Supplementary Information

A Statistics and List of Seed Papers for EHRFlowBench 1
B Human Evaluation Instructions 5
C Task Profiling and Governed Experience Memory 6
D Evaluator-Guided Repair Dynamics and Failure Analysis 7
E LLM Judge Agreement and Expert Calibration 9

A Statistics and List of Seed Papers for EHRFlowBench

Supplementary Table 1. Statistics of the candidate paper corpus for EHRFlowBench construction. A total of 51,280 publications were aggregated from seven major AI conferences based on proceedings publicly available as of July 15, 2025; '–' indicates that the corresponding 2025 proceedings were not yet publicly available at the collection cutoff.

Conference	2020	2021	2022	2023	2024	2025	Total
AAAI	1646	1654	1319	1578	2331	3028	11556
IJCAI	779	723	862	851	1048	–	4263
KDD	217	239	253	310	408	185	1612
WWW	219	355	367	402	428	440	2211
NeurIPS	1898	2334	2671	3218	4035	–	14156
ICML	1085	1183	1234	1828	2610	–	7940
ICLR	687	860	1094	1573	2260	3068	9542
Sum	**6531**	**7348**	**7800**	**9760**	**13120**	**6721**	**51280**

Supplementary Table 2. The complete list of 118 seed papers selected for EHRFlowBench task extraction.

The complete list of 118 seed papers selected for EHRFlowBench task extraction.

1. [AAAI, 2020] Generative Adversarial Regularized Mutual Information Policy Gradient Framework for Automatic Diagnosis
2. [AAAI, 2020] AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration
3. [AAAI, 2020] ConCare: Personalized Clinical Feature Embedding via Capturing the Healthcare Context
4. [AAAI, 2020] DeepAlerts: Deep Learning Based Multi-Horizon Alerts for Clinical Deterioration on Oncology Hospital Wards
5. [AAAI, 2020] Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer
6. [AAAI, 2020] Deep Mixed Effect Model Using Gaussian Processes: A Personalized and Reliable Prediction for Healthcare
7. [AAAI, 2020] Learning Conceptual-Contextual Embeddings for Medical Text
8. [AAAI, 2020] Understanding Medical Conversations with Scattered Keyword Attention and Weak Supervision from Responses
9. [AAAI, 2020] ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network

10. [AAAI, 2020] A System for Medical Information Extraction and Verification from Unstructured Text
11. [AAAI, 2021] GRASP: Generic Framework for Health Status Representation Learning Based on Incorporating Knowledge from Similar Patients
12. [AAAI, 2021] Clinical Risk Prediction with Temporal Probabilistic Asymmetric Multi-Task Learning
13. [AAAI, 2021] ESCAPED: Efficient Secure and Private Dot Product Framework for Kernel-based Machine Learning Algorithms with Applications in Healthcare
14. [AAAI, 2021] MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records
15. [AAAI, 2021] MTAAL: Multi-Task Adversarial Active Learning for Medical Named Entity Recognition and Normalization
16. [AAAI, 2021] Clinical Temporal Relation Extraction with Probabilistic Soft Logic Regularization and Global Inference
17. [AAAI, 2022] Diaformer: Automatic Diagnosis via Symptoms Sequence Generation
18. [AAAI, 2022] Context-Aware Health Event Prediction via Transition Functions on Dynamic Disease Graphs
19. [AAAI, 2022] Clustering Interval-Censored Time-Series for Disease Phenotyping
20. [AAAI, 2023] Heterogeneous Graph Learning for Multi-Modal Medical Data Analysis
21. [AAAI, 2023] KerPrint: Local-Global Knowledge Graph Enhanced Diagnosis Prediction for Retrospective and Prospective Interpretations
22. [AAAI, 2023] Multi-Label Few-Shot ICD Coding as Autoregressive Generation with Prompt
23. [AAAI, 2023] Estimating Treatment Effects from Irregular Time Series Observations with Hidden Confounders
24. [AAAI, 2023] Estimating Average Causal Effects from Patient Trajectories
25. [AAAI, 2023] Causal Recurrent Variational Autoencoder for Medical Time Series Generation
26. [AAAI, 2023] MHCCL: Masked Hierarchical Cluster-Wise Contrastive Learning for Multivariate Time Series
27. [AAAI, 2023] Forecasting with Sparse but Informative Variables: A Case Study in Predicting Blood Glucose
28. [AAAI, 2024] PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation
29. [AAAI, 2024] Automatic Radiology Reports Generation via Memory Alignment Network
30. [AAAI, 2024] KG-TREAT: Pre-training for Treatment Effect Estimation by Synergizing Patient Data with Knowledge Graphs
31. [AAAI, 2024] Inducing Clusters Deep Kernel Gaussian Process for Longitudinal Data
32. [AAAI, 2024] ConSequence: Synthesizing Logically Constrained Sequences for Electronic Health Record Generation
33. [AAAI, 2024] IGAMT: Privacy-Preserving Electronic Health Record Synthesization with Heterogeneity and Irregularity
34. [AAAI, 2024] IVP-VAE: Modeling EHR Time Series with Initial Value Problem Solvers
35. [AAAI, 2024] DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency
36. [AAAI, 2024] Large Language Models Are Clinical Reasoners: Reasoning-Aware Diagnosis Framework with Prompt-Generated Rationales
37. [AAAI, 2024] Collaborative Synthesis of Patient Records through Multi-Visit Health State Inference
38. [AAAI, 2025] DAMPER: A Dual-Stage Medical Report Generation Framework with Coarse-Grained MeSH Alignment and Fine-Grained Hypergraph Matching
39. [AAAI, 2025] Memorize and Rank: Elevating Large Language Models for Clinical Diagnosis Prediction
40. [AAAI, 2025] DECT: Harnessing LLM-assisted Fine-Grained Linguistic Knowledge and Label-Switched and Label-Preserved Data Generation for Diagnosis of Alzheimer's Disease
41. [AAAI, 2025] Medical Manifestation-Aware De-Identification
42. [ICLR, 2024] GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs.
43. [ICLR, 2024] Diagnosing Transformers: Illuminating Feature Spaces for Clinical Decision-Making.
44. [ICLR, 2024] A Flexible Generative Model for Heterogeneous Tabular EHR with Missing Modality.
45. [ICLR, 2024] Graph Transformers on EHRs: Better Representation Improves Downstream Performance.
46. [ICLR, 2025] Small Models are LLM Knowledge Triggers for Medical Tabular Prediction.
47. [ICLR, 2025] Efficiently Democratizing Medical LLMs for 50 Languages via a Mixture of Language Family Experts.
48. [ICML, 2020] BoXHED: Boosted eXact Hazard Estimator with Dynamic covariates.
49. [ICML, 2020] Temporal Phenotyping using Deep Predictive Clustering of Disease Progression.
50. [ICML, 2020] DeepCoDA: personalized interpretability for compositional health data.

51. [ICML, 2021] Neighborhood Contrastive Learning Applied to Online Patient Monitoring.
52. [ICML, 2022] Locally Sparse Neural Networks for Tabular Biomedical Data.
53. [ICML, 2022] Learning of Cluster-based Feature Importance for Electronic Health Record Time-series.
54. [ICML, 2023] Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling.
55. [ICML, 2023] Sequential Multi-Dimensional Self-Supervised Learning for Clinical Time Series.
56. [ICML, 2024] Contrastive Learning for Clinical Outcome Prediction with Partial Data Sources.
57. [ICML, 2024] Reservoir Computing for Short High-Dimensional Time Series: an Application to SARS-CoV-2 Hospitalization Forecast.
58. [ICML, 2024] Exploiting Negative Samples: A Catalyst for Cohort Discovery in Healthcare Analytics.
59. [ICML, 2024] ED-Copilot: Reduce Emergency Department Wait Time with Language Model Diagnostic Assistance.
60. [ICML, 2024] ProtoGate: Prototype-based Neural Networks with Global-to-local Feature Selection for Tabular Biomedical Data.
61. [IJCAI, 2020] A Label Attention Model for ICD Coding from Clinical Text
62. [IJCAI, 2020] The Graph-based Mutual Attentive Network for Automatic Diagnosis
63. [IJCAI, 2020] Automatic Emergency Diagnosis with Knowledge-Based Tree Decoding
64. [IJCAI, 2020] Learning Latent Forests for Medical Relation Extraction
65. [IJCAI, 2020] Generalized Zero-Shot Text Classification for ICD Coding
66. [IJCAI, 2021] Cooperative Joint Attentive Network for Patient Outcome Prediction on Irregular Multi-Rate Multivariate Health Data
67. [IJCAI, 2021] AMA-GCN: Adaptive Multi-layer Aggregation Graph Convolutional Network for Disease Prediction
68. [IJCAI, 2021] Collaborative Graph Learning with Auxiliary Text for Temporal Event Prediction in Healthcare
69. [IJCAI, 2021] Multi-series Time-aware Sequence Partitioning for Disease Progression Modeling
70. [IJCAI, 2021] A Novel Sequence-to-Subgraph Framework for Diagnosis Classification
71. [IJCAI, 2022] Cumulative Stay-time Representation for Electronic Health Records in Medical Event Time Prediction
72. [IJCAI, 2022] ``My nose is running.'' ``Are you also coughing?'': Building A Medical Diagnosis Agent with Interpretable Inquiry Logics
73. [IJCAI, 2022] Chronic Disease Management with Personalized Lab Test Response Prediction
74. [IJCAI, 2022] Data-Efficient Algorithms and Neural Natural Language Processing: Applications in the Healthcare Domain
75. [IJCAI, 2023] Hierarchical Apprenticeship Learning for Disease Progression Modeling
76. [IJCAI, 2023] A Diffusion Model with Contrastive Learning for ICU False Arrhythmia Alarm Reduction
77. [IJCAI, 2023] VecoCare: Visit Sequences-Clinical Notes Joint Learning for Diagnosis Prediction in Healthcare Data
78. [IJCAI, 2024] Predictive Modeling with Temporal Graphical Representation on Electronic Health Records
79. [IJCAI, 2024] Multi-TA: Multilevel Temporal Augmentation for Robust Septic Shock Early Prediction
80. [IJCAI, 2024] MediTab: Scaling Medical Tabular Data Predictors via Data Consolidation, Enrichment, and Refinement
81. [KDD, 2020] Hierarchical Attention Propagation for Healthcare Representation Learning
82. [KDD, 2020] INPREM: An Interpretable and Trustworthy Predictive Model for Healthcare
83. [KDD, 2020] DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation
84. [KDD, 2020] HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records
85. [KDD, 2020] Identifying Sepsis Subphenotypes via Time-Aware Multi-Modal Auto-Encoder
86. [KDD, 2020] HOLMES: Health OnLine Model Ensemble Serving for Deep Learning Models in Intensive Care Units
87. [KDD, 2022] M3Care: Learning with Missing Modalities in Multimodal Healthcare Data
88. [KDD, 2023] Granger Causal Chain Discovery for Sepsis-Associated Derangements via Continuous-Time Hawkes Processes
89. [KDD, 2023] MedLink: De-Identified Patient Health Record Linkage
90. [KDD, 2023] Warpformer: A Multi-scale Modeling Approach for Irregular Clinical Time Series
91. [KDD, 2024] FlexCare: Leveraging Cross-Task Synergy for Flexible Multimodal Healthcare Prediction
92. [KDD, 2024] ProtoMix: Augmenting Health Status Representation Learning via Prototype-based Mixup

93. [KDD, 2024] Synthesizing Multimodal Electronic Health Records via Predictive Diffusion Models
94. [NeurIPS, 2020] Learning to Select Best Forecast Tasks for Clinical Outcome Prediction
95. [NeurIPS, 2021] Medical Dead-ends and Learning to Identify High-Risk States and Treatments
96. [NeurIPS, 2021] Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation
97. [NeurIPS, 2021] Closing the loop in medical decision support by understanding clinical decision-making: A case study on organ transplantation
98. [NeurIPS, 2023] Contrast Everything: A Hierarchical Contrastive Framework for Medical Time-Series
99. [NeurIPS, 2023] Temporal Causal Mediation through a Point Process: Direct and Indirect Effects of Healthcare Interventions
100. [NeurIPS, 2023] Towards Semi-Structured Automatic ICD Coding via Tree-based Contrastive Learning
101. [NeurIPS, 2024] Medformer: A Multi-Granularity Patching Transformer for Medical Time-Series Classification
102. [NeurIPS, 2024] SMART: Towards Pre-trained Missing-Aware Model for Patient Health Status Prediction
103. [NeurIPS, 2024] Knowledge-Empowered Dynamic Graph Network for Irregularly Sampled Medical Time Series
104. [NeurIPS, 2024] Trajectory Flow Matching with Applications to Clinical Time Series Modelling
105. [NeurIPS, 2024] A teacher-teacher framework for clinical language representation learning
106. [NeurIPS, 2024] Automated Multi-Task Learning for Joint Disease Prediction on Electronic Health Records
107. [WWW, 2020] Text-to-SQL Generation for Question Answering on Electronic Medical Records
108. [WWW, 2020] StageNet: Stage-Aware Neural Networks for Health Risk Prediction
109. [WWW, 2020] CLARA: Clinical Report Auto-completion
110. [WWW, 2020] Learning Contextualized Document Representations for Healthcare Answer Retrieval
111. [WWW, 2020] DyCRS: Dynamic Interpretable Postoperative Complication Risk Scoring
112. [WWW, 2021] Online Disease Diagnosis with Inductive Heterogeneous Graph Convolutional Networks
113. [WWW, 2021] Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis
114. [WWW, 2023] SeqCare: Sequential Training with External Medical Knowledge Graph for Diagnosis Prediction in Healthcare Data
115. [WWW, 2023] Cross-center Early Sepsis Recognition by Medical Knowledge Guided Collaborative Learning for Data-scarce Hospitals
116. [WWW, 2025] ColaCare: Enhancing Electronic Health Record Modeling through Large Language Model-Driven Multi-Agent Collaboration
117. [WWW, 2025] MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot
118. [WWW, 2025] Towards Multi-resolution Spatiotemporal Graph Learning for Medical Time Series Classification

B Human Evaluation Instructions

Supplementary Table 3. Instructions for blinded human preference evaluation.

Online instructions for blinded preference evaluation.

We sincerely thank you for contributing your time and expertise to this blinded evaluation. In this study, you will compare anonymous AI-generated reports for EHRFlowBench tasks. Your role is to review the materials for each task and select the method you most prefer overall, or choose None/Others if no displayed report should be preferred.

The evaluation is conducted through our online platform at https://healthflow.medx-pku.com/evaluation and follows a simple workflow:
(1) Review the EHRFlowBench task and the corresponding reference report.
(2) Review the anonymized system reports in their respective tabs.
(3) Select the report you most prefer overall in the decision panel on the right, or choose None/Others if all displayed reports are unsatisfactory.
(4) Save your choice and proceed to the next task. After all assigned tasks are completed, export the recorded results if prompted by the platform.

You will complete the evaluation within a single web interface.

Please open the platform and follow the steps below.

1. Review the task context.
   Start with the Task tab to read the EHRFlowBench task description. Then review the Reference tab so you understand the intended objective and the expected level of methodological rigor.

2. Compare anonymized reports.
   Open each anonymized submission tab one at a time. The reference report and each anonymized baseline report are provided in separate tabs. System identities are hidden and the tab order is randomized to reduce bias. Review the report in each tab before making your decision.

3. Make an overall preference judgment.
   After reviewing the tabs, select the anonymized submission you most prefer in the decision panel on the right. If no displayed report should be preferred, select None/Others.

4. Save and continue.
   After making your selection, save the current task and proceed to the next one until you finish the assigned evaluation set.

When making your overall preference judgment, you may use the following aspects as guidance. These are reference considerations rather than separate scoring dimensions:

1. Factual correctness and safety.
   Consider whether the medical statements, data interpretation, and conclusions are correct, whether the response avoids misleading claims, and whether it remains methodologically sound for the stated task.

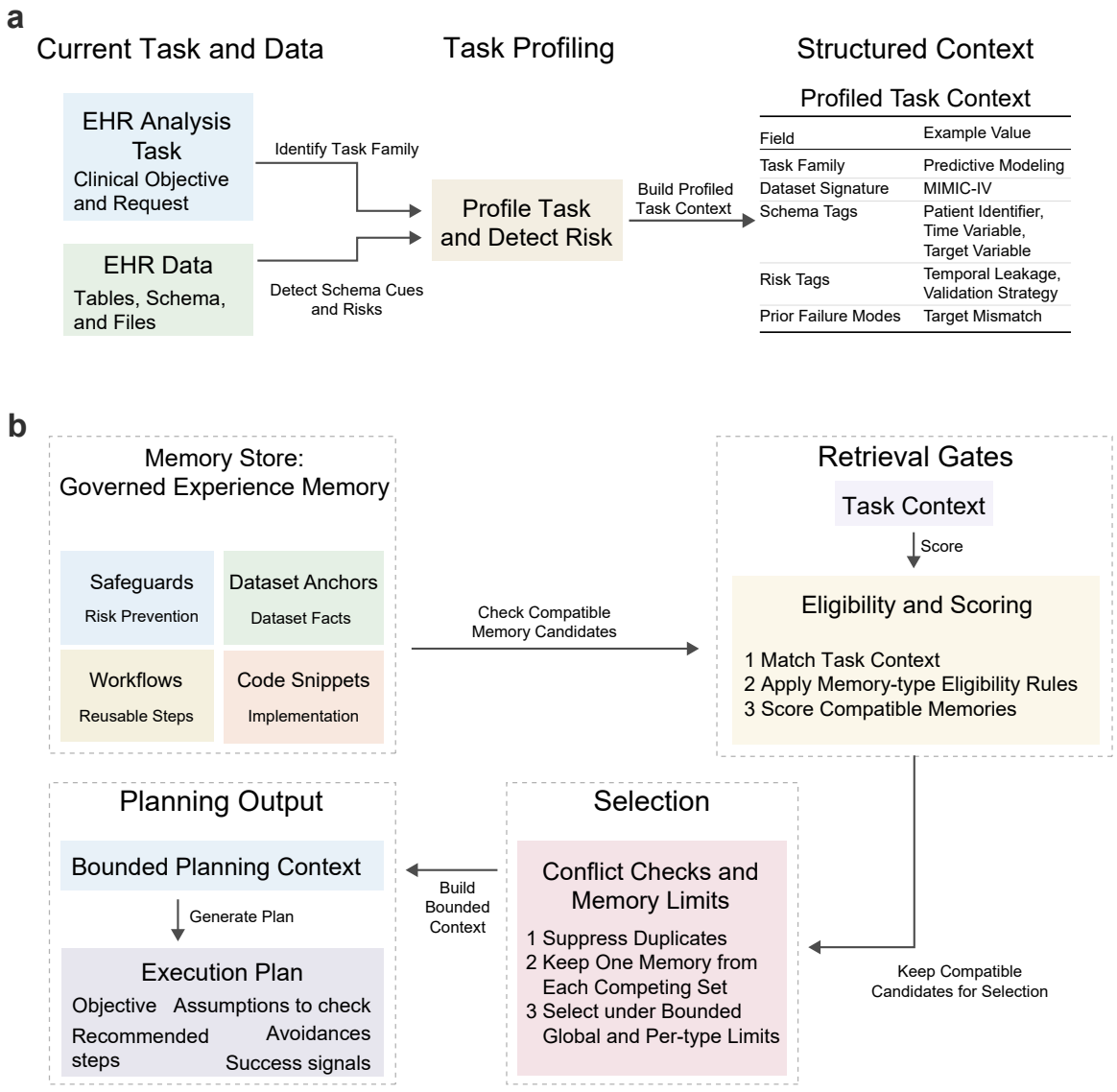
2. Completeness.
   Consider whether the response addresses the main requirements of the task, covers the important analytical steps, and avoids major omissions.

3. Clarity and logical organization.
   Consider whether the response is clear, well-structured, and easy to follow, with a coherent explanation of the analysis or conclusions.

4. Practical clinical and analytical usefulness.
   Consider whether the response is grounded in realistic EHR analysis practice and whether the resulting report or reasoning would be useful to a clinician or health-data researcher.

Please use your expert judgment to balance these considerations and make a single overall preference decision for each task.

C Task Profiling and Governed Experience Memory



Supplementary Figure 1. EHR-aware task profiling and governed experience-memory retrieval in HealthFlow. **(a)** Task profiling converts the current EHR analysis request and available EHR data into a structured context containing the task family, dataset signature, schema tags, risk tags, and prior failure modes. **(b)** Governed experience memory stores safeguards, dataset anchors, workflows, and code snippets, then applies task-context gates, compatibility scoring, conflict checks, and memory limits to construct a bounded planning context for the execution plan.

Supplementary Figure 1 expands the task-context and governed-memory modules summarized in the main framework schematic. The task profile provides dataset- and schema-aware constraints for planning, while the retrieval gates restrict reusable experience to memory items that are compatible with the active task, dataset, schema cues, and risk tags.

D Evaluator-Guided Repair Dynamics and Failure Analysis

To characterize the operational value of evaluator-guided repair beyond aggregate benchmark scores, we analyze the terminal attempt distribution for the OpenCode-based ablation runs. Specifically, we record whether each task is accepted by the evaluator on the first, second, or third attempt, or remains a final failure after the full retry budget is exhausted.

Dataset	Variant	Attempt 1 success	Attempt 2 success	Attempt 3 success	Final fail
EHRFlowBench	Bare	85	–	–	15
	+ Harness	58	22	20	0
	+ Self-Evolving	46	27	27	0
	HealthFlow	39	35	26	0
MedAgentBoard	Bare	100	–	–	0
	+ Harness	98	0	0	2
	+ Self-Evolving	90	3	1	6
	HealthFlow	93	4	3	0
MedAgentsBench	Bare	100	–	–	0
	+ Harness	100	0	0	0
	+ Self-Evolving	100	0	0	0
	HealthFlow	90	10	0	0
HLE	Bare	100	–	–	0
	+ Harness	98	0	2	0
	+ Self-Evolving	97	0	3	0
	HealthFlow	95	1	4	0
CureBench	Bare	100	–	–	0
	+ Harness	100	0	0	0
	+ Self-Evolving	100	0	0	0
	HealthFlow	87	13	0	0

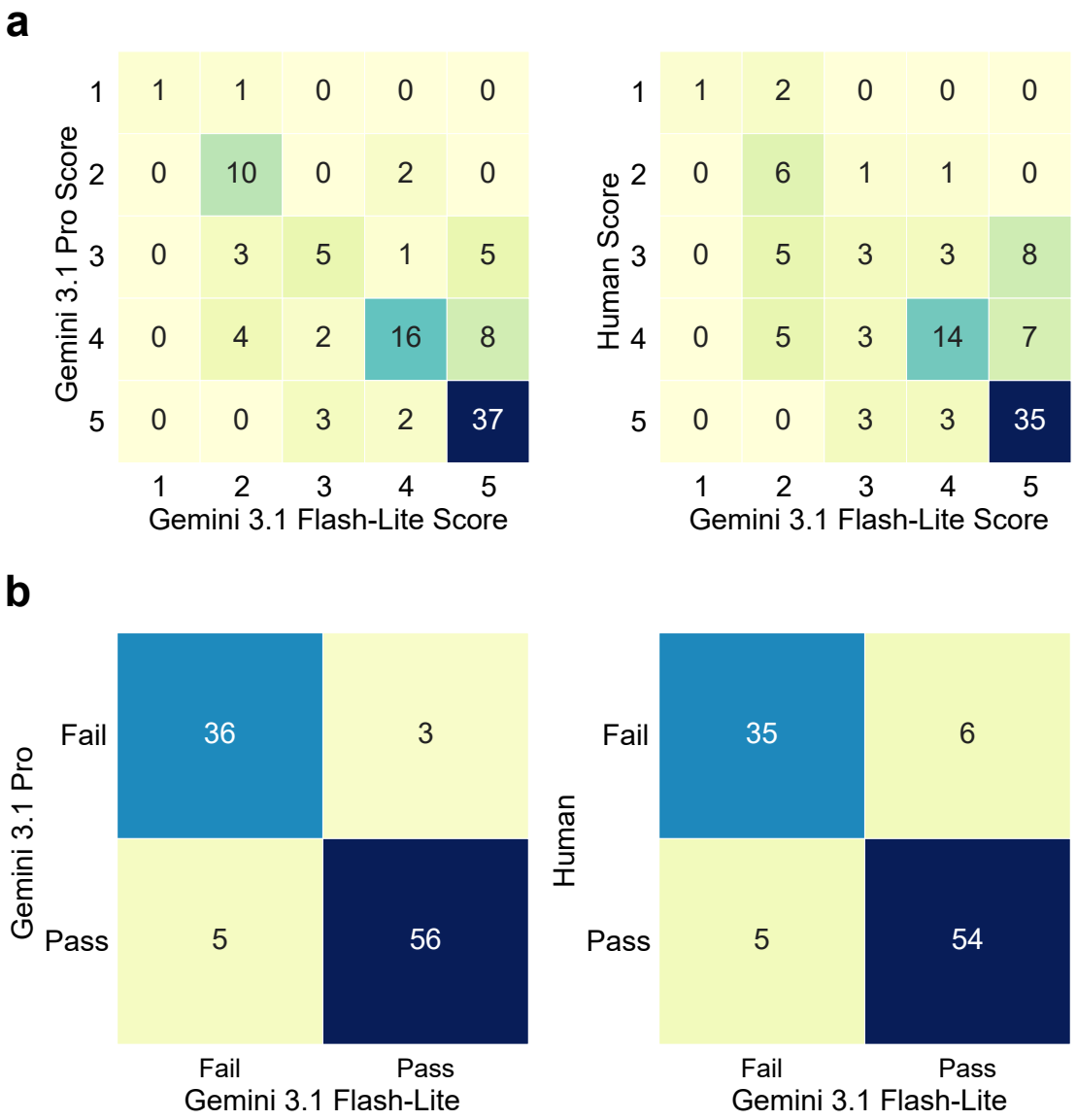
Supplementary Table 4. Attempt-level repair and terminal-failure distribution across five benchmarks, reported as task counts. All variants use OpenCode as the underlying coding agent, and each benchmark–variant row contains 100 tasks. Counts indicate the number of tasks accepted by the evaluator on the first, second, or third attempt, or reaching a terminal failure state. The notation "–" marks retry stages that were not executed, so no performance is reported. Gray rows mark the executor-only Bare setting, and blue rows mark the full HealthFlow configuration.

In Supplementary Table 4, "–" indicates that the corresponding retry stage was not executed and therefore no performance is reported. This occurs for the Bare setting, which runs the OpenCode executor once without evaluator feedback. For harnessed variants, Final Fail includes tasks that

remain below the evaluator's success threshold after three attempts, or tasks where required assets were not generated. If no attempt meets the evaluator's threshold, we evaluate the third-attempt artifacts when available; if required assets such as the report are missing, the task is scored 0.

The clearest effect appears on EHRFlowBench: Bare runs OpenCode once and leaves 15 final failures, while + Harness fixes 42 tasks through the second or third attempt and reduces final failures to 0 (Supplementary Table 4). The later variants add self-evolving memory and task-matched tool use, so the evaluator has more context and can ask for another attempt when an output is missing EHR-specific requirements such as required reports, validity checks, or generated artifacts. This is why some HealthFlow tasks pass on later attempts instead of the first attempt. HealthFlow still ends with 0 final failures on all five benchmarks, including MedAgentBoard, where + Harness and + Self-Evolving leave 2 and 6 final failures.

E LLM Judge Agreement and Expert Calibration



Supplementary Figure 2. Agreement analysis for the Gemini 3.1 Flash-Lite judge across score-based and binary evaluation settings. (a) EHRFlowBench score calibration ($n = 100$). Confusion matrices compare Gemini 3.1 Flash-Lite against Gemini 3.1 Pro and against independent expert ratings on the 1-to-5 report-quality scale. Most disagreements lie on the first off-diagonal bands, indicating that score differences are usually limited to adjacent bins rather than large shifts. **(b) MedAgentBoard binary calibration ($n = 100$).** Confusion matrices compare Gemini 3.1 Flash-Lite against Gemini 3.1 Pro and against independent expert ratings on binary pass/fail judgments. Pairwise agreement is high in both comparisons, supporting Gemini 3.1 Flash-Lite as a lower-cost proxy for the judge role in the binary evaluation setting.

As shown in panel a of Supplementary Figure 2, the judgments assigned by Gemini 3.1 Flash-Lite remain closely aligned with those from Gemini 3.1 Pro in the ordinal EHRFlowBench setting. Disagreements are concentrated on adjacent score bins rather than large score shifts, indicating that

Flash-Lite preserves the same coarse ranking behavior as the stronger Gemini 3.1 Pro judge even when exact scores differ. This supports our use of Gemini 3.1 Flash-Lite as a lower-cost proxy for fine-grained rubric scoring.

As shown in panel b of Supplementary Figure 2, binary agreement on MedAgentBoard reaches 92/100 against Gemini 3.1 Pro and 89/100 against independent expert ratings, with the main diagonal dominating both pass/fail confusion matrices. This indicates that Flash-Lite remains consistent with stronger-model and expert judgments in the binary evaluation setting.

This calibration experiment is separate from the 12-expert blinded preference study reported in the main manuscript, whose purpose is comparative system assessment rather than judge calibration.