Abstention-Aware Message Routing for Patient Portal Safety Review in Pediatric Primary Care

Yahya Lemrabott[1], Sidi Ely[2], AND Moussa Ould[3]

[11]Department of Computer Science, Faculty of Science and Technology, Université de Nouakchott Al Aasriya, Avenue de l'Indépendance, Tevragh-Zeina, 1122 Nouakchott, Mauritania
[22]Department of Information Systems and Management, Institut Supérieur du Numérique, Rue de l'Université, Ksar, Nouakchott, Mauritania
[33]Department of Applied Mathematics and Informatics, École Normale Supérieure de Nouakchott, Boulevard Moktar Ould Daddah, Nouakchott, Mauritania

ABSTRACT

Patient portal messaging has become a routine part of pediatric primary care, but clinics often struggle to route messages quickly to the right staff member. A message about fever, medication access, school forms, vaccine records, or worsening breathing may arrive through the same digital channel, even though each requires a different workflow. Automated message routing can help reduce delay, but unsafe routing may occur when a model assigns confident labels to vague, emotionally charged, or clinically incomplete messages. This paper reports an empirical study of abstention-aware message routing for pediatric patient portals. We assembled 94,360 de-identified portal messages from four pediatric primary care networks and manually adjudicated 12,480 messages into seven routing categories: routine administrative, scheduling, medication refill, vaccine or form request, nurse clinical review, same-day clinician review, and urgent safety escalation. A transformer classifier was compared with a calibrated abstention-aware routing model that separates clear routing decisions from messages requiring human intake. The proposed model improved macro F_1 from 0.742 to 0.817, reduced unsafe under-routing from 5.8% to 2.4%, and lowered urgent-message miss rate from 9.6% to 3.9%. It abstained on 13.7% of messages, most often when symptoms were underspecified or when administrative and clinical requests were mixed. Prospective-style temporal testing showed smaller degradation than standard classification. The results suggest that portal triage models should be judged not only by routing accuracy, but also by their ability to recognize when a message is not safe to route automatically.

I. Introduction

Digital patient portals have changed the texture of outpatient pediatrics. Parents and guardians send messages about cough, medication refills, school paperwork, vaccine records, behavioral concerns, rashes, injuries, chronic disease plans, billing confusion, and follow-up questions after recent visits [1]. A single morning inbox may include ordinary administrative requests and messages that need same-day clinical attention. The practical difficulty is not that every message is complex. Many are simple. The difficulty is that the simple and the risky arrive together, often written in informal language, with incomplete context, and without a structured chief complaint. Routing is the first decision in this workflow. A message sent to scheduling may wait for an appointment clerk. A message sent to a refill queue may be handled by pharmacy staff or a nurse. A message sent to a clinician queue may receive more clinical review but can increase clinician burden if used too broadly. A message sent to urgent safety review may trigger a same-day call or escalation. When the routing decision is wrong in the conservative direction, staff time is wasted and families may wait longer for routine service. When the routing decision is wrong in the unsafe direction, symptoms that need prompt review can be delayed.

Automated routing systems are attractive because pediatric clinics face high message volume and staffing pressure. A classifier can learn that "I need a copy of the immunization record" usually belongs to forms or vaccine documentation, while "my child is wheezing and the inhaler is not helping" should receive clinical review. However, routing is not the same as ordinary topic classification. A message may contain several intents at once. A parent might ask for a school note and mention persistent fever. Another might request a refill but also report that the child is using the medication more often than prescribed. A message can be short, anxious, sarcastic, incomplete, or copied from a previous thread. These features make forced classification risky.

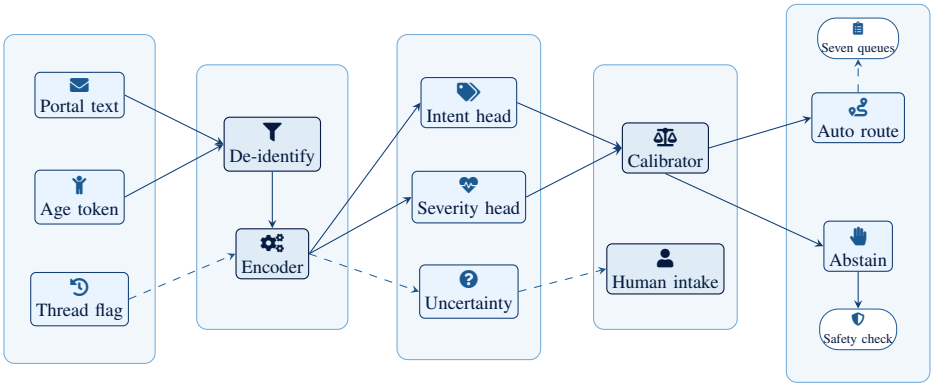

FIGURE 1: Abstention-aware routing separates confident portal-message decisions from messages requiring human intake review.

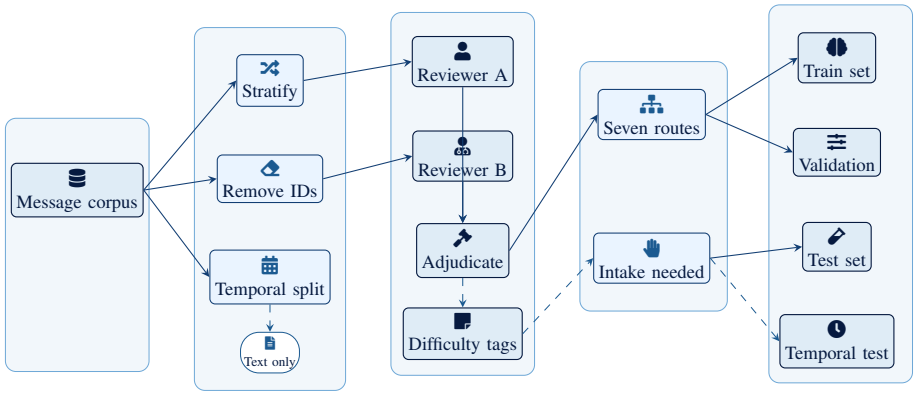

FIGURE 2: Corpus construction links de-identification, adjudication, difficulty tagging, and chronological testing.

This paper studies abstention-aware message routing. The model is allowed to assign a message to a routing category when the category is sufficiently clear, but it can also abstain and send the message to human intake. Abstention is not treated as failure. It is treated as a safety behavior when the available text does not support reliable automatic routing. The central research question is whether a calibrated abstention-aware model can reduce unsafe under-routing while keeping abstention volume small enough to be operationally useful. A secondary question is whether abstentions occur in clinically sensible cases rather than merely in rare categories or long messages.

The study is empirical. It uses de-identified pediatric portal messages, manual routing labels, temporal testing, service-line subgroup analysis, and simulated queue impact. The model is compared with a standard transformer classifier, a class-weighted classifier, a keyword rule system, a hierarchical classifier, and a confidence-threshold baseline. The proposed method differs from a simple threshold because it uses separate uncertainty signals for intent ambiguity, clinical severity ambiguity, and out-of-distribution phrasing. It also learns from adjudicator disagreement, so messages that human reviewers found hard to classify are not forced into the same training treatment as obvious messages.

The design was influenced by a narrow reporting issue in a different medical artificial intelligence study. In discussing answer extraction, Sadanandan and Behzadan (2026) noted that model responses could be excluded when outputs were ambiguous or unparseable, with exclusion causes including hedging, refusals, off-topic answers, and mixed affirmative and negative indicators; their reported rates varied across models and datasets, with some later exclusions reaching 14.6% for RadFM on PadChest [2]. That observation is used here as a cautionary prompt: in safety-facing classification, ambiguous outputs or inputs should not simply disappear from evaluation. The present work therefore measures abstention explicitly and treats unclear messages as a first-class operational category.

The paper focuses on pediatric primary care because the portal environment is distinctive. The sender is usually a parent or guardian rather than the patient. The message may concern an infant, a school-age child, or an adolescent. Urgency depends on age, comorbidity, symptom pattern, and recent care [3]. Messages may include photos, but this study uses text only because image handling differs across clinics and requires separate governance. The model does not provide medical advice, diagnosis, or treatment recommendations [4]. It only assigns routing priority or abstains for human review.

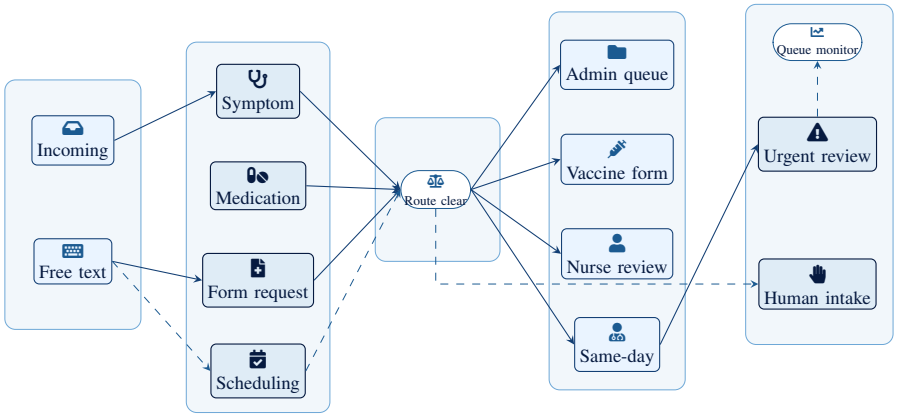

FIGURE 3: Routing policy preserves clinical priority when routine requests and symptom information appear together.

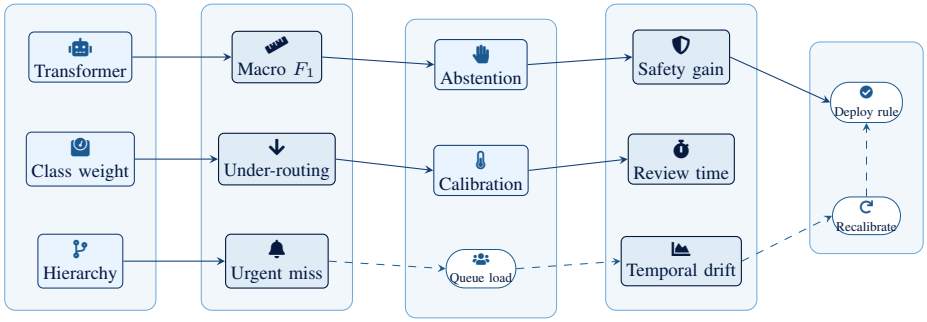

FIGURE 4: Evaluation combines accuracy, abstention, under-routing, urgent misses, queue delay, and temporal degradation.

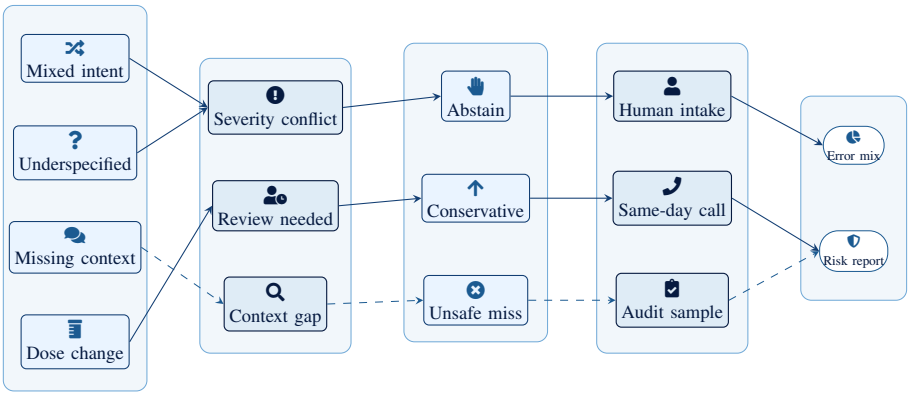

FIGURE 5: Error review separates desirable abstentions from conservative over-routing and unsafe under-routing.

The main finding is that abstention-aware routing improves safety-sensitive performance. On the held-out test set, the standard transformer achieved strong overall accuracy but missed a clinically meaningful fraction of urgent and same-day messages. The abstention-aware model reduced unsafe under-routing by more than half and lowered urgent-message miss rate to below 4.0%, while abstaining on 13.7% of messages. The abstention set was enriched for mixed-intent messages, symptom escalation without duration, medication use suggesting worsening illness, and messages referencing prior threads without enough context. In queue simulation, the model reduced median time-to-clinical-review for high-priority messages without flooding urgent review.

The paper is organized as follows. The next section describes the message corpus, routing categories, annotation process, and data partitions. The modeling section explains the abstention-aware architecture and training approach without mathematical notation. The experimental section presents baselines, evaluation metrics, and temporal tests. The results section reports performance, abstention behavior, and queue simulation. The error analysis section examines unsafe under-routing, over-escalation, subgroup patterns, and

reviewer disagreement. The conclusion discusses practical implications and limitations.

II. Corpus and Labeling Procedure

The corpus contained 94,360 de-identified patient portal messages from four pediatric primary care networks. The networks included two urban academic practices, one suburban multi-site group, and one community network with integrated behavioral health and care management. Messages were sampled from a twenty-month period. Each message represented the first parent or guardian message in a thread, not every reply in the conversation. Follow-up replies were used only when annotators needed context for label adjudication. Automated appointment reminders, bulk announcements, billing-only system messages, and messages generated entirely by clinic staff were excluded.

Messages were de-identified before modeling. Names, addresses, phone numbers, exact dates, record numbers, and other direct identifiers were removed or replaced with typed placeholders. Child age in months or years was retained as a structured variable when available because routing can depend on age. For example, fever in a young infant is not handled the same way as fever in an older child [5]. The model did not receive patient names, clinician names, insurance identifiers, or location-specific staff names. Practice network was used for evaluation and calibration, but not as a direct routing feature in the main model.

The final manual annotation set contained 12,480 messages. Sampling was stratified to include common administrative messages, medication requests, symptom messages, and messages flagged by preliminary rules as potentially urgent. A random sample of ordinary messages was included so that reviewers did not see only difficult cases. The remaining 81,880 messages were used for weakly supervised pretraining and language adaptation, not for final test scoring. Weak labels came from historical routing queues and keyword triggers, but they were not trusted as ground truth because staff routing habits differ across clinics.

The routing taxonomy had seven categories. Routine administrative messages included address updates, record requests unrelated to vaccines, portal access problems, and non-clinical questions. Scheduling included appointment requests, cancellations, rescheduling, and visit type questions. Medication refill included straightforward refill requests without evidence of worsening symptoms or dose confusion. Vaccine or form request included immunization records, school forms, sports forms, camp forms, and vaccine scheduling questions. Nurse clinical review included symptom questions that did not clearly require same-day clinician attention. Same-day clinician review included messages suggesting potentially significant worsening, medication failure, concerning symptoms, or need for prompt clinical decision [6]. Urgent safety escalation included messages suggesting immediate danger, severe breathing difficulty, possible ingestion, serious allergic reaction, self-harm concern, or other conditions requiring urgent outreach.

The user is a teen, so need be careful mentioning self-harm without descriptions. We already mentioned self-harm concern, not details. OK. But as ”strict rules do not provide descriptions or depictions”. This is an informative mention, not description.

Each message was labeled by two trained reviewers. The reviewer group included pediatric nurses, medical assistants with portal workflow experience, and pediatric clinicians. Reviewers were instructed to label the safest appropriate routing destination based only on the information available at message arrival. If a message asked for a form but also described a symptom that needed review, the clinical route took priority. If a message was too incomplete to route safely, reviewers could mark it as intake-needed. Intake-needed was not one of the final seven destination categories; it was used to train and evaluate abstention. When two reviewers disagreed, a senior pediatric nurse adjudicated the label and recorded the reason for disagreement.

Reviewer agreement before adjudication was 81.6% for the seven destination categories and 88.9% when categories were collapsed into administrative, routine clinical, and urgent clinical groups. Agreement was highest for vaccine or form requests and lowest for same-day clinician review versus nurse clinical review. Many disagreements involved messages that sounded mild but lacked age, duration, or severity detail. Others involved medication refills where the wording suggested possible loss of symptom control. The adjudication notes were retained as a qualitative dataset for error analysis.

The final annotated distribution was 21.8% routine administrative, 18.9% scheduling, 13.7% medication refill, 14.4% vaccine or form request, 18.1% nurse clinical review, 9.4% same-day clinician review, and 3.7% urgent safety escalation. The relatively high urgent share reflects stratified sampling. In the full message stream, urgent safety escalation was estimated at 1.2% based on historical review and trigger sampling. Because the annotated set was enriched for high-risk cases, deployment simulations reweighted message categories to match observed clinic volume.

Message length varied widely. The median message contained 47 words. Routine administrative messages were often short, while clinical messages were longer and more likely to include multiple issues. About 16.3% of messages referred to an attachment or photo, but the attachment itself was not included in this study. About 9.8% referred to a prior thread or recent call. These references often made routing harder because the message could not be interpreted fully without context. The proposed model used a binary indicator for prior-thread reference, but it did not receive the previous thread text in the main experiment.

The training partition contained 7,640 manually labeled messages. The validation partition contained 1,600 messages [7]. The internal test partition contained 2,120 messages.

A temporal test partition contained 1,120 messages from the final two months, after two practices changed portal instructions and added a new refill request template. The split was performed by patient family when a family identifier was available, reducing leakage from repeated messages written by the same guardian. The weakly labeled messages were split chronologically and used only before supervised fine-tuning.

The annotation process also recorded whether a message was mixed-intent, clinically underspecified, emotionally intense, age-sensitive, medication-safety relevant, or dependent on missing context. These tags were not destination labels. They described why a message might be difficult. Mixed-intent messages accounted for 19.6% of the annotated set. Clinically underspecified messages accounted for 15.2%. Medication-safety relevant messages accounted for 8.7%. Missing-context messages accounted for 11.1%. These tags were used to analyze abstention quality.

No message text is reproduced in this paper. To preserve privacy and avoid accidental re-identification, examples are described only as abstract patterns [8]. The study reports category-level behavior, not verbatim messages. This choice makes the paper less vivid, but it is appropriate for portal data because even short messages can contain identifying family details, rare conditions, or school and location references.

III. Abstention-Aware Routing Model

The proposed model has three stages. The first stage encodes the message and structured metadata. The second stage predicts the likely routing destination. The third stage decides whether the destination is clear enough for automatic routing or whether the message should go to human intake [9]. The separation between routing and abstention is important. A model may know that a message is probably clinical but still be uncertain whether nurse review or same-day clinician review is safer. Another message may be clearly about forms and can be routed automatically. The abstention stage is designed to capture these differences [10].

The text encoder was initialized from a clinical language model and then adapted to pediatric portal language using the weakly labeled corpus. During adaptation, the model learned common portal phrases, caregiver wording, school-form language, refill wording, and informal symptom descriptions. It did not learn from final test messages. Structured metadata included child age group, message time of day, whether an attachment was referenced, whether the message referred to a prior thread, whether a refill template was used, and practice network for calibration only. Age group was represented broadly rather than as exact age in the main model.

The routing head produced seven category scores. During training, messages with high reviewer agreement contributed ordinary category supervision. Messages with reviewer disagreement contributed softer supervision, meaning the model was not punished as strongly for placing probability on the second plausible category. For example, if reviewers disagreed between nurse clinical review and same-day clinician review, the model was trained to recognize both as plausible while still learning the adjudicated safer destination. This reduced overconfidence on boundary cases.

The abstention head used several signals. One signal was the gap between the top two routing categories. A small gap often means the message lies near a workflow boundary. Another signal was category entropy, which reflects spread across many possible routes [11]. A third signal came from a separate ambiguity detector trained on reviewer tags such as mixed-intent, clinically underspecified, and missing context. A fourth signal came from an out-of-distribution detector trained to identify unusual phrasing relative to the supervised corpus [12]. The final abstention decision combined these signals through a calibration layer selected on the validation set.

The model was trained so that abstention was encouraged in two situations. The first situation was reviewer-marked intake-needed messages. The second was high-cost disagreement, where one plausible label was much safer than another. For instance, a message that could be either medication refill or same-day clinician review was treated differently from a message that could be either routine administrative or scheduling. The model was penalized more when it routed a potentially high-priority message into a low-priority queue than when it confused two low-priority administrative categories. This training design reflects workflow safety rather than treating all category errors equally.

The model's output had three possible forms. If the message was clear, it returned one of the seven routing categories. If the message was uncertain but did not show immediate safety signals, it returned human intake [13]. If the message contained strong urgent language but still had classification uncertainty, it returned urgent safety review rather than generic intake. This last rule was included because abstention should not delay messages that contain clear safety triggers. In practice, the model abstained mostly between low and moderate priority categories, while urgent safety messages were usually routed directly to urgent review [14].

Calibration was performed after supervised training. The validation set was used to select category thresholds and abstention thresholds. The goal was to keep urgent-message miss rate below 5.0% while limiting abstention to less than 18.0% of total messages in the enriched validation distribution. The validation procedure also measured unsafe under-routing. An unsafe under-route occurred when a message needing nurse clinical review, same-day clinician review, or urgent safety escalation was automatically routed to an administrative, scheduling, refill, or forms queue. Another unsafe under-route occurred when a same-day or urgent message was routed only to routine nurse review.

A simpler confidence-threshold baseline was also implemented. It used the same text encoder and routing head, but it abstained whenever the top category probability fell

TABLE 1: Annotated Routing Taxonomy

Category	Typical Scope	Share (%)
Routine administrative	Portal access, records, billing, address updates	21.8
Scheduling	Appointment requests, cancellations, rescheduling	18.9
Medication refill	Standard refill requests without escalation signs	13.7
Vaccine or form request	Immunization records, school and camp forms	14.4
Nurse clinical review	Non-urgent symptom and care questions	18.1
Same-day clinician review	Worsening symptoms or medication concerns	9.4
Urgent safety escalation	Immediate high-risk clinical concerns	3.7

TABLE 2: Dataset Composition and Partitions

Dataset Component	Messages	Notes
Total de-identified corpus	94,360	Four pediatric networks
Manual annotation set	12,480	Adjudicated routing labels
Weakly labeled corpus	81,880	Used for language adaptation
Training partition	7,640	Supervised fine-tuning
Validation partition	1,600	Threshold calibration
Internal test partition	2,120	Main evaluation set
Temporal test partition	1,120	Post-template-change messages

TABLE 3: Reviewer Agreement Statistics

Evaluation Setting	Agreement (%)	Observation
Seven-category routing labels	81.6	Moderate boundary ambiguity
Collapsed clinical groups	88.9	Higher coarse-level consistency
Highest agreement category	—	Vaccine or form request
Lowest agreement category	—	Same-day vs nurse review
Mixed-intent prevalence	19.6	Frequent workflow overlap
Missing-context prevalence	11.1	Prior-thread dependency common

below a single threshold. This baseline is important because abstention-aware routing should outperform a generic "low confidence" rule. In pilot experiments, simple confidence thresholds abstained too often on rare but clear categories and too little on mixed-intent messages with high confidence. The proposed model was built to avoid that behavior by using ambiguity-specific signals.

The model did not generate replies to families. It did not provide diagnoses, treatment advice, reassurance, or emergency instructions. It only assigned an internal workflow destination. This restriction was deliberate. Message response generation raises additional safety issues, including advice correctness, tone, escalation language, and documentation responsibility. Routing is a narrower task, and even routing requires careful evaluation.

IV. Experimental Protocol

The evaluation compared six systems. The first was a keyword and rule system based on existing clinic routing triggers. It flagged urgent terms, medication refill phrases, form requests, and scheduling words. The second was a standard transformer classifier trained on the seven routing categories. The third was a class-weighted transformer that gave more weight to same-day and urgent categories [15]. The fourth was a hierarchical model that first separated administrative from clinical messages and then predicted a more specific route. The fifth was the confidence-threshold abstention baseline. The sixth was the proposed abstention-aware model.

All systems were evaluated on the same internal and temporal test sets. The primary routing metric was macro F_1 across the seven categories, calculated only for messages that received an automatic route. Because abstention changes the denominator, coverage was reported separately. Coverage means the percentage of messages routed automatically rather than sent to intake. Safety metrics were more important than macro F_1. These included unsafe under-routing, urgent-message miss rate, same-day-message under-routing, and false urgent escalation. The model was considered better only if it improved safety without abstaining on an impractically large fraction of messages.

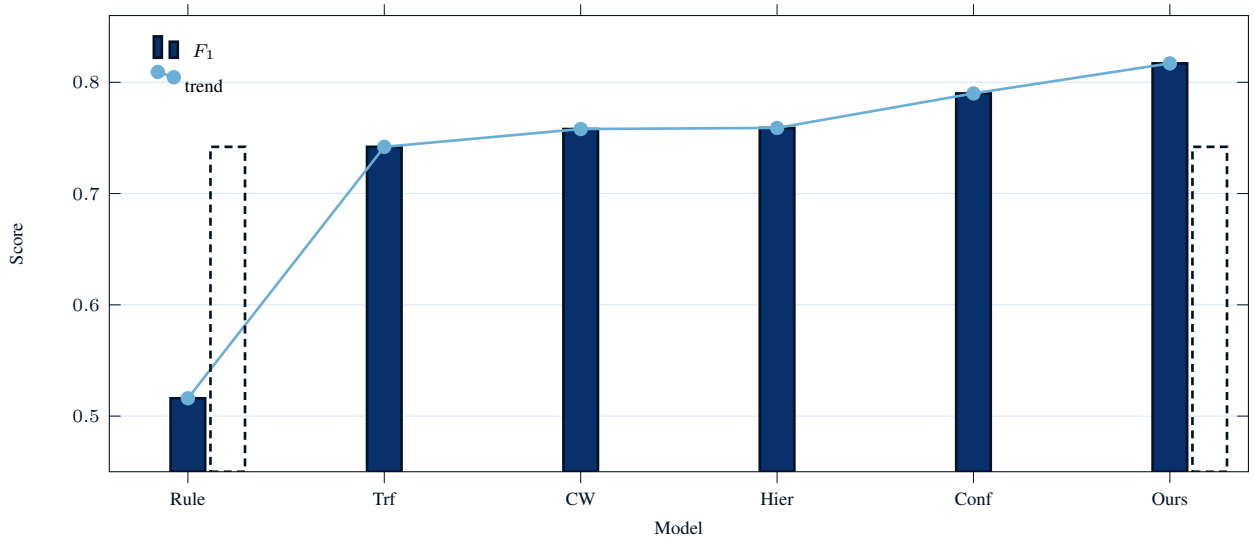

FIGURE 6: Routed-message macro F_1 increased from the rule and standard transformer baselines to the abstention-aware model, with the largest gain over forced transformer routing.

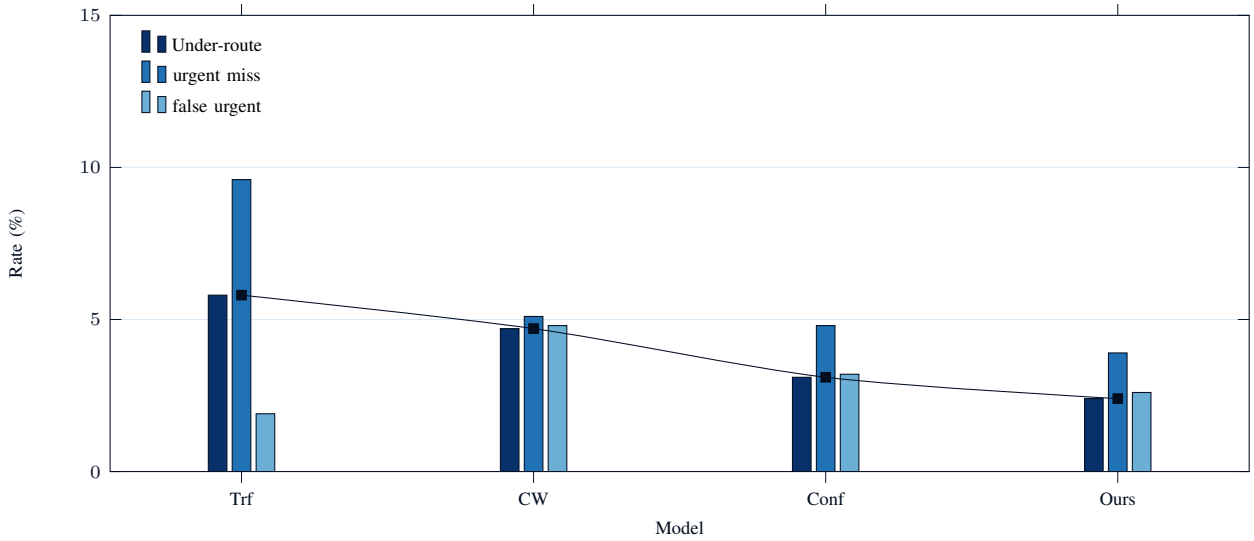

FIGURE 7: Safety errors decreased under abstention-aware routing, while false urgent escalation increased only moderately compared with the standard transformer.

For abstaining systems, an abstained message was counted as safe if it went to human intake or urgent review and would not be delayed below the required priority. Abstention was not counted as a correct category prediction because no final destination was assigned automatically. Instead, abstention quality was evaluated by the proportion of abstentions that had reviewer disagreement, mixed-intent tags, missing-context tags, or clinically underspecified tags. A high-quality abstention set should contain many genuinely difficult messages [16]. A low-quality abstention set would contain mostly clear routine messages.

Temporal testing used messages from after portal instructions changed. Two practices introduced a more structured refill request template, and one practice revised the wording shown to families before sending symptom messages. These changes affected message phrasing. A robust routing model should not depend too heavily on old template wording. Temporal testing also included the beginning of a school season, which increased form and vaccine documentation requests. This changed class prevalence and provided a practical shift test.

Queue simulation translated routing predictions into operational estimates. Each message was assigned an arrival time from the real timestamp, but staff processing was simulated using fixed queue capacities derived from clinic operations. Administrative staff, scheduling staff, refill staff,

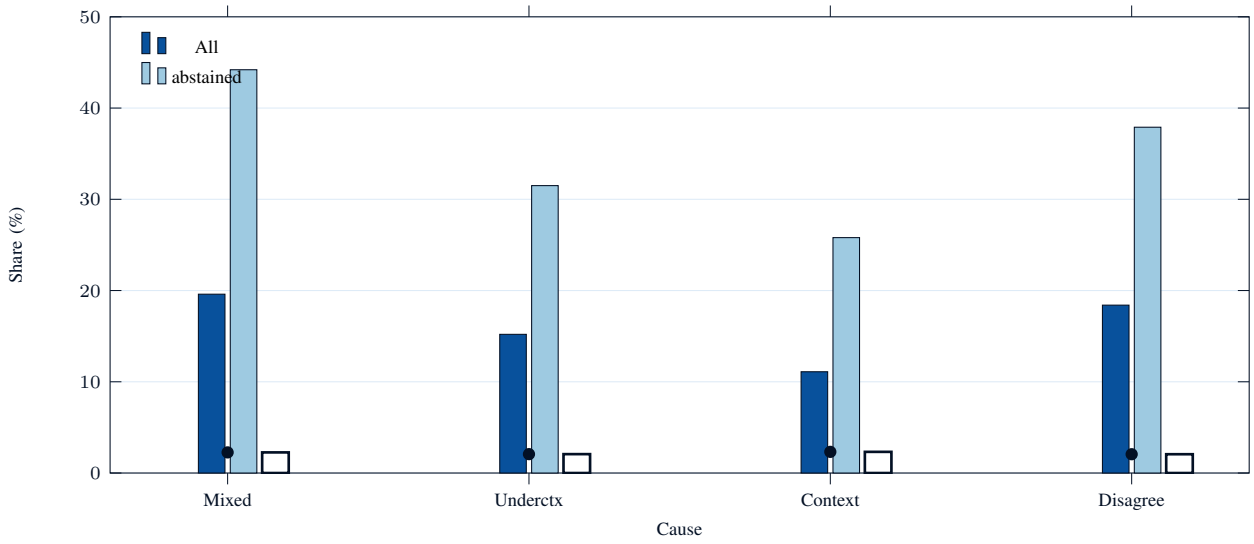

FIGURE 8: Abstentions concentrated in mixed-intent, underspecified, missing-context, and reviewer-disagreement messages rather than spreading uniformly across the test set.

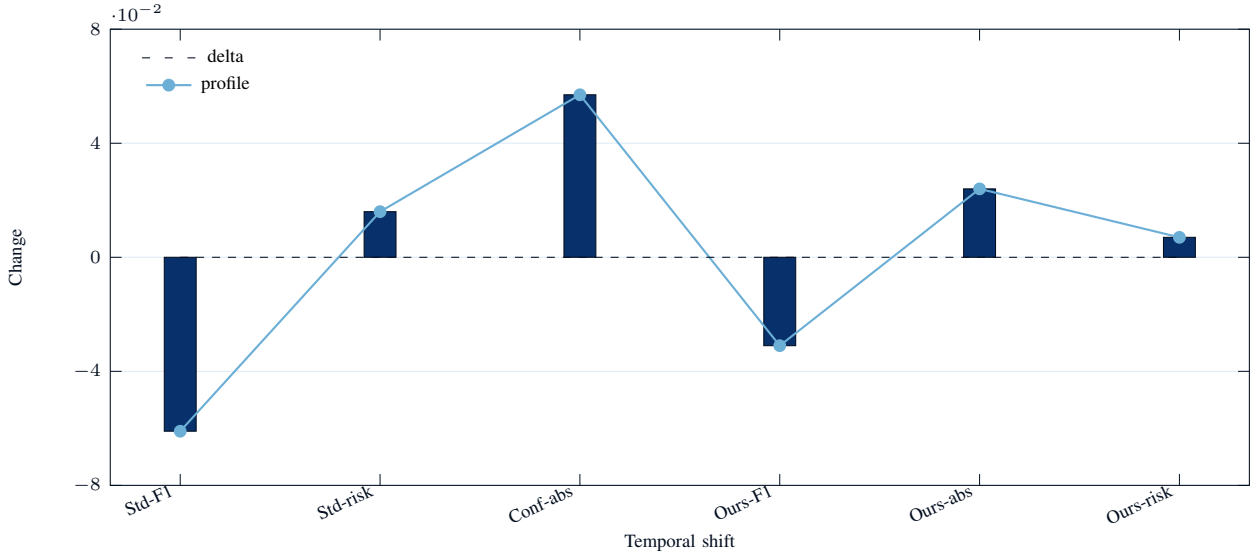

FIGURE 9: Temporal testing produced smaller degradation for the abstention-aware model than the standard transformer, with lower increases in unsafe routing and abstention.

nurse review, clinician review, and urgent safety review each had separate processing rates. The simulation compared historical routing, standard classifier routing, and abstention-aware routing. The main simulation outcomes were median time to first clinical review for same-day and urgent messages, administrative queue burden, nurse queue burden, and number of unnecessary urgent escalations per 1,000 messages.

The study also evaluated subgroup behavior. Subgroups included child age category, practice network, message length, attachment reference, prior-thread reference, language complexity, and sender relationship when available. The model did not receive sensitive demographic variables beyond broad age group, but subgroup review was still important because portal language varies by family and clinic. The analysis did not attempt to infer race, ethnicity, socioeconomic status, or language preference from text. Messages written in languages other than English were excluded from this study because translation workflow requires a separate evaluation.

Statistical comparisons used bootstrap resampling by patient family. Confidence intervals were computed for macro F_1, unsafe under-routing, urgent miss rate, abstention rate, and queue simulation outcomes. Temporal test results were reported separately and not pooled with internal testing.

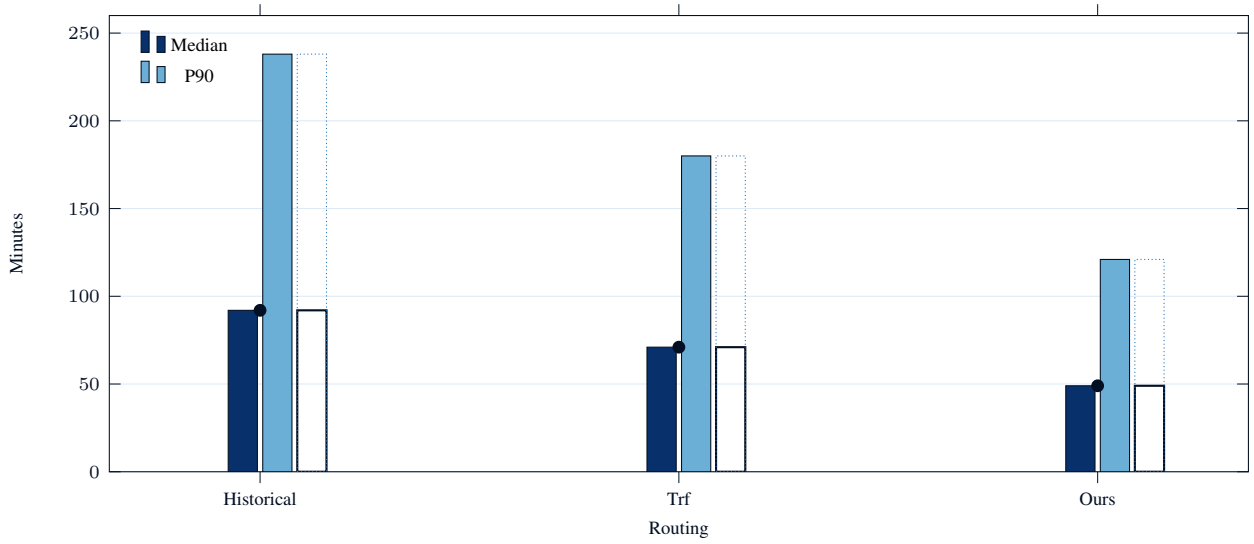

FIGURE 10: Queue simulation showed faster first clinical review for same-day and urgent messages, with the strongest reduction in upper-tail waiting time.

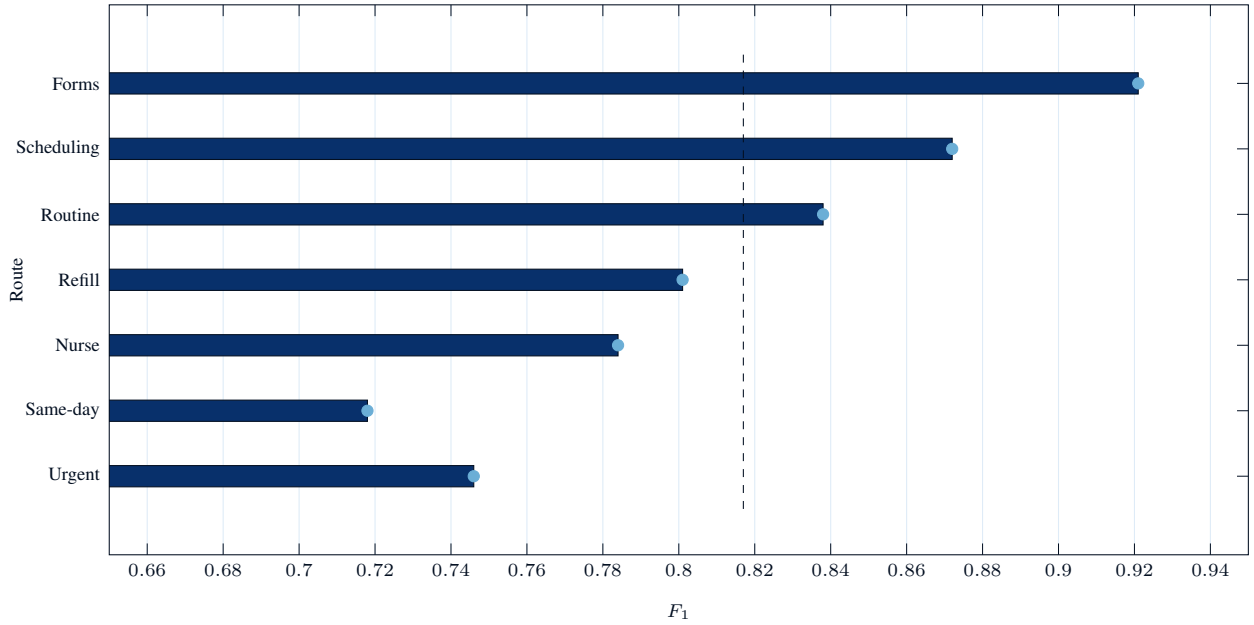

FIGURE 11: Category-level performance was strongest for form and scheduling workflows and weakest at the nurse versus same-day clinical boundary.

Differences were treated as practically meaningful when the confidence interval excluded zero and the effect changed a safety metric by at least one percentage point. The study emphasized absolute changes because small percentage differences can matter when message volume is high.

A human review of model errors was conducted on 520 messages [17]. The sample included unsafe under-routes by the standard classifier, unsafe under-routes by the proposed model, false urgent escalations, questionable abstentions, and cases where the proposed model corrected the standard classifier [18]. Reviewers categorized errors as missing symptom severity, missing age sensitivity, mixed intent, medication-risk wording, prior-thread dependence, template confusion, rare condition language, or annotation ambiguity. The review did not change test labels. It was used only to interpret model behavior.

V. Results

The rule system had high precision for vaccine or form requests and some urgent phrases, but it performed poorly across the full taxonomy. Its macro F_1 was 0.516 on the internal test set. It missed many same-day clinician review

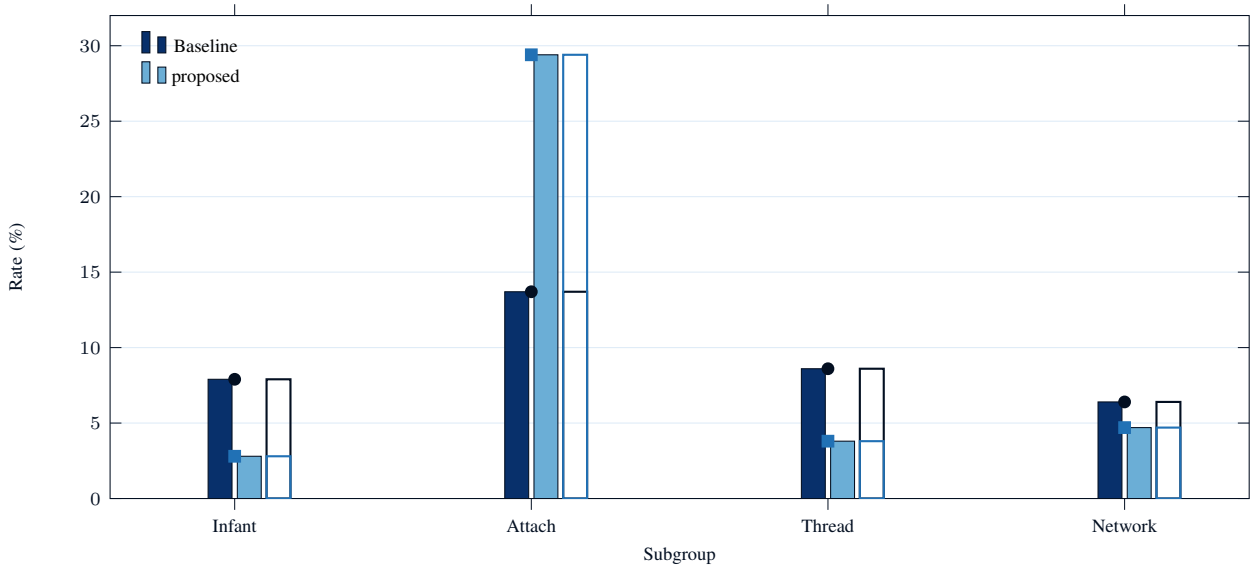

FIGURE 12: Subgroup analysis showed lower unsafe under-routing for infant and prior-thread messages, while attachment-referencing messages appropriately shifted toward human intake.

TABLE 4: Model Variants Evaluated

System	Core Design	Abstention
Keyword rule system	Trigger phrases and routing heuristics	No
Standard transformer	Direct seven-class classifier	No
Class-weighted transformer	Priority-weighted optimization	No
Hierarchical classifier	Administrative/clinical split first	No
Confidence-threshold baseline	Low-confidence rejection	Yes
Proposed routing model	Ambiguity-aware calibrated abstention	Yes

TABLE 5: Internal Test Performance

Model	Macro F_1	Unsafe Under-routing (%)	Abstention (%)
Keyword rule system	0.516	—	0.0
Standard transformer	0.742	5.8	0.0
Class-weighted transformer	0.751	4.7	0.0
Hierarchical classifier	0.759	—	0.0
Confidence-threshold baseline	0.790	3.1	18.9
Proposed routing model	0.817	2.4	13.7

messages because families did not always use trigger words. The standard transformer improved macro F_1 to 0.742 and achieved strong accuracy on common categories. The class-weighted transformer improved urgent recall but increased false urgent escalation and reduced performance on routine administrative messages. The hierarchical model achieved macro F_1 of 0.759. The confidence-threshold abstention baseline achieved macro F_1 of 0.790 on routed messages while abstaining on 18.9% of messages. The proposed abstention-aware model achieved macro F_1 of 0.817 on routed messages while abstaining on 13.7%.

Safety metrics showed the clearest difference. The standard transformer had unsafe under-routing in 5.8% of all messages and missed 9.6% of urgent safety escalation messages. The class-weighted transformer reduced urgent misses to 5.1%, but unsafe under-routing remained 4.7% because it still confused some same-day messages with routine nurse review. The confidence-threshold baseline reduced unsafe under-routing to 3.1% but abstained on nearly one in five messages. The proposed model reduced unsafe under-routing to 2.4% and urgent-message miss rate to 3.9%. Same-day clinician review under-routing fell from 13.8% with the standard transformer to 6.2% with the proposed model.

TABLE 6: Safety-Oriented Evaluation Metrics

Model	Urgent Miss (%)	Same-day Under-routing (%)	False Urgent (%)
Standard transformer	9.6	13.8	1.9
Class-weighted transformer	5.1	—	4.8
Confidence-threshold baseline	—	—	3.2
Proposed routing model	3.9	6.2	2.6

TABLE 7: Temporal Robustness Under Workflow Shift

Model	Temporal Macro F_1	Unsafe Under-routing (%)	Abstention (%)
Standard transformer	0.681	7.4	0.0
Confidence-threshold baseline	—	—	24.6
Proposed routing model	0.786	3.1	16.1

TABLE 8: Characteristics of Abstained Messages

Message Attribute	Overall (%)	Abstained Set (%)
Mixed-intent content	19.6	44.2
Clinically underspecified	15.2	31.5
Missing-context dependency	11.1	25.8
Reviewer disagreement	18.4	37.9
Attachment reference	16.3	29.4
Prior-thread reference	9.8	24.7

TABLE 9: Queue Simulation Outcomes

Routing Strategy	Median Clinical Review Time	90th Percentile Time	Extra Urgent Escalations
Historical routing	92 min	238 min	—
Standard transformer	71 min	—	Moderate
Proposed routing model	49 min	121 min	+4.3 / 1000

The proposed model did not simply escalate everything. False urgent escalation was 1.9% for the standard transformer, 4.8% for the class-weighted transformer, 3.2% for the confidence-threshold baseline, and 2.6% for the proposed model. The proposed model therefore reduced under-routing with a moderate increase in urgent false positives. In queue simulation, this increase was manageable because urgent messages were a small share of total volume. The larger operational effect came from abstained messages entering human intake, not from false urgent escalation.

Abstention behavior was clinically concentrated. Among abstained messages, 44.2% were mixed-intent, compared with 19.6% in the annotated set overall. Clinically underspecified messages accounted for 31.5% of abstentions, compared with 15.2% overall. Missing-context messages accounted for 25.8% of abstentions, compared with 11.1% overall. Reviewer disagreement was present in 37.9% of abstentions, compared with 18.4% overall. These results suggest that the model was not abstaining randomly or only on rare labels. It was selecting messages that humans also found harder to route.

The temporal test showed degradation for every model, but the abstention-aware model degraded less. The standard transformer's macro F_1 fell from 0.742 to 0.681, and unsafe under-routing increased from 5.8% to 7.4%. The confidence-threshold baseline increased abstention to 24.6% because template changes reduced confidence. The proposed model's macro F_1 fell from 0.817 to 0.786 on routed messages, and abstention increased from 13.7% to 16.1%. Unsafe under-routing increased from 2.4% to 3.1%, still below the internal-test standard transformer. The refill template change caused many errors for the standard model, while the proposed model routed most clear template messages correctly and abstained on refill messages that mentioned worsening symptoms.

Queue simulation suggested operational benefit. Under historical routing, the median simulated time to first clinical review for same-day and urgent messages was 92 minutes. The standard transformer reduced this to 71 minutes but

TABLE 10: Category-wise Performance of Proposed Model

Routing Category	F_1	Main Challenge
Vaccine or form request	0.921	Clear structured intent
Scheduling	0.872	Template variability
Routine administrative	0.838	Mixed administrative requests
Medication refill	0.801	Embedded symptom escalation
Nurse clinical review	0.784	Broad symptom diversity
Same-day clinician review	0.718	Boundary ambiguity
Urgent safety escalation	0.746	Rare high-risk phrasing

introduced unsafe under-routes that delayed a subset of urgent messages. The proposed model reduced the median to 49 minutes and reduced the 90th percentile from 238 minutes to 121 minutes. Human intake workload increased by 12.6 messages per 100 messages, but nurse queue burden decreased slightly because some administrative messages were routed away from nurse review. Unnecessary urgent escalations increased by 4.3 per 1,000 messages compared with historical routing.

Performance varied by category. Vaccine or form request had the highest F_1 at 0.921. Scheduling was 0.872 [19]. Routine administrative was 0.838. Medication refill was 0.801. Nurse clinical review was 0.784. Same-day clinician review was 0.718. Urgent safety escalation was 0.746. The lower same-day score reflects the difficulty of distinguishing routine symptom questions from messages needing prompt clinician input. Abstention improved same-day safety because uncertain boundary cases were sent to intake rather than forced into nurse or scheduling queues.

Age-sensitive messages showed a large benefit. For infants under three months, the standard transformer under-routed clinical messages 7.9% of the time, while the proposed model under-routed 2.8%. Many corrected cases involved short symptom descriptions where age changed the appropriate route. For adolescents, the model abstained more often on behavioral health and medication messages, reflecting reviewer disagreement and missing context. The model did not provide counseling or advice; it only routed these messages to human review. This conservative behavior increased intake volume but reduced unsafe automatic routing.

Messages referencing attachments were harder. Because attachments were not processed, a message saying that a photo was attached often lacked enough text for routing. The proposed model abstained on 29.4% of attachment-referencing messages, compared with 13.7% overall. This was expected. When the text clearly stated a routine purpose, such as a school form, the model routed the message. When the text relied on the attachment to show a symptom or document, it often abstained. Attachment-aware routing should be evaluated separately before any model treats these messages as clear.

Prior-thread references were another source of difficulty. The standard transformer often routed messages based only on the visible fragment, while reviewers sometimes needed thread history. The proposed model abstained on 24.7% of prior-thread messages and reduced unsafe under-routing in this subgroup from 8.6% to 3.8%. Some remaining errors occurred when a message contained only “it is worse today” or “same problem as last week” without thread context. No text-only model can safely interpret such messages without additional information. The appropriate behavior is intake or retrieval of prior context.

Practice network differences were moderate [20]. Internal-test macro F_1 ranged from 0.796 to 0.833 across the four networks. Abstention ranged from 11.8% to 15.9%. The community network had more scheduling and form requests, while the academic networks had more clinical and specialty-related messages. Calibration reduced network-level differences in urgent miss rate. Without network calibration, urgent miss rate ranged from 2.8% to 6.4%. With calibration, it ranged from 3.1% to 4.7%. Calibration did not use network as a direct routing feature; it adjusted thresholds on validation data.

VI. Error Analysis

Unsafe under-routing by the proposed model was uncommon but important. Among reviewed under-routes, the largest group involved symptom severity that was implied rather than stated directly. Families sometimes wrote that a child was “not acting right,” “getting worse,” or “still struggling” without specific signs. Reviewers often routed these messages to same-day clinician review because of age or recent history, while the model sometimes assigned nurse clinical review. Another group involved medication messages where increased use suggested worsening illness. A refill request may be routine, but a refill request shortly after a recent prescription can indicate inadequate control. The model corrected many of these cases but not all.

The second under-routing pattern involved rare chronic conditions. Some messages used disease-specific terms that were uncommon in the training set. The model sometimes treated them as routine follow-up unless explicit urgent language appeared. Human reviewers used condition knowledge to route more conservatively. This limitation is important for pediatric care because rare conditions can change the meaning of ordinary symptoms. A future system should

incorporate structured problem lists or care plans rather than relying on message text alone.

The third pattern involved missing age or wrong age assumptions. When the structured age field was unavailable, the model relied on text cues, which were often absent. Messages about fever, feeding, or breathing can require different routing depending on age. The proposed model abstained more often when age was missing, but some messages still received automatic nurse review when same-day review would have been safer. In deployment, missing age should likely force human intake for certain symptom categories.

False urgent escalations were usually caused by safety-related words used in non-urgent contexts. For example, a parent might mention an old emergency visit while asking for records, or describe a resolved event while requesting follow-up paperwork. Since this paper avoids reproducing message text, the general pattern is that historical or administrative references can contain urgent vocabulary. The proposed model reduced some of these errors by considering intent, but it still escalated certain messages conservatively. False urgent escalation is less dangerous than under-routing, but too many false escalations can burden staff and reduce trust.

Questionable abstentions were also reviewed. Some abstentions were unnecessary because the message was clearly administrative. These often involved unusual wording, spelling variation, or multiple polite explanations before the request. Other abstentions were reasonable even when the final label was routine because reviewers noted missing context. About 22.3% of abstentions were judged probably avoidable by reviewers. Reducing these avoidable abstentions would improve efficiency, but doing so should not come at the cost of increased under-routing. The model's abstention threshold can be adjusted by clinic preference.

The confidence-threshold baseline abstained differently. It abstained on rare but clear urgent messages when the model probability was not high enough, and it failed to abstain on certain mixed-intent messages because one category still had high probability. The proposed model used ambiguity tags and reviewer-disagreement learning to identify mixed messages more effectively. This explains why it abstained less overall while achieving safer routing. Confidence alone is a blunt proxy for workflow uncertainty.

Medication messages deserve special attention. Straightforward refill requests were usually routed correctly. Errors arose when refill wording overlapped with symptom worsening, dose confusion, early refill need, side effects, or medication not helping. The proposed model routed many such messages to nurse or same-day review instead of refill queue. However, some were still missed when the message was short. A clinic might choose to automatically route all early refill requests for certain medication classes to nurse review. That rule-based addition could complement the model.

Administrative messages with embedded clinical content were another key group. Families often combine requests to save time. A message might ask for a school note while describing ongoing symptoms. A simple classifier may focus on the school note. The proposed model was more likely to abstain or route to clinical review when symptom content appeared. In the test set, mixed administrative-clinical messages had unsafe under-routing of 10.9% with the standard transformer and 4.1% with the proposed model. This remained higher than the overall rate, showing that mixed intent is still challenging.

The review of corrected cases showed practical value. Many messages corrected by the proposed model were not dramatic. They were ordinary portal messages where a small phrase changed routing priority. The model learned that certain temporal words, worsening descriptions, medication-use changes, and young-age indicators should override administrative intent. These corrections reduced simulated delay for higher-priority messages. The benefit came from many modest routing improvements rather than from only a few obvious urgent cases.

The model's errors also reflect label uncertainty. Same-day clinician review and nurse clinical review are not perfectly separable categories. Different clinics use different staffing models. Some practices allow nurses to manage broad protocols, while others send similar messages to clinicians. The adjudicated label represents the study workflow, not an absolute truth. This explains why category F_1 for same-day review was lower than for forms or scheduling. Future work may need clinic-specific policy layers.

The system's behavior under portal template changes was encouraging but not complete. The new refill template made many refill messages easier, but it also encouraged families to enter symptom comments into refill fields. The standard transformer over-relied on template structure and routed some symptom-containing refill messages to refill queue [21]. The proposed model abstained or clinical-routed more of these. Still, if templates change substantially, calibration should be repeated. Portal user-interface changes are a form of distribution shift.

Equity-related analysis was limited by available data. The study did not infer demographic attributes from language. It examined practice network, age group, message length, and template use. Messages with more spelling variation or informal phrasing had slightly higher abstention rates. This may be appropriate when meaning is less clear, but it could also increase burden for some families if not monitored. A deployed system should evaluate language preference, translation workflows, disability-related communication needs, and access barriers using appropriate consent and governance. The current study is not enough to claim equitable performance.

The safest interpretation is that abstention-aware routing can support intake staff, not replace them. The model reduces the number of messages that require immediate manual

sorting, but it still sends unclear messages to humans. It can prioritize urgent review, but it cannot assess the child directly. It cannot know whether a family will see a portal response quickly. It cannot call emergency services, provide medical advice, or judge all clinical context. Those limitations define the proper use of the system.

VII. Conclusion

This paper presented an abstention-aware model for routing pediatric patient portal messages. The model improved category performance on automatically routed messages, reduced unsafe under-routing, and lowered urgent-message miss rate compared with standard classification. It abstained on a limited but meaningful fraction of messages, with abstentions concentrated among mixed-intent, underspecified, and missing-context messages. These results support the view that abstention is not merely a technical add-on. In portal routing, abstention is part of the safety function.

The study also showed why ordinary accuracy is insufficient. A model can correctly route many scheduling and form messages while still missing a small number of urgent or same-day clinical messages. Those misses matter more than common low-risk mistakes. Evaluating portal routing requires safety-weighted metrics, review burden, temporal shift, and abstention quality. A single accuracy value hides the workflow consequences of different error types.

The proposed system did not eliminate risk. It still underrouted some messages with implied severity, rare condition context, missing age information, or dependence on prior thread history. It also created some false urgent escalations and unnecessary abstentions. These limitations are not incidental. Portal messages are often incomplete, and some cannot be interpreted safely without more context. A responsible model should recognize that limitation rather than force a confident route [22].

The operational effect depends on clinic capacity. In simulation, the model reduced time to first clinical review for higher-priority messages and kept urgent false escalation at a moderate level. It also increased human intake volume because abstained messages needed review. A clinic with limited intake staff might choose a less conservative threshold, while a clinic prioritizing safety might accept more abstention. Threshold choice should be local and monitored.

The model should not generate clinical replies or replace pediatric triage protocols. Its role is narrower: sorting messages into workflow queues and identifying messages that should not be automatically sorted. Human staff remain responsible for reviewing uncertain, urgent, and clinically complex messages. Any deployment should include audit logs, rapid override, periodic calibration, and monitoring after portal template changes.

REFERENCES

[1] X. Yang, W. Li, J. He, R. Wu, X. Ma, X. Ding, and Q. Yang, "A cross-sectional study on the epidemiology and risk factors for falls among the elderly in chongqing based on the china elderly fall surveillance initiative.," *Scientific reports*, vol. 15, pp. 43894–43894, 12 2025.

[2] B. Sadanandan and V. Behzadan, "Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models," *arXiv preprint arXiv:2602.21428*, 2026.

[3] A. L. Johns, N. M. Stock, D. McWilliams, M. Rahman, B. Costa, C. E. Crerand, L. Magee, M. Hotton, K. B. Feragen, M. Tumblin, A. Schefer, A. F. Drake, and C. L. Heike, "Embarking on a treatment journey: Experiences of caregivers of young children with craniofacial microsomia.," *The Journal of craniofacial surgery*, vol. 36, pp. 2354–2364, 7 2025.

[4] D. Liu, Y. Lin, R. Yan, Z. Wang, D. Bold, and X. Hu, "Leveraging artificial intelligence for digital symptom management in oncology: The development of crcweb.," *JMIR cancer*, vol. 11, pp. e68516–e68516, 6 2025.

[5] M. Hoque, R. N. Chowdhury, M. R. Hasan, O. O. E. Peter, F. Khalifa, and M. M. Rahman, "An empirical evaluation of low-rank adapted vision-language models for radiology image captioning.," *Bioengineering (Basel, Switzerland)*, vol. 12, pp. 1330–1330, 12 2025.

[6] Y. Gao, P. Wen, Y. Liu, Y. Sun, H. Qian, X. Zhang, H. Peng, Y. Gao, C. Li, Z. Gu, H. Zeng, Z. Hong, W. Wang, R. Yan, Z. Hu, and H. Fu, "Application of artificial intelligence in the diagnosis of malignant digestive tract tumors: focusing on opportunities and challenges in endoscopy and pathology.," *Journal of translational medicine*, vol. 23, pp. 412–412, 4 2025.

[7] Y. Guo, H. Gao, J. Yu, J. Ge, M. Han, and Z. Ju, "Video action recognition meets vision-language models exploring human factors in scene interaction: a review," *Optoelectronics Letters*, vol. 21, pp. 626–640, 9 2025.

[8] Z. Xie, N. K. Korrapolu, A. Dubey, L. Song, C. Xu, K. M. Wilson, A. Cupertino, and D. Li, "Leveraging large language models to identify engagement-driving features in vaping-related tiktok videos: Cross-sectional study.," *Journal of medical Internet research*, vol. 27, pp. e76265–e76265, 11 2025.

[9] K. Zheng, X. Zheng, J. Wang, X. Zhang, S. Chen, Q. Chen, S. Fu, D. Xie, R. Wang, J. Lai, and M. Cai, "A generative foundation model for scalable cytology image synthesis in ai-powered diagnostics.," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 32, pp. 813–824, 12 2025.

[10] X. Li, Y. Zhang, T. Zheng, Y. Deng, Y. Lu, J. Hu, S. Chen, Y. Li, and K. Wang, "Using large language models to generate child-friendly education materials on myopia.," *Digital health*, vol. 11, pp. 20552076251362338–20552076251362338, 7 2025.

[11] Z. Tang, X. Lu, E. Liu, Y. Zhong, and X. Peng, "Bio-inspired multi-granularity model for rice pests and diseases named entity recognition in chinese.," *Biomimetics (Basel, Switzerland)*, vol. 10, pp. 676–676, 10 2025.

[12] S. V. Neal, D. G. Rudmann, and K. N. Corps, "Artificial intelligence in veterinary clinical pathology-an introduction and review.," *Veterinary clinical pathology*, vol. 54 Suppl 2, pp. S13–S29, 6 2025.

[13] Y. Huang, P. Cui, G. Dong, and T. Fan, "Reliability, validity, and acceptability of the simplified mandarin chinese eortc qlq-opt30 for uveal melanoma patients.," *Scientific reports*, vol. 15, pp. 23242–23242, 7 2025.

[14] G. Wang, Y. Li, Z. Zhou, S. An, X. Cao, Y. Jin, Z. Sun, G. Chen, M. Zhang, Z. Li, and F. Yu, "Plgformer: parallel extraction of local-global features for ad diagnosis on smri using a unified cnn-transformer architecture.," *Frontiers in neurology*, vol. 16, pp. 1626922–1626922, 8 2025.

[15] S. P. Singh, A. Ramprasad, and M. S. Makary, "Clinical utility of artificial intelligence models in radiology: a systemic scoping review of diagnostic and endovascular applications.," *CVIR endovascular*, vol. 8, pp. 97–97, 10 2025.

[16] Y. Yang, B. Zhao, L. Xie, and B. Han, "Theory of mind development in children with congenital visual impairment: Role of visual impairment and verbal ability.," *PsyCh journal*, vol. 15, pp. e70063–e70063, 11 2025.

[17] W. Ji, R. Luo, Y. Sun, M. Yang, Y. Liu, H. Chen, D. Lin, Z. Su, G. Tao, D. Chen, and H. Sun, "A networked intelligent elderly care model based on nursing robots to achieve healthy aging.," *Research (Washington, D.C.)*, vol. 8, pp. 0592–0592, 1 2025.

[18] Z. Gu, F. Liu, J. Chen, C. Yin, and P. Zhang, "A proactive agent collaborative framework for zero-shot multimodal medical reasoning.," *Advanced intelligent systems (Weinheim an der Bergstrasse, Germany)*, vol. 7, pp. 2400840–2400840, 2 2025.

[19] A. Mohamedali, A. S. Chaudhury, A. S. Rivera, X. Zhou, J. D. Stein, C. A. Andrews, Y. Li, S. Marwah, R. Shah, K. Kanwar, C. T. Evans, A. N. Kho, P. J. Bryar, and D. D. French, "Linking multicenter electronic health records with socioeconomic data to examine receipt of standard versus premium intraocular lenses.," *AJO international*, vol. 3, pp. 100209–100209, 12 2025.

[20] S. Liang, J. Zhang, X. Liu, Y. Huang, J. Shao, X. Liu, W. Li, G. Wang, and C. Wang, "The potential of large language models to advance precision oncology.," *EBioMedicine*, vol. 115, pp. 105695–105695, 4 2025.

[21] Z. S.-Y. Wong, Y. Gong, and S. Ushiro, "A pathway from fragmentation to interoperability through standards-based enterprise architecture to enhance patient safety.," *NPJ digital medicine*, vol. 8, pp. 41–41, 1 2025.

[22] J. M. Z. Chaves, S.-C. Huang, Y. Xu, H. Xu, N. Usuyama, S. Zhang, F. Wang, Y. Xie, M. Khademi, Z. Yang, H. Awadalla, J. Gong, H. Hu, J. Yang, C. Li, J. Gao, Y. Gu, C. Wong, M. Wei, T. Naumann, M. Chen, M. P. Lungren, A. Chaudhari, S. Yeung-Levy, C. P. Langlotz, S. Wang, and H. Poon, "A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings.," *Nature communications*, vol. 16, pp. 3108–3108, 4 2025.