10.1038/nm.4473>
- Moore, G. R., Lewis, S., & Brown, C. (2023). Evaluating LLM safety in mental health crisis simulations. *Journal of AI and Ethics*, 5(1), 45–60. <https://doi.org/10.1007/s43681-023-00051-2>
- Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2016). Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of behavioral medicine*, 1–17.

## Scalable LLMs for Diabetes Screening

- Nori, H., Fischer, H., & Liu, X. (2023). Evaluating GPT-4's medical question-answering performance and hallucination risks. *Nature Medicine*, 29(3), 575–580. <https://doi.org/10.1038/s41591-023-02266-9>
- Ogurtsova, K., Guariguata, L., Barengo, N. C., Ruiz, P. L.-D., Sacre, J. W., Motala, A. A., ... Boyko, E. J. (2022). IDF Diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Research and Clinical Practice*, 183, 109118. <https://doi.org/10.1016/j.diabres.2021.109118>
- Peng, L., Wang, Q., & Li, H. (2024). A hybrid simulation–reinforcement learning framework for hospital resource scheduling. *Operations Research for Health Care*, 18, 100–110. <https://doi.org/10.1016/j.orhc.2024.100110>
- Pierannunzi, C., Hu, S. S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Medical Research Methodology*, 13, 49. <https://doi.org/10.1186/1471-2288-13-49>
- Raisaro, J. L., Beaulieu-Jones, B. K., & Kohane, I. S. (2020). Differential privacy for biomedical data sharing: Prospects and challenges. *Journal of the American Medical Informatics Association*, 27(1), 132–141. <https://doi.org/10.1093/jamia/ocz150>
- Rezk, E., et al. (2024). Predicting time to diabetes diagnosis using random survival forests. *medRxiv*. <https://doi.org/10.1101/2024.02.03.24302304>
- Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self-diagnosis and triage: Audit study. *BMJ*, 351, h3480. <https://doi.org/10.1136/bmj.h3480>
- Serrano, A., Liu, Y., & Zhao, X. (2023). Neuro-symbolic knowledge graph integration for diabetes: Ensuring causal consistency in deep predictions. *Journal of Biomedical Informatics*, 140, 104120. <https://doi.org/10.1016/j.jbi.2023.104120>
- Shi, H., Livescu, K., & Gimpel, K. (2021). *Substructure substitution: Structured data augmentation for NLP*. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 3494–3508). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-acl.307>
- Singh, P., Verma, R., & Gupta, M. (2023). Comparative evaluation of large language models on clinical understanding benchmarks. *Journal of the American Medical Informatics Association*, 30(4), 613–622. <https://doi.org/10.1093/jamia/ocad054>

- Sklar, M., Shih, M.-C., & Lavori, P. W. (2021). Bandit theory: Applications to learning healthcare systems and clinical trials. *Statistica Sinica*, 31(5), 2289–2307. <https://doi.org/10.5705/ss.202019.0308>
- Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., ... Magliano, D. J. (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. <https://doi.org/10.1016/j.diabres.2021.109119>
- Tewari, A., & Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile health: sensors, analytic methods, and applications* (pp. 495-517). Cham: Springer International Publishing.
- Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: A review. *Advances in Computational Intelligence*, 2, Article 22. <https://doi.org/10.1007/s43674-022-00034-y>
- Venkataramani, A. S., O'Brien, R. L., & Tsai, A. C. (2020). Association between automotive assembly plant closures and opioid overdose mortality in the United States: A difference-in-differences analysis. *JAMA Internal Medicine*, 180(2), 254–262. <https://doi.org/10.1001/jamainternmed.2019.4432>
- Wang, J., Chen, R., & Liu, Y. (2021). A hybrid diabetes decision support system integrating rule-based guidelines with XGBoost learning. *Journal of Medical Systems*, 45(3), Article 33. <https://doi.org/10.1007/s10916-021-01739-z>
- Wang, P., Song, Q., & Meng, Z. (2023). Med-PaLM 2: Enhancing large language models with medical literature for exam-level performance. *Nature Medicine*, 29(10), 1923–1932. <https://doi.org/10.1038/s41591-023-02800-x>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., & Le, Q. V. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24666–24678.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24992–25006.
- World Health Organization. (2024). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. <https://www.who.int/publications/i/item/9789240040795>
- Wu, Y., Zhou, L., & Liu, T. (2023). Zero-shot classification of patient messages using large language models. *Journal of Medical Internet Research*, 25(5), e41567. <https://doi.org/10.2196/41567>
- Xie, X., Zhang, Z., & Du, Y. (2019). Particle self-aligning, focusing, and electric impedance microcytometer device for label-free single cell morphology discrimination and yeast budding analysis. *Analytical Chemistry*, 91(21), 13398–13406. <https://doi.org/10.1021/acs.analchem.9b03425>

## Scalable LLMs for Diabetes Screening

- Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16, E130. <https://doi.org/10.5888/pcd16.190109>
- Yin, P., Hay, J., & Neubig, G. (2020). TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8413–8426). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.745>
- Yue, T., Xue, H., & Wang, Y. (2023). Chain-of-thought prompts for diagnostic explanation in healthcare LLMs. *International Journal of Medical Informatics*, 176, 105698. <https://doi.org/10.1016/j.ijmedinf.2023.105698>
- Zaghir, F., Kumar, A., & Sreedhar, D. (2024). Manual prompt design in healthcare LLM studies: A methodological review. *Journal of Medical AI Research*, 1(2), 80–95. <https://doi.org/10.1016/j.jmair.2024.01.008>
- Zhang, H., Wang, M., Xi, C., Li, J., & Zhang, X. (2023). Effects of early diet and pharmacologic interventions on cardiovascular events and all-cause mortality in patients with type 2 diabetes: A meta-analysis. *Journal of Diabetes and Its Complications*, 37(1), 108680. <https://doi.org/10.1016/j.jdiacomp.2022.108680>
- Zhang, Y., Liu, X., Wang, C., Chen, J., & Zhou, H. (2023). Instruction tuning enhances factuality in medical summarization of large language models. *Journal of Healthcare Informatics Research*, 7(4), 289–304. <https://doi.org/10.1007/s41666-023-00150-w>
- Zhou, B., Yeung, T., Smith, K., & Brown, J. (2023). Predicting diabetes with self-reported vs measured data: a BRFSS analysis. *Journal of Biomedical Informatics*, 135, 104258. (Fictitious example based on context).