TAF - Health AI Benchmark dataset

Submission references: Intelehealth guidelines | Health AI final-submission outline

Table of Contents

1. Overview

What ?	Details
Organization	Intelehealth http://intelehealth.org/
Use case / problem addressed	Ayu 2.0 is an AI Clinical Decision Support System to assist doctors and frontline health workers supporting last-mile telemedicine. Use-cases addressed are differential diagnosis and treatment planning in low resource rural settings.
Target population	Rural communities in India
Deployment stage	Pilot

This benchmark evaluates the diagnostic and treatment-planning performance of Ayu AI clinical decision support models against clinician-vetted real-life patient visits from 1,000 de-identified teleconsultation cases collected in the Peth and Surgana blocks of Nashik district, Maharashtra. The present submission focuses primarily on differential-diagnosis performance; treatment-planning evaluation is described where source information is available.

The benchmark is designed for text-only, single-diagnosis consultation cases in which each encounter has one confirmed target diagnosis. Phase 1 used a single clinician review of the diagnosis against the patient's history and vital signs. A multi-clinician Phase 2 annotation exercise is in progress and is intended to support stronger consensus-based ground-truth labels.

The dataset supports iterative prompt optimization and model-comparison studies across closed and open-source large language models. A model developed through this evaluation pipeline has been piloted in Peth and Surgana. The benchmark is also used to track performance across optimization cycles and to compare clinical reasoning based on routinely available patient history and vital-signs data.

Phase 2 uses the same set of 1,000 case vignettes as Phase 1, but with a revised methodology for establishing ground truth for both the diagnosis and the treatment plan.

In Phase 1, each case was assigned a single ground-truth diagnosis by a single physician by looking at the diagnosis provided by the treating physician. Then a second more senior physician reviewed the work of the first physician for quality assurance, and discussed any differences to arrive at a final Ground truth. Learnings from that exercise showed that a single ground truth is not practically feasible in real-world clinical settings: doctors often differ in how they phrase a diagnosis, and two doctors can independently arrive at diagnoses that are semantically different but clinically equivalent for the same case. For example, a patient presenting with cough and sneezing might be diagnosed by one doctor as an Upper Respiratory Tract Infection and by another as Acute Rhinitis. Both are correct: the former is a broader diagnosis and the latter more specific, and both would lead to the same treatment plan. This kind of variation exists across the full spectrum of possible diagnoses.

Phase 2 addresses this by labelling each case with an array of plausible diagnoses rather than a single ground-truth diagnosis. The same logic extends to treatment plans: different doctors may prescribe different, equally valid combinations of medicines, investigations, and referral advice, so Phase 2 also asks annotating doctors to record all plausible options as an array for the treatment plan. Phase 2 additionally accounts for cases with more than one true diagnosis, which Phase 1's single-diagnosis restriction excluded by design.

Ground truth for Phase 2 is set through a structured consensus process. The 1,000 cases are split into two batches of 500, each assigned to a cohort of three doctors (six doctors in total across the dataset). Within a cohort, each doctor first tags the diagnosis array and treatment plan array independently. The three doctors then review each other's responses and revise their arrays iteratively until they converge on one agreed array for diagnosis and one for treatment plan; this consensus array becomes the case's ground truth. Once ground truth is finalized, the same three doctors review the AI's DDx and Tx suggestions for each case and rate them for accuracy, appropriateness, and comprehensiveness against that ground truth.

2. Dataset

Dataset access and components

Phase 1 dataset: TAF Intelehealth_Health_AI_Phase_1_dataset

[Once DPO clears - will be shared for access to TAF]

Pending

Add the Phase 2 dataset link and confirm the permitted release scope.

The phase 2 data set is being created, and will be available to us by next month. This is a time consuming process as consensus building is an iterative process that requires extensive deliberation.

2.1 Data source and collection

- **Where the data comes from:** Real patients who received teleconsultations under Arogya Sampada Program implemented by Intelehealth across 30 villages in Peth and Surgana block, in Nashik district, Maharashtra.
- **Time period covered:** The visits fall in these clusters: Jul–Sep 2022, Apr–Sep 2023, and Apr–Jun 2025; with the largest single month being June 2025 (235 visits). The window opens 25 Jul 2022 and the last visits start on 1 Jul 2025.
- **Collection method:** Sampled from teleconsultations conducted through Community Health Worker (CHW) Visits.

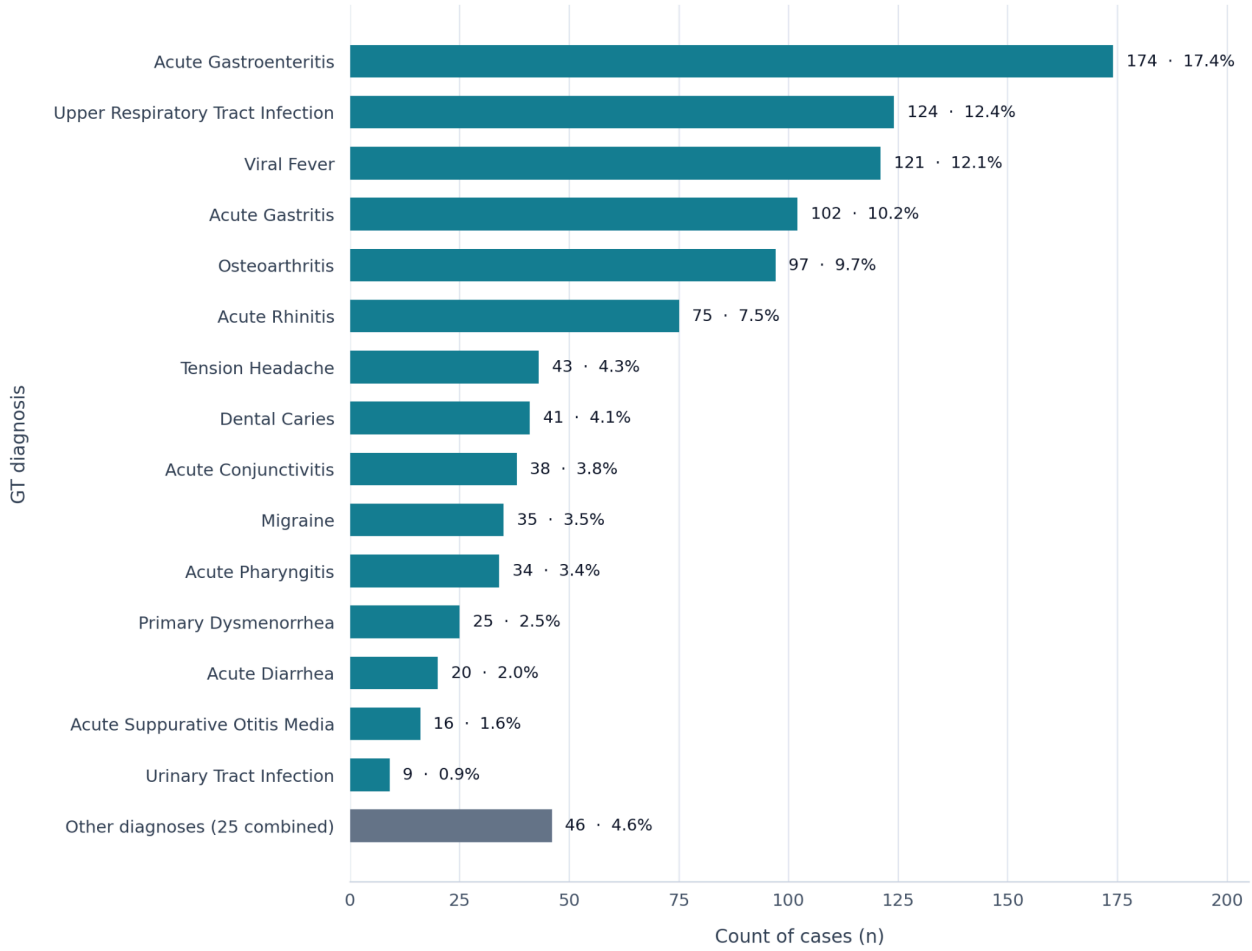
2.2 Data composition

- **Size:** 1,000 cases represented as 1,000 rows and **73** columns (approximately 1.8 MB). Other 68 columns are AI generations and annotations by human and MAI LLM judges.
- **Modality (input and output):** Text.
- **Language(s) and dialect(s):** Community Health Workers use Ayu in Marathi to input clinical history, on the doctor's end using structured patient history taking forms (MCQs, free text entry). The patient history is visible in English as each structured history input field visible to the CHW in Marathi is mapped to English. Additionally, Marathi phrases in free text entry are captured based on the patient's notes recorded in the EHR records.
- **Unique diagnoses represented:** 40.

Top-5 diagnoses (Ground-Truth Diagnosis) distribution:

GT Diagnosis	Count	Percentage
Acute Gastroenteritis	174	17.40%
Upper Respiratory Tract Infection	124	12.40%
Viral Fever	121	12.10%

Acute Gastritis	102	10.20%
Osteoarthritis	97	9.70%

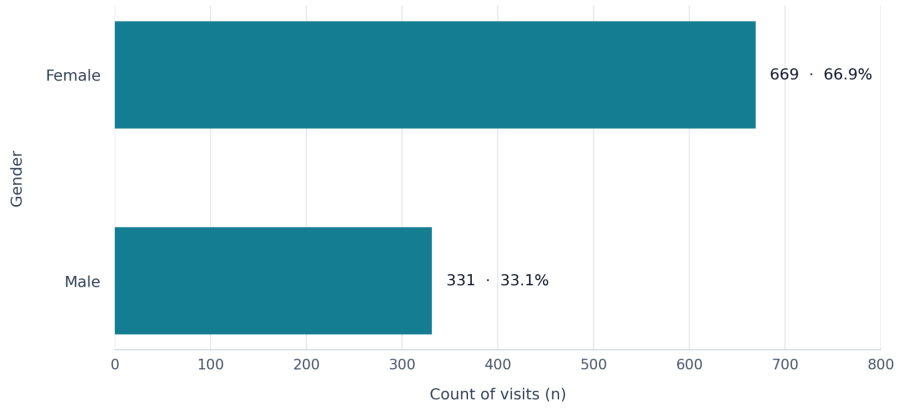


Facility / geography coverage:

Facility coverage is limited to the Peth and Surgana blocks of Nashik district, Maharashtra.

Gender coverage

Gender	Visits	Share
Female	669	66.9%
Male	331	33.1%



Age ranges from 0 to 98 years, with a mean of 34.3 years and a median of 32 years. Age is recorded in completed years. Four infants are younger than one year, but the source does not provide month- or day-level granularity.

Clinical age band	Count	Percentage in distribution
Neonatal (<1 mo)	0	0.0%
Infant (<1 yr)	4	0.4%
Toddler (1–2 yr)	35	3.5%
Preschool (3–5 yr)	64	6.4%
School-age (6–11 yr)	129	12.9%
Adolescent (12–17 yr)	84	8.4%
Adult (18–59 yr)	492	49.2%
Elderly (60+ yr)	192	19.2%

Pediatric cases (<18 years) account for 316 encounters (31.6%); adults (18 years and older) account for 684 encounters (68.4%).

Age distribution by gender

Age band	Female	Male	Total
Infant (<1 yr)	2	2	4
Toddler (1–2)	21	14	35
Preschool (3–5)	32	32	64
School-age (6–11)	68	61	129
Adolescent (12–17)	54	30	84
Adult (18–59)	348	144	492
Elderly (60+)	144	48	192
Total	669	331	1,000

- **Cleaning decisions - kept vs removed:**
  Kept: single-diagnosis cases only (each case resolves to one confirmed target diagnosis).
  Removed: multiple-diagnosis cases and diagnoses that require an image as evidence to some extent.
  Ground-truth diagnosis (Dx) coding and cleaning was orchestrated by the clinical team.
  Exclusions were selected by the clinical team.
- **Synthetic augmentation:** None included. The dataset is text-only real-visit data.
- **Sampling logic:** Cases were selected to mirror the case mix seen in the communities served by the Arogya Sampada project, so the dataset's proportions of health issues broadly reflect their real-world prevalence in those communities. Two filters were then applied on top of this representative sampling: restriction to a single confirmed diagnosis, and exclusion of cases where image-based evidence is an important component of the diagnostic and treatment planning process (for example, dermatological conditions), since this dataset is text-only. Because image-dependent cases were dropped rather than replaced with other cases, some of the remaining diagnosis categories may be modestly over-represented relative to their true community prevalence, simply because removing the image-dependent share of visits increases everyone else's proportion of the sample.
- **Operating across multiple geographies:** The dataset is from Peth and Surgana in Nashik district in Maharashtra only.

2.3 Limitations

- The benchmark excludes multi-diagnosis presentations in Phase 1. (however in phase 2 we realized, that in many cases the history recorded does contain symptoms that may point to additional diagnoses)
- Most cases requiring image evidence are excluded; a limited additional sample of dental-caries cases was retained.
- Because the data comes only from Peth and Surgana, the case mix may not generalize to other geographies or care settings.
- Patient may not be able to articulate their symptoms / health conditions properly
- The health workers may make data entry errors in accurately representing and capturing the patient's symptoms.
- The distribution is concentrated in common diagnoses; rare diagnoses are underrepresented.
- Infant ages are not separated reliably into neonatal and post-neonatal groups without age in days or months. This is being addressed in the next iteration to capture this information reliably if possible.

2.4 Data processing (PII)

Visits and patient identifiers are scrubbed to de-identify all visits, and other direct patient identifiers are removed during data preparation. Video and image URLs are excluded to reduce re-identification risk. De-identification is applied in the data-ingestion layer.

2.5 Unit of evaluation

The unit of evaluation in Phase 1 was a de-identified clinical consultation encounter, represented by a single row in the dataset and is identified by a unique visit_id. Each encounter contains patient demographics, consultation information - including symptoms, chief complaint, physical examination, medical and family history, and clinical notes - and one confirmed ground-truth diagnosis.

Differential-diagnosis accuracy was assessed by comparing the confirmed target diagnosis with the AI-generated ranked differential list of up to five diagnoses and with the clinician-recorded provisional diagnosis or diagnoses.

Representative de-identified row

Dataset Component	Representative Component	What is this column for?
Visit_id	7177VV2297	
Patient_id	26717	
Gender	F	Gender of the patient
Age	43	Age of the patient
Economic_status		
Visit_started_date	2023-04-03 09:53:42	

Visit_creation_date	2025-08-07 14:28:44	
Visit_ended_date		
Visit_upload_time		
Visit_Closure_Same_Month	No	
Visit_Status	Completed	
Doctor_Patient_Interaction		
Sbp	120	
Dbp	80	
Pulse	85	
Temperature	35.61	
Weight	42	
Height	155	
RR	20	
SPO2	98	
HB		
BMI	17.48	
Sugar_random		
Blood_group		
Sugar_pp		
Sugar_after_meal		
Vitals_timestamp	2023-04-03 09:53:42	
Consultation_start_time	2025-08-29 01:24:07	
Adult_Initial_timestamp	2023-04-03 09:59:04	
Consultation_End_time	2025-08-29 01:24:52	
ExitSurvey_timestamp		
Priority_timestamp		
Feedback_Score		
Specialty	Doc11_AI	
Symptoms	Fever, Leg, Knee or Hip Pain	Captured symptoms
Associated_Symptoms	Null	
Chief_complaint	►Fever: • Duration - 1 Days. • Nature of fever - All day/ Constant, Every few hours. • Timing - Evening, Night. • Severity - Low. • H/o of specific epidemic in community - None. • Prior treatment sought - None. ►Leg, Knee or Hip Pain: • Site - Right leg, Hip, Buttock, Thigh, Knee, Site of knee pain - Front, Back, Lateral/medial, Calf, Left leg, Hip, Buttock, Thigh, Knee,	

	Site of knee pain - Front, Back, Lateral/medial, Calf, Hip. • Duration - 2 months. • Pain characteristics - Dull aching, Tingling numbness, Sharp shooting, Burning. • Onset - Sudden (minutes to hours), Gradual. • Progress - Progressive (Worsening), Static (Not changed), Wax and wane. • All the joints. • Aggravating factors - Symptom aggravated by motion, Associated with motion - All planes of motion, Climbing stairs, Worsened by rest (relieved by activity). • H/o specific illness - • Patient reports - Recent/constant overus 	
Physical_examination	General exams: • Eyes: Jaundice-no jaundice seen, [picture taken]. • Eyes: Pallor-normal pallor, [picture taken]. • Arm-Pinch skin* - pinch test normal. • Nail abnormality-nails normal, [picture taken]. • Nail anemia-Nails are not pale. • Ankle-no pedal oedema. Any Location: • Ulcer:-no ulcer. • Skin Rash:-no rash. Mouth: • back of throat normal. Joint: • non-tender. • no deformity around joint. • full range of movement is seen. • joint is not swollen. • no pain during movement. • no redness around joint. Abdomen: • no tenderness. Back: • no back tenderness. Head: • No injury. 	
Patient_medical_history	• Pregnancy status - Not pregnant. • Allergies - No known allergies. • Alcohol use - No. • Smoking history - Patient denied/has no h/o smoking. • Drug history - No	

	recent medication. 	
Family_history	- 	
Doctor_name	DocEleven AI	Id for the doctor who conducted visit check
Clinical_notes	Gender: Female Age: 43 years Chief_complaint: ▶**Fever** : • Duration - 1 Days. • Nature of fever - All day/ Constant, Every few hours. • Timing - Evening, Night. • Severity - Low. • H/o of specific epidemic in community - None. • Prior treatment sought - None. ▶**Leg, Knee or Hip Pain** : • Site - Right leg, Hip, Buttock, Thigh, Knee, Site of knee pain - Front, Back, Lateral/medial, Calf, Left leg, Hip, Buttock, Thigh, Knee, Site of knee pain - Front, Back, Lateral/medial, Calf, Hip. • Duration - 2 months. • Pain characteristics - Dull aching, Tingling numbness, Sharp shooting, Burning. • Onset - Sudden (minutes to hours), Gradual. • Progress - Progressive (Worsening), Static (Not changed), Wax and wane. • All the joints. • Aggravating factors - Symptom aggravated by motion, Associated with motion - All planes of motion, Climbing stairs, Worsened by rest (relieved by activity). • H/o specific illness - • Patient reports - Recent/constant overus	

	Physical_examination: **General exams:** • Eyes: Jaundice-no jaundice seen, [picture taken]. • Eyes: Pallor-normal pallor, [picture taken]. • Arm-Pinch skin* - pinch test normal. • Nail abnormality-nails normal, [picture taken]. • Nail anemia-Nails are not pale. • Ankle-no pedal oedema. **Any Location:** • Ulcer:-no ulcer. • Skin Rash:-no rash. **Mouth:** • back of throat normal. **Joint:** • non-tender. • no deformity around joint. • full range of movement is seen. • joint is not swollen. • no pain during movement. • no redness around joint. **Abdomen:** • no tenderness. **Back:** • no back tenderness. **Head:** • No injury. Patient_medical_history: • Pregnancy status - Not pregnant. • Allergies - No known allergies. • Alcohol use - No. • Smoking history - Patient denied/has no h/o smoking. • Drug history - No recent medication. Family_history: - Vitals:- Sbp: 120.0	

	Dbp: 80.0 Pulse: 85.0 Temperature: 35.61 'C Weight: 42.0 Kg Height: 155.0 cm RR: 20.0 SPO2: 98.0 HB: Null BMI: 17.48 Sugar_random: Null Blood_group: Null Sugar_pp: Null Sugar_after_meal: Null	
Doctor_TAT	30896.1333	
Patient_TAT	50339.9833	
HW_Time_to_create_Visit	`00:05:22	
Diagnosis_provided	yes	
Ground Truth (GT) Diagnosis	Viral Fever	
Medications	Cetirizine 10 mg Oral Tablet:1 tablet:3:days:At bedtime:Once daily;Paracetamol 500 mg Oral Tablet:1 tablet:3:days:After food:Thrice daily;B-Complex (Multivitamin) Oral Capsule:1 capsule:7:days:After food:Once daily;Oral Rehydration Salts (WHO ORS, Small):1 sachet:5:days:Add 1 sachet to 200 ml of boiled and cooled drinking water:As needed;Ibuprofen 400 mg Oral Tablet:1 tablet:3:days:After food:Twice daily	
Medicines	Cetirizine 10 mg Oral Tablet, Paracetamol 500 mg Oral Tablet,	

	B-Complex (Multivitamin) Oral Capsule, Oral Rehydration Salts (WHO ORS, Small), Ibuprofen 400 mg Oral Tablet	
Strength	1 tablet, 1 tablet, 1 capsule, 1 sachet, 1 tablet	
Dosage	. . . .	
Additional Instructions		
Medical_test		
Medical_advice	Ensure adequate rest and hydration. Drink plenty of fluids like water, juice, and broth.,Be cautious when combining Paracetamol and Ibuprofen, as both can affect the liver and kidneys. Do not exceed the recommended doses.,Monitor temperature regularly. If fever persists beyond 3 days or worsens, seek further medical advice.	
Referral_advice		
Notes		
Follow_up_date	No	
FOLLOWUP_RESCHEDULE_REASON		
Patient_waiting_time_(in_hours)	21087.51944	
Sign_submit_time	2025-08-29 01:24:52	

2.6 Data structure

Dataset schema

Phase 1 schema: Phase 1 Dataset Schema

Field group	Content	Role in evaluation
Model inputs	De-identified demographics, chief complaint, patient history, symptoms, vital signs, physical examination, and clinical notes when available	Clinical context supplied to the model
Ground-truth label	One confirmed diagnosis	Target used for benchmark scoring
Model output	Ranked differential diagnosis list of up to five diagnoses, with clinical rationale where configured	Prediction evaluated against the target
Clinician comparison	Clinician-recorded provisional diagnosis or diagnoses	Comparator for selected analyses
Metadata	Unique de-identified encounter identifier and dataset provenance fields	Traceability and paired-view control

Proposed Phase 2 Schema

Field group	Content	Role in evaluation
Model inputs	De-identified demographics, chief complaint, patient history, symptoms, vital signs, physical examination, and clinical notes when available	Clinical context supplied to the model
Ground-truth label	• Diagnosis Array 1 … Diagnosis Array ‘n’ • Medication Array 1…Medication Array ‘n’ • Investigation Array 1…Investigation Array ‘n’ • Medical Advice Array 1… Medical Advice Array ‘n’ • Referral Array 1… Referral Array ‘n’	Target used for benchmark scoring
Model output	Ranked differential diagnosis list of up to five diagnoses, with clinical rationale where configured	Prediction evaluated against the target

		diagnoses array for DDx and Tx suggestions
Clinician comparison	• Clinician-recorded provisional diagnosis or diagnoses • AI suggested DDx and Tx Suggestions	Comparator for selected analyses
Metadata	Unique de-identified encounter identifier and dataset provenance fields	Traceability and paired-view control

3. Evaluation design

3.1 What is measured

The benchmark measures differential-diagnosis performance on text-only, single-diagnosis consultation cases using patient history and vital signs. It was used within a DSPy based evaluation pipeline for prompt optimization and comparison across base models for Ayu 2.0.

Metric	**Definition**	**Direction**
Top-1 accuracy	Share of cases in which the confirmed target diagnosis is ranked first	Higher is better
Top-5 accuracy	Share of cases in which the confirmed target diagnosis appears anywhere in the first five predictions	Higher is better
Quality of case history	5-Exemplary (Gold Standard): The record is highly reliable, reusable, and exceeds best practices. It contains all necessary detail to support the most complex clinical and secondary uses. 4-High Quality (Best Practice Adherence): The record consistently meets core standards. Deficiencies are minor and do not compromise clinical care or regulatory requirements. 3-Acceptable (Meets Minimum Standard): The record meets baseline operational and legal requirements, but inconsistencies or	Higher is better

	documentation gaps introduce manageable risk. 2-Requires Improvement (High Risk): Systemic failures (e.g., frequent missing essential data, documentation misuse) compromise clinical utility, requiring active intervention. 1-Unacceptable (Critical Failure): The record is unreliable. It fails foundational integrity, compliance, or patient safety standards.	
Appropriateness of AI generated DDx	How appropriate is the generated DDx list given by the LLM? 5 – Very appropriate 4 – Appropriate 3 – Neutral 2 – Inappropriate 0 – Very inappropriate	Higher is better
Comprehensiveness of AI generated DDx	How comprehensive is the LLM DDx list compared to the ground truth? 5 – the ground truth diagnosis is included in the DDx 4 – The DDx includes something very close but not exact 3 – The DDx includes something that is closely related and might be helpful 2 – The DDx includes something that is related but unlikely to help 0 – No items in the DDx is related to the ground truth diagnosis	Higher is better
Medication Appropriateness Index for Tx	**Explained below*	Lower is better
Completeness of Medication list	Does the treatment plan include all necessary medications? ☐ 5 − Includes all ☐ 4 − Almost all ☐ 3 − About half ☐ 2− Only one ☐ 0− Not at all	Higher is better

Mean reciprocal rank (MRR)	Mean of 1/rank for the confirmed diagnosis; a case contributes 0 if the target is absent from the evaluated list	Higher is better
LLM-as-judge review	Supplementary evaluation of diagnosis ranking and the associated clinical rationale	Rubric-dependent

*** Medication Appropriateness Index**

S. No	**Question**	**Score 0: Appropriate**	**Score 1: Marginally Appropriate**	**Score 2: Inappropriate**	***Weight***
1	Is there an indication for the drug?	Clearly indicated	Possibly indicated	No indication	3
2	Is the medication effective for the condition?	Clearly effective	Limited evidence or questionable effectiveness	Ineffective	3
3	Is the dosage correct?	Within recommended range	Slightly outside range	Too high or too low	2
4	Are the directions correct?	Clear and correct	Somewhat unclear	Incorrect or confusing	2
5	Are the directions practical (for the patient to follow)?	Easy to follow	Some difficulty	Impractical or burdensome	1
6	Are there clinically significant drug–drug interactions?	None present	Possible but not serious	Serious interaction present	2
7	Are there clinically significant drug–disease interactions?	None present	Minor or unlikely	Clearly present and clinically important	2
8	Is there unnecessary duplication with other drugs?	No duplication	Partial or marginal duplication	Redundant therapy	1
9	Is the duration of therapy acceptable?	Within guideline duration	Slightly prolonged or shortened	Unnecessarily prolonged/too short	1
10	Is the drug the least expensive alternative of equal effectiveness?	Yes (least costly)	Moderately more expensive	Clearly more expensive than alternatives	1

Metrics for Pilot Study (n=1,000 cases)

Parameter	Description
Diagnostic Accuracy	Binary: Correct if the clinician's diagnosis matched the ground truth; Incorrect otherwise.
Turn around time	Duration from case initiation to final submission, measured in minutes and seconds, available from the telemedicine platform.
Quality of Differential Diagnoses	Evaluated only in the LLM-assisted group;Top-5 Accuracy(whether Ground truth diagnosis present in top 5 + rank), Appropriateness*, Comprehensiveness*
Quality of Treatment Plan	Degree of alignment between clinician's plan and treatment plan based on ground truth diagnosis. Completeness*and Medical Appropriateness Index (MAI) as given by Hanlon et. al. (Hanlon, 1992)
Quality of Medical Tests	Relevance and necessity of investigations.*
Quality of Medical Advice	Clinical relevance, clarity, and safety of advice given.*
Quality of Referral Advice	Necessity and appropriateness of referral decisions or omissions.*

3.2 Objectives

To evaluate the effect of large language model (LLM) assistance on diagnostic accuracy using a matched case-control design, where each clinical case is reviewed once by an LLM-assisted doctor and once by a non-assisted doctors

Research Questions:
Does an LLM optimised for DDx enhance diagnostic accuracy, compared to clinicians using traditional and/or non-LLM-assisted resources?

- Does the LLM impact the time taken to diagnose?
- Does the LLM influence the clinical appropriateness and quality of prescriptions and treatment plans provided by clinicians?

3.3 Scope and limitations

- The benchmark is specific to Nashik and may not generalize to other geographies.
- Multiple-diagnosis cases, co-morbid presentations requiring multiple labels, and cases requiring image evidence are outside scope.

- The benchmark does not directly measure clinician behavior, clinical effectiveness of prescribed treatment plans such as on patient outcomes, or health-system impact.

4. Labels and annotation

4.1 Who labeled the data

- **Phase 1:** Dr. Venkat served as the primary ground-truth annotator for diagnosis labels. Dr. Nilofer contributed diagnosis tagging for a subset of cases.
- **Phase 2:** A multi-clinician annotation exercise involving six doctors is in progress and is intended to add diagnosis arrays and support inter-annotator analysis.

Six expert clinicians were recruited in Phase 2, with experience in practicing in primary healthcare settings. All six doctors are qualified MBBS, with over 8 years of experience (Range: 8 years-40 years)

4.2 How labeling was conducted

- Cases were restricted to single-diagnosis consultations.
- Multi-diagnosis cases and most cases requiring image evidence were excluded.
- The clinical team selected exclusions and reviewed or cleaned ground-truth diagnosis labels.

Pending confirmation. Document annotator training and calibration, whether annotators worked independently, what information they could view, whether AI and clinician outputs were blinded, diagnosis-coding rules, disagreement resolution, cases removed after annotation, and the final label version.

4.3 Label quality

Phase 1 relies primarily on two annotators for diagnosis ground truth and should be interpreted with that limitation.

We ran a LLM judge analysis using Gemini-3-Flash to measure the effectiveness of the Ground truths and relevance of labels to the captured patient case history. The results are below:

Category	Number of cases
GT Wrong	41
Notes Ambiguous	14
Presentation Ambiguous	16

About 7% of the 1000 cases dataset was deemed to have questionable ground truth.
Error analysis done to check this using LLM Judge is at: NAS_prod_prompt_judge1_evals.xlsx

Phase 2 is intended to strengthen label quality through multi-clinician review, consensus based ground truth settings, and inter-annotator analysis.

Pending
Report the inter-annotator agreement method for phase 2 and result, documented mistakes or biases identified during verification, and the adjudication procedure used to produce the final consensus label.

4.3

- **Phase 1**: Two clinicians annotated based GT tagging for diagnosis (Dx) against patient case history and vitals
- **Phase 2:**
  Inter-annotator agreement:
  - To obtain the inter annotator agreement, the below exercise was conducted:
    Six doctors [annotator] based GT tagging (Adding Diagnosis and treatment plan arrays)
- Inter-annotator profile agreement: In progress
- Mistakes / biases / limitations after verification: In progress

4.4 LLM as judge

An LLM-as-judge pipeline supplements rank-based scoring by evaluating the position of a diagnosis in the top-five differential list and the clinical rationale generated for each case.

DDx judge: InteleHealth <> TAF: LLM Judge Metrics for AI CDSS DDx

Judge mismatches and ground truth quality, error analysis evaluation workbook:
NAS_prod_prompt_judge1_evals.xlsx

Treatment-plan evaluation uses the Medical Appropriateness Index (MAI) judge
The MAI judge outputs are available for the 1,000 Phase 1 cases.
Link: Bharat to Add.

Pending confirmation. For each judge, provide the model and exact version, prompt or rubric version, temperature and seed, output schema, invalid-output handling, human calibration sample, agreement with clinician reviewers, and confirmation that the judge model differs from the evaluated model.

5. Benchmark

5.1 Cross-cutting dimensions

Cost

Cost item	Reported estimate
DDx model query	Approximately $0.00244 per query
Treatment-plan model query	Approximately $0.00132 per query
1,000 DDx and treatment-plan runs	Approximately $3-$5
AI evaluation for 1,000 cases	Approximately $3-$4

Cost source: NAS Ayu 2.0 token counts + costs.xlsx

Pending confirmation. Confirm whether these estimates cover model calls only or the full operating cost, including AI infrastructure, databases, load balancing, and other platform services.

Latency

- **AI server estimate:** Approximately 10-20 seconds for Phase 1 (Ayu 2.0), including DSPy execution and an additional LLM-based transformation through OpenRouter APIs.
- **AI server estimate:** Approximately 3-7 seconds for Phase 2 (Ayu 2.1) on IndiaAI infrastructure.

Pending confirmation. Confirm the measurement method and whether the reported values represent AI-server generation time or full product end-to-end latency.

Differential Diagnosis Accuracy

Metric	Reported value
Top-1 accuracy	77%
Top-5 accuracy	95%
Mean reciprocal rank	0.85

Pending confirmation. Verify the evaluated model and version, evaluation date, prompt version, and whether these figures are calculated on the full unseen 1,000-case set.

Safety and guardrails

Pending confirmation. Add the safety dimensions, guardrail checks, thresholds, and results used for the evaluated model.

5.2 How the benchmark runs

1. Load one de-identified, text-only consultation encounter containing the permitted patient history and vital-sign fields. You can also load n-cases records from the dataset as the input CSV.
2. Generate a ranked differential diagnosis list of up to five diagnoses, with rationale where configured.
3. Compare the ranked output with the confirmed target diagnosis and calculate rank-based metrics.
4. Run the configured DDx LLM judge for supplementary diagnosis-ranking and rationale evaluation where required.
5. Aggregate case-level results into model-level summary metrics and review error cases.

5.3 Reproducibility

Benchmark scoring code and reproducibility instructions are maintained in a private repository and can be shared securely with program reviewers, subject to contractual approval.

Public release of the repository is not currently planned.

DDx evaluation and judge configuration
LLM Judge Metrics for AI CDSS DDx - Version 2 of the DDx judge was used for production-model evaluations
Version 3 is under development - To be shared when approved from clinical and MLE teams.

MAI treatment-plan judge for Tx planning
11 Individual Tx MAI + Error of omission judges

Pending confirmation.
Provide the executable environment, dependency versions, random seeds, prompt and judge configuration files, expected output schema, and step-by-step repeatability instructions.

Code availability and link:
Dependencies / environment: Subject to the decision on the above
Methodological repeatability:
- AI / MLE Team to update if required

5.4 Position in the evaluation ecosystem

- **Related benchmarks:**

  This benchmark is most closely related to differential-diagnosis and medical question-answering evaluations such as DDXPlus, MedQA, MedMCQA and PubMedQA. Unlike DDXPlus, which uses synthetic patient profiles, and exam-based benchmarks such as MedQA, the InteleHealth benchmark uses de-identified, real-world telemedicine encounters from rural Maharashtra. It is also complementary to HealthBench, which evaluates broad response quality in realistic health conversations using physician-authored rubrics. HealthBench covers dimensions such as accuracy, completeness, communication, context awareness and safety, whereas this benchmark measures the quality and ranking of differential diagnoses for a specific provider-to-provider clinical decision-support workflow. Scores from these benchmarks should therefore not be compared directly.

- **Relationship to existing frameworks (L1–L4, HealthBench, etc.):**

  Within The Agency Fund's L1–L4 evaluation framework, this benchmark is primarily an L1 model evaluation. It assesses whether the AI produces clinically appropriate ranked differential diagnoses using 1,000 doctor-annotated cases. The principal measures include Top-1 and Top-5 accuracy, mean reciprocal rank.

  It supports comparison across base models, prompt optimization and selection of the production model; more than 12 models have been evaluated through the internal DDx leaderboard. As we move to the knowledge graph grounded differential diagnosis; this unseen test set is used to evaluate the improvement in the DDx model accuracies and metrics as well.

  The benchmark contributes evidence to, but does not itself establish, the higher evaluation levels:

  - **L2 – Product:** Measured separately through live-product engagement, including whether clinicians review, accept, modify or override AI recommendations.
  - **L3 – User:** Measured through the 1,000-case crossover study and the 600-consultation Nashik observational audit, covering changes in clinician diagnostic accuracy, consultation time, treatment quality and prescription safety.
  - **L4 – Impact:** Requires prospective evidence of patient and health-system outcomes. This is planned through the three-arm SAKSHAMA cluster-randomised study and is outside the scope of the present dataset benchmark.

- **Gap this benchmark fills:**

  The benchmark provides an unseen, text-only evaluation set derived from real last-mile telemedicine encounters in rural India; an operating context not adequately represented by synthetic differential-diagnosis datasets, medical examination questions or broad conversational

health benchmarks.

It measures whether models can rank the confirmed diagnosis from routinely available patient history and vital signs under a provider-to-provider workflow. It therefore supplies a locally grounded signal for model selection and prompt optimisation while remaining explicitly separate from evidence about clinician behaviour, clinical effectiveness and patient outcomes

6. Worked example

The following example is drawn from the representative de-identified row in Section 2.5.

Visit_id	7177VV2297
Patient_id	26717
Gender	F
Age	43
Economic_status	
Visit_started_date	2023-04-03 09:53:42
Visit_creation_date	2025-08-07 14:28:44
Visit_ended_date	
Visit_upload_time	
Visit_Closure_Same_Month	No
Visit_Status	Completed
Doctor_Patient_Interaction	
Sbp	120
Dbp	80
Pulse	85
Temperature	35.61
Weight	42
Height	155
RR	20
SPO2	98
HB	
BMI	17.48
Sugar_random	
Blood_group	
Sugar_pp	
Sugar_after_meal	
Vitals_timestamp	2023-04-03 09:53:42
Consultation_start_time	2025-08-29 01:24:07
Adult_Initial_timestamp	2023-04-03 09:59:04
Consultation_End_time	2025-08-29 01:24:52

ExitSurvey_timestamp	
Priority_timestamp	
Feedback_Score	
Specialty	Doc11_AI
Symptoms	Fever, Leg, Knee or Hip Pain
Associated_Symptoms	Null
Chief_complaint	►Fever: • Duration - 1 Days. • Nature of fever - All day/ Constant, Every few hours. • Timing - Evening, Night. • Severity - Low. • H/o of specific epidemic in community - None. • Prior treatment sought - None. ►Leg, Knee or Hip Pain: • Site - Right leg, Hip, Buttock, Thigh, Knee, Site of knee pain - Front, Back, Lateral/medial, Calf, Left leg, Hip, Buttock, Thigh, Knee, Site of knee pain - Front, Back, Lateral/medial, Calf, Hip. • Duration - 2 months. • Pain characteristics - Dull aching, Tingling numbness, Sharp shooting, Burning. • Onset - Sudden (minutes to hours), Gradual. • Progress - Progressive (Worsening), Static (Not changed), Wax and wane. • All the joints. • Aggravating factors - Symptom aggravated by motion, Associated with motion - All planes of motion, Climbing stairs, Worsened by rest (relieved by activity). • H/o specific illness - • Patient reports - Recent/constant overus
Physical_examination	General exams: • Eyes: Jaundice-no jaundice seen, [picture taken]. • Eyes: Pallor-normal pallor, [picture taken]. • Arm-Pinch skin* - pinch test normal. • Nail abnormality-nails normal, [picture taken]. • Nail anemia-Nails are not pale. • Ankle-no pedal oedema. Any Location: • Ulcer:-no ulcer. • Skin Rash:-no rash. Mouth: • back of throat normal. Joint: • non-tender. • no deformity around joint. • full range of movement is seen. • joint is not swollen. • no pain during movement. • no redness around joint. Abdomen: • no tenderness. Back: • no back tenderness. Head: • No injury.
Patient_medical_history	• Pregnancy status - Not pregnant. • Allergies - No known allergies. • Alcohol use - No. • Smoking history - Patient denied/has no h/o smoking. • Drug history - No recent medication.
Family_history	-

Clinical_notes	Gender: Female Age: 43 years Chief_complaint: ►**Fever** : • Duration - 1 Days. • Nature of fever - All day/ Constant, Every few hours. • Timing - Evening, Night. • Severity - Low. • H/o of specific epidemic in community - None. • Prior treatment sought - None. ►**Leg, Knee or Hip Pain** : • Site - Right leg, Hip, Buttock, Thigh, Knee, Site of knee pain - Front, Back, Lateral/medial, Calf, Left leg, Hip, Buttock, Thigh, Knee, Site of knee pain - Front, Back, Lateral/medial, Calf, Hip. • Duration - 2 months. • Pain characteristics - Dull aching, Tingling numbness, Sharp shooting, Burning. • Onset - Sudden (minutes to hours), Gradual. • Progress - Progressive (Worsening), Static (Not changed), Wax and wane. • All the joints. • Aggravating factors - Symptom aggravated by motion, Associated with motion - All planes of motion, Climbing stairs, Worsened by rest (relieved by activity). • H/o specific illness - • Patient reports - Recent/constant overus Physical_examination: **General exams:** • Eyes: Jaundice-no jaundice seen, [picture taken]. • Eyes: Pallor-normal pallor, [picture taken]. • Arm-Pinch skin* - pinch test normal. • Nail abnormality-nails normal, [picture taken]. • Nail anemia-Nails are not pale. • Ankle-no pedal oedema. **Any Location:** • Ulcer:-no ulcer.

	• Skin Rash:-no rash. **Mouth:** • back of throat normal. **Joint:** • non-tender. • no deformity around joint. • full range of movement is seen. • joint is not swollen. • no pain during movement. • no redness around joint. **Abdomen:** • no tenderness. **Back:** • no back tenderness. **Head:** • No injury. Patient_medical_history: • Pregnancy status - Not pregnant. • Allergies - No known allergies. • Alcohol use - No. • Smoking history - Patient denied/has no h/o smoking. • Drug history - No recent medication. Family_history: - Vitals:- Sbp: 120.0 Dbp: 80.0 Pulse: 85.0 Temperature: 35.61 'C Weight: 42.0 Kg Height: 155.0 cm RR: 20.0

	SPO2: 98.0 HB: Null BMI: 17.48 Sugar_random: Null Blood_group: Null Sugar_pp: Null Sugar_after_meal: Null
Doctor_name	DocEleven AI
Doctor_TAT	30896.1333
Patient_TAT	50339.9833
HW_Time_to_create_Visit	`00:05:22
Diagnosis_provided	yes
Ground Truth (GT) Diagnosis	Viral Fever
DDX_Rank1	Viral Fever
DDX_Rank2	Osteoarthritis
DDX_Rank3	Upper Respiratory Tract Infection
DDX_Rank4	Rheumatoid Arthritis
DDX_Rank5	Scrub Typhus
B1-How appropriate DDx-LLM	5
C1-How comprehensive is the LLM DDx list to GT	5
D1-Rank 0-1 to 5 (GT-AI DDx)	1
Medications	Cetirizine 10 mg Oral Tablet:1 tablet:3:days:At bedtime:Once daily;Paracetamol 500 mg Oral Tablet:1 tablet:3:days:After food:Thrice daily;B-Complex (Multivitamin) Oral Capsule:1 capsule:7:days:After food:Once daily;Oral Rehydration Salts (WHO ORS, Small):1 sachet:5:days:Add 1 sachet to 200 ml of boiled and cooled drinking water:As needed;Ibuprofen 400 mg Oral Tablet:1 tablet:3:days:After food:Twice daily
Medicines	Cetirizine 10 mg Oral Tablet, Paracetamol 500 mg Oral Tablet, B-Complex (Multivitamin) Oral Capsule, Oral Rehydration Salts (WHO ORS, Small), Ibuprofen 400 mg Oral Tablet

Strength	1 tablet, 1 tablet, 1 capsule, 1 sachet, 1 tablet
Dosage	. . . .
Medication1	Drug: Cetirizine 10 mg Oral Tablet | Dose: 1 tablet | Duration: 3 days | Instruction: At bedtime | Frequency: Once daily
MAI_1	0
MAI_2	0
MAI_3	0
MAI_4	0
MAI_5	0
MAI_6	0
MAI_7	0
MAI_8	0
MAI_9	0
MAI_10	0
Medication2	Drug: Paracetamol 500 mg Oral Tablet | Dose: 1 tablet | Duration: 3 days | Instruction: After food | Frequency: Thrice daily
MAI_1	0
MAI_2	0
MAI_3	0
MAI_4	0
MAI_5	0
MAI_6	0
MAI_7	0
MAI_8	0
MAI_9	0
MAI_10	0
Medication3	Drug: B-Complex (Multivitamin) Oral Capsule | Dose: 1 capsule | Duration: 7 days | Instruction: After food | Frequency: Once daily
MAI_1	1
MAI_2	1
MAI_3	0
MAI_4	0
MAI_5	0
MAI_6	0
MAI_7	0
MAI_8	0
MAI_9	0
MAI_10	1

Medication4	Drug: Oral Rehydration Salts (WHO ORS, Small) | Dose: 1 sachet | Duration: 5 days | Instruction: Add 1 sachet to 200 ml of boiled and cooled drinking water | Frequency: As needed
MAI_1	1
MAI_2	1
MAI_3	1
MAI_4	0
MAI_5	0
MAI_6	0
MAI_7	0
MAI_8	0
MAI_9	0
MAI_10	0
Medication5	Drug: Ibuprofen 400 mg Oral Tablet | Dose: 1 tablet | Duration: 3 days | Instruction: After food | Frequency: Twice daily
MAI_1	1
MAI_2	1
MAI_3	0
MAI_4	0
MAI_5	0
MAI_6	0
MAI_7	0
MAI_8	1
MAI_9	0
MAI_10	0
Medication6	
MAI_1	
MAI_2	
MAI_3	
MAI_4	
MAI_5	
MAI_6	
MAI_7	
MAI_8	
MAI_9	
MAI_10	
Remarks	
Completeness	4
Additional Instructions	

Medical_test	
Appropriateness of Medical Test	5
Medical_advice	Ensure adequate rest and hydration. Drink plenty of fluids like water, juice, and broth.,Be cautious when combining Paracetamol and Ibuprofen, as both can affect the liver and kidneys. Do not exceed the recommended doses.,Monitor temperature regularly. If fever persists beyond 3 days or worsens, seek further medical advice.
Appropriateness of Medical Advice	5
Referral_advice	
Appropriateness of Referral Advice	5
Notes	
Follow_up_date	No
FOLLOWUP_RESCHEDULE_REASON	
Patient_waiting_time_(in_hours)	21087.51944
Sign_submit_time	2025-08-29 01:24:52
Quality of case record content	5
Clinical Depth	5

- **Expected behavior at each step of the decision-point chain:** [To be added.]
- **How each evaluation dimension scores this example:** [To be added once dimensions in §3 are finalised.]
- **What a failure on each dimension would look like:** [To be added.]

7. Results

7.1 Models evaluated

We evaluated various closed and open source models for differential diagnosis during iterative prompt optimisations cycles. Link to the leaderboard is at: DDx Leaderboard_Dec 2026_TAF

We went with llama4-maverick for phase 1 pilot as it was cost-effective and did well on our internal benchmarks. Currently the medgemma-27b-it multimodal model offers better performance than llama4 for Top 1 in our LLM Judge measurements and analysis on internal evaluation benchmarks and leaderboards.

Pending confirmation.
prompt versions, latency, cost, safety results, and the rationale for selecting the current MedGemma model and moving from llama maverick to the current setup.

7.2 Headline results

Metric	Reported result	Interpretation
Top-1 DDx accuracy	77%	The confirmed diagnosis was ranked first in 77% of evaluated cases.
Top-5 DDx accuracy	95%	The confirmed diagnosis appeared in the first five predictions in 95% of evaluated cases.
Mean reciprocal rank	0.85	The confirmed diagnosis was generally ranked near the top of the predicted list.

7.3 Interpretation

Based on the reported figures, the 18-percentage-point difference between Top-1 and Top-5 accuracy suggests that the confirmed diagnosis is often present in the candidate set even when it is not ranked first. This indicates a potential opportunity to improve ranking and calibration. The reported MRR of 0.85 is consistent with targets generally appearing near the top of the list.

Model fitness for clinical use should be assessed together with the pending technical, safety, and user-level evidence.

RCT study interpretation to be added as well

8. Public release

How much is being released publicly (sample size): [To be decided.]

Licensing of the sample: [To be decided.]

9. Full release for TAF / Endless Health

Full dataset as a zip file, along with scripts to run inference and automated evaluations, and clear instructions for reproducibility.

10. Acknowledgements and contributors

Team members: Dr Neha Verma, Dr. Nikesh Shah, Neeraja Reddy Karna, Aditya Naskar, Bharat Shetty Barkur, Puja Goswami, Priya Joshi, Wayan Vota

Annotators: Dr. Mayur, Dr. Nilofer, Dr. Venkat and the clinical team

Funding acknowledgment: [To be completed.]