

# Machine Learning and Large Language Models for Maternal Cardiometabolic Risk Prediction Across Harmonised Public Pregnancy Datasets

**Shaoxun Wen**

The Central Hospital of Yongzhou

**Yaqun Tang**

The Third People's Hospital of Yongzhou

**Binning Ma**

The Central Hospital of Yongzhou

**Bin Cao**

The Central Hospital of Yongzhou

**Zhenxing Deng**

dengzhenxing36@163.com

The Central Hospital of Yongzhou

---

## Article

### Keywords:

**Posted Date:** October 18th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11072651/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

Pregnancy-related cardiometabolic complications remain important contributors to adverse maternal and fetal outcomes. However, risk-prediction models developed from a single dataset often have limited transportability, and the role of large language models in maternal risk stratification remains unclear.

## Methods

We used two publicly available datasets: the Maternal Health Risk dataset and the Preeclampsia in Pregnant Women dataset. Low-risk, mid-risk, and high-risk labels were harmonised, and a binary high-risk versus non-high-risk endpoint was constructed. We compared several machine-learning models and undertook internal cross-validation, external validation, calibration, decision-curve analysis, and sensitivity analyses. Structured records were subsequently converted into label-free case narratives; DeepSeek and Qwen were evaluated under different prompting strategies, and the concordance between keywords in large language model explanations and machine-learning feature importance was examined.

## Results

In internal validation, HistGradientBoosting showed the strongest overall performance, with a macro-averaged F1 score of approximately 0.876. Performance declined in external validation, in which multiclass accuracy was approximately 0.389. Across sensitivity analyses, the area under the receiver operating characteristic curve ranged from 0.437 to 0.528. Among the large language model configurations, Qwen with few-shot chain-of-thought prompting performed best, with an accuracy of approximately 0.583 and a macro-averaged F1 score of approximately 0.467. Explanatory keywords generated by the large language models most often concerned blood glucose or diabetes, blood pressure or hypertension, haemoglobin or anaemia, previous complications or medical history, and proteinuria.

## Conclusion

We developed an integrated framework for maternal cardiometabolic risk prediction that combines feature harmonisation, conventional machine learning, external validation, and large language model explanations. Machine-learning models performed well internally but generalised poorly across datasets. Large language models showed some ability to assess risk from case text and generate readable clinical explanations, but require validation in larger, multicentre, prospective studies.

# 1. Introduction

Pregnancy acts as a dynamic stress test of maternal cardiovascular and metabolic reserve. Hypertensive disorders of pregnancy, including pre-eclampsia, can coexist with dysglycaemia, obesity, and anaemia, cumulatively increasing the risk of preterm birth, abnormal fetal growth, caesarean delivery, severe postpartum haemorrhage, and severe maternal morbidity; these conditions are also associated with a higher risk of cardiovascular disease later in life [1, 2]. A large prospective study showed that even maternal glucose concentrations below conventional diagnostic thresholds for diabetes were continuously associated with several adverse perinatal outcomes [3]. A large population-based study similarly found that moderate-to-severe anaemia during pregnancy was associated with increased risks of placental abruption, preterm birth, severe postpartum haemorrhage, shock, and admission to intensive care [4]. Timely integration of multidimensional information—including blood pressure, blood glucose, haemoglobin, maternal and gestational age, parity, and previous complications—could support early risk identification and stratified management.

Current approaches to maternal risk assessment often rely on individual indicators, fixed thresholds, or disease-specific prediction models and might not adequately capture non-linear relations and interactions among cardiovascular, metabolic, and obstetric factors. Machine learning can identify complex patterns in multidimensional structured clinical data and has increasingly been investigated for the prediction of pre-eclampsia and other perinatal complications [5]. However, many studies are developed in a single institution or dataset, include insufficient external validation, and vary substantially in variable selection, outcome definition, data cleaning, and performance metrics [6]. Some studies primarily report accuracy or the area under the receiver operating characteristic curve, with limited attention to class imbalance, probability calibration, clinical net benefit, and model interpretability. High internal performance therefore does not guarantee safe transfer to new clinical populations.

Large language models (LLMs) offer an alternative approach to processing maternal risk information. LLMs can read case narratives, integrate dispersed clinical signals, and generate risk categories, confidence estimates, and natural-language rationales. Studies suggest that, with task adaptation and careful prompt design, LLMs can retrieve clinical knowledge and show promise in medical question-answering and reasoning [7]. Nevertheless, strong performance on medical benchmarks cannot be equated with reliable patient-level decision making. Prompt sensitivity, factual error, training-data bias, inconsistent outputs, and unclear accountability constrain clinical use [8]. An LLM explanation can also appear clinically plausible without faithfully representing the model's actual decision process [9, 10]. We therefore developed an exploratory framework for intersecting maternal cardiometabolic and obstetric risk.

## 2. Methods

### 2.1 Study design and data sources

This study was a retrospective secondary analysis of publicly available, deidentified data. The Maternal Health Risk dataset was used for model development, and the Preeclampsia in Pregnant Women dataset was used for independent external validation; the datasets were obtained from the UCI Machine Learning Repository and Kaggle, respectively. The analytical framework comprised data and label harmonisation, descriptive analyses, machine-learning modelling, external validation, sensitivity analyses, LLM-based case-text classification, and comparison of model explanations. No new participants were recruited, contacted, or exposed to any intervention.

## **2.2 Data cleaning and outcome definitions**

Before modelling, we reviewed variable names, data types, missing values, duplicate records, conflicting records, and outliers, and excluded observations with missing outcome labels. Original risk labels were harmonised as low, mid, or high risk while preserving their ordinal meaning. Two outcomes were specified a priori: a three-class outcome comprising low, mid, and high risk, and a binary high-risk versus non-high-risk outcome. For the binary outcome, high risk was coded as 1, whereas low and mid risk were combined and coded as 0.

## **2.3 Feature harmonisation**

A feature dictionary was used to align semantically equivalent variables across the two datasets. The common predictors available for external validation were maternal age, systolic blood pressure, diastolic blood pressure, haemoglobin, gestational age, and parity. Units were harmonised as years, mm Hg, g/dL, weeks, and counts, as appropriate, and binary variables were uniformly encoded as 0 or 1. Internal models used all available variables in the development dataset, whereas cross-dataset validation was prespecified to use only the six common predictors. Variables specific to the development dataset were retained for internal modelling and descriptive analyses; variables specific to the external dataset were used only descriptively.

## **2.4 Descriptive analyses and assessment of dataset shift**

Continuous variables were summarised as mean (SD) or median (IQR), as appropriate for their distributions, and visualised with violin plots; categorical variables were reported as counts and percentages. All analyses were stratified by low, mid, and high risk, and no causal inference was attempted. Differences between the development and external datasets were assessed using standardised distributions, absolute standardised mean differences (SMDs), principal component analysis, and uniform manifold approximation and projection (UMAP). SMDs of 0.10 and 0.20 were used as reference thresholds for modest and more marked distributional differences, respectively.

## **2.5 Machine-learning modelling and internal validation**

Six supervised classifiers were compared: HistGradientBoosting, RandomForest, ExtraTrees, GradientBoosting, a support vector machine with a radial basis function kernel (SVM-RBF), and logistic regression. Categorical variables were encoded consistently. Continuous variables were standardised for SVM-RBF and logistic regression, whereas unscaled numerical values were used for the tree-based

models. All preprocessing steps were fitted exclusively within the training fold of each cross-validation iteration to prevent information leakage.

The three-class task was evaluated using accuracy, balanced accuracy, macro-averaged F1 score, macro-averaged precision, and macro-averaged recall; class-specific errors were described with confusion matrices. The binary high-risk task was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). For the best-performing internal high-risk model, uncalibrated probabilities were compared with Platt sigmoid and isotonic calibration. Calibration was performed within a resampling framework to avoid reusing validation data. Decision-curve analysis was used to estimate net benefit across threshold probabilities from 0.00 to 0.80 relative to treat-all and treat-none strategies. Global model interpretation was assessed with permutation importance.

## **2.6 External validation and sensitivity analyses**

The selected algorithm was refitted to the full development dataset using only the six common predictors and applied directly to the external dataset without tuning, threshold optimisation, or recalibration on the external data. For the three-class task, the same performance metrics used in internal validation and a confusion matrix were reported; for the binary task, AUROC, AUPRC, and a confusion matrix were reported. Sensitivity analyses followed prespecified data-cleaning, sample-handling, and feature schemes. External performance was interpreted alongside the SMD, principal component analysis, and UMAP findings.

## **2.7 LLM-based classification of case narratives**

A deterministic template converted structured records into case narratives with a fixed sequence of demographic, cardiovascular, metabolic, obstetric, fetal, and amniotic-fluid variables. True labels and direct label identifiers were removed before submission to the models. When rule-based risk signals were included, they were generated solely from observed predictors and did not use the reference outcome. We compared DeepSeek (deepseek-chat) and Qwen (qwen-plus) under four prompting strategies: zero-shot, chain-of-thought, few-shot chain-of-thought, and guideline-enhanced prompting. All configurations received the same cases, prompt framework, and output schema and were instructed to return a standardised JSON object containing the predicted label, confidence, and a brief clinical rationale.

The LLM evaluation used 12 cases sampled by risk stratum, with four cases from each risk category. After schema validation, model outputs were compared with the hidden labels to calculate accuracy and macro-averaged F1 score, and a confusion matrix was generated for the best-performing configuration. Given the small sample, all LLM comparisons were descriptive and exploratory.

## **2.8 LLM attributions and comparison with machine-learning explanations**

A prespecified dictionary of keywords and synonyms was used to extract the following concepts from the rationale field of valid LLM responses: blood glucose or diabetes, blood pressure or hypertension, haemoglobin or anaemia, previous complications, proteinuria, BMI or obesity, obstetric history, maternal age, distance to hospital, and other clinical concepts. Each concept was counted no more than once per response. The mention rate was defined as the proportion of valid responses containing a given concept and was summarised by LLM and prompting strategy.

Semantically corresponding variables in the two explanation methods were mapped to common clinical concepts, retaining only concepts representable by both machine learning and LLMs. Permutation importance and LLM mention rates were divided by their respective maxima and normalised to a scale from 0 to 1. The resulting values were compared descriptively to identify concordant and divergent explanatory priorities; no causal or biological-mechanistic interpretation was intended.

## 3. Results

### 3.1 Dataset characteristics and harmonised outcomes

After data cleaning and label harmonisation, 1,014 observations were included in the Maternal Health Risk development dataset: 406 (40.0%) were low risk, 336 (33.1%) were mid risk, and 272 (26.8%) were high risk. The Preeclampsia in Pregnant Women external-validation dataset comprised 203 observations: 107 (52.7%) were low risk, 70 (34.5%) were mid risk, and 26 (12.8%) were high risk. A three-class outcome of low, mid, and high risk was defined in both datasets. Low-risk and mid-risk observations were then combined as non-high risk to create the binary high-risk versus non-high-risk outcome.

- After harmonising feature semantics and units of measurement, six comparable predictors were shared by the two datasets: maternal age, systolic blood pressure, diastolic blood pressure, haemoglobin, gestational age, and parity. These common variables were used for cross-dataset external validation, whereas the complete feature set in the development dataset was used for internal model development and explanatory analyses (Fig. 1; Supplementary Figure S1; Supplementary Table S1-S2).

### 3.2 Distribution of clinical characteristics across risk groups

- In the development dataset, maternal age, systolic blood pressure, diastolic blood pressure, blood glucose, gestational age, distance to hospital, and gestational weight gain generally increased across ordered risk strata from low to high, whereas haemoglobin generally decreased. The heart-rate distribution was modestly shifted towards higher values in the high-risk group, but body-temperature distributions overlapped substantially across the three groups. These were descriptive distributional comparisons; no conclusions about statistical significance were drawn (Supplementary Figure S2).

Categorical variables and obstetric history also differed across risk strata. Oedema and previous complications were more common in the higher-risk groups, and the distribution of parity varied by risk level. In the external dataset, the proportions of diabetes, a history of hypertension, and proteinuria differed visibly across risk groups. BMI, fetal weight, and amniotic-fluid level also showed between-group heterogeneity (Supplementary Figure S3).

### **3.3 Distributional shift between the development and external-validation datasets**

After standardisation of the six common variables, the distributions in the two datasets were not fully aligned. Absolute SMDs showed that the largest cross-dataset difference was in gestational age, followed by haemoglobin, diastolic blood pressure, and maternal age; the SMD for each of these variables exceeded 0.20. The SMD for parity was approximately 0.15, between 0.10 and 0.20, whereas that for systolic blood pressure was below 0.10.

Principal component analysis showed partial overlap between the datasets in low-dimensional space, but their overall locations and dispersion still differed. UMAP projections further showed local clustering patterns and separation associated with data source. Collectively, these assessments indicated substantial distributional shift between the development and external-validation datasets (Fig. 2; Supplementary Figure S4).

### **3.4 Internal validation of the machine-learning models**

In internal cross-validation of the six candidate algorithms, HistGradientBoosting had the best overall performance for three-class classification, with an accuracy of approximately 0.878 and a macro-averaged F1 score of 0.876. Its macro-averaged precision and recall were also among the highest of the candidate models. RandomForest, ExtraTrees, GradientBoosting, SVM-RBF, and logistic regression performed slightly less well overall than HistGradientBoosting (Fig. 3A-B; Supplementary Figure S5).

For binary classification of high risk versus non-high risk, the selected HistGradientBoosting model achieved an AUROC of 0.989 and an AUPRC of approximately 0.974, indicating strong internal discrimination of high-risk observations. Compared with uncalibrated predicted probabilities, the Platt sigmoid and isotonic calibration curves were closer to the ideal diagonal in parts of the probability range. Decision curves showed that the models before and after calibration retained a positive net benefit across most of the displayed threshold-probability range of 0.00-0.80 and generally outperformed the treat-all and treat-none strategies (Fig. 3C-E).

Permutation-importance analysis identified distance to hospital, blood glucose, gestational weight gain, systolic blood pressure, and haemoglobin as the five leading variables. Gestational age had lower importance (approximately 0.022), and the contributions of previous complications, body temperature, and oedema were smaller still. The permutation importance of age, parity, heart rate, and diastolic blood pressure was close to zero (Fig. 3F).

## 3.5 External validation and robustness analyses

When the model was restricted to the six common variables and applied directly to the external dataset, 79 of 203 observations were correctly classified in the three-class task, corresponding to an accuracy of 0.389 and a macro-averaged F1 score of approximately 0.27. Of the 107 true low-risk observations, 28 were correctly classified and 79 were classified as mid risk. Of the 70 true mid-risk observations, 51 were correctly classified, 18 were classified as low risk, and one was classified as high risk. None of the 26 true high-risk observations was correctly identified; seven were classified as low risk and 19 as mid risk.

In external binary validation, 174 of 177 non-high-risk observations were correctly identified and three were incorrectly classified as high risk. All 26 high-risk observations were classified as non-high risk, yielding a high-risk recall of 0. The external AUROC was 0.487, and the AUPRC was 0.172, close to the prevalence-based baseline of 0.128. Across different sensitivity-analysis specifications, AUROC ranged from 0.437 to 0.528. Overall, external-validation performance was substantially lower than internal cross-validation performance, and this finding was consistent across sensitivity analyses (Fig. 4; Supplementary Figure S6).

## 3.6 Risk-prediction performance of the large language models

In the exploratory LLM assessment of 12 stratified cases (Fig. 5A-D), DeepSeek achieved accuracies of 0.333–0.417 and macro-averaged F1 scores of 0.222–0.306 across the four prompting strategies. Corresponding values for Qwen were 0.417–0.583 and 0.306–0.467, respectively. Qwen with few-shot chain-of-thought prompting performed best, with an accuracy of 0.583 and a macro-averaged F1 score of 0.467 (Fig. 5E-F; Supplementary Figure S7).

The best configuration correctly classified seven of 12 cases: all four high-risk cases and three of the four mid-risk cases were correctly classified, whereas none of the four low-risk cases was correctly identified. This configuration did not generate any low-risk predictions. In a descriptive comparison, the accuracy and macro-averaged F1 score of the best LLM configuration fell between those of the internal full-feature machine-learning model (approximately 0.878 and 0.876) and the external common-feature machine-learning validation (0.389 and approximately 0.27). Because the feature inputs, evaluation cases, and sample sizes differed across these three settings, this comparison was intended only to show the range of observed performance and does not constitute a formal statistical comparison of model superiority (Fig. 5G-H).

## 3.7 Patterns in large language model explanations

The clinical concepts most frequently mentioned in LLM-generated explanations were blood glucose or diabetes (90.6%), blood pressure or hypertension (88.5%), and haemoglobin or anaemia (86.5%), followed by previous complications or medical history (51.0%) and proteinuria (36.5%). When stratified by model provider, Qwen generally had higher mention rates for these concepts than DeepSeek,

particularly for previous complications or medical history (60.4% vs 41.7%) and proteinuria (50.0% vs 22.9%). Blood glucose, blood pressure, and haemoglobin remained among the most frequently mentioned concepts across all prompting strategies (Fig. 6A-C).

After normalised comparison of LLM keyword frequencies with machine-learning permutation importance, blood glucose, blood pressure, and haemoglobin were prominent in both types of explanation. By contrast, distance to hospital and gestational weight gain ranked highly in the machine-learning model but were mentioned in only 16.7% and 9.4% of LLM explanations, respectively. Previous complications or medical history were frequently mentioned by the LLMs but ranked relatively low in permutation importance. Thus, the two explanation methods showed some concordance for core cardiometabolic indicators but differed in their emphasis on variables related to social access and obstetric history (Fig. 6D).

## 4. Discussion

Using two publicly available maternal-health datasets, we developed an analytical framework that incorporated label and feature harmonisation, machine-learning modelling, external validation, sensitivity analyses, and LLM-based text classification and explanation comparison. Structured clinical variables supported risk stratification within the source dataset, but model performance declined substantially in the external dataset. Although the LLMs showed some capacity to classify narrative cases and produce clinically readable explanations, they also showed class bias. Overall, transportability across data sources remains the principal barrier to clinical translation of these approaches.

Higher-risk groups had higher blood pressure and glucose concentrations and lower haemoglobin, consistent with the broader evidence on hypertensive disorders of pregnancy, maternal hyperglycaemia, and gestational anaemia [4, 11]. Permutation-importance analysis further identified distance to hospital, blood glucose, gestational weight gain, systolic blood pressure, and haemoglobin as influential variables in the internal model. The prominence of distance to hospital suggests that access to health-care services might contribute to maternal risk stratification [12]. However, the outcomes in this study were pre-existing risk labels from public datasets rather than standardised hard clinical endpoints, and permutation importance measures the model's reliance on a variable rather than its biological effect or a causal relation. These findings therefore support the clinical plausibility of the model but require confirmation against real-world clinical outcomes.

HistGradientBoosting performed well in internal validation, with a three-class macro-averaged F1 score of approximately 0.876 and binary high-risk AUROC and AUPRC values of approximately 0.989 and 0.974, respectively. These findings accord with the growing use of machine learning to predict perinatal complications [13]. We also assessed probability calibration and decision curves because discrimination alone is insufficient to establish clinical utility: the accuracy of predicted probabilities and net benefit at clinically relevant thresholds are also important [14]. Although the internal findings suggest potential for

risk stratification, cross-validation within the same source population can overestimate performance; these results should therefore be regarded as development-stage evidence.

In the external dataset, the three-class macro-averaged F1 score fell to approximately 0.27, the binary high-risk AUROC was 0.487, and none of the 26 high-risk observations was identified. Marked cross-dataset shifts were observed in gestational age, haemoglobin, diastolic blood pressure, and maternal age. The datasets might also have differed in variable ranges, risk prevalence, data-collection settings, and labelling rules. Moreover, external validation could use only six common variables, excluding some features that were important in the development dataset. External AUROCs remained between 0.437 and 0.528 across sensitivity analyses, suggesting that the loss of performance was more likely attributable to systematic dataset shift than to any single data-processing choice [14]. The present model therefore lacks an evidential basis for direct use in maternal screening or clinical decision making.

In the LLM analyses, Qwen with few-shot chain-of-thought prompting performed comparatively well, with an accuracy of 0.583 and a macro-averaged F1 score of 0.467, suggesting that examples and structured reasoning prompts might help a model recognise relevant clinical signals in case narratives [15]. However, although this configuration correctly identified all four high-risk cases, it failed to identify any low-risk case, indicating substantial class bias. The evaluation included only 12 cases, and the estimates are therefore highly uncertain. Medical LLMs also remain susceptible to prompt sensitivity, outdated knowledge, inconsistent outputs, and factual errors [16]. At present, LLMs are better positioned as aids for organising case information, summarising risk factors, and supporting clinical communication than as replacements for validated prediction models or clinical judgement.

Frequent LLM references to blood glucose, blood pressure, and haemoglobin overlapped with several important machine-learning features, but the two explanation methods did not place identical emphasis on all variables. Distance to hospital and gestational weight gain, for example, were important to the machine-learning model but seldom appeared in LLM rationales, whereas previous complications were more frequently mentioned by the LLMs. These differences might reflect distinct patterns of information use, but clinical readability does not establish that an explanation faithfully represents the model's decision process [17, 18]. The keyword-concordance analysis should therefore be interpreted as descriptive. Future studies should combine repeated generation, counterfactual testing, measures of explanation faithfulness, and expert clinical assessment to evaluate the stability and clinical value of LLM explanations.

## 5. Limitations

This study has several limitations. First, both public datasets were small, and the external dataset included only 26 high-risk observations. Information on standardised source-data collection, health-care institutions, geographical regions, race and ethnicity, socioeconomic circumstances, and comprehensive obstetric history was unavailable. Second, the risk labels were not standardised hard clinical endpoints; label cleaning and harmonisation into three classes might therefore have obscured differences in the

original outcome definitions. Third, external validation was restricted to six shared variables and could not incorporate all important features from the development dataset. Potential differences in measurement timing and methods might also have introduced informational inconsistency. Fourth, confidence intervals were not available for all performance metrics, and point estimates from the limited samples might be unstable. Future studies should use observed adverse pregnancy outcomes, specify sample sizes a priori, report confidence intervals and calibration measures, undertake multicentre external validation and model updating, and report the development and early clinical evaluation processes in accordance with TRIPOD + AI and DECIDE-AI.

## 6. Conclusion

Structured maternal clinical variables showed potential for risk stratification within the source dataset, whereas external validation revealed substantial dataset shift and limited generalisability. LLMs identified some clinical risk concepts from case narratives and generated readable rationales, but their predictive stability and explanatory faithfulness were insufficient to support autonomous decision making. The principal value of this framework lies not in producing a single high-scoring model, but in placing internal performance, external transportability, and explanatory readability within the same evaluation framework. These tools should be considered for early maternal risk screening and stratified management only after rigorous validation, calibration, and clinical-impact assessment in larger, multicentre, prospective datasets.

## Declarations

## Competing Interests

None declared.

## Funding

None

## Author Contribution

Conceptualization, Zhenxing Deng and Shaoxun Wen; Study design, Zhenxing Deng and Shaoxun Wen; Manuscript draft, Shaoxun Wen; Draft review and editing, All authors.

## Acknowledgments

Not applicable.

# Data Availability

The datasets analysed in this study are publicly available.1. Maternal Health Risk Dataset: available from Kaggle (<https://www.kaggle.com/datasets/csafr12/maternal-health-risk-data>)2. Preeclampsia in Pregnant Women Dataset: available from Kaggle (<https://www.kaggle.com/datasets/brightbuff/preeclampsia-detection>)All data preprocessing, feature harmonisation, model training and evaluation code generated for this study are available upon reasonable request to the corresponding author. No new raw data were generated in this work.

## References

1. Parikh, N. I. et al. Adverse pregnancy outcomes and cardiovascular disease risk: unique opportunities for cardiovascular disease prevention in women: a scientific statement from the American Heart Association. *Circulation* **143** (18), e902–e916 (2021).
2. Garovic, V. D. et al. Hypertension in pregnancy: diagnosis, blood pressure goals, and pharmacotherapy: a scientific statement from the American Heart Association. *Hypertension* **79** (2), e21–e41 (2022).
3. Metzger, B. E. et al. Hyperglycemia and adverse pregnancy outcomes. *N Engl. J. Med.* **358** (19), 1991–2002 (2008).
4. Shi, H. et al. Severity of anemia during pregnancy and adverse maternal and fetal outcomes. *JAMA Netw. Open.* **5** (2), e2147046 (2022).
5. Ranjbar, A. et al. Machine learning models for predicting preeclampsia: a systematic review. *BMC Pregnancy Childbirth.* **24** (1), 6 (2024).
6. Tiruneh, S. A. et al. Externally validated prediction models for pre-eclampsia: systematic review and meta-analysis. *Ultrasound Obstet. Gynecol.* **63** (5), 592–604 (2024).
7. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620** (7972), 172–180 (2023).
8. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med. (Lond)*. **3** (1), 141 (2023).
9. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health.* **3** (11), e745–e750 (2021).
10. Amann, J. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inf. Decis. Mak.* **20** (1), 310 (2020).
11. Parikh, N. I. et al. Adverse pregnancy outcomes and cardiovascular disease risk: unique opportunities for cardiovascular disease prevention in women: a scientific statement from the American Heart Association. *Circulation* **143** (18), e902–e916 (2021).
12. Crear-Perry, J. et al. Social and structural determinants of health inequities in maternal health. *J. Womens Health (Larchmt)*. **30** (2), 230–235 (2021).

13. Sufriyana, H., Wu, Y. W. & Su, E. C. Artificial intelligence-assisted prediction of preeclampsia: development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. *EBioMedicine* **54**, 102710 (2020).
14. Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17** (1), 230 (2019).
15. Liévin, V. et al. Can large language models reason about medical questions? *Patterns* (N Y). ;5(3):100943. (2024).
16. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl. J. Med.* **388** (13), 1233–1239 (2023).
17. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616** (7956), 259–265 (2023).
18. Vasey, B. et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022).

Figures

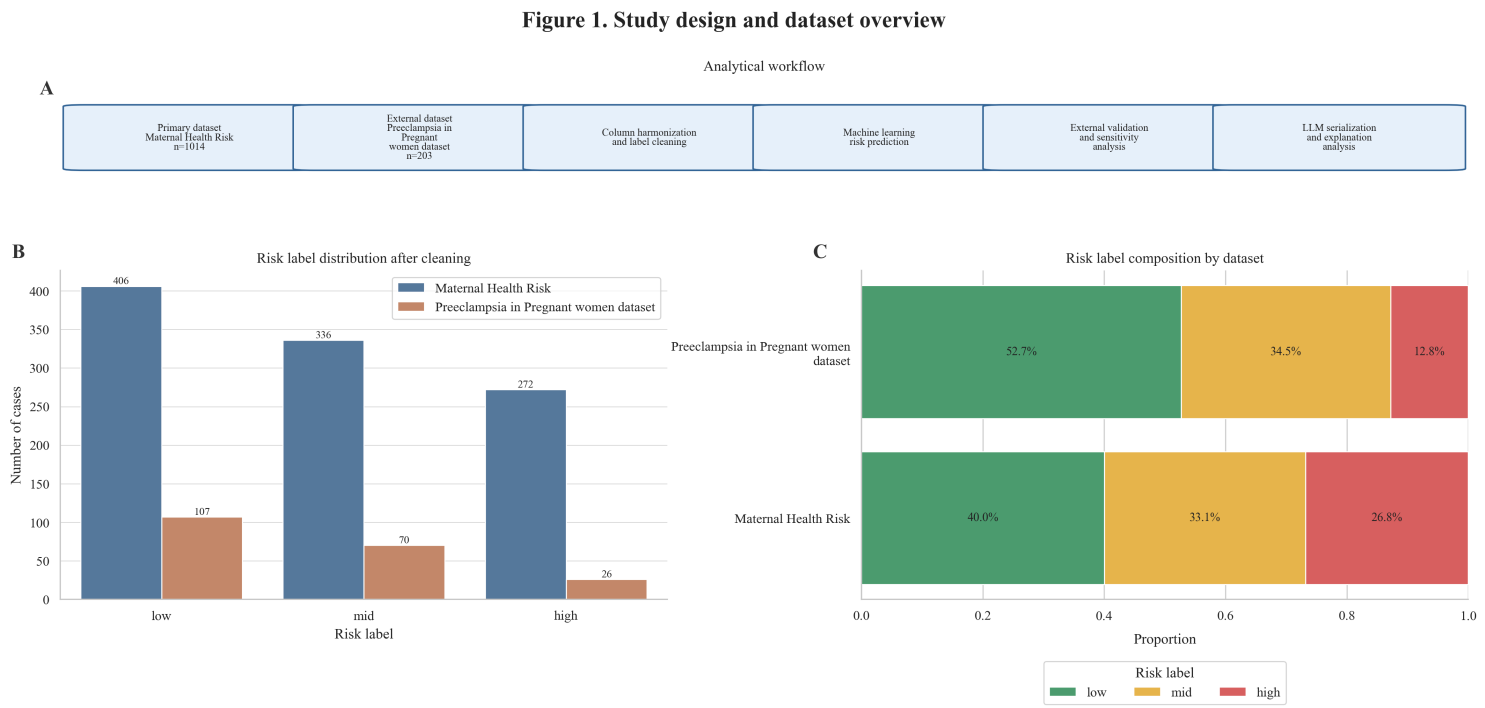


Figure 1

Study design and dataset overview.

(A) Overall analytical workflow integrating the Maternal Health Risk dataset and the Preeclampsia in Pregnant women dataset, including column harmonization, label cleaning, machine learning-based risk prediction, external validation, sensitivity analysis, LLM-based text serialization, and explanation

analysis. (B) Distribution of cleaned low-, mid-, and high-risk labels in the two datasets. (C) Proportional composition of risk labels by dataset.

Figure 2. Feature harmonization and dataset shift assessment

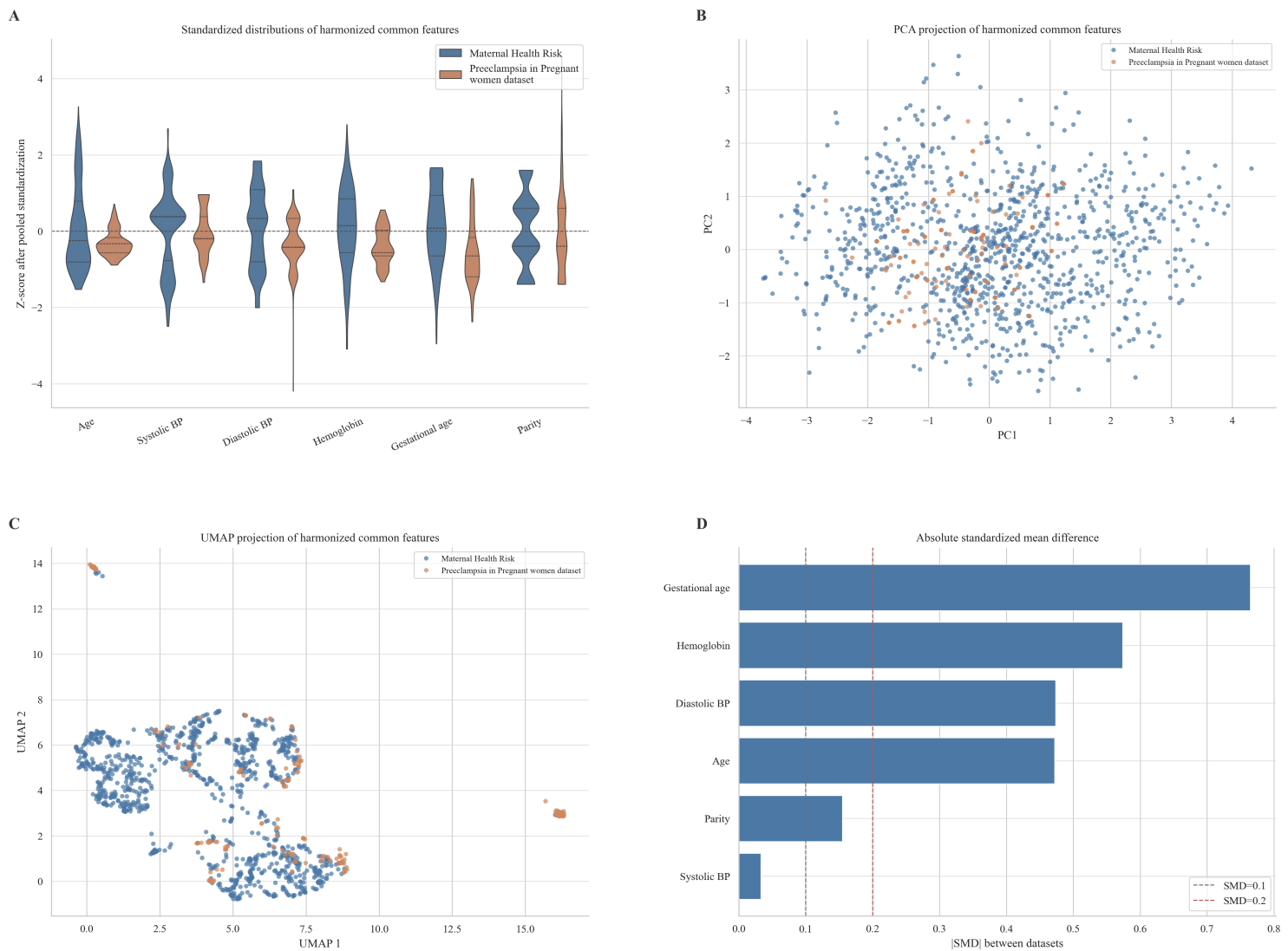


Figure 2

Feature harmonization and dataset shift assessment.

(A) Standardized distributions of harmonized common features between the Maternal Health Risk dataset and the Preeclampsia in Pregnant women dataset. (B) Principal component analysis projection of harmonized common features. (C) UMAP projection of harmonized common features. (D) Absolute standardized mean differences between datasets for common features. Dashed reference lines indicate SMD thresholds of 0.1 and 0.2.

**Figure 3. Internal machine learning validation, calibration and model explanation**

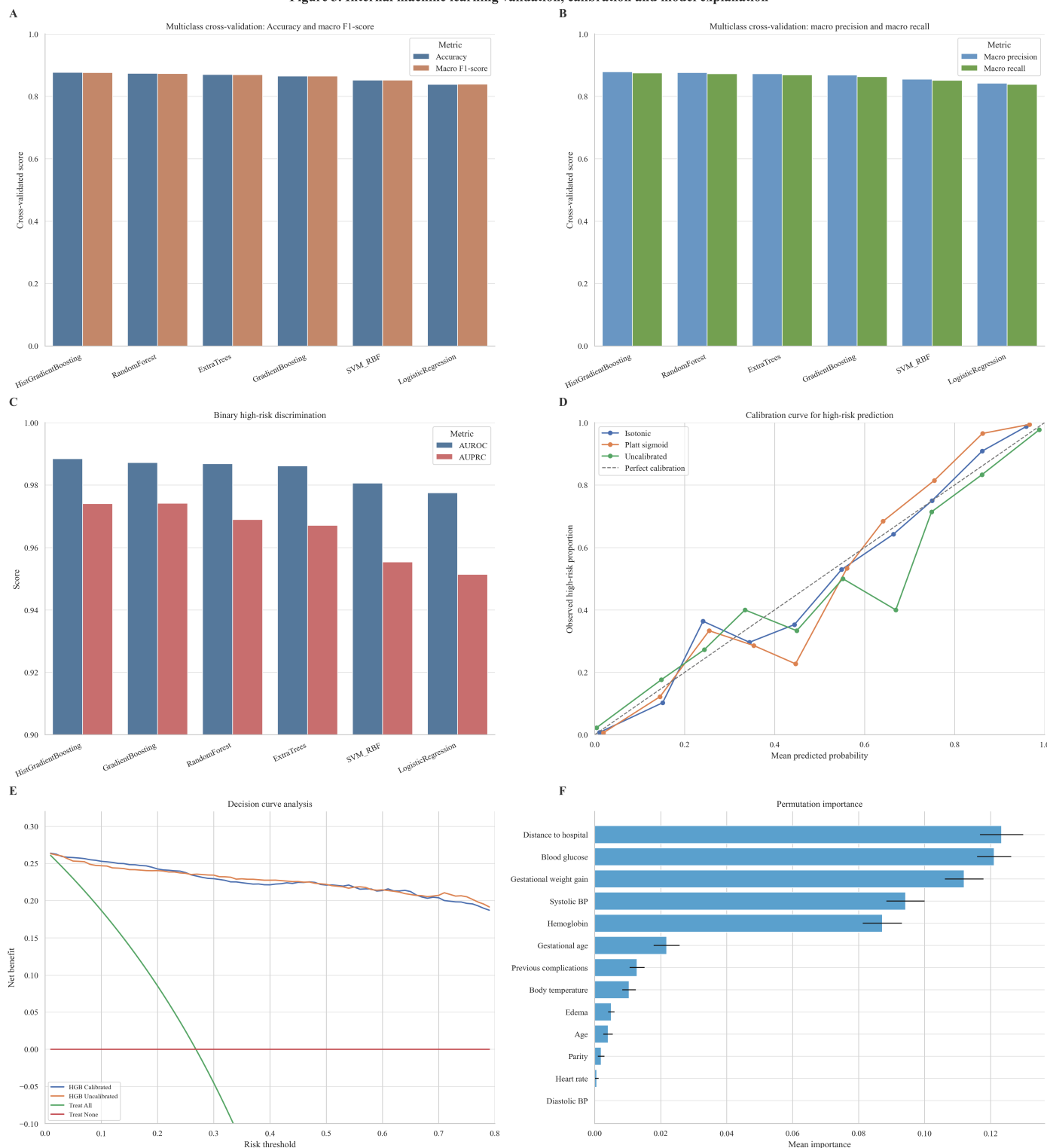


**Figure 3**

### Internal machine learning validation, calibration, and model explanation.

(A) Cross-validated accuracy and macro F1-score for multiclass risk prediction models. (B) Cross-validated macro precision and macro recall for multiclass prediction. (C) AUROC and AUPRC for binary

high-risk discrimination. (D) Calibration curves for high-risk prediction. (E) Decision curve analysis across risk thresholds. (F) Permutation importance of clinical features in the selected model.

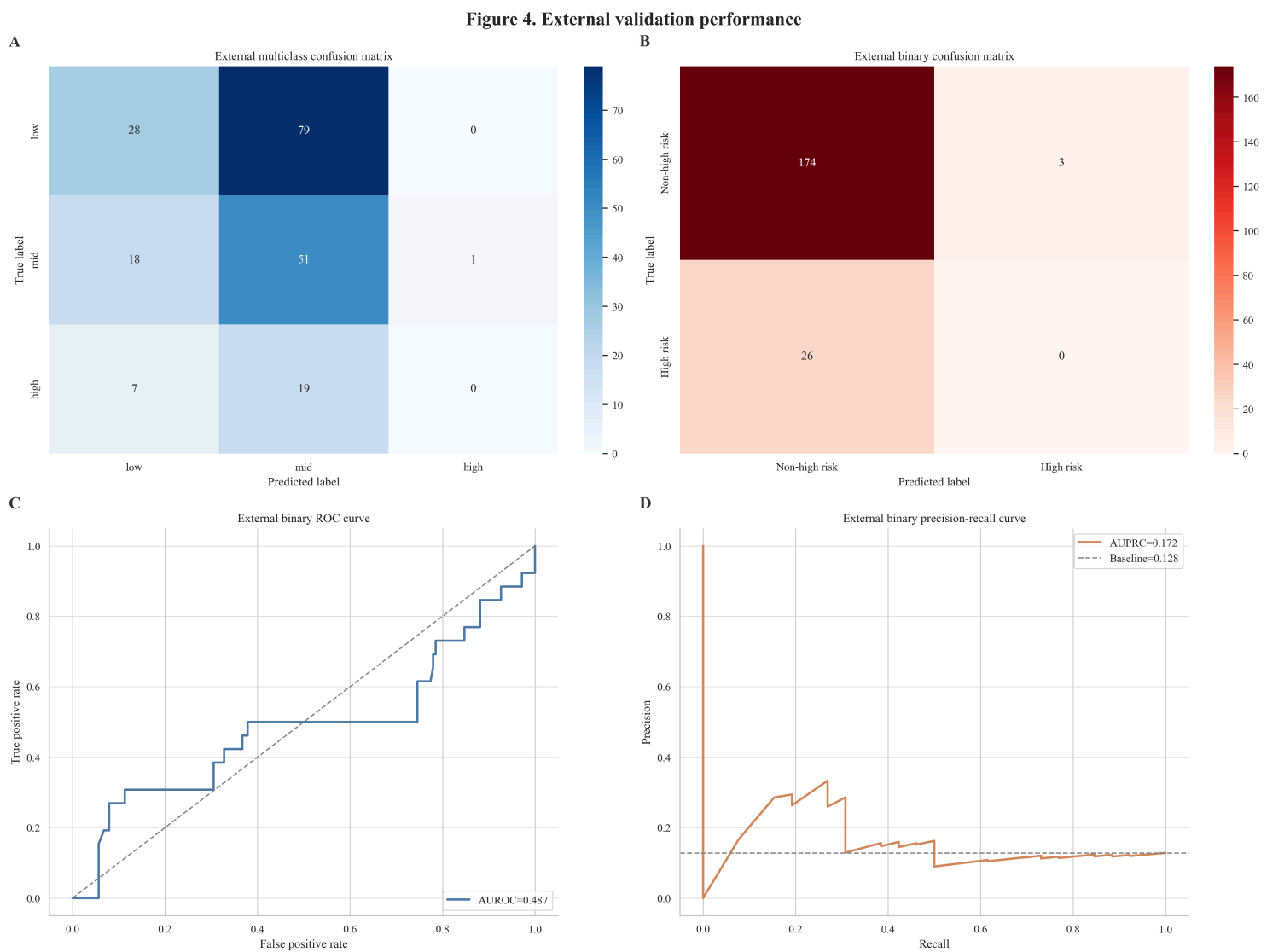


**Figure 4**

**External validation performance.**

(A) Multiclass confusion matrix on the external Preeclampsia in Pregnant women dataset. (B) Binary confusion matrix for high-risk versus non-high-risk classification. (C) Receiver operating characteristic curve for external binary high-risk prediction. (D) Precision-recall curve for external binary high-risk prediction.

Figure 5. LLM text serialization, prompt strategies and predictive performance

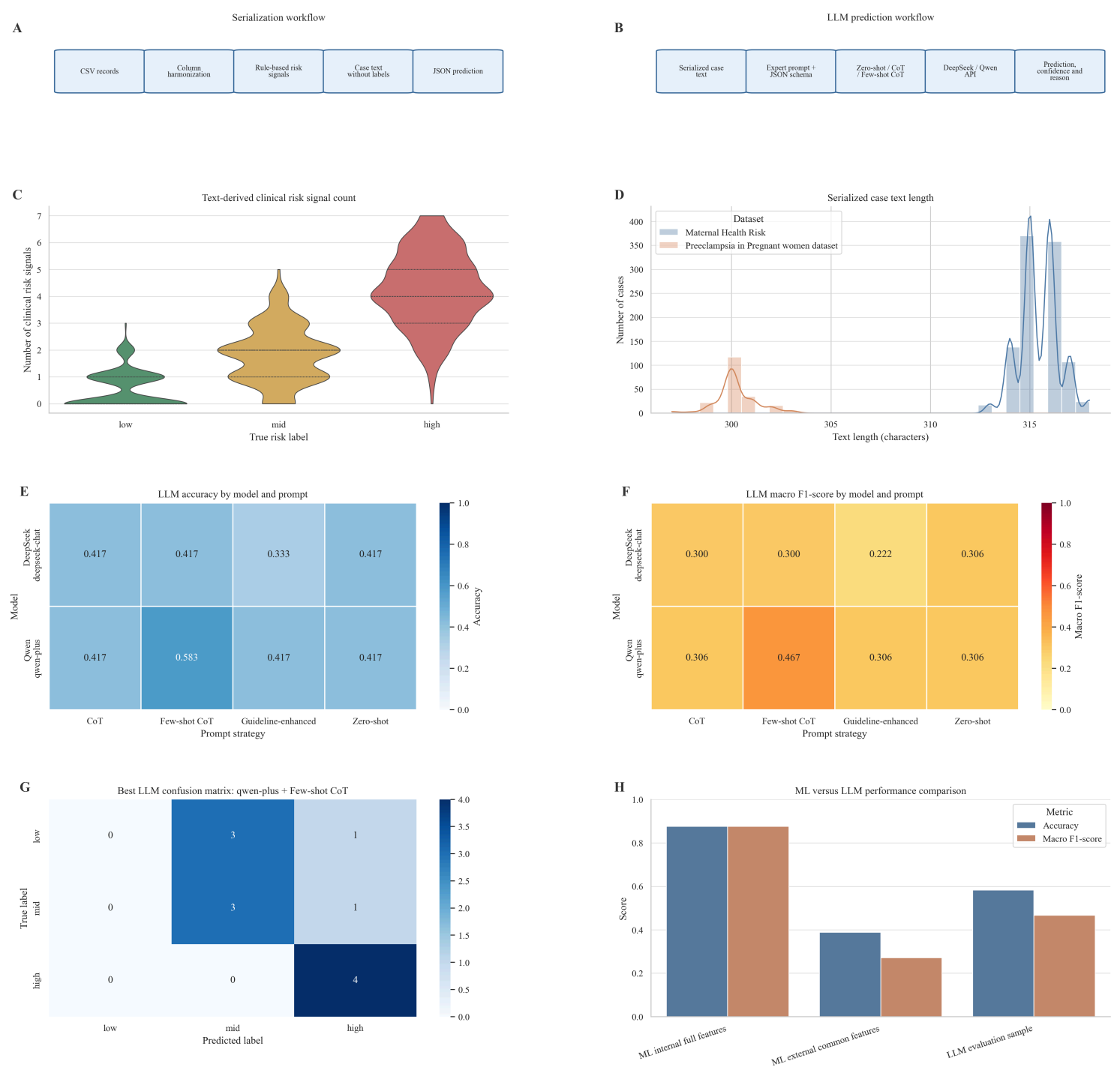


Figure 5

LLM text serialization, prompt strategies, and predictive performance.

(A) Workflow for converting structured clinical records into serialized case text without revealing true labels. (B) LLM prediction workflow using expert prompts, JSON output schema, and different prompt strategies. (C) Distribution of text-derived clinical risk signal counts across risk labels. (D) Serialized case text length by dataset. (E) LLM accuracy across model and prompt combinations. (F) LLM macro

F1-score across model and prompt combinations. (G) Confusion matrix for the best-performing LLM configuration. (H) Performance comparison between machine learning models and LLM-based prediction.

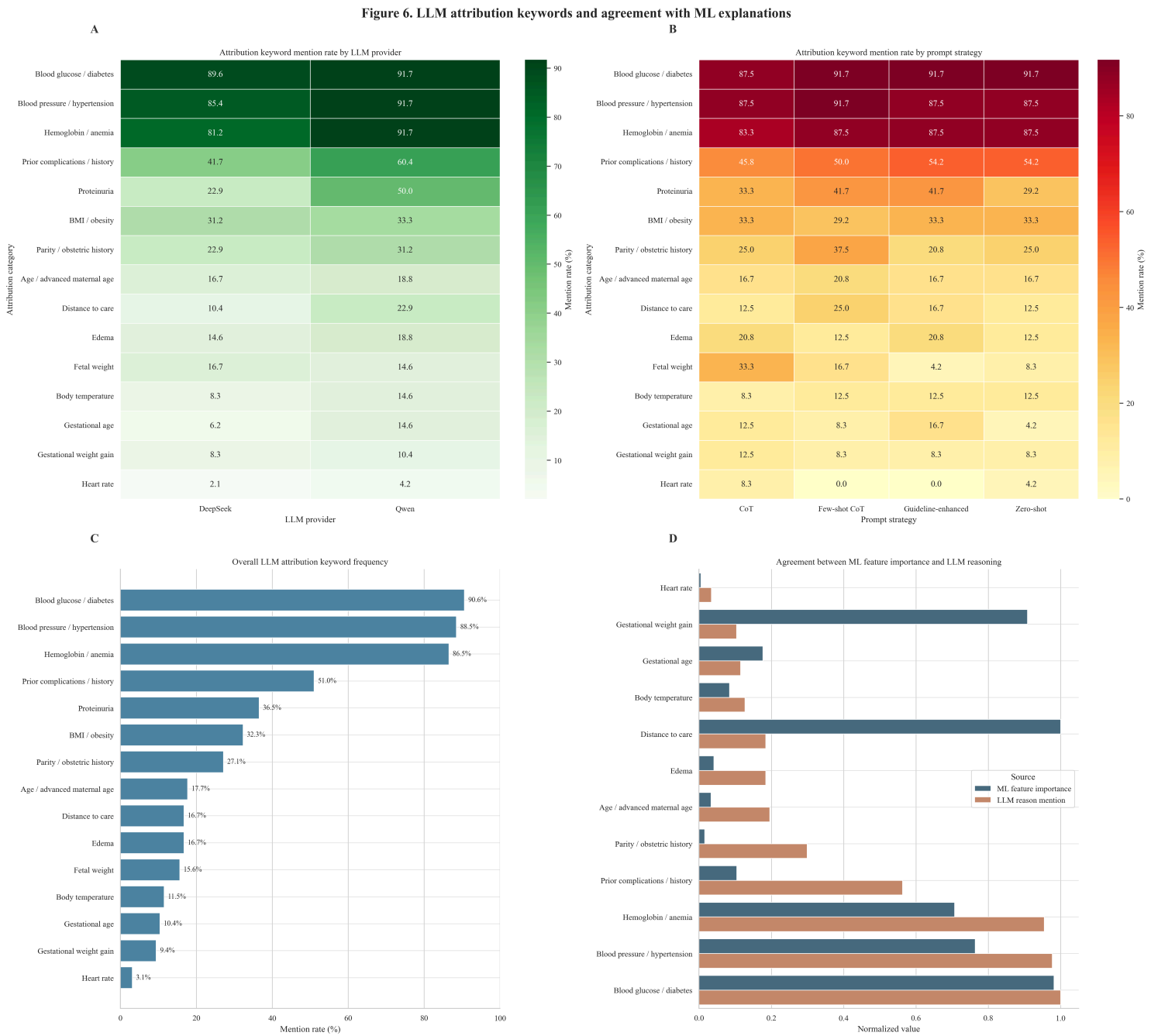


Figure 6

LLM attribution keywords and agreement with machine learning explanations.

(A) Attribution keyword mention rates by LLM provider. (B) Attribution keyword mention rates by prompt strategy. (C) Overall frequency of clinical attribution keywords in LLM-generated explanations. (D) Agreement between normalized machine learning feature importance and LLM reason keyword frequency.

## Supplementary Files

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

- [Supplementary.Figure.S1.png](#)
- [Supplementary.Figure.S2.png](#)
- [Supplementary.Figure.S3.png](#)
- [Supplementary.Figure.S4.png](#)
- [Supplementary.Figure.S5.png](#)
- [Supplementary.Figure.S6.png](#)
- [Supplementary.Figure.S7.png](#)
- [SupplementaryTable.docx](#)
- [MaternalHealthRisk.xlsx](#)
- [PreeclampsiaInPregnantwomendataset.xlsx](#)