International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

MediSentinelAI: A Safety-Critical Multi-Agent LLM Framework for Symptom Triage and Geospatial Healthcare Coordination

Dr. Vathsala M. K., Afwan Saad Ahmed, Arun Kumar B., B. C. Moulya, and Chaitra L.

***Abstract*—MediSentinelAI is an India-oriented progressive web application (PWA) that combines multimodal prescription extraction, symptom triage, medication reminders, sensory alerts, and geospatial emergency coordination. The author-reported implementation uses six role-specialized agents, an event-driven React client, service workers, IndexedDB, Gemini structured outputs, deterministic offline symptom rules, and Google Places with an OpenStreetMap fallback. In a reported 250-image evaluation, exact four-field record accuracy was 79.2%; a 180-vignette triage confusion matrix implies 91.7% accuracy and 45/45 observed critical-case recall; and Web Push succeeded in 488/500 trials (97.6%). These values supersede earlier summary percentages in the project dossier, which conflict with the supplied counts. The datasets, code, model prompts, and execution logs were not available for independent reproduction. Results are preliminary author-reported measurements and do not establish clinical effectiveness, regulatory exemption, or safe autonomous operation.**

***Index Terms*—Clinical decision support, emergency triage, geospatial health, medication adherence, multimodal large language models, progressive web applications.**

I. Introduction

Patients and caregivers often need to organize medication schedules, interpret prescription labels, decide how urgently to seek care, and identify nearby services. MediSentinelAI (hereafter, MediSentinel) is described as an India-oriented PWA integrating these activities through role-specialized agents and client-side services. India-specific framing is relevant to medication and emergency-access workflows; the ICMR-INDIAB national study estimated 101 million adults with diabetes and 315 million with hypertension in 2021, demonstrating substantial chronic-disease burden [3]. The dossier's additional claims about adherence rates, English proficiency, emergency-department occupancy, and non-emergency attendance were not accompanied by supporting sources and are not repeated here.

MediSentinel's scope spans several distinct risk domains. A prescription-image extraction error can create an unsafe reminder; a triage error can delay emergency care; stale facility information can misdirect a patient; and a push-notification failure can interrupt a routine adherence aid. The system therefore must not equate a valid JSON response, a high benchmark percentage, or agreement between agents with clinical correctness. Clinical triage protocols such as the Emergency Severity Index (ESI) require trained application and local governance; an application-specific four-tier label is not, by itself, an ESI classification [2].

This paper presents the architecture and technical claims supplied by the project authors, reports their benchmark values with provenance caveats, and identifies information needed for reproducibility. It also distinguishes implemented features asserted in the dossier from features that could not be verified against source code or raw study materials. In particular, reported internal test results are not represented as independent validation, and product disclaimers do not establish a regulatory classification.

Dr. Vathsala M. K. is with East Point College of Engineering and Technology.
Afwan Saad Ahmed (1EP23IS008), Arun Kumar B. (1EP23IS020), B. C. Moulya (1EP23IS023), and Chaitra L. (1EP23IS035) are with East Point College of Engineering and Technology.
Corresponding authors: Dr. Vathsala M. K. and Afwan Saad Ahmed.

II. System Architecture

A. Agent roles and hierarchy

The dossier describes six role-specialized agents. They communicate through a lightweight event-driven client architecture rather than a server-side multi-agent framework. This separation can support focused interfaces and explicit handoffs, but decentralizing agent roles does not eliminate shared failure modes: the browser runtime, event bus, API gateway, model provider, push service, and underlying safety taxonomy remain potential common dependencies.

B. Event-driven orchestration and service boundaries

The client-side orchestration is described as React state, a custom event bus, IndexedDB, and asynchronous `postMessage` communication with the service worker. Named events include `resync_background_reminders`, `trigger_reminder_alert`, and `SW_REMINDER_MARKED_TAKEN`. REST routes reported in the dossier include `api/analyze-bottle`, `api/analyze-symptoms`, `api/reminders/sync`, `api/push/test`, `api/nearby-hospitals`, and `api/geocode`. These route names and storage identifiers are author supplied; the workspace contained no application source code or API tests with which to verify them.

The reported technology stack is summarized in Table II. Versions are transcribed from the supplied lockfile summary and are not independently confirmed against project source files.

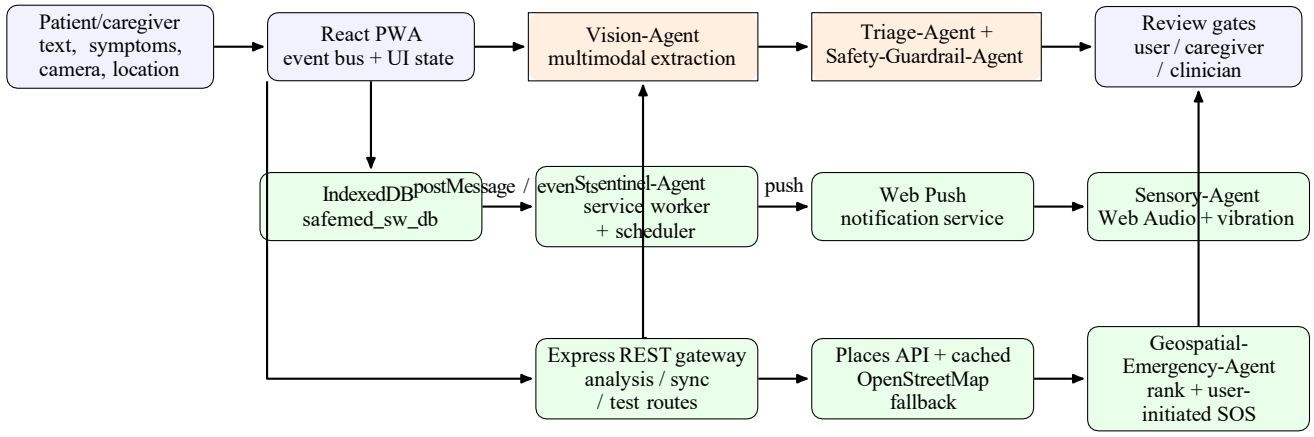


Fig. 1. Architecture reported in the project dossier. The figure is a design-level reconstruction, not a source-code audit. Review gates are shown because extracted prescriptions require confirmation and emergency contacts should remain user initiated.

TABLE I
AGENT ROLES AND PRINCIPAL SAFETY BOUNDARY

Agent	Reported function	Safety boundary / verification requirement
Vision-Agent	Captures camera frames; corrects/handles cylindrical-label distortion; extracts medication name, dosage, frequency, and timing into a constrained schema.	OCR and language interpretation may still be wrong. Do not save a medication schedule until a user verifies each field against the prescription.
Triage-Agent	Conducts symptom inquiry; considers duration and reported comorbidities; proposes Low, Moderate, High, or Critical urgency.	No provided validation of its Bayesian confidence, calibration, prompts, or mapping to clinical protocols. No diagnosis or autonomous treatment.
Safety-Guardrail-Agent	Applies a keyword/regular-expression taxonomy described as covering 48 emergency conditions; intended to intercept red flags and suppress routine advice.	Taxonomy, language coverage, test results, and false-negative rate are not supplied. A deterministic matcher can miss paraphrases and context; it cannot guarantee interception.
Sentinel-Agent	Coordinates IndexedDB, service-worker reminders, synchronization, and server-side push scheduling.	Browser and operating-system lifecycle behavior varies. Scheduled delivery is not guaranteed by a two-second polling interval.
Sensory-Agent	Generates E5–G#5–B5–E6 audio and two vibration patterns using Web Audio and Vibration APIs.	Requires supported hardware/browser state and appropriate user permissions/interaction; accessibility alternatives are needed.
Geospatial-Emergency-Agent	Ranks nearby facilities and presents cellular dialing or preformatted WhatsApp SOS actions using India (+91) number validation.	Ratings and opening status do not establish emergency capability. Location sharing and dialing must be explicit, user-controlled actions.

C. Model prompting and structured outputs

The dossier describes constrained few-shot prompting without fine-tuning. Gemini response schemas are used to constrain output structure, including enumerated schedule fields. This can reduce malformed JSON and simplify validation, but schema adherence does not establish that a medication name, dose, frequency, or triage statement is clinically correct [8]. The phrase "zero hallucination" is therefore not an appropriate system guarantee.

The updated dossier reports `gemini-3-flash-preview` for symptom triage and `gemini-3.5-flash`, falling back to `gemini-2.5-flash`, for prescription vision. It states that unavailable or failing network inference routes to a client-side deterministic rule engine. This replaces the earlier claim of a Gemini 1.5 Flash fallback; that model was shut down on September 29, 2025 [9]. Google documentation lists Gemini 3.5 Flash as deprecated and routed to Gemini 3.6 Flash as of October 2026, while Gemini 3 Flash remains a preview model; therefore, deployed model aliases and response metadata must be logged for reproducibility [9]. Table III records the newly supplied generation parameters. The request timeout, retry delays/jitter constants, and number of consecutive failures that opens the circuit breaker were absent from the supplied text and remain unreported.

The temperature, sampling, seed, and schema configuration are reported settings, not evidence of deterministic or clinically correct output. The dossier's rationale that temperature 0.20 suppresses variation and seed 42 enables reproducibility should be interpreted as an implementation claim; actual repeatability depends on the serving model and API behavior.

D. Operational workflow and failure containment

The safety-relevant path is not a single model call. It comprises input capture, interpretation, validation, user review, persistence, and (where requested) an external action. Each transition can fail independently, so the workflow should preserve the distinction between an untrusted suggestion and a user-confirmed record. Figure 2 summarizes this separation at the design level. It does not imply that the listed controls have been independently tested.

TABLE II
IMPLEMENTATION STACK REPORTED BY THE AUTHORS

Layer	Reported technology	Reported version / standard	Role
Frontend	React, TypeScript, Vite, Tailwind CSS, Motion	React 19.0.1; TypeScript 5.8.2; Vite 6.2.3; Tailwind 4.1.14; Motion 12.23.24	PWA interface and reactive component state
Background client	Service Worker API; IndexedDB	`reminder-sw.js`; database `safemed_sw_db`	Reminder persistence and push-event handling
Backend	Node.js, Express	Node.js 20 LTS; Express 4.21.2	REST gateway and reminder scheduler
AI inference	Gemini API via `@google/genai`	SDK 1.29.0; reported IDs include Gemini 3 Flash Preview, Gemini 3.5 Flash, and Gemini 2.5 Flash fallback	Multimodal bottle parsing and symptom-response generation
Cloud services	Firebase Firestore and Authentication	Firebase JS SDK 12.13.0	Account persistence and synchronization
Push	`web-push`, VAPID	RFC 8030, RFC 8291, RFC 8292	Encrypted Web Push payload delivery and application-server identification
Geospatial	Google Places; browser geolocation; OSM/Nominatim fallback	API versions not specified	Facility lookup, ranking, and fallback
Maps UI / documents	`@vis.gl/react-google-maps`; jsPDF	1.8.3; 4.2.1	Map display and client-side document export
Client alerts	Web Audio API; Vibration API	Browser APIs	Synthesized chime and vibration patterns

TABLE III
INFERENCE AND RETRY SETTINGS REPORTED BY THE AUTHORS

Parameter	Reported setting
Temperature; top-p; top-k	0.20; 0.85; 40
Maximum output tokens; thinking	1024; `ThinkingLevel.MINIMAL`
Response format; schema	`application/json`; strict schema enabled
Seed	42
Retry policy	Up to 3 attempts; truncated exponential backoff with decorrelated jitter
Circuit breaker	HTTP 429/503 or repeated failures in 30 s; open for 60 s; failure-count threshold not supplied
Fallback	Client-side deterministic rule engine

For prescription capture, the model output should remain a draft until the person can compare every extracted field with the source label. Missing or ambiguous fields should remain visibly unresolved rather than being filled by inference. Editing a value should not silently preserve an earlier model confidence or provenance marker. If the source image is unreadable, the safe outcome is a request for a clearer image or manual entry, not a best guess. The dossier reports a confirmation form, but does not provide interaction logs or a usability test showing that users notice and correct errors.

For symptom triage, failure to reach the inference service is not equivalent to a successful offline assessment. An offline rule result should be labeled as such, should expose that its coverage is limited, and should avoid reassuring language when input is incomplete. If the model and the deterministic matcher disagree, the interface should not hide the conflict by selecting whichever tier is less urgent. A conservative escalation can prevent some forms of under-triage, but may also increase unnecessary referrals; both effects require measurement. The dossier does not specify the actual conflict-resolution rule or quantify the resulting workload.

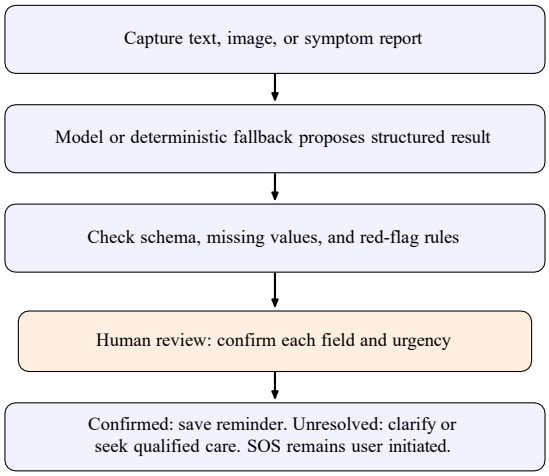


Fig. 2. Recommended safety workflow from model output to user-confirmed action; not an audited application trace.

These boundaries suggest three independent validation targets: semantic accuracy of the generated content, correctness of application state transitions, and successful completion of the user's intended action. A parser test can assess schema validity without assessing clinical meaning; a reminder delivery test can assess notification timing without showing that the underlying schedule is correct; and a facility-search test can assess ranking behavior without establishing that a listed facility can provide emergency care. Keeping these claims separate prevents a technical integration test from being misreported as evidence of clinical safety.

III. CLINICAL AND EMERGENCY SAFETY

A. Target population and protocol review

The stated target population is adults aged 18–85 in rural primary health centres, peri-urban taluk hospitals, and urban emergency-care settings in Bengaluru and Karnataka. The

dossier mentions pediatric and geriatric risk flags but leaves their age cutoffs unspecified; the described benchmark does not establish safety for patients outside the adult range. English, Hindi, and Kannada are reported as supported languages, including code-mixed Hinglish/Kanglish input. Two independent emergency-medicine physicians (MBBS, MD) reportedly reviewed the triage vignettes and reference labels. Unweighted Cohen's kappa is named as the inter-rater measure, but its value and supporting agreement counts are absent; the exact clinical source for each action window was also not supplied.

B. Project-defined urgency tiers

The dossier maps the five-level ESI framework to four application tiers. ESI is a five-level emergency-department triage tool; the application's merged levels and time-to-care instructions are project policy, not official ESI recommendations. Table IV records the mapping as supplied, but its action timings must be clinically reviewed before user-facing use [2].

C. Red flags, uncertainty, and user confirmation

The implemented deterministic fallback is reported in `buildOfflineTriageAnalysis` in `src/services/aiService.ts`; it uses client-side regex matching and emergency gating when network inference is unavailable or fails. The Safety-Guardrail-Agent is described as scanning input and generated text against a 48-condition keyword taxonomy. Prompt calibration reportedly addresses negation, spelling variation, code-mixed language, and contradictory symptoms, but the dossier does not list the patterns, define measurable handling rules, or report performance by language or edge case. A keyword interceptor is not an "un-bypassable" clinical control: it may fail on indirect descriptions or over-trigger on quoted or historical symptoms. The claimed confidence-based clarification route has no supplied formula or numeric threshold. Before clinical use, these components require expert-curated coverage tests, adversarial evaluation, logging, and a fail-safe escalation policy.

The dossier also refers to a Bayesian confidence value and a threshold that would trigger a physician-consultation notice, but the formula, calibration set, and threshold value are missing. No numeric confidence threshold is inserted here. Model self-confidence must not be presented as calibrated risk without a calibration study. The stated non-diagnostic disclaimer is appropriate as a boundary statement, but does not substitute for correct routing, clinical validation, or regulatory assessment.

For medication extraction, the reported workflow auto-fills a reminder form but waits for the patient or caregiver to review and confirm pill name, dosage, and schedule before saving. This is an important human-confirmation boundary. Confirmation should be explicit for each field and should not imply that the application has verified a prescription with a pharmacist or prescriber.

IV. Geospatial Coordination and Reminder Delivery

A. Facility search and ranking

The described primary source is Google Places through backend proxy routes, with browser geolocation and pre-cached OpenStreetMap hospital coordinates as a fallback. The dossier names Haversine distance, road travel time, facility capability, and rating as ranking inputs. It supplies no numeric search radius, minimum facility quorum, emergency/outpatient weights, normalization parameters, rating prior, or iterative radius-expansion increment. The following notation therefore records only the stated design intent, not a reproducible deployed score:

$$S(h) = w_d f_d(d_h) + w_o O_h + w_r R_h + w_e E_h, \quad (1)$$

where d_h is distance, O_h is an open-status indicator, R_h is a rating term, and E_h indicates a stated emergency/trauma capability. The functions, signs, emergency and outpatient weight values, missing-data rules, rating-smoothing equation, and search-radius threshold remain unspecified. The dossier does provide a facility-specialty score:

$$S_{\text{spec}}(f, \sigma) = \begin{cases} 1.0, & \text{tertiary specialty care,} \\ 0.6, & \text{general ED or medicine,} \\ 0.2, & \text{basic primary outpatient.} \end{cases} \quad (2)$$

Travel time is reportedly sourced from Google Maps when available and estimated with an offline urban-speed profile otherwise. Geospatial results are said to be cached in memory and IndexedDB for 24 h, but the geohash precision and invalidation behavior are not supplied; designated 24/7 facilities are reportedly treated as open. The dossier further describes iterative radius expansion if fewer than a minimum number of accredited emergency facilities are found, but the minimum and expansion schedule are missing. A safe implementation should first exclude facilities that do not meet required emergency capability, then rank eligible sites; a high public rating or an "open" listing is not a proxy for staffed emergency care.

For coordinates (ϕ_1, λ_1) and (ϕ_2, λ_2), Haversine distance may be calculated as

$$a = \sin^2\frac{\phi_2 - \phi_1}{2} + \cos\phi_1 \cos\phi_2 \sin^2\frac{\lambda_2 - \lambda_1}{2}, \quad (3)$$

$$d = 2R \arcsin(\sqrt{a}), \quad (4)$$

where angular coordinates are in radians and R is the selected Earth-radius constant. Straight-line distance is not road travel time. Cached coordinates may be stale; the interface should disclose source and timestamp, show uncertainty, and allow the user to select another option. The described SOS workflow requires explicit confirmation before sharing location or medical snippets. It presents a countdown and siren, then allows the user to cancel or proceed; the call action opens the native dialer with 112 or a selected contact prefilled, while a WhatsApp message is prepared for user review. These are user-initiated handoffs, not automated emergency-service dispatch. The sample SOS text and example coordinates in the project dossier are illustrative and are not reproduced as real patient data.

TABLE IV
APPLICATION TRIAGE TAXONOMY REPORTED IN THE DOSSIER (NOT A VALIDATED ESI CROSSWALK)

Application tier	Claimed ESI mapping	Reported example criteria	Reported action / care window
Critical	ESI 1–2	Red flags or unstable vital signs, chest pain, stroke symptoms, respiratory distress	Immediate escalation (target < 15 min), SOS prompt, and facility options; not autonomous dispatch
High	ESI 3	Severe pain, systemic infection, high fever, persistent vomiting	Urgent evaluation within 2–4 h (reported target; validation source not supplied)
Moderate	ESI 4	Mild infection, localized pain, uncomplicated upper respiratory symptoms	Teleconsultation or primary care within 24–48 h (reported target)
Low	ESI 5	Minor self-limiting symptoms and refill requests	Home care or elective follow-up; supportive precautions only

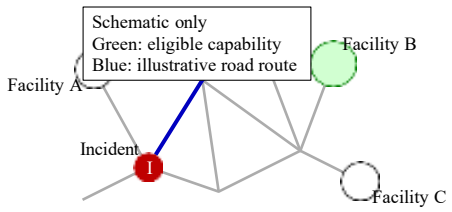


Fig. 3. Schematic geospatial referral view. It contains no real coordinates, facilities, or route-time data.

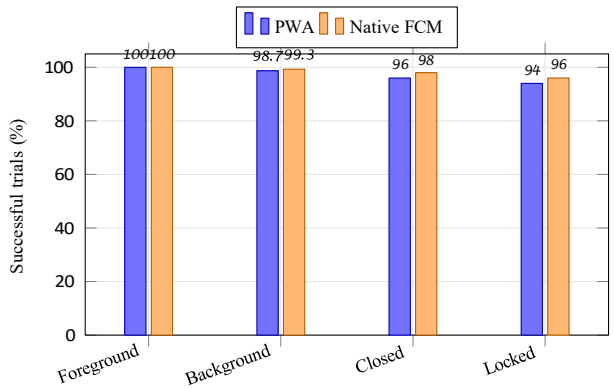


Fig. 4. Notification success by reported device state. The test deadline and per-device allocation are unavailable; percentages summarize the supplied state-level counts and should not be generalized to all browsers or devices.

B. Reminder scheduling, Web Push, and sensory alerts

The project describes client IndexedDB storage, a service worker, a server scheduler that checks pending reminders at a two-second interval, and Web Push delivery. The interval is a reported scheduler configuration, not a measured end-to-end delivery guarantee. W3C service-worker execution is managed by the user agent and may be interrupted; delivery after browser or operating-system termination cannot be generalized across platforms [10], [11]. RFC 8030 specifies the push-service protocol, RFC 8291 specifies encrypted payload processing using P-256 ECDH and AES-GCM, and RFC 8292 specifies VAPID application-server identification; RFC 8291 does not specify Curve25519, as claimed in the dossier [12]–[14].

The updated dossier reports 500 notification trials on a Pixel 7 (Android 14), Galaxy A53 (Android 13), iPhone 14 (iOS 17.4 PWA), and Dell XPS 15 (Windows 11/Chrome 124). A trial counted as successful only if the payload reached the service worker, the system-tray banner appeared with sound/vibration within a strict time limit, and selecting it opened the relevant medication card. The numeric deadline was blank in the supplied text, so the results cannot be interpreted against a reproducible time cutoff. Notifications at or beyond that missing cutoff were classified as late/degraded failures. A unique tag `med-reminder-{id}-{timestamp}` reportedly deduplicated client notifications. The earlier 99.4% aggregate claim conflicts with the newer count of 488/500 (97.6%); the count-based result is used here.

The native comparison used React Native 0.73 with Firebase Cloud Messaging on a Pixel 7 (Android 14), with reported identical network topology. Twelve PWA trials and seven native trials were missed or late. The PWA's aggregate success was 1.0 percentage point lower and median latency 17.6 ms higher than the native result. The study reports these as author-supplied measurements; without the deadline, raw timestamps, trial randomization, and detailed event logs, they do not establish broad parity or production reliability. The other PWA test devices are not part of the stated native comparison.

The Sensory-Agent is described as synthesizing an E5–G#5–B5–E6 chime locally and using gentle or urgent vibration patterns. Where vibration is unavailable, the dossier reports an 85-Hz Web Audio fallback. Client-edge synthesis avoids streaming audio but does not remove device, permission, audio-context, or background-execution constraints. The interface should provide visual and accessible alternatives and must not imply that an alarm has contacted emergency services. Separate compatibility claims list Android 12–14 with Chrome 120–124/Edge Mobile; iOS 16.4 and later with Safari PWA/Chrome iOS; and Windows 11, macOS Sonoma, or Ubuntu 22.04 with Chrome 124, Firefox 125, Edge 124, or Safari 17. Trial-level pairings among these environments were not provided.

V. EVALUATION AND REPORTED RESULTS

A. Datasets and baselines

The prescription evaluation reportedly used 250 real-world images collected from August 2025 through January 2026, spanning 94 active ingredients and 112 commercial brands.

TABLE V
PWA AND NATIVE FCM NOTIFICATION TRIALS (500 TRIALS PER IMPLEMENTATION)

State (trials)	PWA success	PWA median / 95th	Native FCM success	Native median	PWA–native difference
Foreground active (150)	150/150 (100%)	42.1 / 88.0 ms	150/150 (100%)	38.4 ms	0.0 pp; +3.7 ms
Background minimized (150)	148/150 (98.7%)	148.5 / 320.0 ms	149/150 (99.3%)	131.2 ms	−0.6 pp; +17.3 ms
Browser closed / PWA (100)	96/100 (96.0%)	210.0 / 480.0 ms	98/100 (98.0%)	174.0 ms	−2.0 pp; +36.0 ms
Device locked / Doze (100)	94/100 (94.0%)	380.0 / 840.0 ms	96/100 (96.0%)	292.0 ms	−2.0 pp; +88.0 ms
Composite (500)	488/500 (97.6%)	142.4 / 318.5 ms	493/500 (98.6%)	124.8 ms	−1.0 pp; +17.6 ms

The listed image types include blister packs, printed discharge scripts, handwritten prescriptions, syrup bottles, and degraded samples; counts per image type are not supplied. All 250 images are described as a held-out zero-shot set, also divided into five groups of 50, but the role of these groups and any fold-wise analysis are unspecified. Labels were reportedly assigned independently by a licensed pharmacist and an internal-medicine resident, with disagreements referred to a third senior-physician adjudicator. The dossier specifies unweighted Cohen's kappa but the observed agreement, chance agreement, and kappa value were missing from the supplied text. Prescription identifiers were reportedly masked; supporting ethics documentation was not provided.

The triage benchmark reportedly contains 180 synthetic vignettes (45 critical, 55 high, 50 moderate, and 30 low), adapted from the ESI v4 casebook and BMJ Best Practice vignettes and held out from prompt engineering. Two emergency physicians reportedly reviewed the reference labels. The authors state that inter-rater agreement was evaluated with unweighted Cohen's kappa; the number of rated vignettes, observed and chance agreement, and kappa value were absent from the supplied text. Vignette text, source permissions, case-level labels, and adjudication logs are unavailable. Earlier summary percentages for prescription, triage, and push performance conflict with the new count-based tables below. Because the earlier baseline values have no matching numerators, denominators, or confusion matrices, those comparisons are not repeated as verified results.

The reported agreement statistic is Cohen's unweighted kappa,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{5}$$

where p_o is observed agreement and p_e is chance-expected agreement from annotator marginals. The supplied follow-up omitted both terms and the resulting κ values for triage and prescription labels, so inter-annotator reliability cannot be checked or reconstructed from the summary counts.

To make a future replication interpretable, a preregistered protocol should identify the unit of analysis, inclusion/exclusion rules, annotator instructions, adjudication procedure, and whether annotators were blinded to model output. For prescription images, it should state image capture conditions, language and script, whether multiple medicines appear per image, the exact normalization used for each field, and whether the split is performed by image, patient, or prescription. For triage vignettes, it should release the de-identified case text and reference rationale where permissions allow, define urgency direction before scoring, and record the exact prompt, model identifier, API response metadata, and fallback state. The current summary provides insufficient information to determine whether images from the same prescription or near-duplicate vignettes could cross evaluation partitions. No patient-level effectiveness or prospective clinical outcome is inferred from these offline tests.

The earlier dossier names Tesseract 5.3 plus regular expressions, an unconstrained GPT-3.5-Turbo model, and a rule-based symptom tree as baselines. The updated triage matrices include comparator results, but method labels remain inferred from sequence; no new count-based prescription-extraction comparison was supplied. Prompt text, model/API snapshots, decoding parameters, OCR preprocessing, evaluator blinding, and baseline tuning also remain unavailable.

B. Metric interpretation and uncertainty

The reported extraction endpoint is strict exact match after whitespace trimming and case-folding. For field j, the observed exact-match proportion is $\hat{p}_j = c_j/n$, where c_j is the reported number of exact matches and $n = 250$. Complete-record accuracy is the fraction for which all four fields match simultaneously; it is therefore a stricter endpoint than any individual field. These field-level outcomes are paired within images, so comparing their percentages as if they were independent samples would be inappropriate. The partial-match counts use a different criterion and are retained separately. Even a correct text extraction does not verify that the source prescription itself is appropriate or current.

For the triage matrix M, with ground-truth class i and predicted class j, accuracy is $\sum_i M_{ii}/N$. Critical recall is $M_{CC}/\sum_j M_{Cj}$. Under-triage and over-triage are counted by the ordering Critical > High > Moderate > Low: an under-triage error predicts a less urgent class than the reference, and an over-triage error predicts a more urgent one. This directional split is clinically more informative than a single accuracy value but still treats all adjacent-class errors equally and does not encode patient outcomes. Since the matrices were not linked to named methods in the supplied dossier, results tied to those method labels remain conditional on the assumed ordering.

Confidence intervals in the tables describe binomial sampling uncertainty under the stated calculations; they do not account for selection bias, correlated observations, label error,

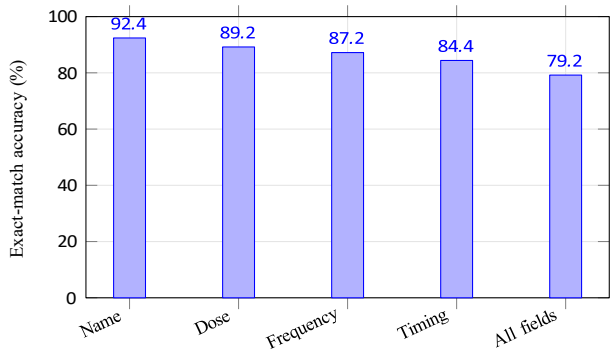


Fig. 5. Exact field and complete-record accuracy on the reported 250-image evaluation set.

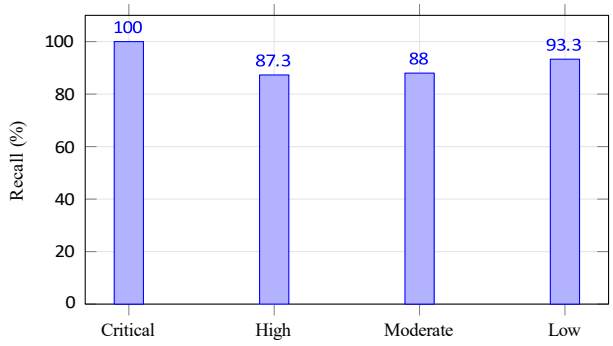


Fig. 6. Per-tier recall derived from the assumed MediSentinel confusion matrix. Critical recall is 45/45; other class recalls are 48/55, 44/50, and 28/30. Labels and method mapping require confirmation from the authors.

repeated prompt development, or dataset shift. The critical-case interval is especially important: 45 correct examples yield an exact two-sided 95% interval whose lower bound is approximately 92%, not certainty of perfect sensitivity. Similarly, 198/250 complete records is a point estimate for the supplied sample, not a guarantee for new handwriting styles, scripts, image quality, or clinical settings. These intervals should be read as descriptive uncertainty around the reported counts, not as a substitute for independent external validation.

C. Prescription-field extraction

Table VI reports exact-match counts and partial-match counts. Exact matching is defined as character equality after whitespace trimming and case-folding; partial matching identifies the active molecule or frequency despite missing salt forms or abbreviated units. The complete-record measure requires all four fields to match exactly. The follow-up specifies 95% Wilson score intervals with continuity correction, but its interval endpoints were absent from the supplied text. Values in the last column are therefore recalculated here from the reported counts using that method; partial-match counts are separate and should not be added to exact counts.

D. Triage and emergency safety

The following confusion matrices transcribe the three matrices in the dossier, whose ordering has now been confirmed as ground-truth rows and predicted-tier columns in the order Critical, High, Moderate, Low. Their method mapping is assumed from their sequence in the dossier (MediSentinel, standard LLM, rule-based tree) because no method labels accompanied the matrices; author confirmation is still required. Rates in Table VII are recalculated from those counts and supersede earlier headline percentages. Under-triage means assignment to any less-urgent tier than the reference; over-triage means assignment to any more-urgent tier.

Using the specified two-sided Clopper–Pearson method, 45/45 critical recall has a recalculated 95% interval of [92.1, 100.0]%; overall-accuracy intervals in Table VII use the stated Wilson score method with continuity correction. The 45/45 result is an observed sample outcome, not evidence that no real-world emergency will be missed. The dossier's earlier values (94.4% system accuracy, 84.4% standard-LLM accuracy, and 76.7% rule-tree accuracy) do not match the supplied matrices (91.7%, 76.1%, and 64.4%, respectively, under the assumed method mapping). The earlier under-/over-triage rates also do not match matrix-derived rates. No prospective clinical study or silent-mode comparison with usual care is reported.

The per-tier view in Fig. 6 makes the class denominator visible, but it does not correct for the synthetic vignette design or establish performance in a population with different prevalence. Critical cases make up 25% of this deliberately stratified test set (45 of 180), and that fraction should not be interpreted as an estimate of emergency prevalence. The matrix-derived 100% critical recall is a count in a finite sample; the lower confidence bound and absence of external replication remain relevant to any safety interpretation.

VI. Ethics, Privacy, and Regulatory Governance

A. Data protection and security

The dossier asserts explicit consent, transient image processing, zero image retention, one-click deletion from IndexedDB and Firestore, TLS 1.3 in transit, AES-256 at rest, and no use of protected health information for model pre-training. Triage determinations, red-flag triggers, and acknowledgements are reportedly written to encrypted client storage and synchronized to `/users/{uid}/triage_logs`. The authors state that the study was exempt under 45 CFR 46.104(d)(4), but no institutional determination or protocol identifier was supplied. These are implementation and ethics claims, not independently verified controls; the application source and data-flow evidence were unavailable. A privacy assessment should document where image bytes and prompts transit, provider-side retention and logging, backup/telemetry behavior, location precision, data-controller/processor roles, retention periods, deletion propagation, incident response, and access logging. TLS between client and backend is transport encryption; it should not be described as end-to-end encryption when the backend or model provider can process the payload.

The Indian Digital Personal Data Protection Act, 2023 and the Digital Personal Data Protection Rules, 2025 should be assessed against the actual deployment, processing purposes,

TABLE VI
PRESCRIPTION IMAGE EXTRACTION RESULTS (N=250; 95% CONFIDENCE INTERVALS RECALCULATED USING WILSON SCORE WITH CONTINUITY CORRECTION)

Field	Exact matches	Partial matches	Exact accuracy	Recalculated 95% CI (%)
Medication / brand name	231/250	244/250	92.4%	[88.2, 95.2]
Dosage / strength	223/250	238/250	89.2%	[84.5, 92.6]
Frequency	218/250	235/250	87.2%	[82.3, 91.0]
Administration timing	211/250	229/250	84.4%	[79.2, 88.5]
Complete four-field record	198/250	222/250	79.2%	[73.5, 84.0]

TABLE VII
TRIAGE CONFUSION COUNTS AND METRICS DERIVED FROM THE SUPPLIED 180-VIGNETTE MATRICES

Actual tier (n)	MediSentinel predicted C/H/M/L	Standard LLM predicted C/H/M/L	Rule tree predicted C/H/M/L
Critical (45)	45/0/0/0	36/7/2/0	31/11/3/0
High (55)	4/48/3/0	3/39/11/2	8/32/12/3
Moderate (50)	0/3/44/3	1/5/37/7	2/7/31/10
Low (30)	0/0/2/28	0/1/4/25	0/2/6/22

Derived metric	MediSentinel	Standard LLM	Rule-based tree
Critical recall	45/45 (100.0%)	36/45 (80.0%)	31/45 (68.9%)
Overall accuracy	165/180 (91.7%)	137/180 (76.1%)	116/180 (64.4%)
Accuracy 95% CI (Wilson CC)	[86.4, 95.1]	[69.1, 82.0]	[56.9, 71.3]
Under-triage	6/180 (3.3%)	29/180 (16.1%)	39/180 (21.7%)
Over-triage	9/180 (5.0%)	14/180 (7.8%)	25/180 (13.9%)

notices, consent flows, retention, and applicable commencement provisions [15], [16]. HIPAA applies only in its statutory covered-entity/business-associate context, and GDPR only where its territorial and material scope is met; generic claims of "HIPAA/GDPR alignment" require evidence and should not be used as blanket compliance statements.

B. Clinical software classification

The dossier calls MediSentinel a non-diagnostic clinical decision-support tool and asserts exemption from U.S. and Indian medical-device oversight. A disclaimer or non-diagnostic label alone does not determine classification. Intended use and actual functions—including patient-specific urgency assessment, emergency escalation, and facility recommendations—must be assessed under applicable jurisdictional rules. FDA issued updated Clinical Decision Support Software guidance in January 2026, and CDSCO published 2026 guidance on medical-device software under India's Medical Devices Rules; the authors should obtain a jurisdiction-specific regulatory assessment before claiming exemption or deployment readiness [17], [18].

C. Human factors and intended use

The dossier reports WCAG 2.1 AA contrast review, ARIA screen-reader labels, and bilingual text-to-speech; no audit report or task-based accessibility study was supplied. The product should be evaluated for automation bias, false reassurance, alarm fatigue, multilingual coverage, and unintended disclosure through prefilled emergency messages. A user should see when information is missing, when a facility listing is stale, and when an answer is uncertain. SOS calls and WhatsApp messages should remain user initiated with an explicit preview of location and message content. No clinical advice in this paper should be used as patient-specific instruction; MediSentinel is not validated for unsupervised care decisions.

VII. RELATED WORK AND SYSTEM POSITIONING

Prior work in this area examines narrower questions than the integrated application workflow described here. Comparative studies have investigated language-model triage and emergency-department recommendation tasks [4], [5]. Such work motivates evaluation of model outputs against clinical references, but the results of those studies cannot be transferred to MediSentinel: model versions, prompts, case distributions, interfaces, and endpoints differ. They are relevant methodological context rather than direct baselines for the author-reported vignettes in this manuscript.

Multi-agent medical reasoning has also been explored as a way to organize specialist roles in zero-shot tasks [6]. MediSentinel uses the "agent" label for role-separated application components that coordinate across browser state, a service worker, and REST services. Without controlled comparisons, it remains unknown whether this decomposition improves accuracy, reduces error propagation, or merely distributes responsibilities across modules. The term should not imply independent evidence, consensus, or fault tolerance.

Clinical triage systems are also shaped by formal operational protocols. The ESI framework is a five-level emergency-department triage tool, while the project defines a four-tier interface with its own action windows [2]. This paper therefore treats the application's tiers as a project taxonomy rather than an ESI implementation. Likewise, ethical and governance guidance for AI in health emphasizes the need to consider

human oversight, context, and accountability beyond model-level performance [7]. MediSentinel's relevant contribution is consequently best assessed as an integrated workflow prototype, and not as a new triage protocol or a general-purpose clinical reasoning benchmark.

The combination of prescription interpretation, reminders, push delivery, and geospatial referral creates cross-component dependencies that are not captured by evaluating a symptom classifier alone. An incorrect extraction could generate a technically successful but clinically incorrect reminder; a correct urgency label could still be undermined by stale facility information or inaccessible notification behavior. A system-level evaluation should therefore measure end-to-end task completion and component-specific failures separately, with denominators and inclusion criteria disclosed for each endpoint.

VIII. Risk Analysis and Failure Containment

Table VIII organizes the principal failure paths as hazards requiring evidence, rather than as a claim that current safeguards have been verified. Severity depends on context: a wrong reminder may be inconsequential in one case and consequential in another, while a failed notification can be masked if users assume delivery. Risk management should focus on both preventing failures and making residual uncertainty observable to the user and evaluator.

The table highlights an important distinction between a local safety feature and system-level risk control. A regular-expression guardrail may be testable as a software component, but its presence alone does not show that every hazardous expression is detected or that users respond appropriately to its output. Similarly, a message stating that a service is unavailable is only useful if the interface avoids presenting stale model output as current. Verification should combine unit tests, integration tests, fault injection, and observation of user behavior; passing one layer cannot replace evidence from the others.

Residual-risk communication should be specific. For example, an uncertain prescription field should be identified by name rather than hidden behind a single confidence badge; an offline result should be labeled as rule-based and possibly incomplete; and a facility result should expose when it depends on cached data. These measures do not make an unsafe output safe, but they can reduce false certainty and give users a meaningful opportunity to seek human assistance. The product should also avoid suggesting that an alert, SOS countdown, opened dialer, or prepared message has completed an external action until that action is confirmed by the relevant service.

IX. Staged Validation and Deployment Plan

The current evidence is retrospective, author reported, and based on supplied aggregate counts. A defensible next step is a staged validation program with explicit stop/go decisions, independent review, and a predeclared analysis plan. The stages in Table IX are proposals, not work claimed to have been completed. Numeric acceptance thresholds should be selected before data collection by clinical, technical, privacy, and accessibility stakeholders; they cannot be selected from the dossier after observing results.

The first stage should resolve the current mismatch between headline summaries and count-derived values, as well as the assumed ordering of the triage matrices. It should also clarify whether the reported 250-image folds were used for development, analysis, or only bookkeeping. Until those facts are resolved, adding more decimal places or aggregating across modalities risks implying a level of precision not supported by the record.

External evaluation should avoid tuning the prompt against the same cases used for final reporting. A clear separation between development and test material, with case-level duplicate checks, helps limit leakage. For clinical labels, the study should retain initial independent ratings and disagreements rather than reporting only adjudicated labels; this enables calculation of agreement and exposes ambiguity in the reference standard. Where the dataset is deliberately enriched for rare critical presentations, stratified performance should be reported while any deployment-relevant predictive values are recalculated or modeled for plausible prevalence settings. No single score should be treated as a universal release criterion.

Any subsequent prospective study should first operate in silent mode, with outputs hidden from care decisions and compared against independently reviewed outcomes. That design can characterize workflow behavior without exposing patients to unvalidated recommendations, though it cannot by itself prove that displaying the system improves outcomes. A later interventional evaluation would require its own protocol, ethical review, monitoring plan, and explicit clinical responsibility model. The staged sequence is therefore not a deployment endorsement; it is a way to define what evidence is missing and to make decisions reversible when new safety information emerges.

X. Discussion

MediSentinel combines an application-level event bus with server services rather than a dedicated distributed-agent platform. This can reduce dependence on continuous server orchestration for local UI interactions, but important operations still rely on external inference, geospatial, authentication, and push services. The term "multi-agent" describes role-specialized software functions in the dossier; independence, fault isolation, and agent-to-agent reliability have not been measured.

The most consequential reported result is 45/45 sensitivity on critical vignettes. Its value depends on representative case construction, independent reference labels, and a sufficiently large test cohort. A regex red-flag layer can improve recall for explicitly represented patterns, but its coverage must be evaluated across synonyms, languages, negation, comorbidities, and malformed input. Similarly, structured output can simplify downstream validation without eliminating semantic errors. The staged evaluation should therefore include adversarial and out-of-distribution cases, subgroup analyses, failure injection, and human-factors observation before any patient-facing deployment.

TABLE VIII
SYSTEM-LEVEL FAILURE SCENARIOS AND EVIDENCE NEEDED BEFORE DEPLOYMENT

Failure scenario	Plausible consequence	Control and verification evidence required
Prescription image is blurred, cropped, or ambiguous	Incorrect medication, strength, or timing is copied into a reminder; a technically valid schedule may appear trustworthy.	Block saving when required fields are unresolved; preserve image-to-field traceability; test by image quality, script, and prescription subtype; measure both exact-field errors and user correction rates.
Triage model misses an indirect or code-mixed red flag	User receives a lower urgency label and may delay care.	Test paraphrases, negation, spelling variation, code switching, comorbidities, and conflicting answers; publish class-specific false-negative counts; require clinician-approved escalation behavior.
Fallback or connectivity state is not visible	User mistakes a limited offline rule result for a complete model assessment, or assumes a failed request was answered.	Show inference and data freshness status; distinguish unavailable, partial, and completed states; inject timeout, rate-limit, malformed response, and repeated-failure conditions.
Facility record is stale or capability is misclassified	A nearby but unsuitable facility is ranked ahead of an emergency-capable site.	Separate capability eligibility from ranking; disclose source and last refresh; test missing, conflicting, and stale attributes; keep the selection and navigation user controlled.
Push delivery is late, duplicated, or suppressed	A reminder is missed, repeated, or interpreted as confirmation that a dose was taken.	Define a measurable deadline and delivery endpoint; test device states and operating-system restrictions; distinguish scheduled, delivered, opened, and user-acknowledged events.
Location, symptoms, or image data are exposed or retained unexpectedly	Sensitive health and location information may be disclosed beyond the user's expectation.	Document data flows and provider retention; verify deletion, access controls, logs, and backup behavior; test that location and emergency-message sharing require explicit consent.

The project dossier proposes two related works—"Building an AI-Driven Symptom Checker Using Python Django for Enhanced Telemedicine Services" and "Medaid: A Safety-First AI Triage System for Rural Healthcare." The author, venue, and DOI metadata supplied for these titles require bibliographic verification before inclusion: the cited identifiers do not reliably establish those title/author combinations. They are therefore not used as evidentiary citations in this draft. The authors should provide publisher or DOI landing-page records before the related-work comparison is finalized.

A. Limitations

This manuscript is based on an author-provided architecture and evaluation dossier; the workspace contained the LaTeX manuscript but no application code, images, annotated vignettes, model prompts, event logs, or study protocol. Consequently, implementation claims and numeric performance values could not be independently audited or reproduced. Key open issues include the unlabeled method mapping for the triage matrices, conflicts with earlier headline metrics, missing Cohen's kappa values and agreement counts, the rationale for five image folds, prescription-subtype counts and comparison baselines, the numeric push success deadline and per-device allocation, actual model-serving IDs, API timeout and retry-delay/circuit-breaker thresholds, and geospatial weights, normalization parameters, initial radius, minimum quorum, and expansion schedule. The reported IRB exemption, security controls, and accessibility checks also lack supporting documentation. These gaps preclude exact replication or claims of broad reliability and clinical effectiveness.

Several distinct validity threats should be kept in view. First, construct validity is limited by incomplete endpoint definitions: "success" for push combines delivery, presentation, and user navigation, while its deadline is missing; triage accuracy compresses clinically directional errors into a single count; and extraction exact match treats all character-level discrepancies alike even when their practical consequences differ. The figures and percentages are therefore descriptive summaries of the supplied definitions, not interchangeable measures of safety or usefulness.

Second, internal validity and reproducibility are constrained by missing artifacts. The named model aliases, prompts, output schema versions, OCR settings, annotation instructions, adjudication records, and event timestamps were not supplied. An API alias can route to a changed underlying model, while minor prompt or preprocessing changes can affect output. Without immutable versions and executable analysis code, it is not possible to determine whether a difference arose from a model, a software change, a change in test data, or a calculation error. Earlier summary figures conflict with the available counts, underscoring the need for a single versioned results manifest.

Third, external validity is uncertain. The image set's language, script, device, handwriting, and prescription-type composition is not enumerated; the triage set is synthetic and enriched for critical cases; and the notification sample has an unspecified allocation across device states and hardware. Results may differ in other regions, care pathways, languages, network conditions, operating-system versions, or populations with different symptom prevalence. Aggregate figures conceal these variations, and no subgroup claim should be made until denominators and error cases are available.

Finally, the paper cannot establish usability, clinical utility, or net benefit. A benchmark does not show whether people understand the difference between an application tier and a clinical diagnosis, correct a wrong medication field, locate an appropriate facility, or act safely after an alert fails. Nor does a simulated or offline result measure whether the system improves health outcomes compared with usual care. Those questions require independent human-factors research and appropriately governed prospective studies, as described in the proposed validation plan. The current manuscript should be read as a technical description and evidence inventory, not a clinical performance claim.

TABLE IX
PROPOSED STAGED EVIDENCE PROGRAM

Stage	Evaluation objective	Required evidence artifact	Gate before proceeding
1. Reproducibility	Reconstruct reported image, vignette, and notification results.	Frozen datasets or access procedure; annotation guide; prompt and model identifiers; environment manifest; raw event logs; executable metric scripts.	Independent analyst reproduces counts and resolves discrepancies; protocol and data provenance are documented.
2. Technical verification	Test application logic and dependencies outside the happy path.	Unit and integration tests; fault-injection report; schema and storage migration tests; retry, offline, stale-cache, and duplicate-notification traces.	No unresolved critical software hazard under the risk controls approved for the next stage.
3. External retrospective evaluation	Estimate performance on independently collected, held-out material.	Prespecified sampling and labeling plan; subgroup breakdowns; adjudication records; confidence intervals; error review stratified by severity and input quality.	Clinical reviewers judge residual error patterns acceptable for a supervised research setting; no threshold is inferred post hoc.
4. Human-factors and accessibility study	Observe comprehension, correction, and action selection.	Approved usability protocol; representative language and accessibility needs; task completion, error, and comprehension measures; adverse-event and distress monitoring.	Users understand uncertainty, confirmation, and SOS boundaries; material usability hazards are mitigated and retested.
5. Silent prospective evaluation	Compare system suggestions with usual care without influencing decisions.	Ethics and governance approval; consecutive or otherwise justified sampling; clinician reference review; drift monitoring; prespecified stopping rules.	Independent oversight reviews safety signals and determines whether a separately governed interventional study is justified.
6. Controlled deployment and surveillance	Monitor performance under defined operational constraints.	Versioned release and rollback plan; incident reporting; monitoring for latency, subgroup drift, and stale data; periodic privacy/security review.	Ongoing authorization, clinically governed thresholds, and reliable rollback remain in place; stop criteria trigger suspension and review.

XI. CONCLUSION

This paper describes MediSentinelAI as a reported PWA combining multimodal prescription extraction, four-tier symptom triage, offline red-flag matching, medication reminders, sensory alerts, and geospatial emergency coordination. The updated count-based dossier reports 79.2% exact complete-record accuracy on 250 prescription images, 91.7% triage accuracy with 45/45 observed critical-case recall on an assumed method mapping of 180 vignettes, and 97.6% notification success across 500 trials compared with 98.6% for the reported native FCM baseline. The PWA-native difference was 1.0 percentage point in delivery and 17.6 ms in aggregate median latency under the stated test, but the success deadline and raw logs remain unavailable. Results conflict with earlier headline summaries and are not independently verified. Future deployment requires reproducible evaluation records, verified privacy controls, and independent clinical and regulatory review. EHR integration using FHIR/HL7 and hospital bed-occupancy streaming remain future work.

REFERENCES

[1] World Health Organization, *Emergency Triage Assessment and Treatment (ETAT) Course: Manual for Participants*. Geneva, Switzerland: WHO, 2005.
[2] N. Gilboy, T. Tanabe, D. Travers, and A. M. Rosenau, *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4*. Rockville, MD, USA: Agency for Healthcare Research and Quality, 2012.
[3] R. M. Anjana *et al.*, "Metabolic non-communicable disease health report of India: The ICMR-INDIAB national cross-sectional study," *The Lancet Diabetes & Endocrinology*, vol. 11, no. 7, pp. 474–489, 2023.
[4] L. Masanneck *et al.*, "Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study," *Journal of Medical Internet Research*, vol. 26, Art. no. e53297, 2024, doi: 10.2196/53297.
[5] C. Y. K. Williams, B. Y. Miao, A. E. Kornblith, and A. J. Butte, "Evaluating the use of large language models to provide clinical recommendations in the emergency department," *Nature Communications*, vol. 15, Art. no. 8236, 2024, doi: 10.1038/s41467-024-52415-1.
[6] X. Tang *et al.*, "MedAgents: Large language models as collaborators for zero-shot medical reasoning," in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 2024, pp. 599–621, doi: 10.18653/v1/2024.findings-acl.33.
[7] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva, Switzerland: WHO, 2021.
[8] Google, "Structured outputs," *Gemini API Documentation*. [Online]. Available: https://ai.google.dev/gemini-api/docs/structured-output. Accessed: Oct. 23, 2026.
[9] Google, "Gemini API deprecations," *Gemini API Documentation*. [Online]. Available: https://ai.google.dev/gemini-api/docs/deprecations. Includes Gemini 1.5 Flash shutdown information and current model routing; accessed Oct. 23, 2026.
[10] World Wide Web Consortium, "Service Workers," W3C Recommendation, Nov. 2022.
[11] World Wide Web Consortium, "Push API," W3C Recommendation, 2023.
[12] M. Thomson and P. Beverloo, "Generic event delivery using HTTP," RFC 8030, Internet Engineering Task Force, Oct. 2016.
[13] M. Thomson, "Message encryption for Web Push," RFC 8291, Internet Engineering Task Force, Nov. 2017.
[14] M. Thomson, "Voluntary application server identification (VAPID) for Web Push," RFC 8292, Internet Engineering Task Force, May 2018.
[15] Government of India, *The Digital Personal Data Protection Act, 2023*, Act No. 22 of 2023, Aug. 11, 2023.

[16] Ministry of Electronics and Information Technology, Government of India, *Digital Personal Data Protection Rules, 2025*, Nov. 2025.
[17] U.S. Food and Drug Administration, *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff*, Jan. 2026.
[18] Central Drugs Standard Control Organization, Government of India, *Guidance Document on Medical Device Software under the Medical Devices Rules, 2017*, 2026.