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.

Clinically grounded multi-agent artificial intelligence for preventive health management

Hao Lin

hlin@uestc.edu.cn

University of Electronic Science and Technology of China https://orcid.org/0000-0001-6265-2862

Yang Zhang

Chengdu University of Traditional Chinese Medicine

Dongxin Ye

University of Electronic Science and Technology of China

Sicheng He

University of Electronic Science and Technology of China

Zhaowu Du

Chengdu University of Traditional Chinese Medicine

Yang Yu

University of Electronic Science and Technology of China

Xiao Yu

University of Electronic Science and Technology of China

Liping Ren

Chengdu University of Traditional Chinese Medicine

Nanqing Dong

Shanghai Innovation Institute

Fang Hu

Shanghai Innovation Institute

Jinsong Su

Chengdu University of Traditional Chinese Medicine

Jie Zhang

Safety Environment Quality Surveillance & Inspection Research Institute, CNPC Chuanqing Drilling Engineering Company Ltd

Yun Tan

Safety Environment Quality Surveillance & Inspection Research Institute, CNPC Chuanqing Drilling Engineering Company Ltd

Li Zhao

Safety Environment Quality Surveillance & Inspection Research Institute, CNPC Chuanqing Drilling Engineering Company Ltd

Biological Sciences - Article

Keywords:

Posted Date: April 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9150301/v1

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

Additional Declarations: There is **NO** Competing Interest.

Clinically grounded multi-agent artificial intelligence for preventive health management

Dongxin Ye[1,2], Sicheng He[1], Zhaowu Du[3], Yu Yang[1], Yu Xiao[1], Liping Ren[4], Nanqing Dong[2,5], Fang Hu[2], Jinsong Su[3], Jie Zhang[6], Yun Tan[6], Li Zhao[6], Yang Zhang[3,*], Hao Lin[1,*]

[1] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, 610054, China

[2] Shanghai Innovation Institute, Shanghai 200231, China

[3] State Key Laboratory of Southwestern Chinese Medicine Resources, Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, 611137, China

[4] School of Healthcare Technology, Chengdu Neusoft University, Chengdu, 611844, China

[5] Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China

[6] Safety Environment Quality Surveillance & Inspection Research Institute, CNPC Chuanqing Drilling Engineering Company Ltd., Sichuan, 618300, China.

Abstract

Routine health examinations generate dense, heterogeneous data, yet their preventive value depends on consistent interpretation, calibrated risk stratification, and actionable follow-up. In practice, these tasks are distributed across clinicians and time, leading to variability in the detection of subtle abnormalities and in decisions about when and how to intervene. Such variability reflects the difficulty of maintaining consistent, high-quality preventive decision-making at scale. Here we present G-Health, a clinically grounded multi-agent artificial intelligence framework that translates examination reports into structured preventive action. The system combines three-stage clinical alignment of large language models with specialist quantitative risk models and guideline-informed retrieval to stabilize reasoning under uncertainty. Trained on large-scale medical dialogue data and further specialized on multi-center real-world examination reports, the framework integrates 20 quantitative risk models that provide calibrated multi-disease estimates with feature-level interpretability. Across 13 medical and general benchmarks, the aligned models achieve the best overall average rank among strong baselines. In a fully blinded evaluation involving 79 medically trained assessors, G-Health reports were consistently preferred over outputs from three other large language models and 12 senior practicing physicians across five clinical dimensions. Together, these findings establish a deployable paradigm that transforms routine examinations into structured and scalable preventive decision-making.

Main

Routine health examinations are central to preventive medicine, especially for non-communicable diseases (NCDs) that account for most global morbidity and mortality[1]. Population-level screening can improve detection, risk-factor control, and uptake of preventive services[2,3], but its clinical benefit depends on whether abnormal findings are clearly explained and acted upon. Many patients struggle with densely formatted reports, and clinicians must interpret multiple results, weigh comorbidities, and provide individualized counselling in short consultations. Workforce shortages, increasing case complexity, and widespread burnout further constrain this process across health systems[4-6]. As a result, much of the information contained in health examination reports is not fully used in routine care, not because individual physicians lack ability, but because preventive medicine demands sustained cognitive effort that is hard to deliver consistently at scale.

Recent advances in artificial intelligence (AI), including medical large language models (LLMs), show that these models can encode clinical knowledge and answer medical queries[7-9]. Systematic evaluations, however, continue to report hallucinated or non-guideline-concordant recommendations, weak calibration and bias, and limited prospective validation in real clinical workflows[10-12]. A more basic limitation is that most systems are built for isolated question answering or diagnosis, rather than for supporting the longitudinal, multi-step decisions typical of preventive care. They also

seldom combine structured health examination data with patient dialogue, and few offer calibrated, multi-disease risk assessment within workflows that clinicians can audit[13-15].

To address this gap, we design and evaluate an end-to-end, clinically grounded multi-agent framework for health examination analysis and preventive risk stratification (**Fig. 1**). The backbone LLMs, based on the Qwen3 family, are adapted using a three-stage clinical alignment strategy[16]. We first apply supervised fine-tuning (SFT) on more than 2.8 million medical question--answer pairs to build a stable medical knowledge base and more consistent clinical reasoning. We then perform Direct Preference Optimization (DPO) on 1.7 million preference-labeled instances to encourage safer[17], more conservative behavior, including acknowledging uncertainty, avoiding unsupported claims, and following clinical guidelines. Finally, we conduct task-specific adaptation using real-world health examination reports and report-grounded instruction data from 3,503 reports spanning 123 hospitals, which improves the model's ability to handle heterogeneous report formats and tailor explanations to diverse patient contexts[16]. In parallel, we develop 20 disease-specific risk assessment models trained on structured health examination datasets with longitudinal outcomes. These models map heterogeneous examination indicators to calibrated, multi-disease quantitative risk estimates with explicit decision thresholds, enabling standardized and auditable risk stratification[18,19].

Building on these components, we construct G-Health, a multi-agent framework that supports multimodal inputs, including structured indicators, dialogue, speech, and imaging. A conversation agent first infers user intent and the relevant clinical context, then routes each request either to open-domain dialogue or to task-specific pipelines. For disease risk assessment, the system calls the appropriate specialist risk models to produce calibrated estimates, along with interpretation and actionable recommendations. For health examination report analysis, it extracts and standardizes key findings, grades abnormalities, and generates structured follow-up and referral guidance. Across workflows, intermediate outputs from language reasoning and tool models are consolidated into a single clinical response.

To ground recommendations in evidence, the system integrates a retrieval corpus of clinical guidelines, textbooks, and consensus statements[20]. With retrieval-augmented generation (RAG), the model cites this corpus when generating recommendations, reducing reliance on model memory alone. For each quantitative prediction, SHAP-based feature attributions provide both local and global explanations, supporting clinician oversight and interpretability[21-23].

To evaluate whether the framework improves clinical decision quality, we used a multi-layered evaluation protocol that combines benchmark testing with blinded human assessment. Across 13 general and medical benchmarks, the aligned G-Health models achieved the best average rank among 13 strong medical and general-purpose baselines, and reached at least 0.80 accuracy on eight datasets. We then conducted a fully blinded

human evaluation in which 79 medically trained raters compared structured reports produced by four LLMs, including G-Health, with reports written by 12 senior practicing physicians. In total, we collected 6,955 evaluation records across five clinical dimensions: medical correctness, comprehensiveness, risk identification, department recommendation, and expression clarity. Across all five dimensions, G-Health-32B-Agent obtained higher aggregate scores than the other LLMs and the physician cohort. The gains were largest for risk identification and comprehensiveness, consistent with stronger detection of subtle abnormalities and more systematic follow-up planning. Overall, these results suggest that the multi-agent, tool-grounded framework yields clinically meaningful improvements and, under blinded evaluation, outperforms both strong general-purpose models and experienced physicians for structured health examination report analysis in this study setting.

Finally, we embedded the G-Health workflow and explanation layer into an interactive interface, establishing a preventive-care paradigm that translates routine examination data into actionable follow-up. Heterogeneous indicators were standardized into a unified clinical representation, routed to calibrated multi-disease predictors, and interpreted under retrieved evidence with transparent feature-level attributions. By making intermediate reasoning steps explicit and transferable across settings, the framework moves health examinations beyond static numerical summaries toward a more auditable and continuously improvable preventive decision process.

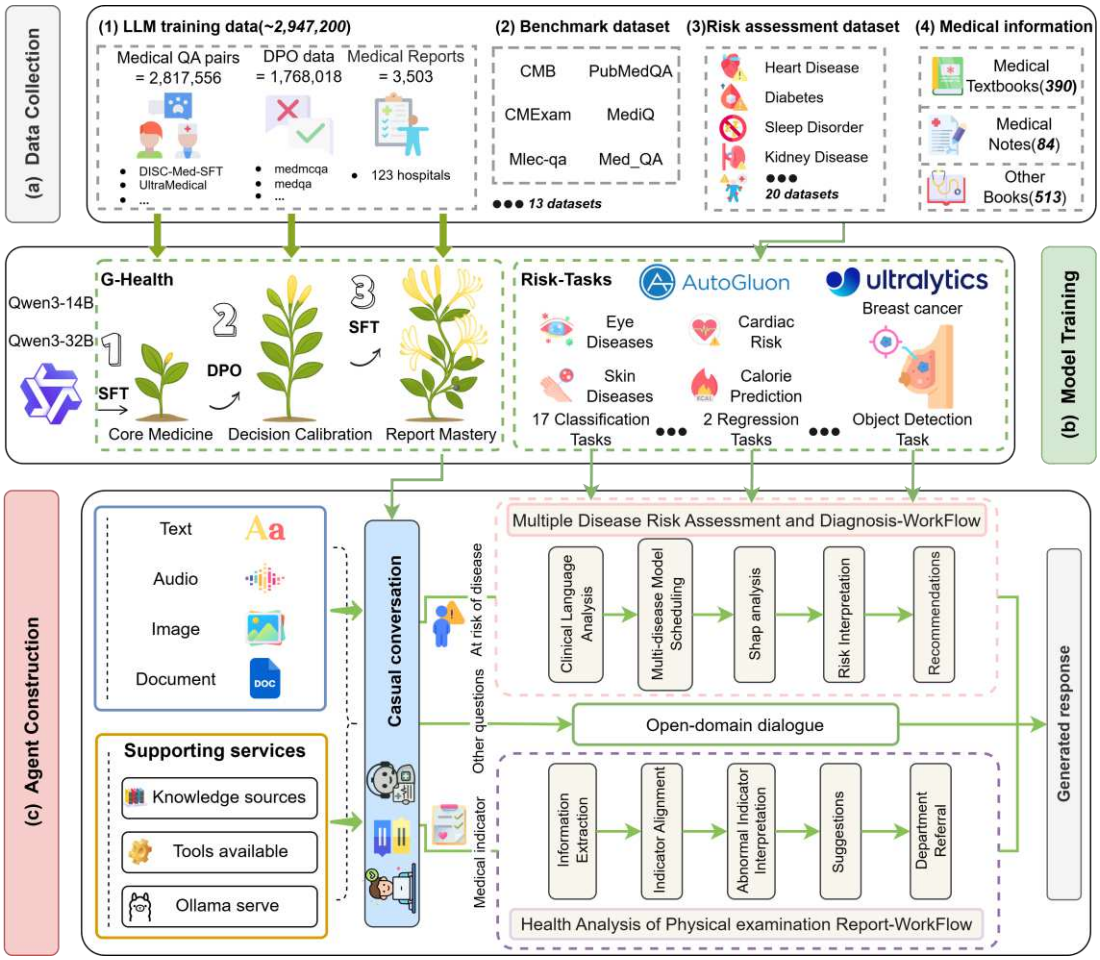


Fig. 1 | Overall architecture of the proposed system. The framework comprises four main components. **(a) Data collection and corpus construction.** We collect 2,817,556 medical dialogue pairs for first-stage supervised fine-tuning, 1,643,350 and 124,668 dialogue pairs as the training and validation sets for direct preference optimization, and 3,503 health examination dialogues generated from real-world health examination reports through a structured dialogue synthesis pipeline. In addition, we collect 13 benchmark datasets to evaluate our medical LLMs, 20 datasets for training disease-specific risk assessment models, and 987 documents for the RAG knowledge base. **(b) Model training.** We perform three-stage fine-tuning of Qwen3-14B and Qwen3-32B, and train a panel of risk assessment models using AutoGluon for tabular and image classification and regression tasks, and YOLO-based detectors for medical image object detection. **(c) Agent Construction.** We design a multi-agent architecture that covers the full pipeline from input to output, including three major workflows: risk assessment, health examination report analysis and open-domain question answering, each of which can invoke external tools and the RAG knowledge base when needed.

Results

Improved medical reasoning performance through multi-stage clinical alignment

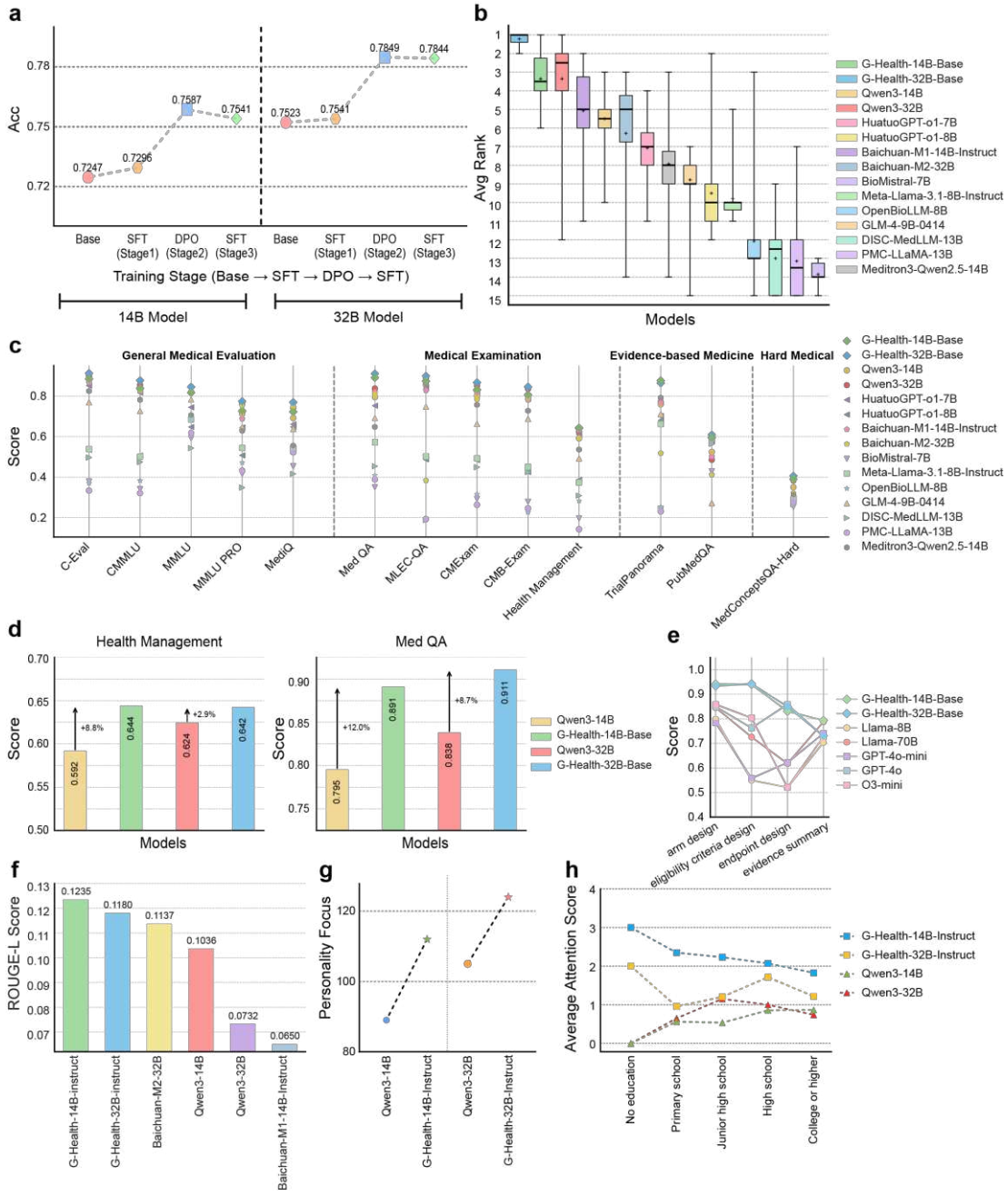


Fig. 2 | Overall evaluation results of the proposed models. a, Mean accuracy across multiple benchmark datasets after three-stage fine-tuning for the 14B and 32B models. **b**, Average rank of each model over all evaluation tasks (lower is better). **c**, Accuracy

of all models on each dataset, ordered by the performance of G-Health-32B-Base from high to low. **d**, Relative performance gains (%) brought by fine-tuning on the Health Management and Med QA tasks. **e**, Comparison of different models on the TrialPanorama dataset. **f**, ROUGE-L scores on 100 held-out health examination reports after the third-stage fine-tuning. **g**, Change in the model's attention to personality-related information after the third-stage fine-tuning. **h**, Average attention scores assigned to patients with different education levels after the third-stage fine-tuning.

We quantify how staged clinical alignment and model scale affect performance across 13 general and medical benchmarks. We compare the aligned G-Health models with strong medical and general-domain baselines to situate their performance across diverse tasks. We also examine report-oriented fine-tuning for health examination report generation, asking whether specialization improves structural fidelity and patient-aware personalization while preserving general medical competence.

Progressive gains from staged clinical alignment and robustness after report-specific specialization. To validate the backbone models that serve as the cognitive core of our multi-agent system, we evaluated medical reasoning and general-domain capability on 13 general and medical benchmarks, including MedQA, PubMedQA, and TrialPanorama (**Fig. 2a; Supplementary Table 1-4**) [24-35]. Across both model scales, staged alignment led to consistent improvements, with the largest gain coming from preference optimization. For the 14B model, accuracy increased from 0.7247 to 0.7296 after supervised alignment and to 0.7587 after preference optimization; the 32B model rose from 0.7523 to 0.7541 and then to 0.7849. This 3 percentage-point boost from preference optimization narrowed the scale gap and allowed the aligned 14B model to slightly outperform the unaligned 32B original model. We then performed report-

specific tuning to strengthen health-examination report interpretation and generation. This specialization largely preserved broad medical competence: overall benchmark accuracy decreased slightly, from 0.7587 to 0.7541 for 14B (0.46 percentage points) and from 0.7849 to 0.7844 for 32B (0.05 percentage points). Together, these results suggest that the alignment pipeline improves general medical reasoning beyond what is explained by parameter scale alone, and that report-specific adaptation can be introduced with minimal forgetting, providing a stable backbone for downstream multi-agent health management workflows. Throughout this paper, we refer to the model after stage-2 preference optimization as the Base model, and the model after stage-3 report-specific supervised fine-tuning as the Instruct model.

Consistent superior performance across diverse medical benchmarks. We compared our models with 13 strong medical and general-domain baselines on the same benchmark suite (**Fig. 2b, c**). The comparison set includes G-Health-14/32B-Base (our two-stage aligned variants), the original Qwen3-14B/32B backbones, HuatuoGPT-o1-7B/8B[36], Baichuan-M1-14B-Instruct and Baichuan-M2-32B[37,38], BioMistral-7B[39], Meta-Llama-3.1-8B-Instruct[40], OpenBioLLM-8B[41], GLM-4-9B[42], DISC-MedLLM-13B[43], PMC-LLaMA-13B[44] and Meditron3-Qwen2.5-14B[45]. Detailed results for each dataset are reported in **Supplementary Table 1-4**. For each dataset, we ranked models by accuracy and summarized rank distributions across tasks, where lower values indicate better performance. G-Health-14B-Base and G-Health-32B-Base achieve the best or near-best ranks on almost all benchmarks. G-Health-32B-Base has the best

average rank overall, while G-Health-14B-Base performs slightly better on TrialPanorama and the Health Management datasets. Both G-Health models consistently outperform the original Qwen3 backbones without medical specialization, and G-Health-32B-Base reaches at least 0.80 accuracy on eight datasets.

Strong improvements on high-stakes clinical reasoning tasks. These gains in general capability also translate to better performance on clinically consequential benchmarks. On Health Management, a benchmark derived from national qualification examinations, G-Health-14B-Base improves over its backbone by 8.8% (relative) and exceeds the 0.60 passing threshold (**Fig. 2d**). On the USMLE-style MedQA benchmark, G-Health-14B-Base and G-Health-32B-Base improve over their respective baselines by 12.0% and 8.7%, reaching accuracies of 0.891 and 0.911. We further assessed structured reasoning with TrialPanorama, which focuses on clinical trial design (**Fig. 2e**). The G-Health models score higher on arm design and eligibility criteria generation, outperforming general-purpose baselines such as GPT-4o[46]. Performance on the more challenging endpoint_design subtask remains competitive, suggesting that the training pipeline strengthens structured clinical reasoning rather than only improving biomedical fact recall.

Report-oriented fine-tuning improves structural fidelity and personalization. We evaluated models after task-specific fine-tuning on health examination reports. Using a held-out test set of 100 reports generated by the PPMSA-Factory framework (a pipeline that converts physical examination materials into structured patient records and

automatically generates multi-role clinical dialogues with built-in quality control; **Extended Data Fig. 1**), we first computed ROUGE-L to measure lexical overlap with PPMSA-derived reference interpretations (**Fig. 2f**). G-Health-14B-Instruct and G-Health-32B-Instruct achieved ROUGE-L scores of 0.1235 and 0.1180, respectively, outperforming all comparator models. This indicates that domain-adapted G-Health models match the references more closely at the lexical level and better preserve the structural characteristics of PPMSA-style interpretations. We then assessed whether fine-tuned models respond to patient personality and educational background. After health examination-specific fine-tuning, sensitivity to personality traits increased substantially, and G-Health-14B-Instruct outperformed Qwen3-32B (**Fig. 2g**). This pattern suggests improved tailoring to different personality profiles, supporting the type of interpersonal calibration that can help build trust and encourage adherence in preventive settings. Attention to educational level also increased relative to the backbone models. In particular, G-Health-14B-Instruct placed greater emphasis on patients with lower educational attainment, indicating a stronger tendency to adjust explanations and recommendations to users who may need more support (**Fig. 2h**). This behavior is valuable in preventive care, where lower educational attainment is often associated with greater barriers to understanding and acting on health information. By reducing these barriers, the framework can help make high-quality, actionable guidance more accessible across socioeconomic backgrounds.

Quantitative and interpretable risk assessment using task-specific models

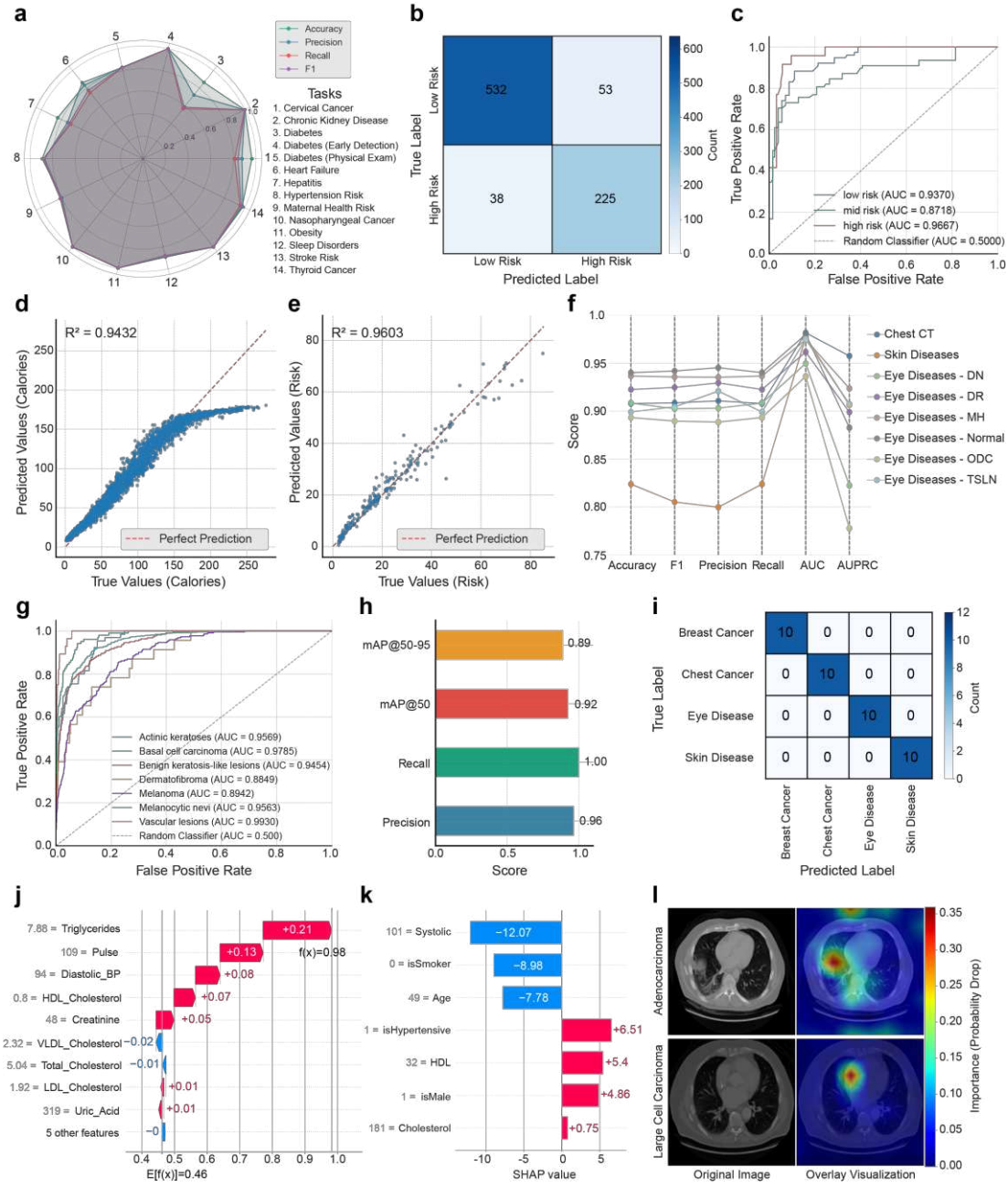


Fig. 3 | Key results of the risk assessment models. a, Performance across 14 tabular classification tasks, summarized by accuracy, precision, recall and F1 score. **b**, Confusion matrix for the hypertension risk classification task, showing model performance in distinguishing low- and high-risk individuals. **c**, Receiver operating characteristic curves for the maternal health risk classification task, showing high discriminative performance across low-, mid- and high-risk groups (AUC = 0.9370,

0.8718 and 0.9667, respectively). **d**, Comparison of true and predicted calorie values in the calorie prediction task, showing near-perfect agreement along the identity line (R^2 = 0.9432). **e**, Comparison of true and predicted cardiac risk scores in the cardiac risk prediction task, demonstrating close agreement along the identity line (R^2 = 0.9603). **f**, Performance of all image classification tasks across multiple evaluation metrics, including accuracy, F1 score, precision, recall, AUC and AUPR, for chest CT, skin disease and various eye disease datasets. **g**, Receiver operating characteristic curves for the skin disease classification task, showing high discriminative performance across seven lesion subtypes (AUC ≥ 0.8850 for all classes). **h**, Performance of the breast cancer lesion detection task, showing high precision (0.96), recall (1.00), mAP@0.50 (0.92) and mAP@0.50–0.95 (0.89). **i**, Confusion matrix for the image routing module, demonstrating perfect assignment of breast cancer, chest cancer, eye disease and skin disease images to their corresponding specialist models. **j**, SHAP-based interpretability analysis for the diabetes (physical examination) risk model, illustrating how individual examination indicators (notably triglycerides, pulse and diastolic blood pressure) contribute to the predicted risk for a representative patient. **k**, SHAP-based feature importance for the cardiac risk model. **l**, Interpretability analysis for the chest CT image classification model, showing original scans and overlay visualizations that highlight the most influential regions for the predicted labels (adenocarcinoma and large cell carcinoma).

To support risk assessment in our health management setting, we trained a set of task-specific "tool models" that the LLM can invoke within the multi-agent system. When the conversational agent identifies potential disease risk from structured examination indicators, imaging findings, or patient narratives, it calls the corresponding risk model to produce quantitative risk estimates and explanatory outputs, which are then used to guide follow-up and management recommendations. These tool models span four task types: tabular classification, tabular regression, image classification, and object detection. For image classification, we also trained a routing classifier that assigns each input image to the appropriate specialist model. For selected

tasks, we use SHAP to provide post-hoc explanations that help users and clinicians interpret the predicted risks.

Performance of tabular disease risk classification models. For tabular classification, we trained 14 disease-related risk assessment models (**Fig. 3a**). Eleven tasks achieve scores of at least 0.8 across accuracy, precision, recall, and F1. Six tasks (thyroid cancer, stroke risk, obesity, nasopharyngeal cancer, diabetes (early detection), and chronic kidney disease) approach 1.0 on all metrics, indicating strong discriminative performance. Hypertension risk provides a representative example. Its confusion matrix (**Fig. 3b**) shows that, among 263 truly high-risk individuals, the model correctly identifies 225, reflecting clear separation between low- and high-risk groups. For maternal health risk (**Fig. 3c**), the model attains AUC values of 0.9370, 0.8718, and 0.9667 for the low-, mid-, and high-risk strata, respectively. AUC curves for the remaining models are reported in **Extended Data Fig. 2**. The high-risk curve lies closest to the upper-left corner, consistent with the strongest discrimination for high-risk individuals. Performance is slightly lower for the low-risk group but remains strong, while the mid-risk group is lower yet still within a favorable range. Together, these results indicate that the model can distinguish maternal health risk levels reliably and provide quantitative support for antenatal follow-up and intervention planning.

Performance of continuous risk regression tasks. For regression, we considered calorie prediction and cardiac risk scoring (**Fig. 3d**, **e)**. In calorie prediction (**Fig. 3d**), the predicted values closely match the references across the observed range, with R^2 =

0.9432, indicating little systematic deviation. In cardiac risk scoring (**Fig. 3e**), predicted and true scores also agree closely ($R^2 = 0.9603$), with only small deviations in a limited number of samples. These high-precision regression tools can be invoked by the LLM when relevant risks are detected, enabling continuous risk estimates and more fine-grained follow-up recommendations.

Image-based risk assessment across multiple clinical modalities. In the image classification branch, we trained three families of models for chest CT, skin diseases, and eye diseases, with the eye disease category further divided into six sub-tasks (**Fig. 3f**). With the exception of the skin disease task, most chest CT and eye disease models reach accuracy, F1, precision, and recall at or above 0.90. On AUC, which better reflects class separation, nearly all tasks reach 0.95 or higher, indicating strong discrimination between positive and negative cases. AUPR values are typically lower than AUC, and some long-tail sub-tasks (for example, specific eye disease categories) are affected by class imbalance, but overall performance remains high and suggests that these classifiers are approaching clinically useful operating points. The skin disease multi-class model (**Fig. 3g**) provides a detailed view: all seven lesion subtypes achieve AUC values of at least 0.8850. Vascular lesions perform best (AUC=0.9930). Basal cell carcinoma and actinic keratoses, which are malignant or pre-malignant, both exceed 0.9500. Benign keratosis-like lesions, melanocytic nevi, and melanoma are generally above 0.9000. Dermatofibroma is the weakest class but still attains an AUC

of 0.8850, well above chance. Overall, these results indicate that the model can distinguish multiple common and clinically important skin lesions with high reliability.

Lesion detection performance in medical imaging. For object detection, we trained a breast cancer lesion detector, with results summarized in **Fig. 3h**. Precision is 0.96, indicating that most positive detections correspond to true lesions, and recall is 1.00, suggesting few missed lesions in the test set. The standard detection metrics mAP@0.50 and mAP@0.50-0.95 are 0.92 and 0.89, respectively, reflecting accurate localization over a range of intersection-over-union thresholds. Overall, the detector delivers strong performance for breast cancer image recognition and lesion localization and can serve as a specialized tool that the LLM invokes when breast imaging requires analysis.

Automated image routing for specialist model selection. To route multimodal images to the appropriate specialist models, we trained an image routing classifier. As shown in **Fig. 3i**, all 10 test images in each category (breast cancer, chest tumor, eye disease, and skin disease) are assigned to the correct specialist model, yielding 100% routing accuracy. This design lets the LLM determine that an image requires analysis, while the routing model automatically selects the most suitable downstream classifier or detector.

Model interpretability and feature attribution analysis. We also include an interpretability module to improve the transparency of risk assessment outputs. For the

diabetes (physical examination) risk model, SHAP analysis (**Fig. 3j**) suggests that triglycerides, pulse, and diastolic blood pressure contribute most strongly to the individual risk prediction, while several lipid-related variables act as relatively protective factors for the representative patient. For the cardiac risk model, SHAP feature importance (**Fig. 3k**), highlights systolic blood pressure, smoking status, age, history of hypertension, HDL level, sex, and total cholesterol as the primary drivers of the risk scores. For the chest CT classification model (**Fig. 3l**), the most influential regions overlap closely with adenocarcinoma and large-cell carcinoma lesions, suggesting that the model relies on tumor-related image features rather than background cues. Together, these tool models, combined with feature-level explanations, provide the LLM-based system with callable and more transparent risk assessment components for health examination and broader health management workflows.

Agent performance surpassing human expert benchmarks in health examination report analysis

Automated and human evaluation framework. To assess the health management agent comprehensively, we evaluated its generated reports from two complementary perspectives. For LLM-based evaluation, we asked multiple models to produce structured interpretations and recommendations for 100 test health examination reports using a unified template, and used GPT-OSS as an automatic judge to conduct pairwise comparisons and select the better output. For human evaluation, we built a fully blinded web platform. Twelve senior practicing physicians from multiple large tertiary hospitals,

together with the candidate models, generated structured reports for 10 health examination cases. All reports were then anonymized and randomly shuffled. We invited 18 practicing clinicians, 15 resident trainees, and 46 pre-clinical medical students to log into the platform and score the reports along predefined dimensions (**Extended Data Fig. 3**).

Results of automated pairwise comparisons. Using GPT-OSS as the judge, we first compared the G-Health-Agent models with their baselines (**Fig. 4a–d**). With a 14B backbone, G-Health-14B-Agent records 48 wins versus 42 for the single-model G-Health-14B-Base, with 10 ties, suggesting that multi-agent orchestration provides a stable but moderate gain at the same parameter scale. Against Qwen3-14B, G-Health-14B-Agent wins 46 cases versus 41 for Qwen3-14B, with 13 ties, indicating that at 14B our medically tailored agent is preferred over a strong general-purpose baseline. Scaling the backbone to 32B further increases the benefit of the multi-agent design: G-Health-32B-Agent wins 54 cases against G-Health-32B-Base's 26, with 20 ties. Against Qwen3-32B, G-Health-32B-Agent wins 46 cases, while Qwen3-32B wins 18 and 36 cases are ties. Overall, these results indicate that a health-check oriented multi-agent architecture improves generation quality when paired with the same backbone, with larger gains at higher model scales. Relative to Qwen3 baselines without scenario-specific alignment, the G-Health-Agent models are preferred by GPT-OSS in most cases, suggesting that task decomposition, tool-model invocation, and medical data

alignment jointly contribute to stronger performance in health examination interpretation and risk assessment.

Overall human evaluation across clinical dimensions. We aggregated scores from all human raters to obtain an overall evaluation (**Fig. 4e**). The panel reports each system's mean score on five dimensions: medical correctness, comprehensiveness, risk identification, appropriateness of department recommendations, and clarity of expression. Across all five dimensions, G-Health-32B-Agent achieves the highest mean score in each group. Medical correctness is close to 4.0, comprehensiveness and risk identification both exceed 4.0, and department recommendations and expression clarity are higher than those of the other models and the physician group. This pattern indicates that G-Health-32B-Agent is rated most favorably by evaluators with diverse clinical backgrounds. General-purpose LLMs such as DeepSeek, GPT-5, and Gemini form a middle tier, scoring below G-Health-32B-Agent but generally above most individual clinicians. The 12 participating physicians show a lower mean score and larger dispersion, consistent with substantial variation in report quality. Overall, under this protocol, G-Health-32B-Agent attains the highest aggregate quality among the evaluated systems, with strong ratings for medical accuracy, coverage, risk signaling, triage guidance, and clarity.

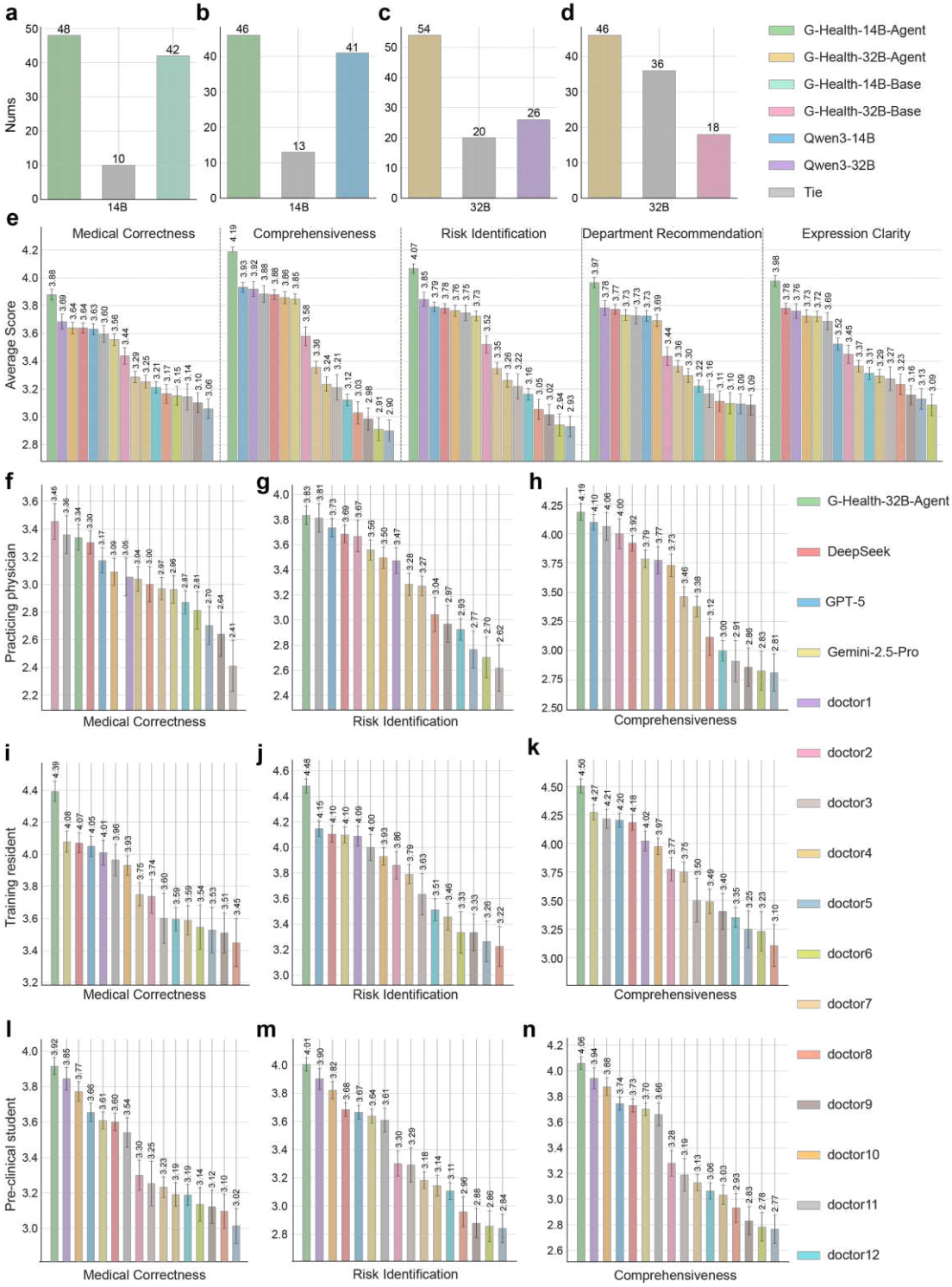


Fig. 4 | Results on the health examination report analysis task. a-d, pairwise comparisons between G-Health-14B/32B-Agent, G-Health-14B/32B-Base and Qwen3-14B/32B, with GPT-OSS as the judge, showing the number of wins and ties for each model. **e**, summarizes the overall human evaluation across all raters, showing average scores for medical correctness, comprehensiveness, risk identification, department recommendation and expression clarity. **f-n**, Scores given by evaluators at

different seniority levels (practicing physicians, training residents and pre-clinical students) for model-generated reports and physician-written reports in terms of medical correctness, risk identification and comprehensiveness, using the same legend as panels **e**.

Evaluation results stratified by rater seniority. Beyond the aggregated analysis, we stratified scores by rater group (**Fig. 4f-n**). Practicing physicians, resident trainees, and pre-clinical students independently evaluated each "answerer" (G-Health-32B-Agent, several general-purpose LLMs, and 12 clinicians) on three dimensions: medical correctness, risk identification, and comprehensiveness, using the same color scheme as in **Fig. 4e**. The pattern is consistent across all nine subpanels. For every dimension and in each rater group, G-Health-32B-Agent achieves the highest mean score and remains the top-ranked system. DeepSeek, GPT-5, and Gemini-2.5-Pro typically form the next tier. The 12 human clinicians score below the general-purpose LLMs in most cases and show greater dispersion. Overall, this stratified analysis supports the conclusion that G-Health-32B-Agent outperforms both general-purpose models and the clinician cohort under the same evaluation protocol.

Differences in rating distributions among evaluator groups. Comparing score ranges across rater groups, practicing physicians are the most stringent, with scores mostly between 2.48 and 4.14 (**Fig. 4f–h**). Resident trainees assign higher scores overall (**Fig. 4i–k**), while pre-clinical students fall in between (**Fig. 4l–n**). Results for Department Recommendation and Expression Clarity are reported in **Extended Data Fig. 4**. Despite differences in absolute scoring levels, the relative ordering of systems

remains consistent across subpanels. G-Health-32B-Agent ranks first throughout, followed by general-purpose LLMs, while the physician group falls in the middle-to-lower range. This consistency indicates that raters with different levels of training reach similar qualitative conclusions, which supports the effectiveness of G-Health in producing higher-quality preventive-care outputs under heterogeneous clinical scrutiny.

Performance across individual evaluation dimensions. Across individual dimensions, G-Health-32B-Agent shows the largest and most consistent advantages in medical correctness and risk identification, outperforming general-purpose LLMs and most clinicians. This pattern suggests that, within this study, the system captures key medical facts, diagnostic reasoning, and warning signals at a level comparable to, and in many cases higher than, the average clinician. For comprehensiveness (**Fig. 4h, 4k and 4n**), the advantage is even more pronounced. Relative to clinician reports, G-Health-32B-Agent more consistently covers disease explanations, risk factors, follow-up planning, and lifestyle recommendations. This behavior is consistent with the multi-agent design and the explicit integration of specialized risk-assessment tools.

Taken together, the GPT-OSS-based automatic comparisons and the multi-layered human evaluations point to the same conclusion. The G-Health-Agent family, especially G-Health-32B-Agent, achieves substantially better performance on health examination interpretation and disease risk assessment than strong general-purpose LLM baselines. Under this evaluation protocol, its ratings are comparable to, and in some cases higher than, the mean score of the participating clinicians, with the largest

gains in risk identification and comprehensiveness. These results support the case for deployment in real-world health management settings.

Interactive web-based deployment of the G-Health health management system

To support adoption by both clinicians and patients, we developed a user-friendly web interface for G-Health. Users can submit health-related questions or engage in free-form natural-language dialogue. A multi-agent backend parses the conversation and constructs structured inputs for downstream models. The system also integrates a medical RAG knowledge base and Model Context Protocol (MCP)-based services for medication lookup and diet recommendations. When appropriate, the model queries guidelines, textbooks, and consensus statements to provide evidence-supported responses, and can also return drug information and personalized dietary advice. Results are delivered through an interactive front end designed to fit into existing clinical workflows.

When the Clinical Language Analyst agent identifies a potential disease risk in the dialogue (for example, diabetes), the system automatically launches a corresponding risk-assessment questionnaire for the patient to complete. The responses are then passed to a backend risk model to produce quantitative risk scores and explanatory outputs. Based on these predictions, the nurse agent provides lay-friendly explanations and personalized lifestyle and follow-up recommendations (**Fig. 5**).

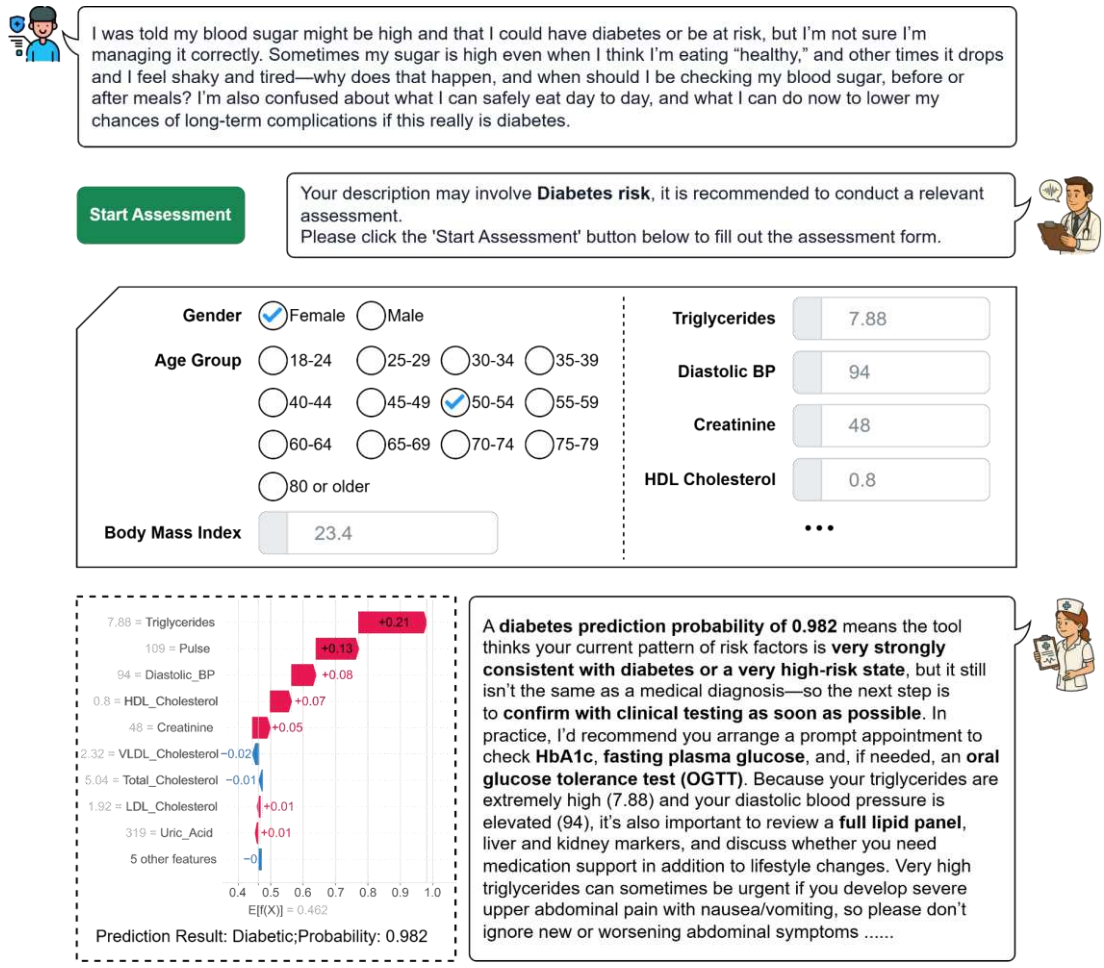


Fig. 5 | Diabetes risk-assessment workflow in the G-Health health management system. The Clinical Language Analyst agent first interprets the user's free-text dialogue, identifies a potential risk of diabetes and triggers a diabetes-specific questionnaire. The system then presents a structured assessment form that is completed by the patient, and the collected responses are fed into the risk model to generate a quantitative risk estimate with SHAP-based explanations. The nurse agent uses these risk estimates and explanations to provide personalized feedback and recommendations to the patient.

When a health examination report is uploaded, the platform launches a dedicated report-analysis workflow that runs a six-step structured pipeline. This pipeline includes indicator extraction, loading reference ranges, aggregating structured indicators, identifying and grading abnormal findings, interpreting abnormalities with lifestyle guidance, recommending additional diagnostic tests, and suggesting appropriate

referral departments, culminating in a patient-oriented summary (**Extended Data Fig. 5)**. For medical images uploaded by the user, the system first applies an image-routing model to infer the modality and anatomical region, and then dispatches the image to the corresponding image-based risk assessment model for lesion recognition and risk stratification. The consolidated results are returned to the front end for display (**Extended Data Fig. 6**).

After the dialogue and multimodal analyses are complete, users can generate a comprehensive diagnostic and management report with a single click. The system formats the content using a predefined template and exports it as a PDF document, supporting integration with electronic health record systems and routine archival workflows.

The G-Health web system has been piloted in a large health examination center, where it helps clinicians rapidly generate structured interpretations of health examination reports and quantitative disease risk estimates. In this setting, the system improves the efficiency and consistency of communicating examination results and helps examinees better understand their health status and follow recommended follow-up and interventions.

Discussion

This study focuses on a common but often under-supported step in preventive care: turning complex health examination reports into clear interpretation, calibrated risk

estimates, and practical follow-up. We frame this as a report-to-action problem and implement it through G-Health, an end-to-end framework that combines a multi-source data ecosystem, a three-stage medical alignment pipeline built on Qwen3-14B and Qwen3-32B, and a coordinated multi-agent workflow integrating structured indicators, free-text dialogue, and multimodal findings. In this system, the language model mainly coordinates reasoning, retrieves relevant guideline-based evidence, and produces patient-facing explanations, while quantitative risk estimation is performed by specialist models.

A key finding is that preference-based alignment drives the largest gains in decision quality. Across diverse medical benchmarks, supervised fine-tuning produces only modest improvements, whereas Direct Preference Optimization leads to a clear performance increase. This suggests that preventive-care behaviors such as focusing on clinically important findings, remaining cautious when evidence is uncertain, and following established guidelines are better shaped by preference signals than by imitation alone. Report-specific tuning further improves the structural quality of outputs and increases sensitivity to users' education level and communication needs, which is particularly relevant in preventive care where understanding varies widely.

Our results also highlight the value of separating generation from quantitative prediction. G-Health uses specialist tool models to generate calibrated disease risk estimates from tabular and multimodal inputs, supported by SHAP-based explanations, while the language model concentrates on synthesizing evidence and translating

findings into practical recommendations. Grounding recommendations in explicit model outputs and retrieved guidelines reduces reliance on free-form generation and mitigates a known weakness of medical LLMs, namely unsupported or guideline-inconsistent advice. The multi-agent workflow further improves consistency in report interpretation, as shown by both automated comparisons and blinded human evaluation. The wide variability observed in clinician-written summaries reinforces the need for a standardized and auditable interpretation process supported by tools and evidence.

Beyond benchmark results, G-Health addresses a practical constraint in real-world preventive care: limited time for detailed explanation and counseling in increasingly busy clinical settings. By converting heterogeneous report data into structured interpretation, calibrated risk stratification, and transparent follow-up guidance, while keeping reasoning steps inspectable, the framework proposes a scalable model for preventive health management. In this view, routine examinations are not static collections of numbers but inputs to a traceable decision process that can be monitored and continuously improved. Such an approach may be especially helpful for non-specialist providers and front-line services and may improve access to preventive guidance for individuals with lower health literacy.

Limitations

Several limitations warrant careful consideration. Although the report corpus spans multiple hospitals and organizations, regional, institutional, and demographic heterogeneity may introduce bias and distribution shifts that affect robustness and

calibration. Variation in report templates, measurement conventions, and clinical practice patterns could also influence downstream recommendations. Future work should evaluate subgroup performance and calibration stability across institutions, demographic strata, and report formats to better characterize generalizability. Second, this study emphasizes benchmark performance and preference-based assessments rather than prospective clinical outcomes. External validation in unseen institutions, calibration and decision-curve analyses for risk tools, and workflow-level endpoints such as time to interpretation, department recommendation, and patient comprehension are needed to determine real-world impact. These limitations point to opportunities for iterative refinement rather than fundamental barriers. They also motivate decision-impact studies that test whether agentic assistance improves clinical reasoning and reduces avoidable harm under realistic operational constraints.

Main references

1 Muller, S. A. *et al.* The burden and treatment of non-communicable diseases among healthcare workers in sub-Saharan Africa: a multi-country cross-sectional study. *Front Public Health* **12**, 1375221 (2024). https://doi.org/10.3389/fpubh.2024.1375221

2 Al-Abdulla, S. A., Haj Bakri, A., Mansaray, M. A. & Al-Kuwari, M. G. Assessing the impact of annual health screenings in identifying noncommunicable disease risk factors within Qatar's primary health care corporation Qatari registered population. *Front Public Health* **12**, 1305636 (2024). https://doi.org/10.3389/fpubh.2024.1305636

3 Liss, D. T., Uchida, T., Wilkes, C. L., Radakrishnan, A. & Linder, J. A. General Health Checks in Adult Primary Care: A Review. *JAMA* **325**, 2294-2306 (2021). https://doi.org/10.1001/jama.2021.6524

4 Agyeman-Manu, K. *et al.* Prioritising the health and care workforce shortage: protect, invest, together. *Lancet Glob Health* **11**, e1162-e1164 (2023). https://doi.org/10.1016/S2214-109X(23)00224-3

5 Mohr, D. C., Elnahal, S., Marks, M. L., Derickson, R. & Osatuke, K. Burnout Trends Among US Health Care Workers. *JAMA Netw Open* **8**, e255954 (2025). https://doi.org/10.1001/jamanetworkopen.2025.5954

6 West, C. P. The Paradox of Burnout From Physicians' Most Meaningful Roles. *JAMA Netw Open* **7**, e2452106 (2024). https://doi.org/10.1001/jamanetworkopen.2024.52106

7 Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172-180 (2023). https://doi.org/10.1038/s41586-023-06291-2

8 Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nature Medicine* **31**, 943-950 (2025).

9 Peng, C. *et al.* A study of generative large language model for medical research and healthcare. *NPJ Digit Med* **6**, 210 (2023). https://doi.org/10.1038/s41746-023-00958-w

10 Zhou, S. *et al.* Large language models for disease diagnosis: a scoping review. *NPJ Artif Intell* **1**, 9 (2025). https://doi.org/10.1038/s44387-025-00011-z

11 Lawrence, H. R. *et al.* The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Ment Health* **11**, e59479 (2024). https://doi.org/10.2196/59479

12 Meng, X. *et al.* The application of large language models in medicine: A scoping review. *iScience* **27**, 109713 (2024). https://doi.org/10.1016/j.isci.2024.109713

13 Chen, X. *et al.* Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ Digit Med* **8**, 159 (2025). https://doi.org/10.1038/s41746-025-01550-0

14 Kim, Y. *et al.* Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems* **37**, 79410-79452 (2024).

15 SHENG, R. A Survey of LLM-based Multi-agent Systems in Medicine. *Authorea Preprints* (2025).
16 Yang, A. *et al.* Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
17 Rafailov, R. *et al.* Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* **36**, 53728-53741 (2023).
18 Erickson, N. *et al.* Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505* (2020).
19 Khanam, R. & Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* (2024).
20 Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459-9474 (2020).
21 Sadeghi, Z. *et al.* A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering* **118**, 109370 (2024).
22 Salih, A. M. *et al.* A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems* **7**, 2400304 (2025).
23 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).
24 Huang, Y. Z. *et al.* C-EVAL: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *Adv Neur In* (2023).
25 Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**, 6421 (2021).
26 Li, J., Zhong, S. & Chen, K. in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* 8862-8874.
27 Li, H. *et al.* in *Findings of the Association for Computational Linguistics: ACL 2024.* 11260-11285.
28 Liu, J. *et al.* Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems* **36**, 52430-52452 (2023).
29 Wang, Z. *et al.* TrialPanorama: Database and Benchmark for Systematic Review and Design of Clinical Trials. *arXiv preprint arXiv:2505.16097* (2025).
30 Wang, X. *et al.* in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).* 6184-6205.
31 Hendrycks, D. *et al.* Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
32 Wang, Y. *et al.* Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems* **37**, 95266-95290 (2024).
33 Li, S. *et al.* Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems* **37**, 28858-28888 (2024).

34 Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP).* 2567-2577.

35 Shoham, O. B. & Rappoport, N. MedConceptsQA: Open source medical concepts QA benchmark. *Computers in Biology and Medicine* **182**, 109089 (2024).

36 Zhang, H. *et al.* in *Findings of the Association for Computational Linguistics: EMNLP 2023.* 10859-10885.

37 Wang, B. *et al.* Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671* (2025).

38 Dou, C. *et al.* Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208* (2025).

39 Labrak, Y. *et al.* Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373* (2024).

40 Dubey, A. *et al.* The llama 3 herd of models. *arXiv e-prints*, arXiv: 2407.21783 (2024).

41 Pal, A. & Sankarasubbu, M. *OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences*, <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B> (2024).

42 GLM, T. *et al.* Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).

43 Bao, Z. *et al.* Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346* (2023).

44 Wu, C. *et al.* PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association* **31**, 1833-1843 (2024).

45 Initiative, O. M. *Qwen2.5 Meditron-3[14B]*, <https://huggingface.co/OpenMeditron/Meditron3-Qwen2.5-14B> (2025).

46 Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

Methods

Data ecosystem construction

To more comprehensively cover the diverse needs and task settings in real-world clinical scenarios, we systematically collected and integrated data from multiple sources, including medical dialogues, preference alignment, health examination reports, and risk assessment. These data were further organized into a unified corpus to support our staged training and evaluation. An overall summary of the collected data is provided in **Extended Data Fig. 7**.

Medical dialogue datasets

As the stage-1 training data, we aggregated multi-source Chinese medical dialogue and question-answering datasets, whose overall composition is shown in **Extended Data Fig. 7a**, totaling 2,964,200 dialogue samples. The two largest components are Chinese-medical-dialogue (799,743 samples, ~27.0%) and huatuo_knowledge_graph_qa (798,444 samples, ~26.9%), together accounting for ~53.9%[47]. The next major sources include DISC-Med-SFT (464,882 samples, ~15.7%), UltraMedical (409,593 samples, ~13.8%), and huatuo_encyclopedia_qa (364,420 samples, ~12.3%)[43,47,48]. Collectively, these five datasets comprise ~95.7% of the SFT corpus and form its primary backbone. We further incorporated smaller but more targeted datasets to strengthen specific capabilities: medical-o1-reasoning-SFT (90,120 samples, ~3.0%) and Medical-R1-Distill-Data-Chinese (17,000 samples, ~0.57%) emphasize long-chain reasoning and learning verifiable reasoning traces, while

XunYiWenYao (19,998 samples, ~0.67%) provides additional QA examples that more closely reflect real patient-style queries[49]. After data cleaning and deduplication, we obtained 2,817,556 dialogue samples in total, including 342,562 multi-turn samples with dialogue history and 2,474,994 single-turn samples without dialogue history.

Preference alignment datasets

In stage 2, our goal is to maximally strengthen the model's preference for correct answers. To this end, we construct DPO dialogue samples starting from the datasets used in our evaluation, using only their training and validation splits, and we enforce a strict data-isolation protocol to ensure that the train/validation data do not overlap with the test set, thereby preventing data leakage. Concretely, for each multiple-choice question, we treat the response corresponding to the correct option as the "rewarded" sample (chosen) and responses corresponding to incorrect options as the "penalized" samples (rejected), forming preference pairs. Therefore, a question with four options yields three DPO dialogue samples. The overall data scale is summarized in **Extended Data Fig. 7b**: the training set contains 1,643,350 dialogue samples and the validation set contains 124,668 samples. Among the training data, 41.6% of samples include chain-of-thought content, whereas this proportion is 24.5% in the validation data. More detailed statistics and breakdowns are provided in **Supplementary Table 5**.

Health examination report datasets

In stage 3, we conducted task-specific fine-tuning for health examination report interpretation and generation. We collected 3,661 health examination reports from 123

hospitals across three institutions. For well-structured Word reports, we used template-based matching to extract key fields and align them into a consistent format. For PDF reports, we applied MinerU for layout parsing and text structuring, and then ran the same matching procedure to keep processing consistent across formats[50]. We removed reports with empty content, substantial missing information, or clearly invalid fields, yielding 3,503 valid reports (see **Supplementary Table 6**). To meet privacy and compliance requirements, we de-identified all reports by removing personal identifiers such as names, contact information, and national ID numbers. We also standardized measurements across hospitals and report templates by mapping heterogeneous items to a unified schema with consistent naming, so that all reports share the same structured representation. In parallel, we compiled auxiliary knowledge for each examination metric, including normal reference ranges, common factors that influence the measurement, and brief explanatory notes, covering 199 metrics to support downstream interpretation and risk alerts.

Next, we quantified and visualized the most frequent health issues mentioned in the reports (**Extended Data Fig. 7c**). Metabolism-related abnormalities dominate the distribution, including dyslipidemia, fatty liver, hyperuricemia, and hypertension. This pattern suggests that chronic diseases and their preclinical states, such as metabolic syndrome and elevated cardiovascular risk, are common in the checkup population. Imaging and organ examinations also yield many incidental findings, including nodules, cysts, stones, and polyps, which underscores the need for risk-stratified follow-up and, when appropriate, additional diagnostic workups. The sizable “Others” category points

719 to a long tail of less frequent findings that may still carry meaningful risk, motivating
720 coverage that extends beyond the most common patterns.

721 To better reflect real-world user variation and support personalized health
722 education in the PPMSA framework, we assigned each report a synthetic education
723 background and personality type as controllable conditional inputs. The education
724 distribution follows the Seventh National Population Census of China, and the
725 personality distribution is based on a public Kaggle dataset. **Extended Data Fig. 7d**
726 shows that education levels are centered on mid-level attainment: junior high school
727 (1,307) is most common, followed by primary school (938), high school (571), and
728 college or higher (586), with a smaller group reporting no formal education (101). This
729 setup helps the model address typical differences in health literacy and supports
730 stratified generation, ranging from plain-language, step-by-step guidance to more
731 mechanistic explanations with supporting details. **Extended Data Fig. 7e** shows that
732 personality types are concentrated in INFP (740), INFJ (594), INTP (526), and INTJ
733 (440), followed by ENTP (277), ENFP (273), ISTP (136), ISFP (109), ENTJ (93), ISTJ
734 (83), ENFJ (77), ISFJ (67), ESTP (36), ESFP (19), ESFJ (17), and ESTJ (16)[51]. These
735 profiles are not inferred attributes of real examinees; they are introduced solely to
736 encourage PPMSA to learn and generate differentiated communication styles and health
737 education strategies under varying user characteristics.

738 To assess generalization after fine-tuning, we split the 3,503 reports into a training
739 set (n = 3,403) and a test set (n = 100). Both splits show a broad and largely comparable

distribution of indicator counts per report (**Extended Data Fig. 7f**), suggesting that the test set is representative in terms of metric coverage and complexity.

Risk assessment model datasets

For risk-assessment model development, we organized the training data into four task types: tabular classification, regression, image classification, and object detection. The overall scale and per-subtask sample counts are summarized in **Extended Data Fig. 7g** (x-axis in log scale). Tabular classification is the largest component, combining public and in-house datasets that span a wide range of diseases and risk themes. Using diabetes as an example, and to reflect the temporal order and heterogeneous data sources typical of health examinations, we further split diabetes-related data into three complementary settings: pre-examination risk screening, classification based on examination results, and early-warning identification. This design supports end-to-end modeling from risk alerts to result interpretation and early screening reminders[52]. Beyond diabetes, the classification tasks cover chronic disease and key-population risks such as stroke risk, hypertension risk, obesity, and maternal health risk, and extend to additional conditions relevant to screening and long-term health management, including nasopharyngeal cancer, cervical cancer, chronic kidney disease, thyroid cancer, sleep disorders, heart failure, and hepatitis[53-59]. Together, these tasks cover both high-prevalence chronic risks and cancer-screening themes, enabling the model to address common metabolic and cardiovascular risks while remaining sensitive to less frequent but potentially high-risk conditions.

For regression, we focus on two types of continuous prediction. The first is cardiac risk assessment, which maps multiple examination indicators to a continuous risk score (or an ordinal risk level). The second covers calorie-related prediction, where the model estimates lifestyle variables such as energy intake and expenditure to provide quantitative support for personalized health management recommendations.

For image classification, we include three visual modalities that are central to health examinations: ophthalmic imaging, dermatology images, and chest CT. These data complement tabular indicators by providing imaging and phenotypic evidence that is not well captured by laboratory values alone[60]. The ophthalmic task is not framed as a simple "diseased vs.\ normal" classification. Instead, we use a multi-label, multi-class formulation to detect distinct lesion cues and structural abnormalities, including diabetic retinopathy changes (DR), diabetes-associated neovascular or proliferative changes (DN), macular-region abnormalities (MH), optic disc or cup abnormalities (ODC), signs related to retinal nerve fiber layer or structural thinning (TSLN), and normal findings. This design supports more fine-grained risk alerts and follow-up recommendations.

Finally, for object detection, we primarily include breast cancer detection data to learn lesion localization and highlighting, enabling risk alerts and referral suggestions that better match real clinical workflows[61]. Task definitions and label taxonomies for each dataset are summarized in **Supplementary Table 7**

Model evaluation datasets

To assess model capability broadly, we curated evaluation datasets along four dimensions: General Medical Evaluation, Medical Examination, Evidence-based Medicine, and Hard Medical (**Fig. 2c**). For general-domain benchmarks such as C-Eval, CMMLU, MMLU, and MMLU-Pro, we used only the medical-related subsets to keep the evaluation clinically relevant.

To evaluate performance in health examination settings, we additionally constructed a dedicated "Health Management" dataset from past national qualification examinations for health managers. The dataset contains 615 questions and is dominated by multiple-answer items, which are substantially more challenging than standard single-choice questions and place greater demands on model reasoning.

RAG datasets

To support evidence-grounded medical question answering in G-Health QA, we built a retrieval corpus for RAG. The corpus is centered on authoritative medical references and is complemented by widely used clinical study notes and key-point summaries, as well as supplementary medical textbooks. This design aims to improve retrieval precision, evidence coverage, and traceability across common use cases, including disease explanation, interpretation of medications and tests, and health management guidance.

Extended Data Fig. 7h summarizes the 987 curated document entries by source type and disciplinary domain. The corpus spans a broad range of clinical and biomedical areas, including clinical medicine, organ systems, pharmacy, nursing, dentistry, and

traditional Chinese medicine. The source types play complementary roles: textbooks provide standardized definitions and diagnostic or therapeutic workflows, books extend coverage to long-tail specialty topics beyond standard curricula, and study notes offer dense, quickly accessible key points.

For preprocessing, we used MinerU to convert heterogeneous documents (e.g., PDF and Word) into searchable text while preserving hierarchical structure, including chapter, section, and subsection headings as well as item numbering. This structure supports precise provenance tracking and citation. We removed low-value noise such as headers and footers, page numbers, duplicated tables of contents, and reference lists, and normalized line-break artifacts, irregular characters, and encoding issues. To improve corpus reliability, we merged duplicate references and, when multiple editions of the same book were available, kept the most recent edition as the primary version. We also performed integrity checks and excluded truncated or corrupted files, ensuring that all indexed passages are complete and traceable to their original sources.

Personalized Perception-based Medical Scenario Architecture Factory (PPMSA-Factory)

To construct high-quality training and evaluation samples for medical question answering, we developed PPMSA-Factory. The system converts real-world physical examination materials into structured patient information and, conditioned on this representation, automatically generates multi-role clinical dialogues with an integrated quality-control process. As illustrated in **Extended Data Fig. 1**, the pipeline has four

stages: data processing and structuring, scenario setup, dialogue generation, and automated quality inspection and filtering.

The data processing stage follows the procedure described in the Data section. During structuring, we parse the cleaned text into a standardized patient record with two categories of key information. The first captures basic demographics and visit context, including age, sex, and the department or institution. The second encodes medical content, including examination indicators (e.g., blood glucose and blood pressure), imaging or laboratory conclusions (e.g., fatty liver or suspected hypertension), primary health issues and their associated attributes (diagnoses, potential causes, and salient features), and secondary or comorbid risks with their supporting indicators (e.g., LDL/HDL/TG for dyslipidemia). We represent this information using unified fields and a hierarchical schema, which supports scenario instantiation and controllable dialogue generation in later stages.

During scenario setup, PPMSA-Factory instantiates patient profiles from real-world seed data and assigns each profile a personality type and education level. By pairing the structured patient record with scenario parameters, the system defines dialogue-generation conditions with clear context, traceable evidence, and consistent objectives, which improves controllability and supports reuse across samples.

In the dialogue generation stage, we create samples that approximate real clinical communication through multi-role interaction. At minimum, each dialogue includes a patient who states the chief complaints and key report information, a health

communicator who asks clear and medically accurate follow-up questions to fill missing details, and a doctor who interprets the report and provides recommendations. During generation, the model is restricted to facts present in the structured record to avoid unsupported inferences. It also systematically elicits and explains key indicators, abnormal findings, and risk factors, ensuring that the dialogue remains consistent with the underlying patient data.

To control data quality, we apply an automated quality inspection and filtering module after generation. Each dialogue is scored along multiple dimensions, including coverage of required questions, factual accuracy, and coherence or clinical usefulness of the responses. We retain a dialogue only if all three scores are ≥ 4; otherwise, the system regenerates it. This closed loop of generation, evaluation, and filtering allows PPMSA-Factory to scale the production of high-quality and diverse medical dialogues while preserving traceability and consistency, providing a stable data foundation for downstream model training and evaluation.

Model training

G-Health model training

For LLM training, we selected four representative open-weight, instruction-tuned backbones from three model families: Qwen3 (Qwen3-14B and Qwen3-32B), Llama (Meta-Llama-3.1-8B-Instruct), and GLM (GLM-4-9B-0414). We compared these backbones under a unified evaluation protocol (**Fig. 2b**). The Qwen3 models achieved the strongest overall performance, so we chose Qwen3 as the baseline for G-Health QA

and initialized all subsequent training from this backbone. To progressively introduce medical capability, align preferences, and specialize for physical-examination scenarios, we adopted a three-stage training pipeline implemented in LLaMA-Factory and applied consistently across stages[62].

For input construction, we followed the Qwen3 conversational template and introduced explicit control tokens to toggle a thinking mode. For samples with intermediate reasoning, we prepend "\ think" to the dialogue; for non-thinking samples, we prepend "\ no think". To keep outputs consistent, we require every response to contain a "<think> </think>" block. When a sample is in thinking mode, the "<think>" field contains the corresponding reasoning; otherwise, the block is left empty. This unified format provides a stable interface for distinguishing thinking and non-thinking behavior during both training and inference, and it improves alignment robustness across heterogeneous data sources.

We trained the model using a progressive three-stage strategy. In stage one, we performed supervised fine-tuning (SFT) on a large-scale medical dialogue corpus and used LoRA for parameter-efficient adaptation, producing a strong baseline for medical question answering[63]. In stage two, we applied DPO to align responses with preferences, improving factual accuracy and consistency in clinical settings[17]. In stage three, we carried out task-specific fine-tuning on health examination data constructed with PPMSA-Factory, again using LoRA, to further improve robustness for target applications such as health examination report interpretation and health management recommendations. Hyperparameter settings are reported in **Supplementary Table 8**.

Risk assessment model training

For training the risk-assessment tool models, we adopted modelling choices that are matched to each data modality and amenable to practical deployment. For tabular classification and regression, we used AutoGluon's TabularPredictor to perform automated machine learning, which under a fixed training budget automatically carries out candidate model selection, hyperparameter optimization and ensemble learning[18]. We further employed five-fold cross-validation to obtain more robust estimates of generalization performance (AutoGluon also supports fold-based validation together with bagging and stacking to improve stability).

For image classification, we trained models with AutoGluon AutoMM's MultiModalPredictor and used five-fold cross-validation to keep training and evaluation protocols consistent across datasets with different class distributions[64].AutoMM offers an end-to-end training interface for images and other modalities, which reduces implementation overhead when developing multi-task image classifiers.

For object detection, we used Ultralytics YOLO11 to balance detection accuracy and inference efficiency, meeting the real-time and deployment requirements of health examination settings[65]. For the image routing model, the goal is coarse-grained identification of imaging modality and anatomical region, where class boundaries are clear and lightweight inference is important. We therefore trained a ResNet-18 backbone to minimize additional latency and improve overall system throughput[66].

Model interpretability

To improve the interpretability of model outputs for medical decision-making, we developed an explanation framework that covers both structured health examination indicators and medical images. For tabular inputs such as health examination reports, we used SHAP (SHapley Additive exPlanations), which is grounded in cooperative game theory, and applied the model-agnostic Kernel SHAP explainer to attribute an individual risk prediction to the marginal contribution of each feature[67]. In practice, we used the training-set median of each feature as the background reference to quantify how a subject deviates from a "typical" population level, yielding individualized local explanations. We further visualized contribution pathways for key indicators, such as blood pressure, blood glucose, and lipids, using waterfall plots to make the primary drivers of predicted risk easy to trace.

For medical imaging, we used an occlusion-sensitivity analysis based on spatial perturbations to localize regions that the model relies on. Given an input image I and the predicted probability for a target class $p_y(I)$, we partitioned the image into an 8×8 grid Ω_i[68]. For each region Ω_i, we generated a perturbed image $I^{(i)}$ by replacing pixels in that region with the global pixel mean μ:

$$I^{(i)}(u) = \begin{cases} \mu, & u \in \Omega_i \\ I(u), & u \notin \Omega_i \end{cases} \quad (1)$$

We then defined regional importance as the change in target probability before and after occlusion:

$$s_i = p_y(I) - p_y\left(I^{(i)}\right) \quad (2)$$

Larger s_i values indicate that the region is more influential for the model's decision. Normalizing s_i and mapping the scores back to the image space yields a saliency heatmap, providing a quantitative characterization of the model's spatial sensitivity.

For visualization, feature importance in tabular tasks is presented by ranking SHAP values, whereas for imaging tasks we upsampled the grid-level importance matrix to the original resolution using bilinear interpolation and overlay a pseudo-color heatmap onto the input image to align salient regions with anatomical structures. By combining structured feature attribution with spatial localization, the proposed framework improves the readability and auditability of risk prediction and image recognition outputs without modifying the underlying models, providing more reliable evidence to support clinical use.

Evaluation metrics

In designing evaluation metrics, we assessed performance at two levels: generation quality for the LLMs and predictive performance for the downstream risk-assessment tool models. For single-answer multiple-choice questions, we used conditional log-likelihood scoring (CPL). Given a question prompt and context x, we compute the conditional probability of each candidate option o_k and select the option with the highest score as the prediction:

$$\hat{k} = arg\max_{k} \log p(o_k \mid x) \tag{3}$$

Here, $\log p(o_k \mid x)$ is computed by summing token-level log probabilities over the option sequence. A prediction is counted as correct if the selected option matches the gold answer. For multiple-answer questions, we require the model to output a structured response (e.g., a set of options). We then normalize the output to a canonical representation and perform exact matching against the gold option set; a prediction is considered correct if and only if the two sets are identical.

To evaluate health examination report generation, we used a held-out test set of 100 reports generated by the PPMSA framework. We first computed ROUGE-L to quantify overlap between the generated text and the reference at the level of the longest common subsequence (LCS). Let the reference be r and the generated text be c; a commonly used formulation is:

$$R = \frac{\mathrm{LCS}(r,c)}{|r|} \quad P = \frac{\mathrm{LCS}(r,c)}{|c|} \quad F_\beta = \frac{(1+\beta^2)PR}{R+\beta^2 P} \tag{4}$$

We further assessed how well the models attend to individual differences. For personality sensitivity, we constructed a lexicon $\mathcal{V}_{\mathrm{MBTI}}$ based on MBTI-related descriptions. If the response contains any term from this lexicon, the personality-attention count for that sample increases accordingly:

$$S_{\mathrm{pers}}(g) = \sum_{w \in \mathcal{V}_{\mathrm{MBTI}}} \mathbf{1}, w \in g \tag{5}$$

For educational-level sensitivity, we used an analogous lexicon-based proxy. We constructed an education-related lexicon $\mathcal{V}_{\mathrm{edu}}$ (e.g., terms indicating simplified

explanations or explicit references to educational attainment) and defined the education-attention score as:

$$S_{\text{edu}}(g) = \sum_{w \in \mathcal{V}_{\text{edu}}} \mathbf{1}, w \in g \tag{6}$$

A higher $S_{\text{edu}}(g)$ indicates that the model more explicitly tailors its explanations to the user's educational background. Detailed lexicon construction and hyperparameters are provided in **Supplementary Table 9**.

For the risk-assessment tool models, we used standard task-specific metrics. For classification tasks, we report accuracy, precision, recall and F1 score:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

For regression tasks, we report the coefficient of determination R^2:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{11}$$

where y_i denotes the ground-truth value, $\hat{y}_i$ the prediction and $\bar{y}$ the mean of the ground-truth values. For object detection, we report precision, recall and mean average precision (mAP), where mAP is computed by aggregating average precision across intersection-over-union (IoU) thresholds (for example, mAP@0.50 and mAP@0.50–0.95) to capture overall localisation and classification performance.

Agent design

In this study, we develop a multi-agent collaborative architecture that combines heterogeneous medical knowledge resources with callable external tools, supporting an end-to-end workflow from health consultation and information collection to personalized clinical recommendation generation. The system is grounded by a RAG module whose retrieval corpus includes nearly one thousand authoritative medical textbooks and clinical notes, providing traceable evidence for key conclusions and recommendations at inference time. In parallel, we connect an extensible tool suite via MCP to enable task-specific structured computation and querying. Concretely, Nexonco-mcp supports symptom- and disease-related information retrieval, HowtoCook-mcp provides dietary guidance and recipe recommendations, and G-Health-mcp performs automated calculation and normalization of health indicators such as BMI.

Operationally, the report-analysis and risk-assessment modules work together to parse user inputs, retrieve supporting evidence, and synthesize actionable recommendations. This division of labor enables the system to switch components based on user intent and input modality, and to execute two core workflows automatically (**Extended Data Fig. 8**).

Health examination report analysis workflow

The health examination report analysis workflow uses a multi-step heuristic pipeline that transforms unstructured user dialogue or OCR-extracted text into a

structured medical assessment report with traceable provenance. The workflow starts with a Medical Data Extraction Specialist that identifies examination-related entities in the input. An LLM then normalizes indicator names and extracts values and units, writing them into a unified schema. To reduce noise and formatting variability in free-text inputs, the parsing module combines regular-expression checks with JSON-repair strategies to constrain outputs and correct key fields. It also computes derived indicators such as BMI to complete core health features, so downstream analyses operate on standardized data.

After structuring and normalization, the workflow performs a knowledge-driven abnormality adjudication step. The agent loads a predefined reference-standard library and uses a unit-conversion engine to map heterogeneous input units into a unified reference system. Each indicator is then compared against its corresponding reference range, enabling automatic detection of abnormal findings across domains such as metabolic and hematological panels. This process produces a traceable evidence chain, including the original value, the normalized value, the reference range, and the abnormality verdict, which provides stable inputs for downstream interpretation.

For abnormal findings, the system provides multi-level clinical interpretation, progressing from explanation to follow-up recommendations and referral, and is implemented by specialized agents, including an Indicator Interpretation agent, a Follow-up Recommendation Generator, and a Medical Department Referral Specialist. The system first explains the potential clinical implications and risks associated with each abnormal item, then recommends repeat testing and targeted follow-up

examinations, and finally suggests relevant clinical departments and care pathways when appropriate. A summarization agent then consolidates key abnormalities, risk signals, and management recommendations into a structured overview, producing a consistently formatted electronic health report that supports reading, archiving, and follow-up management.

Risk assessment workflow

The risk-assessment workflow targets early screening and stratification in health management settings. It combines multimodal routing with coordinated tool-model invocation, allowing the system to handle both structured questionnaire data and medical images and to return explainable risk estimates together with actionable intervention recommendations within a unified framework. The workflow starts with a Clinical Language Analyst agent that performs semantic parsing and intent recognition over the user's natural-language description. By extracting symptoms, medical history, and risk factors, the agent produces disease-specific triggers across multiple conditions. When the estimated risk for a disease exceeds a predefined threshold (e.g., diabetes or hypertension), the system selects the corresponding risk-assessment task and generates or retrieves the appropriate questionnaire and structured form, prompting the user for key physiological measures and behavioral information. This process converts conversational input into structured features that can be directly consumed by the risk models.

In image-input scenarios, the system first uses a lightweight routing classifier to identify the imaging modality and anatomical region (e.g., breast, chest, skin, or fundus) and to select the corresponding specialist model. It then invokes the appropriate image-based risk model to perform classification or risk prediction. For tasks that require object detection, such as breast lesion identification, the system also returns lesion localization from YOLO-style detectors along with overlaid visualizations, enabling intuitive inspection of the evidence supporting the model's decision.

After the tabular and imaging models return risk probabilities, predicted labels, and associated explanations, the Nurse Agent enters a closed-loop intervention step. The agent integrates the user's chief complaint and history with current health examination indicators and explainability evidence from the risk models (e.g., SHAP attributions) to produce lay-friendly interpretations and actionable guidance, including lifestyle interventions, follow-up testing, and care-seeking recommendations. Together, this workflow implements an end-to-end process that links risk identification, evidence-grounded explanation, and health management recommendations into a deployable risk warning and management pipeline.

Human evaluation of report quality via blinded web platform

To assess the quality of model-generated health examination reports, we built a web-based anonymous evaluation platform (workflow shown in **Extended Data Fig. 3**) to compare LLM-generated reports with physician-written reports under a unified protocol. The platform was implemented with Flask, and SQLite was used to store

evaluation records and associated metadata.

For data preparation and report generation, we manually selected 10 medical examination records (MERs). For each MER, reports were independently written by four LLMs (GPT-5, Gemini-2.5-Pro, DeepSeek-R1, and the proposed model) and by 12 physicians from seven tertiary hospitals. To ensure a fair and controlled comparison between human experts and AI systems, all participants received the same input materials. Each report was produced solely from the original MER content, without access to any additional clinical information. For the LLMs, we used a single prompt template that defined the task, specified the intended scope of reasoning, and enforced a standardized output format, so that all models generated reports under the same instructions and formatting constraints. Physicians completed the task through a structured questionnaire that provided the identical MER input and required reports to follow a predefined template aligned with the model output format. This protocol ensured comparable conditions for human- and AI-generated reports, with consistent input information and a standardized reporting structure, enabling a fair assessment of content quality and clinical usefulness. In total, we collected 111 reports, including 40 generated by LLMs and 71 written by physicians.

To reduce rater bias, we used a strict double-blind design. For each MER, all available reports were randomized, de-identified, and presented only with numeric identifiers, with no indication of authorship. After reading each report, evaluators rated quality and accuracy on a 1--5 Likert scale across five dimensions: medical correctness, comprehensiveness, risk identification, department recommendation, and clarity of

expression. We collected only anonymized evaluation records and did not store any personally identifiable information at any stage.

In total, the platform collected 6,955 evaluation records. These were provided by 46 pre-clinical medical students (3,613 ratings) from two medical universities, and by 33 clinical evaluators, including 18 practicing physicians (1,754 ratings) and 15 resident trainees (1,588 ratings). Clinical evaluators were recruited from multiple tertiary hospitals. Participants represented six departments: Internal Medicine, Pediatrics, Health Examination, Neurology, Cardiology, and Obstetrics and Gynecology.

Web development

We implemented an integrated web system for health examination report parsing and health consultation. The backend is built with Flask to handle request routing, session management, and orchestration of model and tool calls, while the frontend uses Bootstrap to provide a responsive interface for conversational interaction, form entry, result presentation, and export. System data and runtime logs are stored in SQLite to support lightweight deployment and experiment traceability. The tool layer is implemented with MCP and supports streaming over HTTP for tool invocation, enabling real-time responses in multi-step tasks such as retrieval, parsing, and generation. Model inference is served via Ollama, which provides a unified deployment interface for both question answering and report generation.

Data availability

In this study, we used several datasets for medical dialogue training, preference alignment, health examination report specialization, and health management evaluation. The publicly available datasets used in training and evaluation are as follows. The SFT dataset for stage 1 training is available at https://huggingface.co/datasets/YDXX/G-Health-sft-data. The DPO dataset for stage 2 alignment is available at https://huggingface.co/datasets/YDXX/G-Health-dpo-data. The instruction-tuning dataset specialized for health examination reports, generated using the PPMSA-Factory pipeline, is available at https://huggingface.co/datasets/YDXX/PPMSA-Factory-data. The health management benchmark used for evaluation is available at https://huggingface.co/datasets/YDXX/Health-management-questions.In addition, we curated a collection of datasets used for training disease-specific risk assessment models, compiled from multiple publicly available medical and clinical datasets. These risk assessment datasets are available at https://huggingface.co/datasets/YDXX/Risk-assessment-datasets.The anonymized physical examination report dataset used for task-specific fine-tuning was derived from real-world records collected across multiple institutions. To protect participant privacy and comply with institutional ethical requirements, all reports were irreversibly de-identified before use. Access to the anonymized report dataset is provided through a controlled request process at https://huggingface.co/datasets/YDXX/Health-examination-report. Source data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The source code for the G-Health web application is publicly available on GitHub (https://github.com/QIANJINYDX/G-Health). The online instance used in this study is accessible at https://www.jishi.chat/.

Methods references

17 Rafailov, R. *et al.* Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* **36**, 53728-53741 (2023).

18 Erickson, N. *et al.* Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505* (2020).

43 Bao, Z. *et al.* Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346* (2023).

47 Li, J. *et al.* Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526* (2023).

48 Zhang, K. *et al.* Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems* **37**, 26045-26081 (2024).

49 Chen, J. *et al.* Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925* (2024).

50 Wang, B. *et al.* Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839* (2024).

51 Ontoum, S. & Chan, J. H. Personality type based on myers-briggs type indicator with text posting style by using traditional and deep learning. *arXiv preprint arXiv:2201.08717* (2022).

52 Islam, M. F., Ferdousi, R., Rahman, S. & Bushra, H. Y. in *Computer Vision and Machine Intelligence in Medical Image Analysis: International Symposium, ISCMM 2019.* 113-125 (Springer).

53 Palechor, F. M. & De la Hoz Manotas, A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in brief* **25**, 104344 (2019).

54 Ahmed, M., Kashem, M. A., Rahman, M. & Khatun, S. in *InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019.* 357-365 (Springer).

55 Fernandes, K., Cardoso, J. S. & Fernandes, J. in *Iberian conference on pattern recognition and image analysis.* 243-250 (Springer).

56 Rubini, L., Soundarapandian, P. & Eswaran, P. (UCI Machine Learning Repository, 2015).

57 Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R. & Tarokhian, A. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. *European Archives of Oto-Rhino-Laryngology* **281**, 2095-2104 (2024).

58 Mao, R. A Multidimensional Statistical Analysis of Lifestyle Determinants Affecting Sleep Quality in Adults. *Journal of Student Research* **14** (2025). https://doi.org/10.47611/jsrhs.v14i1.8839

59 Chicco, D. & Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making* **20**, 16 (2020).

60 Codella, N. *et al.* Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint*

arXiv:1902.03368 (2019).

61 Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data in brief* **28**, 104863 (2020).

62 Zheng, Y. *et al.* Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372* (2024).

63 Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).

64 Tang, Z. *et al.* Autogluon-multimodal (automm): Supercharging multimodal automl with foundation models. *arXiv preprint arXiv:2404.16233* (2024).

65 Ultralytics YOLO11 (2024).

66 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770-778.

67 Lee, D. D., Pham, P., Largman, Y. & Ng, A. Advances in neural information processing systems 22. *Tech Rep* (2009).

68 Zeiler, M. D. & Fergus, R. in *European conference on computer vision.* 818-833 (Springer).

Acknowledgements

We thank clinicians from multiple tertiary hospitals for contributing to clinician-written report drafting and blinded evaluation. Twelve physicians independently drafted reports based on ten medical examination records, and 33 clinical evaluators (18 practicing physicians and 15 resident trainees) provided blinded ratings on the evaluation platform, yielding 6,955 evaluation records. Participating clinicians were from Sichuan Provincial People's Hospital, Zigong First People's Hospital, Chengdu Women's and Children's Central Hospital, Guangyuan Central Hospital, Liuzhou Maternal and Child Health Hospital, Chongqing Banan District People's Hospital, Longquan Maternal and Child Health Hospital, the Affiliated Hospital of the University of Electronic Science and Technology of China, the TCM Affiliated Hospital of Southwest Medical University, and the Affiliated Hospital of Southwest Medical University. These clinicians represented multiple clinical specialties, including Internal Medicine, Pediatrics, Health Examination, Neurology, Cardiology, and Obstetrics and Gynecology. We also acknowledge the participation of 46 pre-clinical medical students from the Class of 2023 Nine-Year Clinical Medicine Program at Chengdu University of Traditional Chinese Medicine and the Class of 2023 Clinical Integrative Medicine Program at Southwest Medical University. Computing resources were provided by the QiZhi Platform at the Shanghai Institute of Intelligent Computing. This work was supported by the National Natural Science Foundation of China (62471071, 62501117), the Natural Science Foundation of Sichuan Province (2026NSFSC0422), Joint Innovation Fund of Health Commission of Chengdu and Chengdu University of

Traditional Chinese Medicine(WXLH202402041，WXLH202501011), and Shanghai Innovation Institute.

Contributions

All listed authors meet the ICMJE four criteria for authorship. All authors (D.Y., S.H., Z.D., Y.Y., Y.X., L.R., N.D., F.H., J.S., J.Z., Y.T., L.Z., Y.Z. and H.L.) contributed to the conception and design of the study. D.Y. led the overall framework design and computational methodology. S.H. and Z.D. contributed to the system design and experimental implementation. D.Y., S.H., Y.Y. and Y.X. performed data processing, model development and experimental evaluation. L.R., J.Z., Y.T. and L.Z. contributed to data acquisition and dataset organization. and F.H. provided technical and infrastructure support. Z.D. and J.S. contributed clinical perspectives and domain knowledge to ensure the medical validity of the framework and evaluation protocoled. drafted the manuscript. S.H., Z.D., Y.Y., Y.X., L.R., N.D., F.H., J.S., J.Z., Y.T. and L.Z. contributed to manuscript revision and provided critical feedback. Y.Z. and H.L. supervised the project and provided overall guidance on the study design and interpretation of results. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Ethics declarations

All research procedures were conducted in accordance with the Declaration of Helsinki. Ethical approval was granted by the Ethics Committee of the University of Electronic Science and Technology of China (No.106142025082834175). This study involved a retrospective analysis of health examination reports and related clinical data. Prior to computational analysis and model development, all internal digital data were irreversibly anonymized. The requirement for individual informed consent was waived, as the study did not involve direct patient participation or prospective recruitment.

Additional information

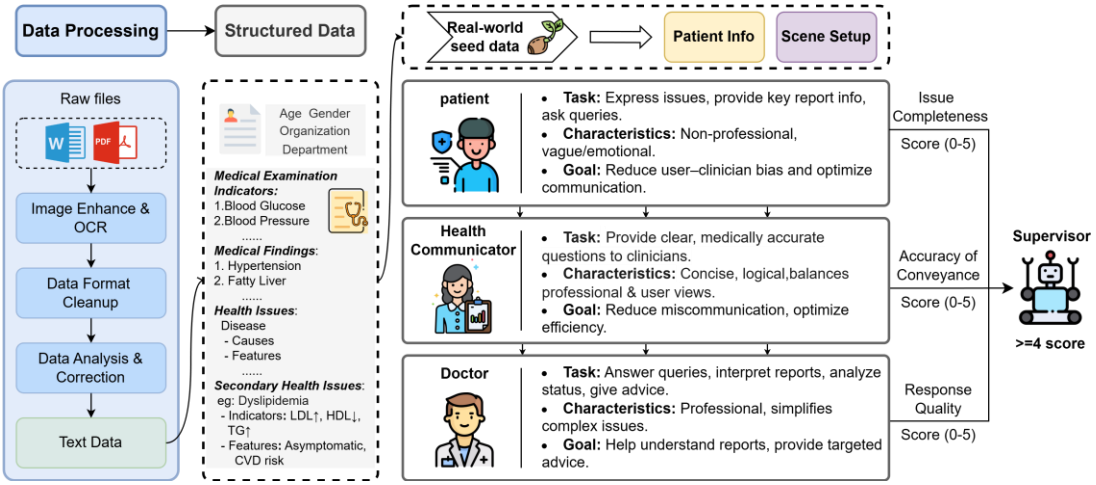
Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Yang Zhang (zhy1001@alu.uestc.edu.cn) or Hao Lin (hlin@uestc.edu.cn).

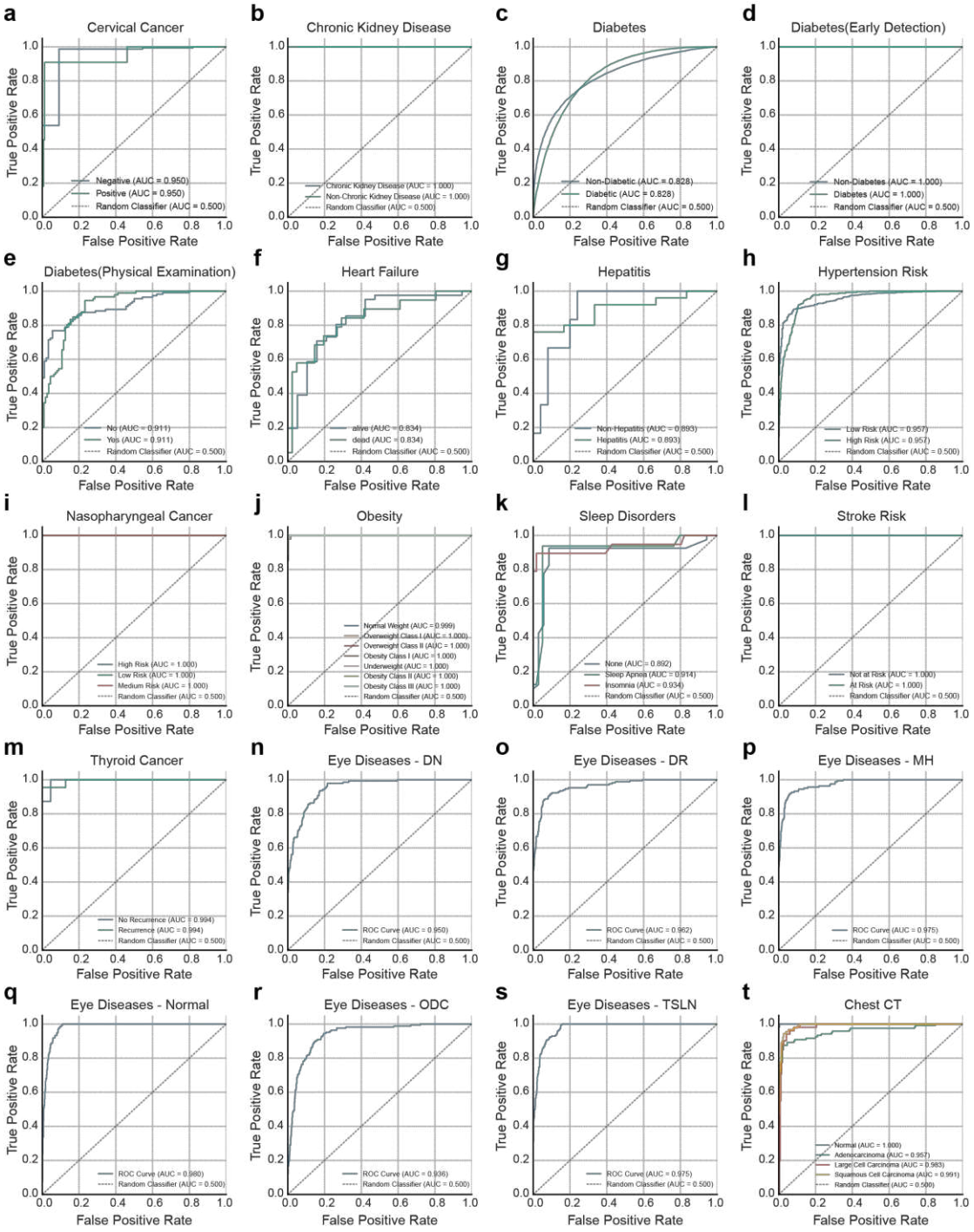
Reprints and permissions information is available at www.nature.com/reprints.

Extended data

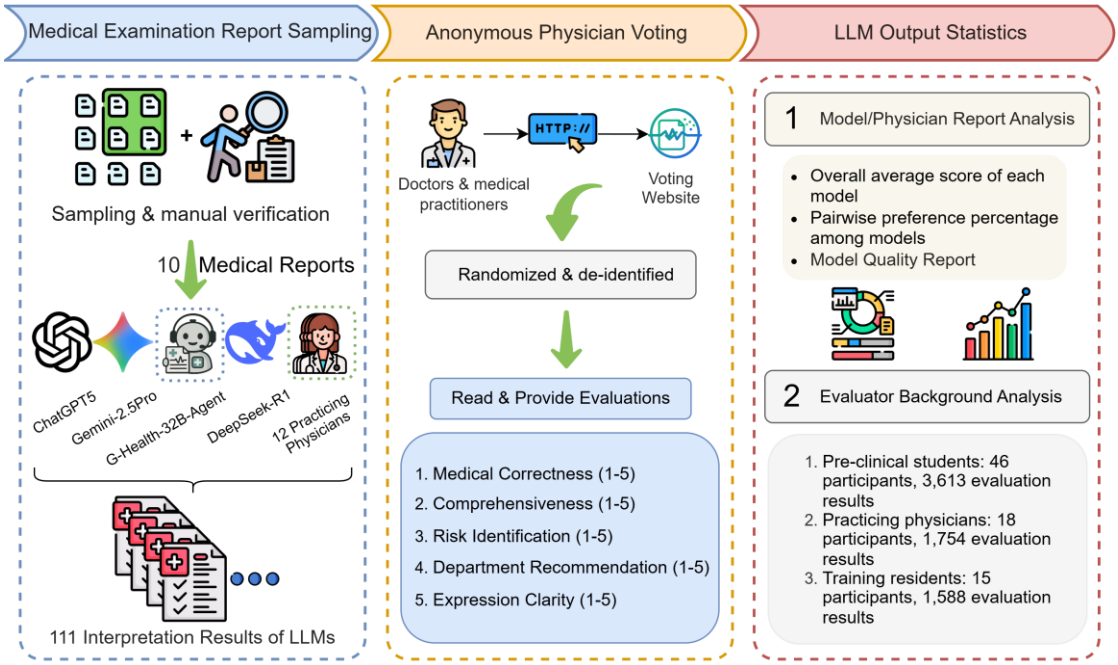
Extended data fig



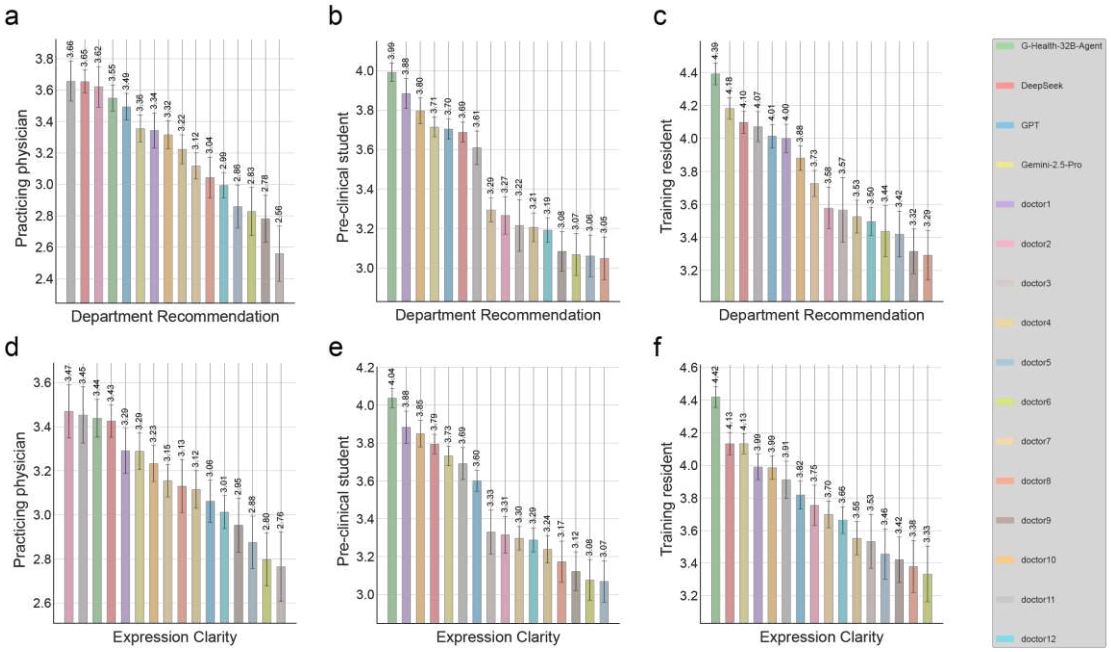
Extended Data Fig. 1 | Overview of the PPMSA-Factory pipeline. Raw clinical files are converted into structured patient records via OCR and text cleaning, then combined with real-world seed data to configure medical scenarios. The system generates multi-role dialogues among the patient, communicator, and doctor, and applies an automatic quality-scoring step to filter for high-quality outputs.



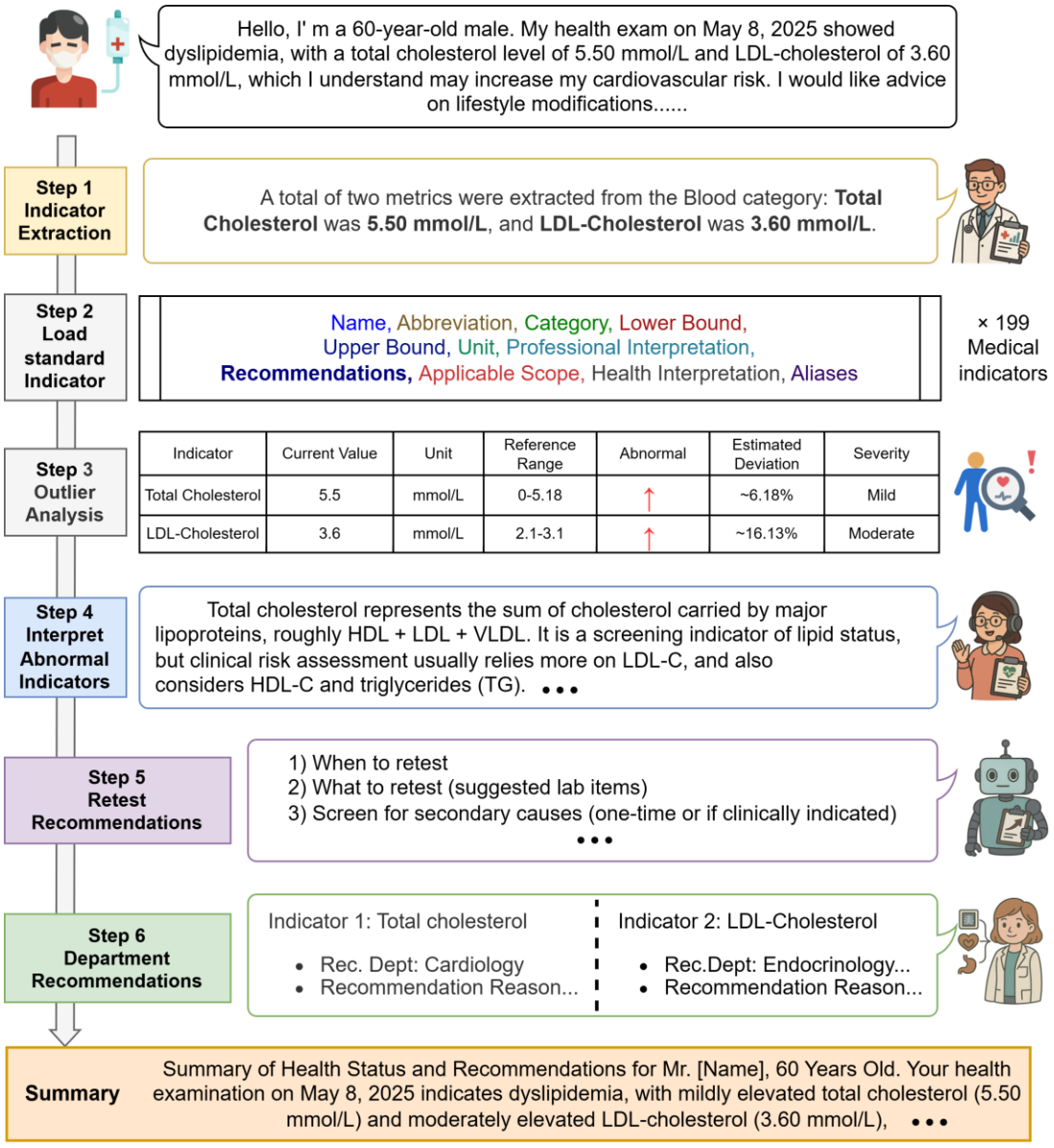
Extended Data Fig. 2 | ROC curves for multi-task disease risk assessment classifiers. Each panel corresponds to one disease/task (for example, cervical cancer, chronic kidney disease, diabetes and its sub-tasks, heart failure, hepatitis, hypertension risk, nasopharyngeal cancer, obesity, sleep disorders, stroke risk, thyroid cancer, ophthalmic disease tasks, and chest CT). For multi-class tasks, one-vs-rest ROC curves are plotted for each class, with the corresponding AUC reported in the legend; the grey diagonal dashed line denotes the random baseline (AUC = 0.5).



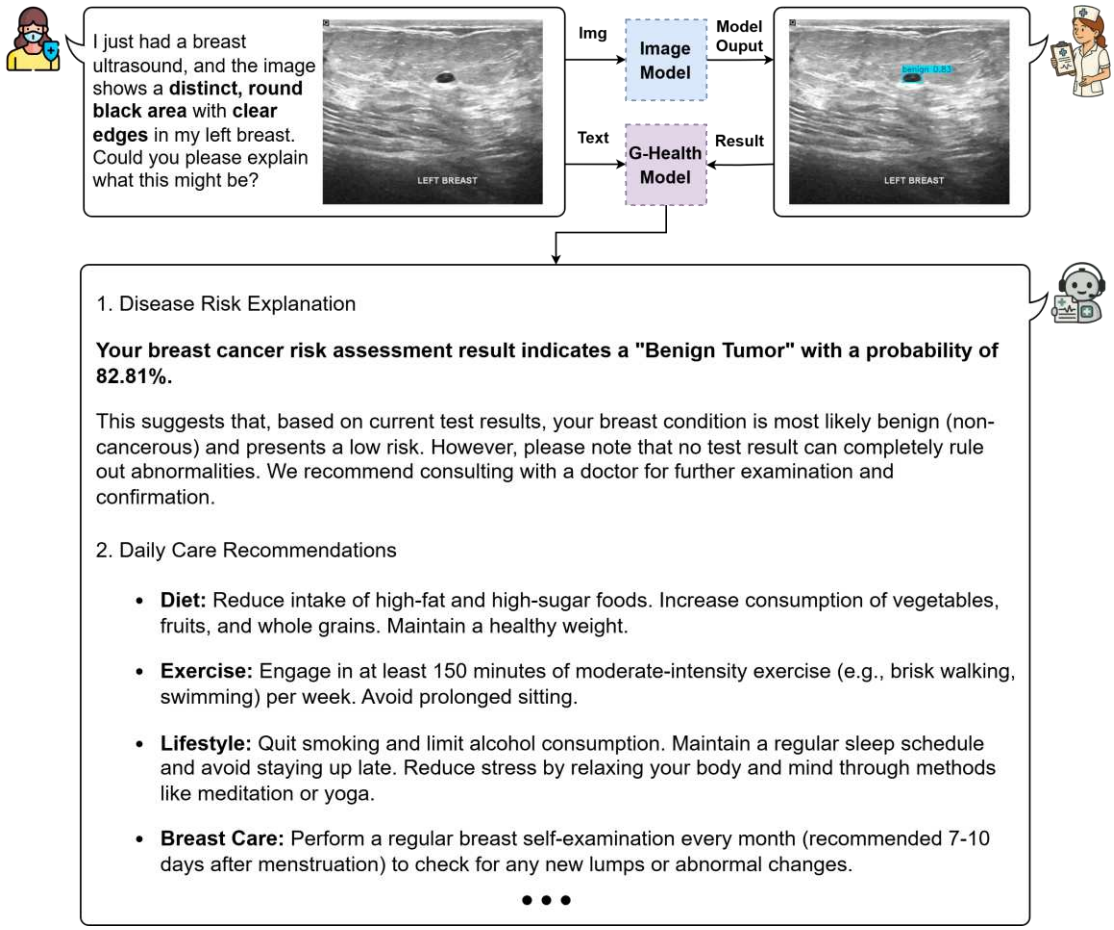
Extended Data Fig. 3 | Human evaluation workflow for health examination report generation. Reports are sampled and manually verified, then randomized and de-identified for blinded physician voting across five criteria (medical correctness, comprehensiveness, risk identification, department recommendation, and expression clarity), with aggregated statistics used to derive model-level performance summaries.



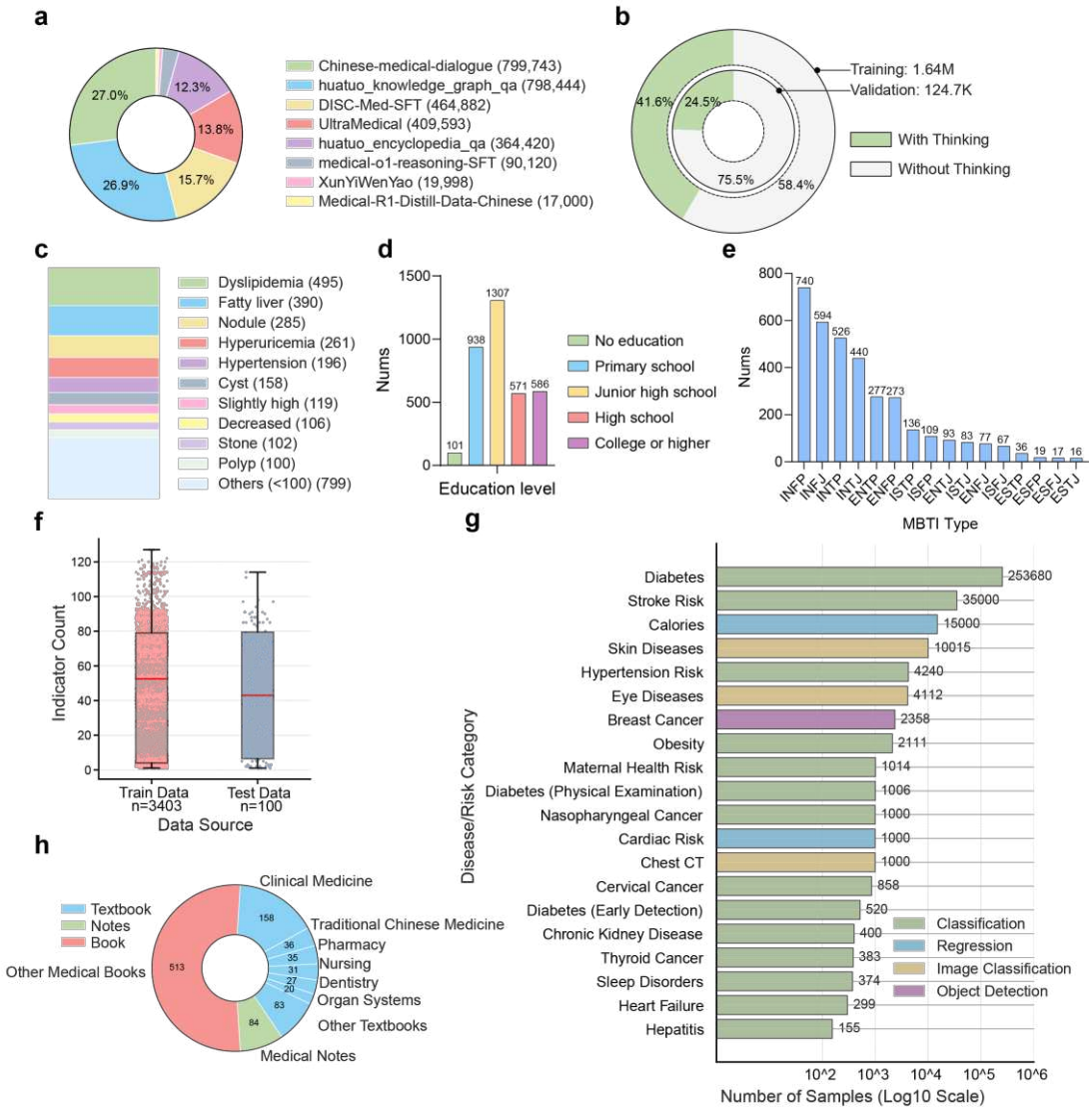
Extended Data Fig. 4 | Rater-group scoring distributions for Department Recommendation and Expression Clarity. Top row shows Department Recommendation and bottom row shows Expression Clarity. Columns correspond to practicing physicians (left), pre-clinical students (middle), and resident trainees (right). Bars indicate the mean Likert score (1–5) for reports produced by each model and by individual physicians; error bars reflect variability across evaluated cases/ratings. Colors denote different models and physician authors.



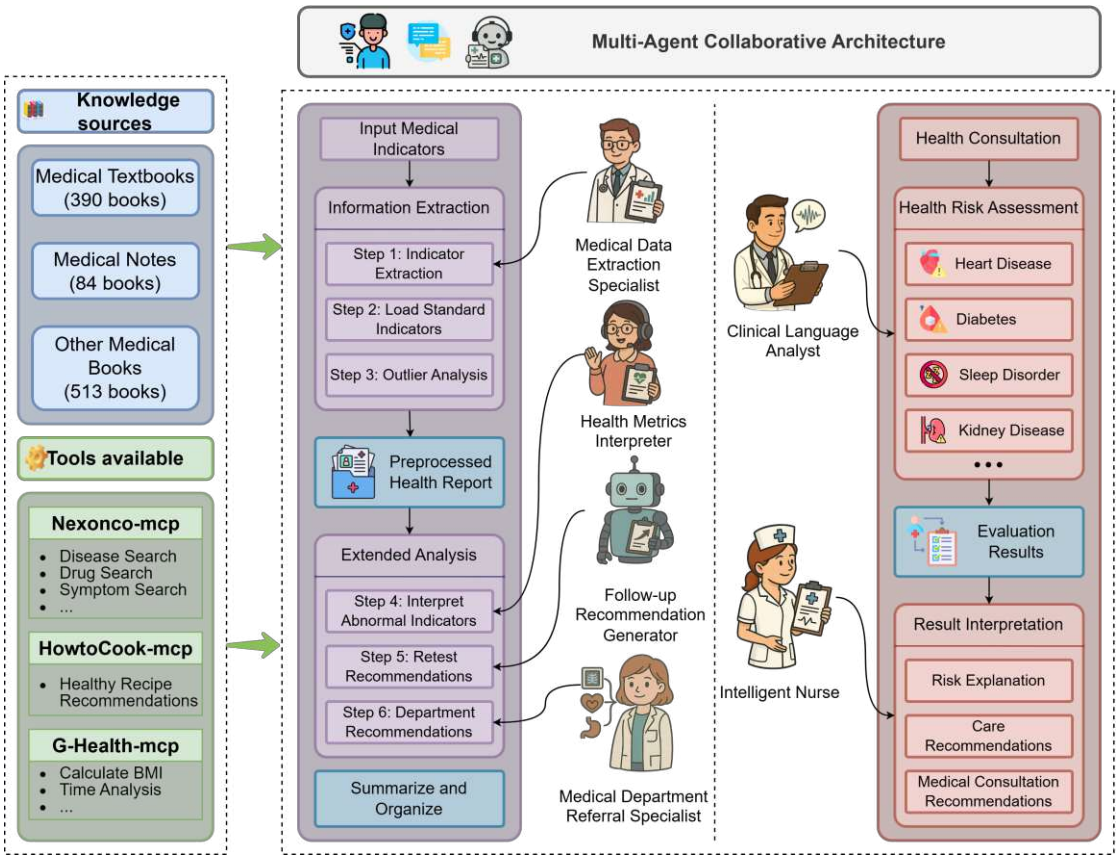
Extended Data Fig. 5 | Six-step pipeline for health examination report understanding and recommendation generation. The system starts from a user query and extracts key examination metrics (Step 1), loads standardized physical examination data (Step 2), performs metric-level analysis (Step 3), and generates interpretations for abnormal indicators with deviation and severity estimates (Step 4). It then proposes follow-up tests (Step 5), recommends appropriate clinical departments with rationales (Step 6), and produces a final patient-oriented summary of health status and actionable recommendations.



Extended Data Fig. 6 | Image-routing and image-based risk assessment pipeline for user-uploaded medical images. After a user uploads an image, an image-routing model first infers the imaging modality and anatomical region (example: breast ultrasound), and then dispatches the input to the corresponding specialized image risk-assessment model (e.g., the yolo model) for lesion localization and risk stratification. The predicted findings (e.g., bounding box and benign/malignant probability) are consolidated and returned to the front-end for visualization and user-facing feedback.



Extended Data Fig. 7 | Overview of data sources and dataset statistics used in G-Health and the PPMSA architecture. a, composition of datasets used for stage-1 SFT. **b**, composition of datasets used for stage-2 DPO training. **c**, distribution of high-frequency health issues appearing in health examination reports. **d**, predefined education-background settings in PPMSA and their sample distribution. **e**, predefined personality-type settings in PPMSA and their sample distribution. **f**, distribution of the number of evaluation metrics across tasks in stage-3 health examination report specific fine-tuning (training and test sets). **g**, dataset size for the risk assessment model (training and validation sets). **h**, category distribution of books used to construct the medical knowledge base.



Extended Data Fig. 8 | Multi-agent collaborative architecture for health management. The system integrates retrieval-augmented medical knowledge sources and MCP-based external tools, and orchestrates specialized agents to support two core workflows: a multi-step pipeline for structured health examination report processing (indicator extraction, standardization, outlier analysis and recommendations) and a risk-assessment pipeline (health consultation, disease-specific assessment, result interpretation and care recommendations).

Supplementary Files

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

- SupplementaryInformation.docx