Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diabetes-specialized versus general-purpose large language models for 5-year mortality prediction in adults with diabetes: a validation study using NHANES

Hari Prakash G
M S Ramaiah University of Applied Sciences
Sunil Kumar D
sunilkumard@jssuni.edu.in

JSS Academy of Higher Education and Research
Tejaswini BD
M S Ramaiah University of Applied Sciences
Pramod Kumar A
M S Ramaiah University of Applied Sciences
Anush Nayak
M S Ramaiah University of Applied Sciences
Vanishri Arun
B.M.S. College of Engineering

Research Article

Keywords: Large language models, Clinical prediction, Diabetes mellitus, Mortality, Fine-tuning, Calibration, Model validation, NHANES

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11663085/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models (LLMs) are increasingly proposed for clinical prediction, but their prognostic accuracy, calibration and reliability relative to conventional models remain unclear. We evaluated whether diabetes-specific fine-tuning improves LLM prediction of mortality.

Methods

We included 6,242 adults with diabetes from NHANES 2005 to 2018 with linked National Death Index follow-up, split 70:30 into development and validation sets. Diabot, a Llama 3.1 8B model fine-tuned on diabetes guidelines, and its base model estimated 5-year all-cause mortality from text summaries of 19 routine predictors under seven prompting conditions. Comparators were Cox, logistic regression and XGBoost models. We assessed discrimination, calibration, net benefit, output format adherence and run-to-run stability.

Results

Among 1,307 validation participants (202 deaths), the Cox model achieved an AUC of 0.787 (95% CI 0.752 to 0.820). Zero-shot Diabot achieved 0.755 (0.720 to 0.790) with 98.6% valid outputs, versus 0.721 (0.679 to 0.759) and 76.7% for base Llama. Diabot outperformed base Llama with few-shot and stepwise prompts (AUC difference 0.03; $p < 0.05$). Both LLMs over-predicted mortality (expected/observed 1.20 and 1.43) and produced few distinct risk values. Recalibration in the development set corrected calibration (slope 1.01). Adding Diabot's score to the Cox model did not improve discrimination (AUC change + 0.001).

Conclusions

Domain fine-tuning improved output reliability and modestly improved discrimination, but neither LLM matched a Cox model using the same routine data. LLM risk estimates require recalibration before any clinical use.

Background

Diabetes affects an estimated 589 million adults worldwide and contributes to more than 3.4 million deaths each year [1]. Knowing an individual's risk of death helps clinicians set the intensity of treatment, prioritize follow-up and discuss prognosis with patients. Conventional prediction models built from routinely collected clinical measurements, such as the RECODe equations and other mortality models for

type 2 diabetes, achieve moderate to good discrimination [2, 3]. Still, they require structured inputs and dedicated implementation in clinical systems.

Large language models (LLMs) accept free-text input and can be queried with a description of a patient, which has prompted interest in using them directly as prediction engines. A language model pretrained on the clinical notes of a single health system predicted in-hospital mortality, readmission and other outcomes well after task-specific fine-tuning [4]. Most clinicians and health systems, however, are more likely to use general-purpose or lightly adapted open models without outcome-specific training, and the evidence for this zero-shot or few-shot use is less encouraging. General-purpose LLMs were less accurate and less well calibrated than locally trained machine learning models for hospital prediction tasks [5, 6], did not reliably follow instructions in simulated clinical decision-making [7], and produced probabilities that were unreliable for risk estimation whether these were read from token probabilities or stated in words [8]. LLM outputs are also sensitive to apparently trivial changes in prompt wording and format [9].

Domain adaptation, in which an open model is fine-tuned on specialty-specific material, is a common way to build medical LLMs. Yet, systematic comparisons have found that medically adapted models often do not outperform their general-purpose base models [10]. Whether fine-tuning on clinical guidelines, which teaches management rather than prognosis, transfers to individual risk prediction is unknown. Reporting guidance for studies of LLMs also emphasizes that evaluations should describe not only accuracy but also calibration, reproducibility and the handling of invalid outputs [11], aspects that are seldom examined together.

We therefore evaluated Diabot, a diabetes-specialized LLM created by fine-tuning Llama 3.1 8B on diabetes guideline material, for prediction of 5-year all-cause mortality in a national sample of US adults with diabetes. We compared it with its own base model to isolate the effect of fine-tuning, and with conventional Cox, logistic regression and gradient-boosted models using the same predictors. We assessed discrimination, calibration, clinical utility, output format adherence, stability across repeated runs and prompts, and subgroup performance, and in post hoc analyses we examined strategies to improve LLM risk estimates.

Methods

Study design and reporting

We conducted a prognostic model validation study comparing a diabetes-specialized large language model (LLM), its general-purpose base model, and three conventional prediction models for 5-year all-cause mortality in adults with diabetes. All models were evaluated on the same held-out validation set. Reporting follows the TRIPOD-LLM guideline [11] and, for the conventional models, TRIPOD + AI [12]. The primary comparison (Diabot versus base Llama 3.1 8B, zero-shot prompt, greedy decoding) and the secondary prompting conditions were fixed before any LLM inference was run. The analyses of

alternative risk extraction, recalibration and incremental value were defined after the primary results were known and are reported as post hoc.

Data source and participants

We used the US National Health and Nutrition Examination Survey (NHANES), a continuous, cross-sectional survey of the civilian non-institutionalized population that combines household interviews with standardized physical examinations and laboratory tests in mobile examination centres [13]. We pooled seven 2-year cycles from 2005–2006 to 2017–2018. The 2019–2020 cycle was not used because data collection was interrupted by the COVID-19 pandemic and because mortality follow-up ended in 2019.

We included adults aged 20 years or older with diabetes, defined as a self-reported physician diagnosis, current use of insulin or oral glucose-lowering drugs, or an HbA1c of 6.5% or higher without a prior diagnosis (previously undiagnosed diabetes) [14]. We excluded participants who were not eligible for mortality linkage and those without follow-up time measured from the examination.

Outcome

The outcome was all-cause death within 5 years of the NHANES examination. Vital status was ascertained through the NCHS public-use Linked Mortality Files, which link NHANES participants to the National Death Index with follow-up to 31 December 2019 [15]. Follow-up time was measured in months from the examination. Participants who died within 60 months were classified as events and participants alive with more than 60 months of follow-up as non-events. Participants alive but followed for less than 60 months could not be classified and were excluded from the 5-year analyses; they were retained in the development of the Cox model, which uses all available follow-up. Cause of death (heart disease or cerebrovascular disease) was recorded for descriptive purposes only.

Predictors

Nineteen candidate predictors were selected a priori on clinical grounds and because they are routinely available in primary care: age, sex, HbA1c, estimated glomerular filtration rate (eGFR), urine albumin-to-creatinine ratio (UACR, log-transformed), systolic blood pressure (mean of up to three readings), body mass index, total and HDL cholesterol, family income-to-poverty ratio, diabetes duration, current and former smoking, atherosclerotic cardiovascular disease (self-reported coronary heart disease, angina, myocardial infarction or stroke), heart failure, insulin use, oral glucose-lowering drug use, antihypertensive medication use and previously undiagnosed diabetes. eGFR was calculated with the race-free 2021 CKD-EPI creatinine equation [16]. Serum creatinine values from 2005–2006 were recalibrated to the standardized assay using the NCHS-recommended equation (standardized creatinine = -0.016 + 0.978 x measured creatinine) [17]. Race and ethnicity were not used as a predictor in any

model and were retained only for the subgroup analysis. All predictors were measured at the same examination visit, before the outcome period.

Data split and sample size

Eligible participants were randomly split 70:30 into a development set and a validation set, stratified by vital status at the end of follow-up (random seed 2026). The development set was used to fit the conventional models, select few-shot examples, and fit the post hoc recalibration and incremental value models. The validation set was used only for evaluation. We assessed the adequacy of the development sample against the criteria of Riley and colleagues [18], assuming an outcome proportion of 0.172 (observed in the development set) and an anticipated C statistic of 0.78 (Cox-Snell R-squared 0.14). For the largest model (logistic regression with 39 parameters after spline expansion), the minimum required sample was 2,245 participants with 386 events; the development set contained 3,013 evaluable participants with 519 events. The validation set contained 202 events, exceeding the commonly recommended minimum of 100 events for validation studies [19].

Missing data

For the conventional models, missing predictor values were imputed with a single chained-equations imputation (iterative Bayesian ridge regression, 20 iterations) [20] fitted on the development set only and then applied unchanged to the validation set, so that no information from the validation set entered model development. Imputed binary values were rounded to 0 or 1. For the LLMs, missing values were not imputed: a measurement that was missing was omitted from the patient text, as it would be absent from a real clinical note. The proportion of missing values in the 5-year analysis sample was 21.9% for diabetes duration, 9.4% for income-to-poverty ratio, 7.0% for eGFR, 6.3% for cholesterol and below 5% for all other predictors.

Conventional comparator models

We developed three comparators on the development set using the same 19 predictors. (1) A Cox proportional hazards model with a small ridge penalty (0.01), fitted on all development participants using their full follow-up; 5-year risk was calculated as one minus the predicted survival at 60 months [21]. (2) A logistic regression model for the binary 5-year outcome, with cubic B-spline terms (four knots) for age, eGFR, HbA1c, body mass index and log UACR, linear terms for the remaining predictors and the default L2 penalty. (3) A gradient-boosted tree model (XGBoost) [22] with 400 trees, learning rate 0.03, maximum depth 3, row and column subsampling of 0.8 and minimum child weight 5. Hyperparameters were set in advance and not tuned on the validation data. The Cox model was designated the reference comparator.

Large language models

Diabot

Diabot is a diabetes-specialized LLM developed by our group (Fig. 1A). It was created by parameter-efficient fine-tuning of Meta Llama 3.1 8B Instruct [23], a 32-layer decoder-only transformer with 8.03 billion parameters, using quantized low-rank adaptation (QLoRA) [24, 25]. The base model was loaded in 4-bit NF4 precision with double quantization and kept frozen. Trainable low-rank adapters (rank 16, alpha 32, dropout 0.05) were inserted into all seven projection matrices of every layer (query, key, value and output projections of the attention block and the gate, up and down projections of the feed-forward block). This added 41,943,040 trainable parameters, 0.52% of the base model, stored as a 168 MB adapter file.

The training corpus contained 1,544 instruction-response pairs. Of these, 896 were template-generated guideline vignettes covering combinations of age (8 levels), eGFR (8 levels), HbA1c (7 levels) and blood pressure (2 levels), with responses encoding individualized glycaemic targets, metformin and SGLT2 inhibitor use by kidney function and blood pressure management; 90 were comorbidity and emergency vignettes (atherosclerotic cardiovascular disease, heart failure, diabetic kidney disease with albuminuria, high hypoglycaemia risk, diabetic ketoacidosis and gestational diabetes, each across 5 ages and 3 treatment regimens); and 558 were diabetes-related question-answer pairs selected by keyword filtering from the MEDIQA dataset [26, 27]. Responses were grounded in the ADA Standards of Care 2024, KDIGO 2022, RSSDI 2022 and ICMR 2018 guidelines [28–31]. Examples were formatted with the Llama 3.1 chat template and split 90:10 into training (n = 1,389) and validation (n = 155) sets.

Training was performed in Google Colab on a single NVIDIA Tesla T4 GPU (16 GB) with the Hugging Face transformers, peft and bitsandbytes libraries. The model was trained for 2 epochs (348 optimization steps) with a causal language-modelling loss, a per-device batch size of 2 with 4 gradient accumulation steps (effective batch size 8), the AdamW optimizer, a learning rate of 0.0002 with 10 warm-up steps and cosine decay, fp16 mixed precision and a maximum sequence length of 1,024 tokens. The full training and inference configuration is given in Supplementary Table S2. Diabot was trained for treatment recommendation and was not trained on any prognostic task, mortality outcome or NHANES data, so the present study evaluates its transfer to an unseen prognostic task. Only the fine-tuned language model was evaluated; no retrieval augmentation or rule-based post-processing was used.

Base model comparator

The general-purpose comparator was the same Llama 3.1 8B Instruct model with the Diabot adapter disabled. Both models were therefore run from identical quantized weights, and any difference between them is attributable to the fine-tuning alone.

Inference environment

Inference was run on 5 November 2026 on a single NVIDIA L4 GPU (24 GB) in Google Colab, using the Hugging Face transformers and peft libraries, 4-bit NF4 quantization with double quantization and

bfloat16 computation. Prompts were formatted with the Llama 3.1 chat template and processed in batches of 32 (8 for few-shot prompts) with left padding.

Patient text and prompting conditions

Each participant's record was converted to a short clinical summary using a fixed template, an approach described as text serialization of tabular data [32]. The summary (mean 50 words) stated age, sex, diabetes status and duration, glucose-lowering treatment, cardiovascular history, smoking status, HbA1c, eGFR, UACR, total and HDL cholesterol, blood pressure, body mass index, antihypertensive treatment and income band. Race and ethnicity were not included. Because NHANES does not reliably distinguish diabetes type, the text referred to diabetes without specifying type. An example is: *"70-year-old man with diabetes, diagnosed 7 years ago. Diabetes treatment: none reported. Cardiovascular history: none reported. Smoking: never smoked. Labs: HbA1c 7.1%, eGFR 68 mL/min/1.73m2, urine albumin-creatinine ratio 4 mg/g, total cholesterol 147 mg/dL, HDL 59 mg/dL. Examination: blood pressure 131/57 mmHg, BMI 24.7 kg/m2. Taking antihypertensive medication. Household income: high income.*

A system prompt asked the model to act as an experienced endocrinologist estimating prognosis for adults with diabetes in a general US population setting. Each model was run under seven conditions on all 1,307 evaluable validation participants: (1) the main zero-shot prompt, which asked for the probability of death from any cause within 5 years and required the answer as JSON in the form {"probability": number between 0 and 100}; (2) a few-shot prompt that prefixed the main prompt with 10 labelled development-set cases (2 deaths and 8 survivors, approximating the observed event rate, selected at random); (3) a reworded prompt with the same request phrased differently; (4) a stepwise prompt that asked the model to consider named risk factors before answering; and (5 to 7) three repeated runs of the main prompt with stochastic sampling (temperature 0.7, top-p 0.9, different random seeds). Conditions 1 to 4 used greedy decoding. The maximum output length was 24 tokens. In total, 18,298 LLM predictions were generated (2 models x 7 conditions x 1,307 participants).

Output parsing

A prediction was considered valid only if the output contained an explicit probability field ("probability": value) with a value between 0 and 100, which was divided by 100. Any other output, including free-text answers that were truncated at the token limit, was classified as a format failure and treated as missing for that condition. A lenient parser that extracted the first number in any free-text answer was examined and rejected before the final analysis because it misread phrases such as "5-year risk" as a 5% probability.

Statistical analysis

Discrimination was measured with the area under the receiver operating characteristic curve (AUC), with 95% confidence intervals from 1,000 bootstrap resamples of the validation set (percentile method) [33]. **Calibration** was assessed with the calibration slope and calibration-in-the-large from logistic recalibration models, the ratio of expected to observed deaths (E/O), and calibration plots of observed

against predicted risk by decile of predicted risk [34, 35]. Overall accuracy was summarised with the Brier score. **Clinical utility** was assessed by decision curve analysis over risk thresholds from 5% to 40% [36].

LLM performance was computed on participants with a valid output for that condition. Because format failures could select a non-random subset, every comparison between two models used only participants with valid outputs from both (paired analysis), and the proportion of valid outputs was reported alongside performance, separately for participants who died and survived. Differences in AUC were estimated with paired bootstrap resampling; two-sided p-values were derived from the bootstrap distribution of the difference. Output resolution was described as the number of distinct risk values produced across the 1,307 participants.

Run-to-run stability was assessed in the three sampled runs among participants with valid outputs in all three, using the mean within-participant standard deviation of predicted risk, the minimum pairwise Spearman correlation, the range of AUCs and the proportion of participants whose predicted risk fell on different sides of the 10% and 20% thresholds across runs. **Prompt sensitivity** was assessed by comparing each alternative prompt with the main prompt using the Spearman correlation and the mean absolute difference in predicted risk. **Subgroup performance** was examined by sex, race and ethnicity and income-to-poverty ratio (below 1.0, 1.0 to 1.99, 2.0 to 3.99, 4.0 or higher), reporting the AUC and E/O ratio in subgroups with at least 10 events. Analyses were unweighted, because the aim was to estimate predictive performance in the sample rather than population prevalence.

Post hoc improvement analyses

Three strategies were evaluated after the primary results showed coarse and miscalibrated LLM outputs. (1) **Token probability scoring**: instead of asking the model to write a number, we asked whether the person would die within 5 years (answer Yes or No) and took the normalized probability of the Yes token, P(Yes) / (P(Yes) + P(No)), from the model's next-token distribution in a single forward pass [32]. This produces a continuous score for every participant and cannot fail to parse. (2) **Logistic recalibration** [35, 37]: an intercept and slope on the logit scale were fitted in the development set (Yes/No scores for all 3,013 evaluable development participants, and generated risks from the main zero-shot prompt) and applied unchanged to the validation set. (3) **Incremental value**: the logit of each model's Yes/No score was added as a twentieth predictor to the Cox model in the development set, and the extended and original Cox models were compared in the validation set using the change in AUC, the continuous net reclassification improvement (NRI) [38] and the integrated discrimination improvement (IDI) [39]. Because the continuous NRI can overstate the value of weak markers, it is reported as supportive information only [40].

Software, ethics and data availability

Analyses used Python 3.13 with scikit-learn 1.9.1, lifelines 0.30.3, xgboost 3.4.1, statsmodels 0.15.0, pandas 2.3.3 and numpy 2.5.3. NHANES protocols were approved by the NCHS Research Ethics Review Board, and all participants gave written informed consent. This secondary analysis used publicly

available, de-identified data and did not require additional ethical approval. NHANES data and the public-use Linked Mortality Files are freely available from the NCHS. Analysis code, prompts, the Diabot adapter configuration, and all model outputs are available from the corresponding author on reasonable request.

Results

Participants

Of 70,190 NHANES participants, 6,450 were adults with diabetes. After excluding 10 participants not eligible for mortality linkage and 198 without examination follow-up time, 6,242 were included and split into a development set of 4,369 and a validation set of 1,873 (Fig. 2). Over a median follow-up of 6.0 years (interquartile range 3.2 to 9.5), 1,340 participants died, 445 of them from cardiovascular causes. A total of 4,320 participants could be classified for the 5-year outcome: 3,013 in the development set (519 deaths, 17.2%) and 1,307 in the validation set (202 deaths, 15.5%). Because participants examined in the most recent cycles had less than 5 years of follow-up, 96.5% of the 5-year analysis sample came from the 2005–2006 to 2013–2014 cycles.

Baseline characteristics were similar in the development and validation sets (Table 1). Participants who died within 5 years were older than survivors (mean age 70.8 versus 59.8 years) and more often had atherosclerotic cardiovascular disease (44.9% versus 20.6%), heart failure (26.1% versus 7.5%), an eGFR below 60 mL/min/1.73 m2 (41.2% versus 15.3%) and albuminuria (51.7% versus 27.6%).

Table 1
Baseline characteristics of participants evaluable for the 5-year outcome, by data set and by vital status at 5 years

Characteristic	Development set	Validation set	Died within 5 years	Alive at 5 years
Participants, n	3,013	1,307	721	3,599
Age, years, mean (SD)	61.6 (13.4)	61.9 (13.5)	70.8 (10.7)	59.8 (13.1)
Male sex, n (%)	1,540 (51.1)	659 (50.4)	422 (58.5)	1,777 (49.4)
Race and ethnicity, n (%)				
Non-Hispanic White	1,045 (34.7)	502 (38.4)	355 (49.2)	1,192 (33.1)
Non-Hispanic Black	856 (28.4)	381 (29.2)	208 (28.8)	1,029 (28.6)
Mexican American	547 (18.2)	223 (17.1)	70 (9.7)	700 (19.4)
Other Hispanic	289 (9.6)	119 (9.1)	43 (6.0)	365 (10.1)
Other or Multiracial	276 (9.2)	82 (6.3)	45 (6.2)	313 (8.7)
Family income-to-poverty ratio, mean (SD)	2.2 (1.5)	2.2 (1.5)	1.9 (1.3)	2.3 (1.5)
Smoking status, n (%)				
Never	1,484 (49.3)	637 (48.7)	280 (38.8)	1,841 (51.2)
Former	1,033 (34.3)	434 (33.2)	321 (44.5)	1,146 (31.8)
Current	493 (16.4)	236 (18.1)	120 (16.6)	609 (16.9)
Body mass index, kg/m2, mean (SD)	32.4 (7.6)	32.6 (7.7)	30.8 (8.1)	32.8 (7.5)
Systolic blood pressure, mmHg, mean (SD)	132.1 (20.7)	132.4 (20.8)	136.9 (24.5)	131.3 (19.8)
HbA1c, %, mean (SD)	7.4 (1.8)	7.4 (1.7)	7.3 (1.7)	7.4 (1.8)
Previously undiagnosed diabetes, n (%)	519 (17.2)	250 (19.1)	94 (13.0)	675 (18.8)
Diabetes duration, years, mean (SD)	12.0 (11.3)	11.8 (11.4)	16.1 (13.4)	11.1 (10.7)
Insulin use, n (%)	717 (23.8)	283 (21.7)	249 (34.5)	751 (20.9)
Oral glucose-lowering drug use, n (%)	1,190 (39.5)	506 (38.7)	262 (36.3)	1,434 (39.8)
Antihypertensive medication, n (%)	1,771 (58.8)	811 (62.1)	497 (68.9)	2,085 (57.9)

Characteristic	Development set	Validation set	Died within 5 years	Alive at 5 years
eGFR, mL/min/1.73 m2, mean (SD)	81.4 (25.8)	81.2 (25.3)	63.3 (25.7)	84.8 (24.1)
eGFR below 60 mL/min/1.73 m2, n (%)	582 (19.3)	265 (20.3)	297 (41.2)	550 (15.3)
UACR, mg/g, median (IQR)	14.9 (7.0 to 51.8)	13.0 (7.0 to 44.7)	43.6 (12.5 to 162.7)	12.6 (6.6 to 36.6)
UACR 30 mg/g or higher, n (%)	974 (32.3)	394 (30.1)	373 (51.7)	995 (27.6)
Total cholesterol, mg/dL, mean (SD)	186.1 (47.6)	186.7 (49.5)	173.9 (48.5)	188.7 (47.7)
HDL cholesterol, mg/dL, mean (SD)	47.8 (13.8)	48.3 (14.6)	48.5 (16.1)	47.8 (13.7)
Atherosclerotic cardiovascular disease, n (%)	750 (24.9)	316 (24.2)	324 (44.9)	742 (20.6)
Heart failure, n (%)	324 (10.8)	133 (10.2)	188 (26.1)	269 (7.5)
Death within 5 years, n (%)	519 (17.2)	202 (15.5)	721 (100.0)	0 (0.0)

Values are n (%), mean (SD) or median (IQR) of observed values. The vital status columns combine both data sets. Proportions of missing values: diabetes duration 21.9%, income-to-poverty ratio 9.4%, eGFR 7%, cholesterol 6.3%, HbA1c 4.6%, systolic blood pressure 4.4%, UACR 3.8%, body mass index 2.9%. eGFR, estimated glomerular filtration rate (2021 CKD-EPI creatinine equation); HDL, high-density lipoprotein; UACR, urine albumin-to-creatinine ratio.

Performance of the conventional models

In the validation set, the three conventional models discriminated similarly (Table 2). The Cox model had an AUC of 0.787 (95% CI 0.752 to 0.820) and was well calibrated (calibration slope 0.92, calibration-in-the-large 0.10, E/O 0.94). Logistic regression (AUC 0.786) and XGBoost (AUC 0.789) did not differ from the Cox model (differences of -0.001 and + 0.002, both with confidence intervals including zero).

Primary comparison: Diabot and base Llama 3.1 8B

With the main zero-shot prompt, Diabot returned a valid prediction for 1,289 of 1,307 participants (98.6%) and achieved an AUC of 0.755 (95% CI 0.720 to 0.790). In contrast, base Llama 3.1 8B returned a valid prediction for only 1,003 participants (76.7%) and achieved an AUC of 0.721 (0.679 to 0.759) (Table 2). Both LLMs discriminated less well than the Cox model when compared on the same participants: the AUC difference was − 0.033 (-0.056 to -0.008) for Diabot and − 0.049 (-0.083 to -0.012) for base Llama. Receiver operating characteristic curves are shown in Fig. 3.

In the paired comparison restricted to the 986 participants with valid outputs from both models, the AUC was 0.744 for Diabot and 0.720 for base Llama (difference 0.024, 95% CI -0.011 to 0.061, p = 0.20); the Cox model achieved 0.771 in the same participants. Diabot discriminated significantly better than base Llama with the few-shot prompt (0.745 versus 0.715 in 1,259 participants; difference 0.030, 0.001 to 0.062, p = 0.048) and with the stepwise prompt (0.775 versus 0.746 in 1,307 participants; difference 0.029, 0.001 to 0.052, p = 0.030).

The best-performing LLM condition was Diabot with the stepwise prompt (AUC 0.775, 0.740 to 0.808), which did not differ significantly from the Cox model (difference − 0.012, -0.036 to 0.011). Few-shot examples did not improve discrimination for either model (Diabot 0.747; base Llama 0.715). Base Llama with the reworded prompt produced no valid predictions and could not be evaluated.

Table 2
Discrimination and calibration of the conventional models and the large language models for 5-year all-cause mortality in the validation set

Model and condition	Valid predictions, n (%)	AUC (95% CI)	AUC difference vs Cox (95% CI)	Brier score	Cal. slope	CITL	E/O ratio
Cox proportional hazards	1,307 (100.0)	0.787 (0.752 to 0.820)	Reference	0.109	0.92	0.10	0.94
Logistic regression (splines)	1,307 (100.0)	0.786 (0.751 to 0.820)	-0.001 (-0.013 to 0.012)	0.111	0.83	-0.13	1.08
XGBoost	1,307 (100.0)	0.789 (0.753 to 0.822)	+ 0.002 (-0.016 to 0.020)	0.110	0.81	-0.07	1.04
Diabot, zero-shot	1,289 (98.6)	0.755 (0.720 to 0.790)	-0.033 (-0.056 to -0.008)	0.116	1.29	-0.24	1.20
Llama 3.1 8B, zero-shot	1,003 (76.7)	0.721 (0.679 to 0.759)	-0.049 (-0.083 to -0.012)	0.123	1.46	-0.47	1.43
Diabot, few-shot	1,307 (100.0)	0.747 (0.708 to 0.781)	-0.040 (-0.066 to -0.016)	0.119	1.64	-0.30	1.26
Llama 3.1 8B, few-shot	1,259 (96.3)	0.715 (0.680 to 0.745)	-0.072 (-0.105 to -0.040)	0.141	1.12	-0.93	1.93
Diabot, reworded prompt	1,281 (98.0)	0.761 (0.724 to 0.794)	-0.031 (-0.054 to -0.009)	0.134	0.73	-0.87	1.69
Llama 3.1 8B, reworded prompt	0 (0.0)	Not evaluable	Not evaluable				
Diabot, stepwise prompt	1,307 (100.0)	0.775 (0.740 to 0.808)	-0.012 (-0.036 to 0.011)	0.135	0.71	-1.06	1.82
Llama 3.1 8B, stepwise prompt	1,307 (100.0)	0.746 (0.714 to 0.777)	-0.041 (-0.068 to -0.015)	0.129	1.59	-0.72	1.71

Validation set: 1,307 participants, 202 deaths. LLM conditions used greedy decoding. LLM metrics are computed on participants with a valid output; differences from the Cox model are paired and computed on the same participants. The ideal calibration slope is 1 (a slope above 1 indicates predictions that are too narrow in range, below 1 too extreme), the ideal calibration-in-the-large is 0 (negative values indicate

over-prediction), and the ideal E/O ratio is 1. AUC, area under the receiver operating characteristic curve; Cal. slope, calibration slope; CITL, calibration-in-the-large; E/O, expected over observed number of deaths.

Calibration

Both LLMs over-predicted mortality (Table 2, Fig. 4). With the main prompt, Diabot predicted 20% more deaths than observed (E/O 1.20; calibration-in-the-large − 0.24) and base Llama 43% more (E/O 1.43; -0.47). Calibration slopes above 1 (Diabot 1.29; base Llama 1.46) indicated that the spread of LLM predictions was too narrow rather than too extreme. Over-prediction was larger with the alternative prompts, reaching an E/O ratio of 1.82 for Diabot with the stepwise prompt and 1.93 for base Llama with the few-shot prompt. The calibration curves of the LLMs were step-shaped (Fig. 4) because their outputs clustered on a small number of values: across 1,307 participants, Diabot produced 32 distinct risk values with the main prompt and base Llama 20 (Table 3). The most frequent Diabot outputs were 12.5% (258 participants), 34.5% (210 participants) and 14.5% (185 participants).

Format adherence and output resolution

Adherence to the required output format differed markedly between the models (Table 3). Diabot returned valid JSON in 98.0% to 100% of participants under all greedy conditions. In contrast, base Llama returned valid JSON in 76.7% of cases with the main prompt and 0% with the reworded prompt, where it answered every case with free text that described a risk-scoring method and was truncated before giving a number. Under stochastic sampling, adherence fell for both models, to 65.1% to 66.3% for Diabot and 48.9% to 52.6% for base Llama. With sampling, adherence was consistently higher among participants who died than among survivors (for example, 71.8% versus 63.9% for Diabot in the second run, and 60.4% versus 51.1% for base Llama in the first run), so that format failures were not random with respect to the outcome.

Table 3
Output format adherence and resolution of large language model predictions across all 14 conditions (validation set, n = 1,307 per condition)

Model	Prompt	Decoding	Valid JSON, n (%)	Valid among deaths, %	Valid among survivors, %	Distinct risk values
Diabot	Main (zero-shot)	Greedy	1,289 (98.6)	99.5	98.5	32
Llama 3.1 8B	Main (zero-shot)	Greedy	1,003 (76.7)	75.7	76.9	20
Diabot	Few-shot (10 examples)	Greedy	1,307 (100.0)	100.0	100.0	22
Llama 3.1 8B	Few-shot (10 examples)	Greedy	1,259 (96.3)	97.0	96.2	17
Diabot	Reworded	Greedy	1,281 (98.0)	99.0	97.8	54
Llama 3.1 8B	Reworded	Greedy	0 (0.0)	0.0	0.0	0
Diabot	Stepwise	Greedy	1,307 (100.0)	100.0	100.0	42
Llama 3.1 8B	Stepwise	Greedy	1,307 (100.0)	100.0	100.0	15
Diabot	Main (zero-shot)	Sampled, run 1	853 (65.3)	67.3	64.9	191
Llama 3.1 8B	Main (zero-shot)	Sampled, run 1	687 (52.6)	60.4	51.1	123
Diabot	Main (zero-shot)	Sampled, run 2	851 (65.1)	71.8	63.9	189
Llama 3.1 8B	Main (zero-shot)	Sampled, run 2	669 (51.2)	52.5	51.0	125
Diabot	Main (zero-shot)	Sampled, run 3	866 (66.3)	71.3	65.3	187
Llama 3.1 8B	Main (zero-shot)	Sampled, run 3	639 (48.9)	54.0	48.0	118

A valid output contained an explicit probability between 0 and 100. Greedy decoding is deterministic; sampled runs used temperature 0.7 and top-p 0.9. Distinct risk values are the number of different probabilities produced across the participants with valid outputs.

Run-to-run stability and prompt sensitivity

Only 407 participants for Diabot and 188 for base Llama had valid outputs in all three sampled runs. Among them, the mean within-participant standard deviation of predicted risk across runs was 4.9 percentage points for Diabot (maximum 18.9) and 5.1 for base Llama (maximum 16.2). The minimum pairwise Spearman correlation between runs was 0.72 for Diabot and 0.66 for base Llama, and the AUC ranged from 0.688 to 0.700 for Diabot and from 0.610 to 0.671 for base Llama. Repeating the same prompt moved 26.3% of Diabot predictions and 12.2% of base Llama predictions across the 10% risk threshold, and 21.1% and 34.0% across the 20% threshold.

Changing the prompt wording altered the level of predicted risk more than its ranking. For Diabot, the Spearman correlation with the main prompt was 0.84 for the reworded prompt and 0.83 for the stepwise prompt. Still, the mean predicted risk rose from 18.8% to 26.6% and 28.4%, respectively (mean absolute differences of 10.0 and 11.4 percentage points). Few-shot examples changed Diabot predictions less (Spearman 0.77; mean absolute difference 5.1 percentage points). For base Llama, the correlation with the main prompt was 0.86 for the stepwise prompt and 0.75 for the few-shot prompt.

Clinical utility

In decision curve analysis (Fig. 5), the conventional models provided the highest net benefit across most thresholds. At a 10% risk threshold, net benefit was 0.089 for the Cox model, 0.070 for Diabot, 0.064 for base Llama and 0.061 for treating everyone. At a 20% threshold, Diabot (0.050) was close to the Cox model (0.052), whereas base Llama (0.023) was not. At 30%, the net benefit of base Llama fell below zero (-0.005), while Diabot retained a positive net benefit (0.027) that remained below the Cox model (0.037).

Subgroup performance

Discrimination was lowest in participants with income below the poverty line for all three models (AUC 0.667 for Cox, 0.646 for Diabot and 0.606 for base Llama) (Supplementary Table S1). Both LLMs over-predicted mortality in every subgroup except Diabot in the low-income band (E/O 0.99), with the greatest over-prediction among Mexican American participants (E/O 2.16 for Diabot and 2.55 for base Llama, compared with 1.60 for Cox) and participants in the highest income band (1.99 and 2.11, compared with 1.20 for Cox). Diabot discriminated less well than the Cox model in every subgroup. Several subgroups had few events (11 to 16), so these estimates are imprecise.

Post hoc improvement analyses

Token probability scoring produced a valid, continuous risk score for every participant, increasing the number of distinct values to 327 for Diabot and 71 for base Llama (Table 4). Its discrimination was slightly lower than that of the generated risks (AUC 0.741 for Diabot and 0.704 for base Llama). Raw Yes/No probabilities were poorly calibrated, with E/O ratios of 2.88 and 3.95.

Logistic recalibration fitted in the development set corrected most of the miscalibration in the validation set without changing discrimination (Table 4, Fig. 6). After recalibration, the Diabot Yes/No score had a

calibration slope of 1.01, calibration-in-the-large of -0.17 and E/O of 1.14, and the recalibrated Diabot generated risk had a slope of 0.86 and E/O of 1.14. Base Llama reached a slope of 1.04 and E/O of 1.09 after recalibration of its Yes/No score. Recalibration also reduced the Brier score of the Diabot Yes/No score from 0.219 to 0.116, close to that of the Cox model (0.109).

Table 4
Post hoc analyses: performance of token probability scores and recalibrated large language model outputs in the validation set

Analysis	n	AUC (95% CI)	Brier score	Cal. slope	CITL	E/O ratio	Distinct values
Cox proportional hazards (reference)	1,307	0.787 (0.752 to 0.820)	0.109	0.92	0.10	0.94	1041
Diabot: Yes/No probability, raw	1,307	0.741 (0.702 to 0.777)	0.219	0.79	-1.88	2.88	327
Diabot: Yes/No probability, recalibrated	1,307	0.741 (0.705 to 0.778)	0.116	1.01	-0.17	1.14	212
Diabot: generated risk, raw (valid JSON only)	1,289	0.755 (0.717 to 0.788)	0.116	1.29	-0.24	1.20	32
Diabot: generated risk, recalibrated (valid JSON only)	1,289	0.755 (0.718 to 0.790)	0.115	0.86	-0.18	1.14	32
Llama 3.1 8B: Yes/No probability, raw	1,307	0.704 (0.662 to 0.743)	0.366	0.46	-3.23	3.95	71
Llama 3.1 8B: Yes/No probability, recalibrated	1,307	0.704 (0.667 to 0.746)	0.121	1.04	-0.12	1.09	66
Llama 3.1 8B: generated risk, raw (valid JSON only)	1,003	0.721 (0.681 to 0.758)	0.123	1.46	-0.47	1.43	20
Llama 3.1 8B: generated risk, recalibrated (valid JSON only)	1,003	0.721 (0.677 to 0.761)	0.119	0.95	-0.12	1.09	20

Yes/No probability: P(Yes) / (P(Yes) + P(No)) from the next-token distribution after asking whether the person would die within 5 years. Generated risk: the probability written by the model with the main zero-shot prompt and greedy decoding. Recalibration intercepts and slopes were estimated in the development set and applied unchanged to the validation set. AUC confidence intervals are from separate bootstrap runs and may differ slightly between raw and recalibrated scores that have identical

ranking. Cal. slope, calibration slope; CITL, calibration-in-the-large; E/O, expected over observed number of deaths.

Adding the Diabot score to the Cox model showed that it carried independent prognostic information (hazard ratio 1.18 per unit increase in the logit of the score, 95% CI 1.06 to 1.32, p = 0.002) but did not materially improve discrimination (AUC 0.787 to 0.788; difference + 0.001, -0.002 to 0.004) (Table 5). The IDI was 0.004, and the continuous NRI was 0.24 (0.11 for events and 0.13 for non-events). The corresponding results for base Llama were weaker (AUC difference − 0.001, IDI 0.0006, continuous NRI 0.12, with a negative component for events).

Table 5
Post hoc analysis: incremental value of adding the large language model score to the Cox model (validation set, n = 1,307, 202 deaths)

LLM score added	AUC, Cox	AUC, Cox + LLM	AUC difference (95% CI)	Continuous NRI (events; non-events)	IDI	Hazard ratio per logit unit (95% CI)
Diabot	0.787	0.788	+ 0.001 (-0.002 to 0.004)	0.240 (0.109; 0.131)	0.0044	1.18 (1.06 to 1.32); p = 0.002
Llama 3.1 8B	0.787	0.786	-0.001 (-0.004 to 0.003)	0.119 (-0.059; 0.178)	0.0006	1.15 (1.05 to 1.25); p = 0.002

The LLM score is the logit of the Yes/No token probability. Both Cox models were fitted in the development set (n = 4,369) with the same 19 clinical predictors. The continuous NRI is reported as supportive information because it can be inflated by weak predictors. IDI, integrated discrimination improvement; NRI, net reclassification improvement.

Discussion

Principal findings

In this validation study of 1,307 adults with diabetes, a diabetes-specialized LLM predicted 5-year all-cause mortality more reliably and somewhat more accurately than its general-purpose base model. Still, neither LLM matched a conventional Cox model built from the same routinely collected variables. The main benefit of fine-tuning was reliability: Diabot almost always returned a usable risk estimate, whereas the base model failed to do so for nearly a quarter of patients with the main prompt and for every patient with a reworded prompt. The gain in discrimination was modest (about 0.03 in AUC) and was statistically significant with two of the three prompts tested. Both LLMs produced coarse, overestimated risks that shifted with prompt wording and varied between repeated runs. Recalibration in the development data corrected most of the miscalibration, and token probability scoring removed the coarseness and the format failures. Still, neither strategy closed the gap in discrimination, and adding the LLM score to the Cox model did not improve it.

Comparison with other studies

Our results are consistent with studies that found general-purpose LLMs to be less accurate than conventional or locally trained models for clinical prediction [5, 6]. The gap we observed (0.033 in AUC for Diabot) was smaller than the differences reported for hospital prediction tasks, where LLM AUCs were around 0.52 to 0.64 against 0.85 to 0.89 for gradient-boosted trees [5]. A likely reason is that the predictors of mortality in diabetes, such as age, kidney function, albuminuria and cardiovascular disease, are well described in the medical literature that LLMs are trained on, so a model can rank patients sensibly from general knowledge. The base model's frequent failure to follow the answer format agrees with reports that LLMs do not reliably follow instructions in clinical tasks [7], and the large shifts in predicted risk between equivalent prompts agree with evidence that LLM outputs are sensitive to superficial features of prompt design [9].

Medically adapted LLMs often do not outperform their base models on medical benchmarks [10]. Our comparison, which used identical quantized weights with and without the adapter, suggests that adaptation on guideline material can still help, mainly by making the model a more dependable source of structured output. Diabot was trained on treatment decisions that depend on eGFR, albuminuria, cardiovascular disease and age, which are also the strongest predictors of mortality; this overlap may explain why its ranking of patients improved slightly even though it was never trained on outcomes.

Calibration, resolution and stability

The written probabilities from both LLMs behaved more like a small number of risk categories than continuous estimates. Diabot used only 32 distinct values for 1,307 patients, and three values accounted for half of its predictions. This explains the step-shaped calibration curves and the calibration slopes above 1, which indicate that the predicted risks were too similar between patients rather than too extreme. The level of predicted risk also depended on the prompt: asking the model to reason through risk factors improved discrimination but raised the mean predicted risk from about 19% to 28%. Similar unreliability of both stated probabilities and token probabilities has been reported for diagnostic probability estimation [8]. In our data, the token probability approach produced a continuous score for every patient but was itself poorly calibrated until recalibrated, after which calibration was close to ideal (slope 1.01).

Stochastic sampling, which is the default in many LLM interfaces, made matters worse: valid outputs fell to about two-thirds or less, and between a fifth and a third of patients moved across a 20% risk threshold when the same prompt was repeated. Any clinical use of LLM-generated risk would therefore need deterministic decoding, fixed prompts and recalibration in the target population, and these settings should be reported in full [11].

Format failures as informative missing data

Format failures were not random. Under sampling, both models were more likely to give a valid answer for patients who later died than for survivors. Evaluating an LLM only on its valid outputs can therefore give a biased picture of its performance, and comparisons between models with different failure rates are not like-for-like unless they are restricted to patients with valid outputs from both, as in our paired analyses. We recommend that LLM prediction studies report the proportion of valid outputs, the parsing rule used and the failure rate by outcome status. Our experience also shows that the parsing rule needs care: a lenient parser that took the first number from free text would have recorded base Llama's truncated answers as a 5% risk for every patient and produced spurious results.

Incremental value

Diabot's score was associated with mortality independently of the clinical predictors (hazard ratio 1.18 per logit unit). Still, it did not improve discrimination when added to the Cox model, and the integrated discrimination improvement was negligible. The LLM therefore appears to re-express, in a noisier form, information that is already contained in the structured predictors. The continuous NRI of 0.24 should not be interpreted as evidence of useful added value, because this measure can be large for predictors that do not change clinical decisions [40].

Implications

For clinicians and health systems that already hold structured data on age, kidney function, glycaemia, blood pressure and cardiovascular history, a conventional regression model remains the better choice for predicting mortality in diabetes: it was more accurate, better calibrated, fully reproducible, transparent and far cheaper to run. LLMs are more likely to add value where important prognostic information is held only in unstructured text, such as clinical notes, which our design did not test. For developers of specialized medical LLMs, our results show that fine-tuning should be evaluated against the base model with identical settings and against conventional models, and that output reliability is an outcome in its own right. For researchers, LLM risk estimates should be assessed for calibration and resolution, not only for discrimination, and should not be presented to clinicians without recalibration.

Strengths and limitations

This study has several strengths. Using the same quantized weights with the adapter switched on or off isolated the effect of fine-tuning from all other differences between models. The data came from a national survey with standardized measurements and linkage to death registrations, and Diabot had never seen NHANES data or outcome labels, which excludes leakage. The primary analysis and prompting conditions were fixed before inference, the development sample met formal sample size criteria, and we assessed discrimination, calibration, clinical utility, output reliability, stability, prompt sensitivity and subgroup performance together, as recommended for LLM prediction studies [11]. Both LLMs are open so that the analysis can be reproduced.

There are also important limitations. First, we used a single random split of one dataset; the validation set is internal, and performance in other countries, health systems and periods, including Indian populations, remains to be established. Because recent NHANES cycles had insufficient follow-up, 96.5% of the 5-year sample was examined between 2005 and 2014. Second, the patient summaries were generated from structured variables, so the LLMs could use only the information that the conventional models also had; real clinical notes may contain additional prognostic information but also more noise. Third, we evaluated one fine-tuned model of 8 billion parameters in 4-bit precision and did not test larger open models or commercial models, which may perform differently. Fourth, Diabot was trained for treatment recommendation rather than prognosis. Hence, the results describe the transfer of guideline fine-tuning to an unseen task and not the performance achievable by training on outcomes. Fifth, the same validation set was used for all prompting conditions; the primary comparison was fixed in advance, but the secondary prompting conditions and the post hoc analyses should be regarded as exploratory. Sixth, the 24-token output limit truncated free-text answers; a longer limit might have rescued some of base Llama's invalid outputs, although the instruction asked for a number only. Finally, diabetes type could not be distinguished reliably, comorbidities were self-reported, analyses were unweighted, and several subgroups had few events, so the subgroup estimates are imprecise.

Future research

Future work should evaluate LLMs on unstructured clinical text, where they may add information that structured models cannot use; fine-tune models directly on outcome labels and compare them with conventional models of similar complexity; test larger and commercial models; and validate any LLM-based risk tool temporally and externally, including in low- and middle-income settings, with recalibration and full reporting of output reliability.

Conclusions

Fine-tuning Llama 3.1 8B on diabetes guideline material produced a model that gave valid risk estimates far more reliably and ranked patients slightly better than its base model, but neither LLM matched a conventional Cox model using the same routinely collected data, and both produced coarse, overestimated, and prompt-dependent risks. Recalibration and token probability scoring corrected calibration and resolution but not discrimination. For predicting mortality in adults with diabetes from structured clinical data, conventional models should remain the standard, and LLM-derived risk estimates should be recalibrated and fully validated before any clinical use.

Abbreviations

AUC
area under the receiver operating characteristic curve
CI
confidence interval

CITL
calibration-in-the-large
eGFR
estimated glomerular filtration rate
E/O
ratio of expected to observed deaths
IDI
integrated discrimination improvement
JSON
JavaScript Object Notation
LLM
large language model
LoRA
low-rank adaptation
NCHS
National Center for Health Statistics
NHANES
National Health and Nutrition Examination Survey
NRI
net reclassification improvement
QLoRA
quantized low-rank adaptation
UACR
urine albumin-to-creatinine ratio.

Declarations

Ethics approval and consent to participate

NHANES protocols were approved by the National Center for Health Statistics Research Ethics Review Board, and all participants provided written informed consent. This secondary analysis used publicly available, de-identified data and did not require further ethical approval.

Consent for publication

Not applicable.

Availability of data and materials

The NHANES data and the public-use Linked Mortality Files analysed in this study are publicly available from the National Center for Health Statistics (https://wwwn.cdc.gov/nchs/nhanes/ and https://www.cdc.gov/nchs/data-linkage/mortality-public.htm). The analysis code, prompts, patient text template, Diabot adapter configuration and model outputs are available from the corresponding author on reasonable request.

Competing interests

G.H.P. developed Diabot, the fine-tuned model evaluated in this study. The other authors declare that they have no competing interests.

Funding

Nil

Authors' contributions

G.H.P. conceived and designed the study, developed and trained Diabot, curated and analyzed the data, prepared the figures and wrote the first draft of the manuscript. D.S.K. contributed to conceptualization, supervision, interpretation of findings and review and editing of the manuscript. T.B.D. contributed to methodology, validation and review and editing. A.P.K. contributed to the clinical interpretation of diabetes pharmacotherapy content and review and editing. A.N. contributed to data curation, validation of the analysis and review and editing of the manuscript. V.A. contributed to the machine learning and language model methodology, validated the modelling pipeline and reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

We thank the National Center for Health Statistics and the NHANES participants.

References

1. International Diabetes Federation. IDF Diabetes Atlas. 11th ed. Brussels: International Diabetes Federation; 2025.
2. Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from randomized trials. Lancet Diabetes Endocrinol. 2017;5(10):788–98.
3. Wells BJ, Jain A, Arrigain S, Yu C, Rosenkrans WA Jr, Kattan MW. Predicting 6-year mortality risk in patients with type 2 diabetes. Diabetes Care. 2008;31(12):2301–6.

4. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023;619(7969):357–62.
5. Brown KE, Yan C, Li Z, Zhang X, Collins BX, Chen Y, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. J Am Med Inf Assoc. 2025;32(5):811–22.
6. Chen C, Yu J, Chen S, Liu C, Wan Z, Bitterman D et al. ClinicalBench: can LLMs beat traditional ML models in clinical prediction? arXiv. 2024. 10.48550/arXiv.2411.06469
7. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613–22.
8. Gao Y, Myers S, Chen S, Dligach D, Miller T, Bitterman DS, et al. Uncertainty estimation in diagnosis generation from large language models: next-word probability is not pre-test probability. JAMIA Open. 2025;8(1):ooae154.
9. Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. In: International Conference on Learning Representations; 2024.
10. Jeong DP, Garg S, Lipton ZC, Oberst M. Medical adaptation of large language and vision-language models: are we making progress? In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami: Association for Computational Linguistics; 2024.
11. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–9.
12. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
13. Zipf G, Chiappa M, Porter KS, Ostchega Y, Lewis BG, Dostal J. National Health and Nutrition Examination Survey: plan and operations, 1999–2010. Vital Health Stat. 2013;1(56):1–37.
14. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: Standards of Care in Diabetes-2024. Diabetes Care. 2024;47(Suppl 1):S20-42.
15. National Center for Health Statistics, Office of Analysis and Epidemiology. Public-use Linked Mortality Files, 2019. Hyattsville, MD: National Center for Health Statistics. 2022. Available from: https://www.cdc.gov/nchs/data-linkage/mortality-public.htm
16. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med. 2021;385(19):1737–49.
17. National Center for Health Statistics. National Health and Nutrition Examination Survey 2005–2006 data documentation: Standard Biochemistry Profile (BIOPRO_D). Hyattsville, MD: National Center for Health Statistics.
18. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.

19. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35(2):214–26.
20. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67.
21. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. 2019;4(40):1317.
22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. pp. 785 – 94.
23. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A et al. The Llama 3 herd of models. arXiv. 2024. 10.48550/arXiv.2407.21783
24. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S et al. LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations; 2022.
25. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient fine-tuning of quantized LLMs. In: Adv Neural Inf Process Syst 36; 2023.
26. Ben Abacha A, Shivade C, Demner-Fushman D. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Florence: Association for Computational Linguistics; 2019. pp. 370-9.
27. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Loser A, et al. MedAlpaca: an open-source collection of medical conversational AI models and training data. arXiv. 2023. 10.48550/arXiv.2304.08247.
28. American Diabetes Association Professional Practice Committee. Standards of Care in Diabetes-2024. Diabetes Care. 2024;47(Suppl 1):S1–321.
29. Kidney Disease: Improving Global Outcomes (KDIGO) Diabetes Work Group. KDIGO 2022 clinical practice guideline for diabetes management in chronic kidney disease. Kidney Int. 2022;102(5S):S1–127.
30. Chawla R, Madhu SV, Makkar BM, Ghosh S, Saboo B, Kalra S, et al. RSSDI-ESI clinical practice recommendations for the management of type 2 diabetes mellitus 2022. Int J Diabetes Dev Ctries. 2022;42(Suppl 1):1–143.
31. Indian Council of Medical Research. ICMR guidelines for management of type 2 diabetes 2018. New Delhi: Indian Council of Medical Research; 2018.
32. Hegselmann S, Buendia A, Lang H, Agrawal M, Jiang X, Sontag D. TabLLM: few-shot classification of tabular data with large language models. In: Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. PMLR 206; 2023. pp. 5549-81.
33. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall; 1993.
34. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.

35. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Cham: Springer; 2019.
36. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565–74.
37. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45(3–4):562–5.
38. Pencina MJ, D'Agostino RB, Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30(1):11–21.
39. Pencina MJ, D'Agostino RB, Sr, Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157–72.
40. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology. 2014;25(1):114–21.

Figures

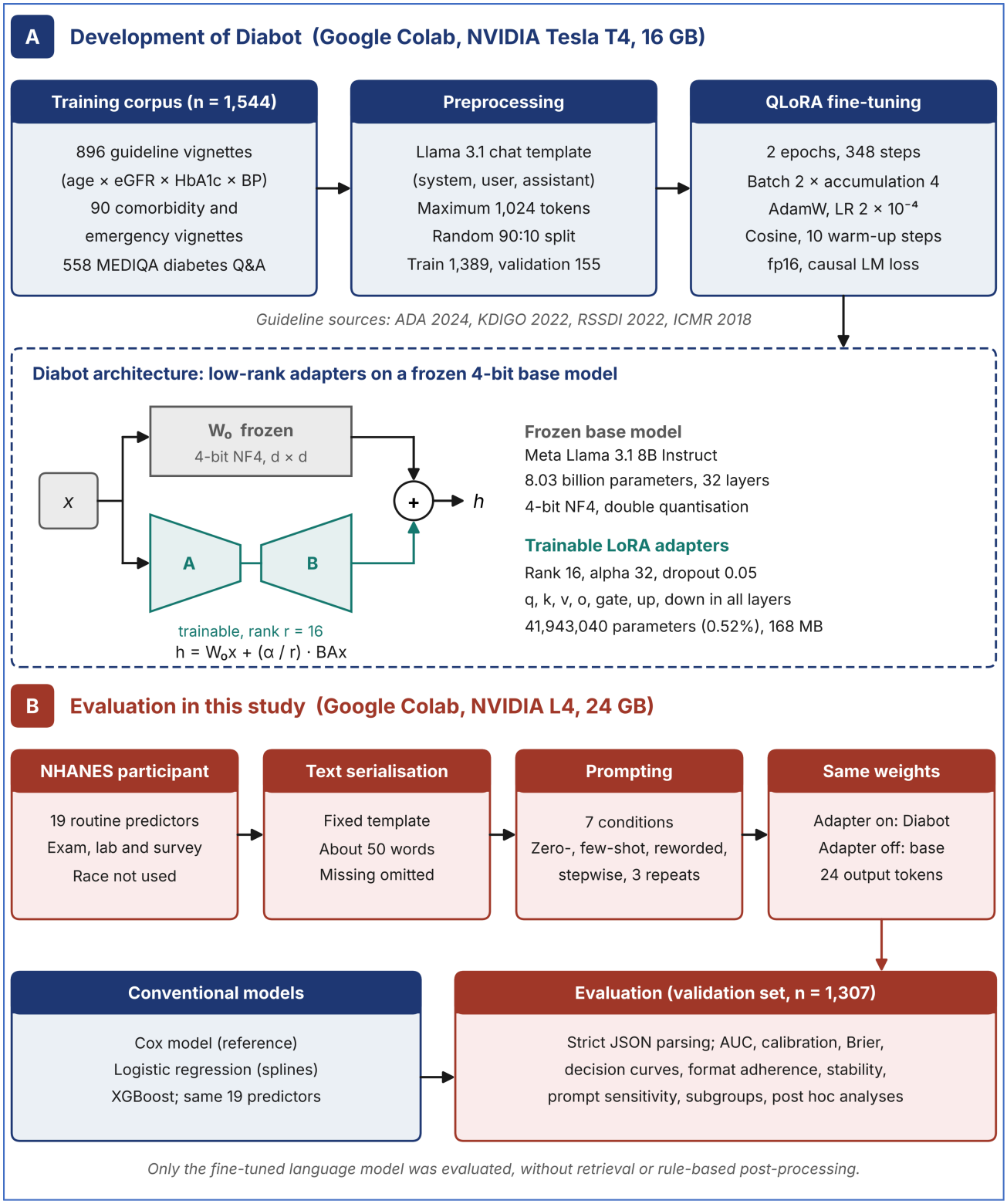

Figure 1

Overview of Diabot and the evaluation. (A) Development of Diabot: training corpus, preprocessing, QLoRA fine-tuning settings and model architecture, with a frozen 4-bit Llama 3.1 8B base model and trainable low-rank adapters. (B) Evaluation pipeline: NHANES records were converted to text, presented under seven prompting conditions to the same quantized model with the adapter switched on (Diabot) or off (base Llama 3.1 8B), parsed and evaluated against conventional models. ADA, American Diabetes

Association; eGFR, estimated glomerular filtration rate; ICMR, Indian Council of Medical Research; KDIGO, Kidney Disease: Improving Global Outcomes; LoRA, low-rank adaptation; QLoRA, quantized LoRA; RSSDI, Research Society for the Study of Diabetes in India.

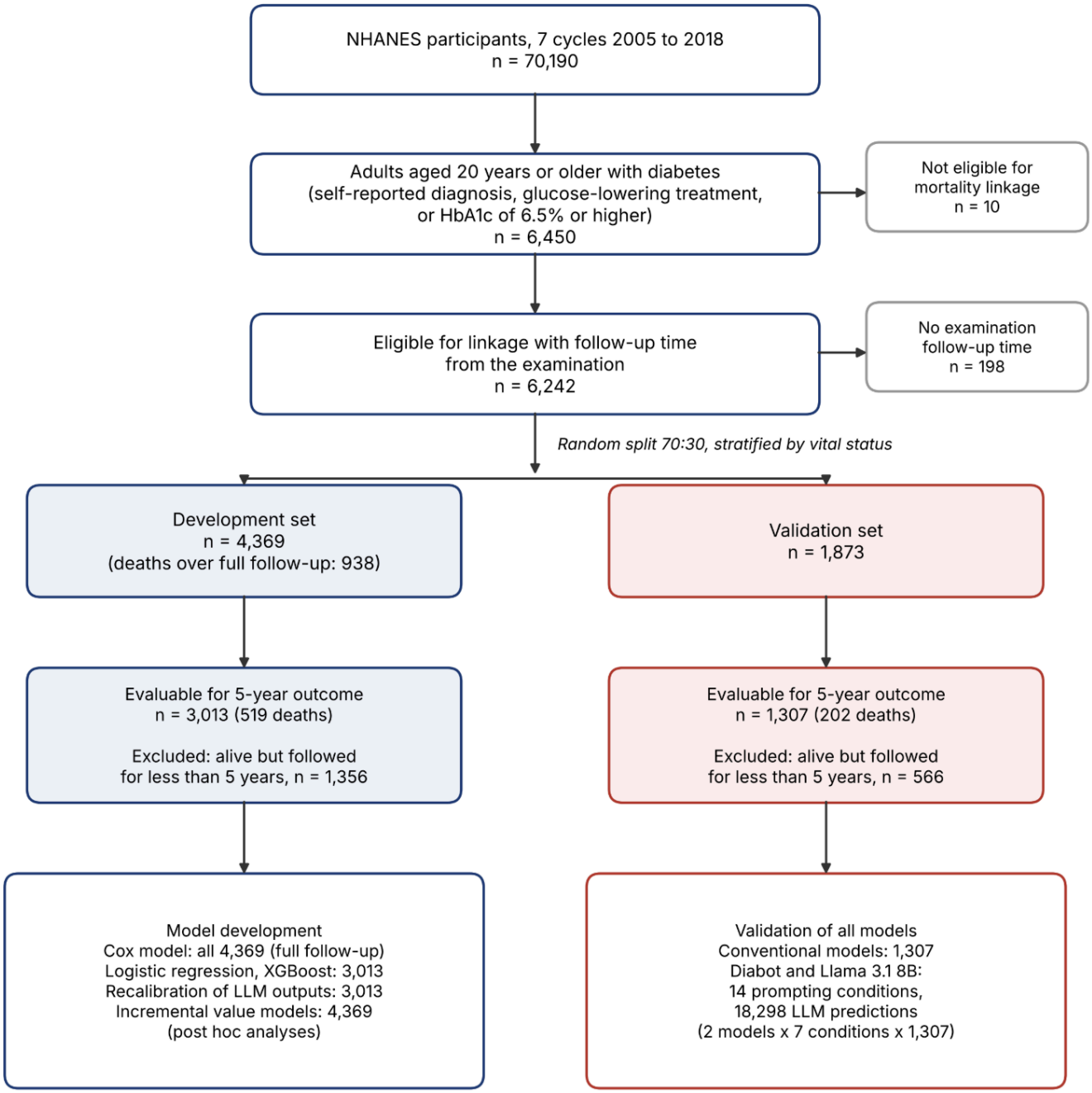

Figure 2

Participant flow. NHANES, National Health and Nutrition Examination Survey; LLM, large language model.

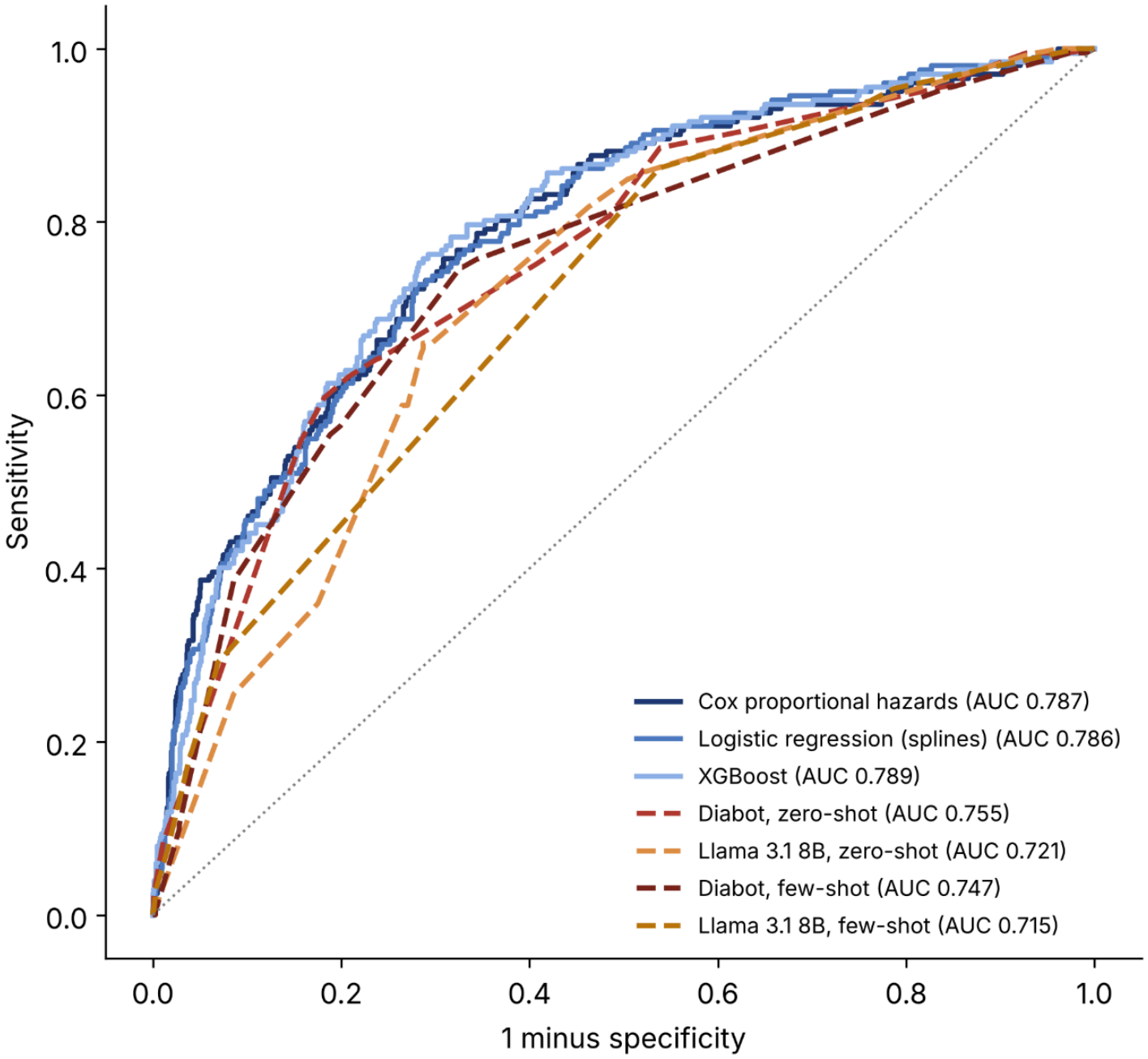

Figure 3

Receiver operating characteristic curves for 5-year all-cause mortality in the validation set. Solid lines show the conventional models; dashed lines show Diabot and base Llama 3.1 8B with zero-shot and few-shot prompts. LLM curves are based on participants with valid outputs.

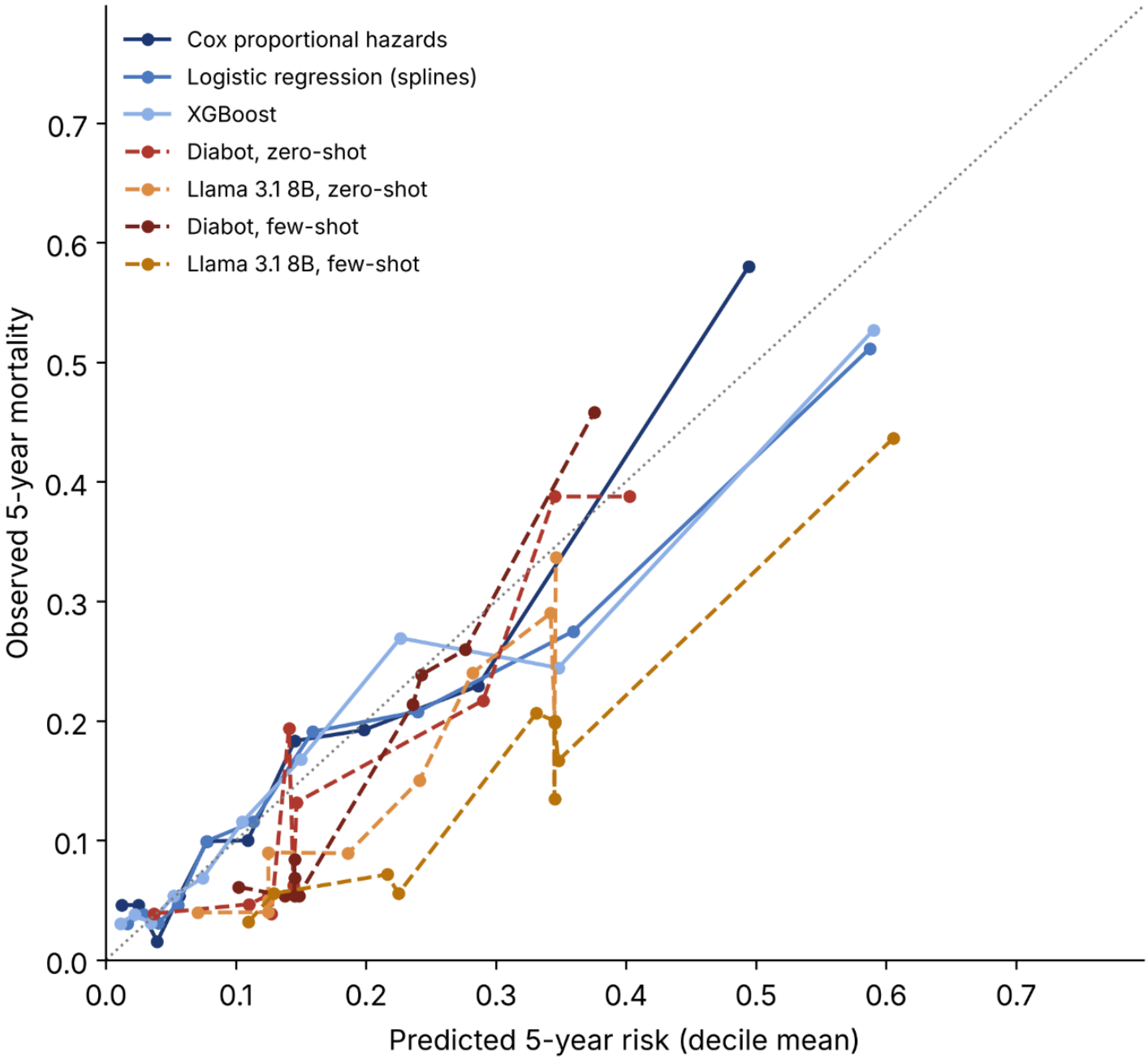

Figure 4

Calibration of predicted 5-year mortality risk in the validation set. Points show the mean predicted risk and the observed mortality within deciles of predicted risk; the dotted diagonal indicates perfect calibration. Solid lines show the conventional models; dashed lines show the LLMs.

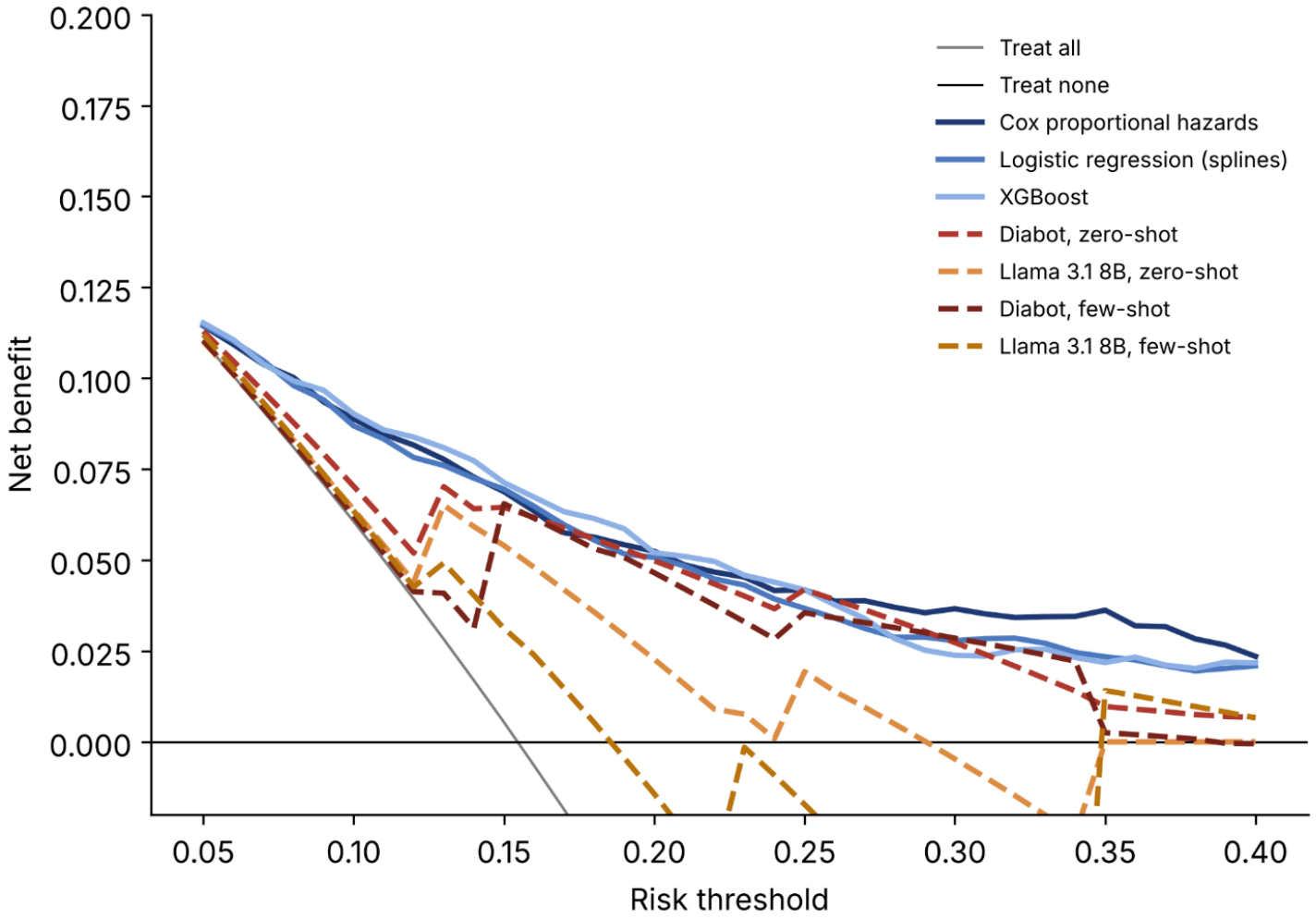

Figure 5

Decision curves for 5-year all-cause mortality in the validation set. Net benefit is plotted against the risk threshold for each model and for the strategies of treating all or no patients. LLM net benefit is computed on participants with valid outputs.

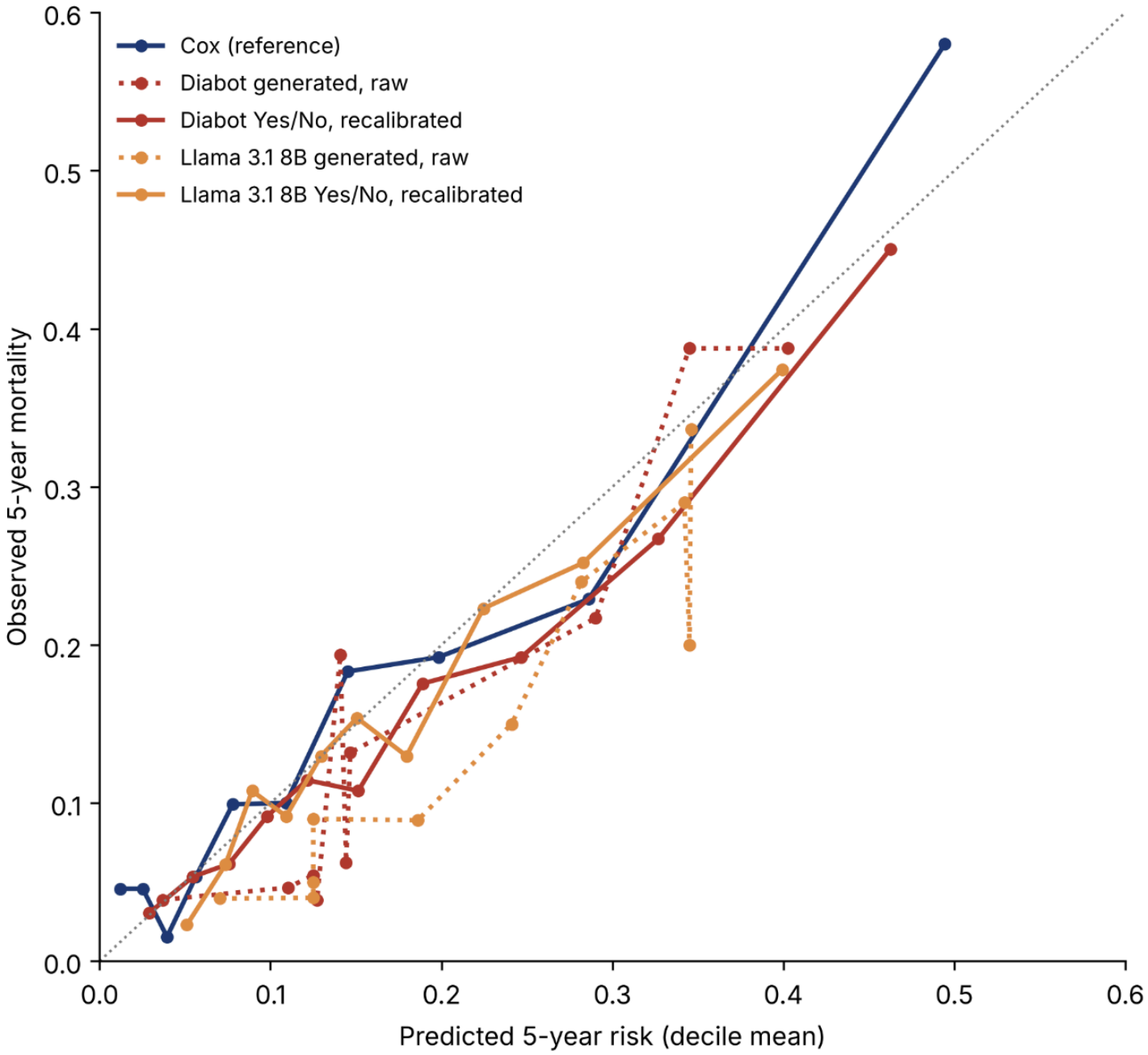

Figure 6

Calibration of large language model risk predictions before and after post hoc improvement. Dotted lines show the raw generated risks with the main prompt; solid coloured lines show token probability (Yes/No) scores after logistic recalibration fitted in the development set. The Cox model is shown for reference.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx