

ORIGINAL RESEARCH

Open access

Large Language Model Agent for Drafting Structured Responses to Non-Urgent Patient Portal Messages Using Local Clinical Protocols, Physician-Approved Templates, Patient History, and Message Intent Classification

Victor Hugo^{1*}, Daniel Cruz¹, Javier Salazar²

Abstract

Non-urgent patient portal messages consume a substantial fraction of clinician time and can shift attention away from higher-acuity care. As portal communication becomes a routine part of ambulatory medicine, inbox management increasingly functions as an additional clinical workload. Current inbox workflows often depend on manual review, interpretation, and free-text reply by clinicians. This creates delays, variability, and cognitive burden, especially for routine requests that could be answered through standardized guidance. This article proposes a large language model agent that classifies incoming message intent, identifies non-urgent queries suitable for templated responses, and drafts a structured reply. The draft is grounded in local clinical protocols, physician-approved templates, and patient-specific information from the electronic health record. The proposed agent includes a message intent classifier, a protocol-and-template retrieval module, a retrieval-augmented drafting model, a rule-based safety filter, and a human-verification interface. These components work together to keep generated text within a clinically approved and auditable workflow. By preparing protocol-adherent draft responses for clinician review, the agent could reduce routine documentation effort while preserving clinician judgment. Its value depends on safety boundaries, template quality, patient-data integration, and seamless fit within the existing portal workflow. The agent represents a pragmatic near-term use of large language models in clinical administration. It supports automation without removing human oversight from patient-facing communication.

Keywords Large language models, Retrieval-augmented generation, Human-in-the-loop, Patient portal messages, Clinical communication, Intent classification

*Correspondence:

Victor Hugo

victor.hugo@gmail.com

¹ Department of Clinical Informatics and Digital Health, Faculty of Medicine, National University of Colombia, Bogota, Colombia

² Department of Intelligent Healthcare Engineering, Faculty of Engineering, University of Antioquia, Medellin, Colombia

Introduction

Patient portal messaging has become a central channel for outpatient communication, but its growth has also intensified the burden of asynchronous clinical work. Studies of artificial intelligence-generated draft replies and in-basket communication describe a workflow in which clinicians must interpret patient concerns, retrieve context, compose an appropriate response, and maintain a professional tone under time pressure [1, 2]. The burden is asymmetric because portal messages often arrive outside scheduled encounters, yet they still require clinical judgment and documentation. A

drafting agent for non-urgent messages should therefore be designed not as an autonomous clinician, but as a structured support tool embedded in the communication workflow [3, 4].

A key design distinction is between urgent messages requiring immediate clinical attention and non-urgent messages that can be handled through protocol-driven guidance. Prior work on message prioritization, emergency detection, and patient-message classification shows that portal content can be categorized by intent and routed according to safety-relevant features [5, 6]. Requests such as prescription refills, appointment logistics, routine follow-up questions, and result inquiries often have predictable response structures, whereas new or worsening symptoms require escalation. This distinction supports selective automation in which only non-urgent, template-covered messages enter a drafting pathway [7, 8].

Large language models can produce coherent and empathic text, but clinical communication requires stronger constraints than fluency alone. Evaluations of patient-message replies emphasize factual correctness, tone, clinical appropriateness, and alignment with the intended care context [9, 10]. For a portal-message drafting system, the model should be grounded in local protocols and physician-approved templates rather than relying on open-ended generation. Retrieval-augmented drafting, structured templates, and explicit safety filters can help translate language-generation capacity into a clinically bounded workflow [11, 12].

This article conceptualizes an emerging artificial intelligence agent that drafts structured responses to non-urgent patient portal messages using local clinical protocols, physician-approved templates, patient history, and message intent classification. The agent classifies message intent, retrieves relevant protocol and template content, incorporates structured patient data, generates a draft, and presents that draft for clinician verification before any message is sent. This framing preserves the clinician as the accountable sender while reducing repetitive composition work. The goal is not to replace clinical communication, but to make routine communication more consistent, reviewable, and efficient.

Background

Patient portal messaging burden

Patient portal messaging has expanded the reach of ambulatory care, but it has also created a persistent stream of work that may be difficult to schedule, delegate, or measure. Research on artificial intelligence-generated replies and clinical messaging describes inbox management as a source of cognitive load because clinicians must move between message interpretation, chart review, and response writing [2, 3]. The content of these messages spans administrative questions, medication requests, symptom follow-up, and interpretation of results, creating both routine and clinically sensitive work within the same channel. An agent-based approach must therefore treat portal messaging as a heterogeneous workflow rather than a single text-generation task [13, 14].

Large language models for clinical communication

Large language models are relevant to clinical communication because they can transform structured or semi-structured inputs into coherent, patient-facing prose. Prior studies have examined their use for drafting replies to patient messages, with attention to readability, empathy, and physician review requirements [1, 9]. However, the same generative flexibility that makes these systems useful also creates risks when the model is insufficiently grounded in clinical policy, patient context, or approved wording. For this reason, an LLM agent for portal replies should combine generation with retrieval, templates, and human oversight rather than operate as a free-form response engine [10, 15].

Intent classification and message triage

Intent classification provides the routing layer that determines whether a message is appropriate for drafting, escalation, or ordinary manual handling. Earlier work on patient portal message classification used rule-based, machine learning,

convolutional neural network, transformer-based, and pretrained language-model approaches to identify patient concerns and message categories [16-19]. These approaches are conceptually important because a drafting agent must first determine whether a message is about a refill, symptom change, laboratory result, appointment request, administrative issue, or mixed concern. The classifier should also distinguish non-urgent templatable intents from messages that require direct clinical assessment [6, 7].

Protocol- and template-guided generation

Protocol- and template-guided generation constrains a model's response to institutionally approved clinical language. In patient-message drafting, templates can define the required structure, disclaimers, escalation instructions, and patient-specific fields that should appear in a response [4, 11]. Local protocols can further define which requests are eligible for routine handling and which must be redirected to a clinician, nurse triage process, or emergency pathway. This design reduces reliance on unconstrained model judgment and makes the generated draft more auditable for the reviewing clinician [12, 20].

Human-in-the-loop and safety for patient communication

Human-in-the-loop design is essential for patient-facing clinical communication because the generated text can influence care decisions, medication use, and patient understanding. Studies and commentaries on AI-drafted patient messages emphasize clinician review, transparency, safety monitoring, and patient or clinician acceptance as central deployment concerns [21-23]. In the proposed agent, the clinician remains responsible for approving, editing, or rejecting every draft before it reaches the patient. This approach supports efficiency while preserving professional accountability and a clear escalation route for uncertain or unsafe cases [14, 24].

Prior work on automated message drafting

Prior work has explored AI-generated draft replies, smart-reply style workflows, and integrated electronic-health-record communication support. These studies demonstrate that draft responses can be embedded into existing message workflows, but they also show the need for careful design around source context, clinician trust, and editing burden [25, 26]. A comprehensive agent differs from a generic drafting assistant because it explicitly combines intent classification, protocol retrieval, template selection, patient-history integration, safety checks, and mandatory human verification. This broader architecture addresses gaps that remain when drafting is treated only as a text-completion task [8, 13].

System Overview

High-level agent architecture

The proposed agent begins when a patient portal message enters the clinical inbox and is passed through an intent classifier. If the message is non-urgent and matches an approved template category, the system retrieves the relevant local protocol, identifies patient-specific context, and generates a draft for clinician review [5, 6]. If the message is urgent, ambiguous, or outside template scope, the agent bypasses drafting and routes the message to standard clinician review. This architecture positions the agent as a selective workflow layer rather than a universal responder [7, 8].

Figure 1 illustrates the proposed human-verified LLM agent workflow, showing how portal-message intent classification, protocol-and-template retrieval, patient-history grounding, safety filtering, and clinician approval interact before any patient-facing response is sent.

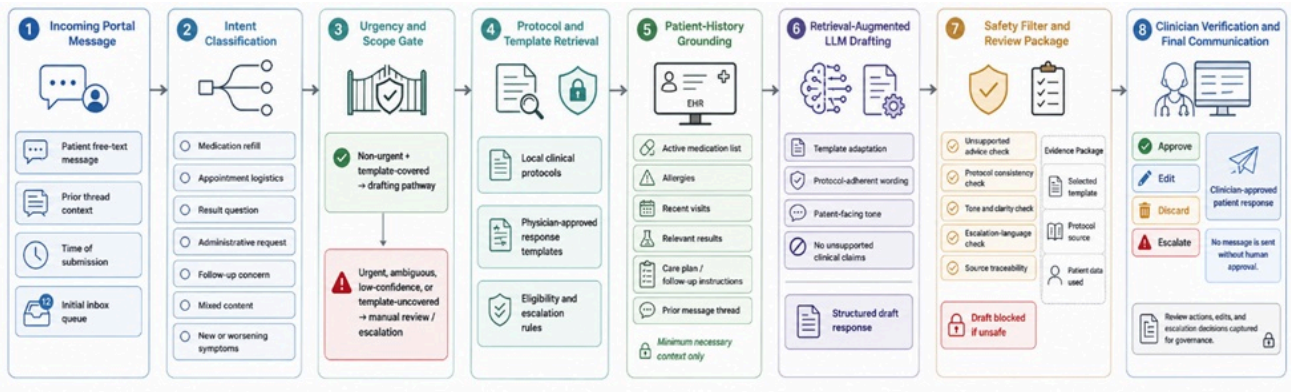


Figure 1. Human-Verified LLM Agent Workflow for Drafting Structured Responses to Non-Urgent Patient Portal Messages

Core inputs and outputs

The agent's inputs include the patient's message, prior thread context, structured patient data, local protocols, and the physician-approved template store. Relevant structured data may include active diagnoses, medications, allergies, recent visits, and recent results, depending on the classified intent [11, 20]. The output is not a sent message but a draft reply ready for clinician sign-off, accompanied by visible evidence of which template, protocol, and patient data informed the draft. This makes the output reviewable, editable, and traceable within the clinical workflow [4, 12].

Table 1 defines the functional architecture of the proposed LLM agent by linking each system component to its required inputs, operational task, safety function, and clinician-facing output.

Table 1. Functional Architecture of the LLM Agent for Protocol-Grounded Portal Message Drafting

Agent Component	Primary Inputs	Core Operational Function	Safety or Reliability Constraint	Clinician-Facing Output	Manuscript Contribution
Portal message intake layer	Patient free-text message; prior thread context; message timestamp; inbox metadata	Captures the patient's message and prepares it for downstream routing without altering the original content	Original patient wording must remain visible and preserved for clinician review	Original message displayed beside any generated draft	Frames the system as an inbox-support layer rather than an autonomous communication channel
Intent classification module	Portal message text; prior thread context; predefined intent taxonomy	Classifies the message into categories such as refill request, appointment logistics, result inquiry, administrative request, follow-up concern, symptom concern, or mixed content	Classification must be conservative; low-confidence or mixed-intent messages should not be forced into a single templated pathway	Intent label with confidence indicator and routing recommendation	Establishes classification as the first safety-critical decision point in the agent workflow

Urgency and scope gate	Intent label; urgency terms; symptom indicators; escalation rules; classification confidence	Determines whether the message is eligible for automated drafting or should be routed to manual review	Urgent, ambiguous, safety-sensitive, low-confidence, or template-uncovered messages must bypass drafting	"Draft eligible," "Manual review," or "Escalate" routing status	Separates selective automation from unsafe general-purpose response generation
Protocol retrieval module	Classified intent; local clinical protocol library; eligibility rules; escalation criteria	Retrieves institution-specific rules that define whether and how the message may be answered	Retrieved protocol must be current, locally approved, and relevant to the classified intent	Protocol snippet or policy reference visible to reviewer	Grounds the agent in local practice rather than generic medical advice
Physician-approved template library	Intent category; approved response templates; required disclaimers; standardized wording	Selects the appropriate response structure for routine, non-urgent message types	Template must define required content, optional fields, tone expectations, and escalation language	Selected template shown with editable draft	Converts LLM generation into controlled template adaptation
Patient-history grounding layer	EHR data; medication list; allergies; recent results; recent visits; care plan; prior messages	Retrieves only patient-specific facts needed to personalize the draft	Minimum necessary data principle; no inferred facts; all patient-specific statements must be source-traceable	Patient-data panel showing fields used in the draft	Makes personalization auditable and reduces unsupported clinical inference
Retrieval-augmented LLM drafting engine	Patient message; selected template; retrieved protocol; verified patient-history elements	Produces a structured draft response in patient-facing language	Draft must remain within retrieved sources and approved template boundaries	Draft reply ready for clinician review	Positions the LLM as a constrained drafting engine, not a clinical decision-maker
Rule-based safety and content filter	Draft text; retrieved protocol; template requirements; escalation rules; unsupported-claim checks	Screens the draft for unsafe advice, missing escalation language, unsupported claims, inappropriate tone, or protocol inconsistency	Unsafe drafts should be blocked or routed to manual handling rather than repaired through unconstrained regeneration	"Pass," "Block," or "Manual review required" safety status	Adds a second safety layer after generation and before clinician review

Clinician review interface	Original message; draft; protocol source; selected template; patient-history panel; safety status	Allows clinician to approve, edit, discard, or escalate the draft within the existing portal workflow	No patient-facing response may be sent without clinician approval	Editable draft with source context and action buttons	Preserves clinician accountability and supports efficient verification
Audit and governance layer	Clinician edits; approvals; rejections; escalations; safety-filter triggers; template usage	Captures implementation data for quality improvement, monitoring, and template refinement	Feedback should inform governed updates, not uncontrolled model behavior changes	Audit log and performance-monitoring dataset	Links local deployment to continuous oversight, safety review, and system maintenance

Design principles

The agent should be designed around safety first, meaning that no message is sent without clinician approval. It should use a templated foundation for predictable intents, personalize only with verified patient information, and expose the reasoning context needed for rapid review [9, 15]. It should also fit naturally into the existing patient portal interface so that reviewing, editing, and sending the draft does not create a separate workflow burden. These design principles reflect lessons from early deployments and evaluations of AI-generated patient-message replies [21, 25, 26].

Intent Classification and Message Triage

NLP-based intent classifier

The intent classifier categorizes incoming portal messages into predefined classes such as medication refill, appointment question, laboratory result query, new symptom, follow-up concern, administrative request, or mixed-content message. Prior classification studies show that patient messages can be represented with traditional machine learning, neural networks, and transformer-based models, making intent detection a feasible foundation for workflow routing [16-18]. For the proposed agent, classification should be conservative because an incorrect assignment may retrieve the wrong template or miss a safety-relevant concern. The classifier’s role is therefore to support triage and template matching, not to make final clinical decisions [19, 27].

Urgency detection and escalation

Urgency detection is the safety gate that prevents the drafting pipeline from processing messages that may need immediate clinical attention. Work on emergency detection and patient-message prioritization indicates that language models and structured routing approaches can help identify concerning content for escalation [5, 6]. Messages describing new or worsening symptoms, mental health concerns, medication reactions, or other safety-sensitive issues should be diverted to an urgent or manual review queue. The drafting agent should generate responses only after the message has cleared this urgency filter and matched a non-urgent, approved workflow [7, 28].

Matching non-urgent intents to templates and protocols

For messages classified as non-urgent, the agent maps the intent to a physician-approved template and any associated clinical protocol. For example, a refill-related message would retrieve a refill response template and the local rules governing which medications can be handled through portal communication [20]. Administrative questions, routine result

inquiries, and follow-up logistics would similarly map to templates that define tone, required content, and escalation language. This mapping transforms classification into a practical workflow action and limits drafting to institutionally defined response pathways [4, 11].

Retrieving relevant patient history

After a message has been classified and matched to a template, the agent retrieves patient-history elements relevant to the specific intent. A medication-related request may require active medication information, allergies, prior prescribing context, or recent monitoring data, whereas an appointment follow-up may require recent encounter details and the prior message thread [11, 20]. The agent should retrieve only information needed for the draft, reducing irrelevant context and making clinician verification easier. Patient-specific retrieval also helps the final response read as individualized communication rather than a generic macro [12, 15].

Response Generation: Template Selection, Protocol Grounding, and LLM Drafting

Curating physician-approved templates

Physician-approved templates define the canonical structure of responses for each routine intent. These templates can include greeting style, acknowledgment of the patient's request, required clinical content, protocol-based instructions, escalation language, and optional personalization fields [4, 9]. Curation should involve clinicians who understand both local policy and patient communication norms, because template wording must be clinically safe and acceptable to the care team. The LLM's task is then to adapt a controlled template into a polished draft rather than inventing the response structure [22, 23].

Encoding local clinical protocols

Local clinical protocols provide the rules that determine whether and how a request may be answered through a portal-message draft. These rules may specify which medication requests require manual clinician assessment, which laboratory explanations can use standardized educational language, and which symptoms should trigger escalation rather than routine reply [5, 20]. Encoding protocols as retrievable constraints gives the agent a bounded operational space and supports consistent application of institutional policy. The resulting draft should reflect local practice rather than generic medical advice [12, 29].

Retrieval-augmented LLM drafting

Retrieval-augmented drafting allows the model to generate a response using the selected template, relevant protocol snippets, and patient-specific structured data as grounding material. This approach differs from unconstrained prompting because the model receives approved source content and is expected to remain within that content while producing fluent patient-facing language [11, 12]. The draft can include the patient's concern, a protocol-consistent explanation, and any clinician-verifiable next step. Because the output remains a draft, the clinician can confirm whether the retrieved context and generated wording are appropriate before sending [1, 10].

Personalization with patient history

Personalization should be limited to verified patient-history elements that are directly relevant to the message intent. For example, the agent may incorporate the patient's medication name, recent encounter context, or relevant follow-up plan when those data are retrieved from structured records or prior messages [11, 15]. The goal is to make the response clinically specific and relationally appropriate without allowing the model to infer facts that are absent from the record. This patient-history layer should therefore be transparent to the clinician and traceable to source data [14, 24].

Handling template-uncovered message sub-content

Patient portal messages often contain multiple requests, some of which may be covered by templates and others that require clinician judgment. When the agent detects sub-content outside approved template scope, it should avoid fabricating advice and instead prepare a draft that clearly leaves that portion for clinician review [6, 19]. The draft can address the templatable portion while marking the remaining concern for manual completion, preserving workflow efficiency without overstating automation. This conservative behavior is especially important for ambiguous symptoms, medication concerns, and mixed clinical-administrative messages [27, 28].

Safety, Human Oversight, and Feedback Loop

Clinician review interface

The clinician review interface should display the generated draft beside the original portal message, relevant thread history, selected template, protocol source, and patient-data elements used to populate the response. Prior evaluations of AI-generated patient-message replies emphasize that clinician trust depends not only on the draft text but also on whether the reviewer can verify its grounding quickly [21, 25]. The interface should allow the clinician to send, edit, discard, or escalate the draft without leaving the ordinary inbox workflow. By making source context visible, the agent supports rapid verification while preserving the clinician's role as accountable communicator [12, 26].

Safety guardrails and content moderation

Safety guardrails should operate before a draft reaches the clinician and again at the point of review. Automated checks can screen for contraindicated advice, inappropriate tone, unapproved references, unsupported clinical claims, missing escalation language, or inconsistency with retrieved protocols [5, 14]. Content moderation should also identify drafts that appear overly definitive, insufficiently empathic, or misaligned with the patient's expressed concern. If a safety rule is triggered, the agent should block the draft and route the original message for manual handling rather than attempting to repair the response through additional generation [23, 29].

Escalation for ambiguous or unrecognized requests

The agent should treat ambiguity as a reason to withhold drafting rather than as an invitation to improvise. Patient-message classification studies show that messages may contain overlapping concerns, indirect symptom descriptions, and nonstandard phrasing that can complicate intent assignment [16, 17, 19]. When classification confidence is low, urgency is uncertain, or no approved template applies, the system should route the full message to the clinician's usual inbox. This conservative escalation policy aligns with the principle that the agent should support routine communication but not replace clinical triage for unclear or safety-sensitive requests [6, 7].

Learning from clinician edits and feedback

Clinician edits, rejections, and escalation decisions should be captured as structured feedback for improving templates, protocols, routing rules, and drafting behavior. Studies of AI-assisted message workflows highlight the importance of monitoring how clinicians actually use, modify, or decline generated drafts after deployment [13, 21, 26]. Feedback should be reviewed through governance processes rather than automatically changing clinical behavior without oversight. In this way, the agent becomes a continuously refined communication support system while remaining anchored to physician-approved content and institutional safety standards [4, 24].

Integration into Clinical Portal Workflow

Embedding in the patient portal messaging interface

The agent should be embedded directly within the patient portal messaging interface so that draft review occurs where clinicians already manage inbox work. Prior implementations of generated draft replies in electronic health record workflows suggest that adoption depends on minimizing additional clicks, context switching, and separate documentation steps [2, 3]. The draft should appear inline, with clear labeling that it is AI-generated and pending clinician review. This integration allows the clinician to treat the draft as editable support rather than a parallel system that competes with established communication habits [8, 25].

Workload reduction and clinician adoption

The expected workload benefit comes from reducing repetitive composition, standardizing routine language, and presenting a near-complete response for clinician verification. Clinician adoption would likely depend on whether drafts are concise, correct, empathic, and easier to edit than writing from scratch [9, 22]. If the draft frequently requires substantial rewriting, clinicians may perceive the agent as a burden rather than a support tool. Therefore, deployment should prioritize high-confidence, high-volume, non-urgent use cases where templated language and patient-specific context can meaningfully reduce cognitive effort [13, 26].

Evaluation Strategy

Draft quality and accuracy

Evaluation should assess whether generated drafts are factually correct, clinically appropriate, protocol-adherent, readable, and empathic. Expert review can compare the draft against the original message, retrieved template, local protocol, and patient-history elements used during generation [1, 15]. Reviewers should also examine whether the draft avoids unsupported claims and whether it clearly addresses the patient's intent. Because the agent is conceptualized as a human-reviewed drafting tool, evaluation should focus on draft usefulness and safety rather than autonomous response performance [10, 12].

Efficiency and time savings

Efficiency evaluation should examine whether the agent reduces clinician effort in reviewing and finalizing non-urgent responses. Relevant outcomes can include time required to review, edit, and send a draft, perceived cognitive load, workflow fit, and acceptability among clinicians who manage high portal-message volume [2, 25]. These assessments should be interpreted cautiously because efficiency gains depend on message mix, template coverage, interface design, and clinician trust. The central question is whether the agent makes routine communication easier while maintaining the clinician's ability to exercise judgment [21, 26].

Safety and adverse event monitoring

Safety evaluation should track erroneous, incomplete, misleading, or potentially harmful drafts that reach clinician review, as well as any failure modes that could affect patient understanding. Prior work on emergency detection, patient-message prioritization, and generative AI safety underscores the need to monitor both routing errors and draft-content errors [5, 6]. Evaluation should also examine whether urgent or ambiguous messages are appropriately withheld from the drafting pipeline. Ongoing safety monitoring should be part of operational governance, not a one-time predeployment assessment [14, 29].

Table 2 consolidates the safety, evaluation, and implementation requirements that determine whether an LLM drafting agent can reduce inbox burden without weakening clinical accountability or patient communication quality.

Table 2. Safety, Evaluation, and Implementation Framework for Human-Supervised LLM Drafting of Patient Portal Replies

Evaluation Domain	Key Question	Suggested Measures	Failure Modes to Detect	Governance Response	Practical Implementation Implication
Intent classification safety	Does the system correctly identify which messages are eligible for drafting?	Intent classification accuracy; low-confidence rate; mixed-intent detection rate; manual override frequency	Symptom messages labeled as routine; mixed-content messages oversimplified; incorrect template category assigned	Review misclassified cases; refine taxonomy; adjust confidence thresholds; expand escalation rules	Conservative routing is more important than maximizing automation volume
Urgency and escalation performance	Are urgent, ambiguous, or safety-sensitive messages withheld from drafting?	Sensitivity for urgent-message detection; false-negative urgent routing rate; escalation appropriateness; clinician safety review	New or worsening symptoms entering the drafting pipeline; mental health concerns missed; medication reactions treated as routine	Strengthen escalation rules; retrain urgency detector; introduce mandatory manual review for high-risk terms	The agent should default to manual handling when safety status is uncertain
Protocol adherence	Do drafts follow local clinical protocols rather than generic medical advice?	Protocol-consistency review; proportion of drafts with correct protocol source; unsupported recommendation rate	Draft contradicts local protocol; outdated policy used; response gives advice beyond approved scope	Update protocol library; retire outdated templates; require source-version control	Content maintenance is a core operational requirement, not a technical afterthought
Template fidelity	Does the draft preserve the required structure and wording of physician-approved templates?	Required-field completion; disclaimer inclusion; template-deviation rate; clinician correction frequency	Missing escalation language; altered approved wording; excessive free-form generation	Lock required template sections; revise prompt constraints; conduct periodic template audits	The LLM should adapt approved language, not replace institutional templates
Patient-history accuracy	Are patient-specific details correct, relevant, and source-traceable?	Accuracy of medication, allergy, result, visit, and care-plan references; source-traceability rate; irrelevant-context rate	Wrong medication named; outdated result referenced; inferred patient fact included;	Improve EHR retrieval rules; limit data fields by intent; require visible source display	Personalization should be narrow, verified and easy for clinicians to check

			excessive chart context used		
Draft clinical appropriateness	Is the reply clinically safe, complete, and suitable for clinician review?	Expert review score; clinical appropriateness rating; harmful or misleading draft rate; completeness score	Overly definitive advice; incomplete answer; inappropriate reassurance; unsupported next steps	Block unsafe drafts; add safety rules; revise templates; require specialty-specific review	Draft quality should be judged by usefulness for clinician review not by fluency alone
Patient-facing communication quality	Is the draft understandable, respectful, concise, and empathic?	Readability level; tone rating; empathy score; patient-centered wording assessment; clinician acceptability	Generic language; overly technical wording; insensitive tone; failure to acknowledge patient concern	Revise template tone; add communication-quality checks; incorporate clinician feedback	Efficient drafts must still preserve the relational quality of clinical communication
Clinician workload impact	Does the agent reduce effort rather than create additional review burden?	Time to finalize response; edit distance; discard rate; perceived cognitive load; clinician satisfaction	Drafts require extensive rewriting; review interface causes extra clicks; clinicians distrust source context	Improve interface design; restrict deployment to high-confidence message categories; retire low-yield templates	Adoption depends on whether reviewing is faster than writing from scratch
Human accountability	Is clinician control preserved before the response reaches the patient?	Approval-before-send compliance; edit/approve/discard/escalate rates; audit completeness	Draft sent without review; unclear clinician ownership; automation bias in approval	Enforce mandatory approval; display AI-draft label; audit all final actions	The final message must remain a clinician-approved communication
Postdeployment monitoring	Does the system remain safe as message patterns, protocols, and clinical workflows change?	Drift in intent distribution; template coverage; safety-trigger frequency; complaint or incident reports; longitudinal edit patterns	Degrading performance over time; protocol changes not reflected; increased unsafe draft blocks	Establish governance committee review; schedule template/protocol updates; monitor safety dashboards	Deployment requires continuous operational governance rather than one time validation
Equity and communication consistency	Does the agent produce consistent,	Draft quality by language complexity, demographic subgroup when appropriate,	Less helpful responses for complex	Conduct fairness review; improve language	Standardization should improve consistency

	appropriate drafts across patient groups and message styles?	health-literacy proxy, and message type	wording; biased tone; inconsistent escalation across groups	handling; add subgroup monitoring safeguards	without masking inequitable performance
Legal, privacy, and audit readiness	Can the institution reconstruct why a draft was generated and approved?	Source logging completeness; template version tracking; reviewer action history; data-access auditability	Missing audit trail; unclear source of recommendation; excessive patient-data retrieval	Maintain versioned template/protocol records; log retrieved sources; restrict data access	Traceability is essential for clinical governance, compliance, and trust

Limitations

Dependency on template coverage and protocol completeness

The agent's usefulness depends heavily on the completeness, accuracy, and maintenance of physician-approved templates and local clinical protocols. If protocol content is outdated, incomplete, or inconsistently translated into machine-readable constraints, the agent may generate drafts that are polished but operationally misaligned [4, 20]. Template coverage also determines which message categories can be safely supported, leaving mixed, nuanced, or unusual requests for manual handling. This limitation means that institutional governance and content maintenance are as important as the language model itself [12, 29].

Acceptability and trust

Acceptability depends on whether clinicians and patients view AI-drafted communication as safe, respectful, and aligned with the therapeutic relationship. Studies of patient and clinician perspectives suggest that trust may be influenced by transparency, perceived empathy, editing burden, and confidence that a human clinician remains accountable for the final message [22, 23]. Clinicians may resist adoption if drafts feel generic, inaccurate, or inconsistent with their communication style. For that reason, implementation should emphasize visible human review, local customization, and feedback mechanisms that allow the agent to improve without weakening clinician ownership [13, 24].

Conclusion

The proposed agent uses large language models to support, rather than replace, clinical communication in the patient portal. It classifies incoming messages, identifies non-urgent and template-covered requests, retrieves local protocols and patient-specific context, and prepares a structured draft for clinician review.

Its principal strengths are selective automation, protocol-grounded generation, mandatory human verification, and integration into existing inbox workflows. These design choices keep the agent within a safe administrative support role while preserving clinician accountability for every patient-facing message.

Important challenges remain in template maintenance, protocol encoding, nuanced message handling, and clinician trust. The system must avoid overgeneralization, make its source context visible, and route uncertain or urgent messages away from automated drafting.

Pilot implementations should focus on large ambulatory networks with high portal-message volume and mature clinical protocol libraries. Such settings are well positioned to evaluate whether a carefully bounded drafting agent can reduce

routine inbox burden while maintaining safe, consistent, and patient-centered communication.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 29 Aug 2025

Revised: 26 Sep 2025

Accepted: 09 Nov 2025

Published online: 25 February 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Liu S, McCoy AB, Wright AP, Carew B, Jenkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages: A subjective analysis. *J Am Med Inform Assoc.* 2024;31(6):1367–79.
- Garcia P, Ma SP, Shah S, Smith M, Jeong Y, Devon-Sand A, et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw Open.* 2024;7(3):e243201.
- Tai-Seale M, Baxter SL, Vaida F, Walker A, Sitapati AM, Osborne C, et al. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw Open.* 2024;7(4):e246565.
- Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: Early lessons learned. *JAMIA Open.* 2024;7(2):ooae028.
- Liu S, Wright AP, McCoy AB, Huang SS, Steitz B, Wright A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *J Am Med Inform Assoc.* 2025;32(6):1032–9.

Yang J, So J, Zhang H, Jones S, Connolly DM, Golding C, et al. Development and evaluation of an artificial intelligence-based workflow for the prioritization of patient portal messages. *JAMIA Open*. 2024;7(3):ooae078.

Anderson BJ, Zia ul Haq M, Zhu Y, Hornback A, Cowan AD, Mott M, et al. Development and evaluation of a model to manage patient portal messages. *NEJM AI*. 2025;2(3):Aloa2400354.

Nguyen D, Lee S, La K, Kellogg M, Synghal R, Anwar B, et al. Performance of an intelligent messaging tool for clinical communications. *JAMA Netw Open*. 2026;9(1):e2553174.

Yan S, Knapp W, Leong A, Kadkhodazadeh S, Das S, Jones VG, et al. Prompt engineering on leveraging large language models in generating responses to InBasket messages. *J Am Med Inform Assoc*. 2024;31(10):2263–70.

Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, et al. The effect of using a large language model to respond to patient messages. *Lancet Digit Health*. 2024;6(6):e379–e381.

Liu S, Wright AP, McCoy AB, Huang SS, Jenkins JZ, Peterson JF, et al. Using large language models to guide patients to create efficient and comprehensive clinical care messages. *J Am Med Inform Assoc*. 2024;31(8):1665–70.

Hong C, Chowdhury A, Sorrentino AD, Wang H, Agrawal M, Bedoya A, et al. Application of unified health large language model evaluation framework to In-Basket message replies: Bridging qualitative and quantitative assessments. *J Am Med Inform Assoc*. 2025;32(4):626–37.

Bootsma-Robroeks CM, Workum JD, Schuit SC, Hoekman A, Mehri T, Doornberg JN, et al. AI-generated draft replies to patient messages: Exploring effects of implementation. *Front Digit Health*. 2025;7:1588143.

Biro JM, Handley JL, Malcolm McCurry J, Visconti A, Weinfeld J, Trafton JG, et al. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *NPJ Digit Med*. 2025;8(1):222.

Lee NS, Richards N, Grandominico J, Cronin RM, Hendricks AK, Tripathi RS, et al. Use of a medical communication framework to assess the quality of generative artificial intelligence replies to primary care patient portal messages: Content analysis. *JMIR Form Res*. 2025;9:e71966.

Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int J Med Inform*. 2017;105:110–20.

Suliman L, Gilmore D, French C, Cronin RM, Jackson GP, Russell M, et al. Classifying patient portal messages using convolutional neural networks. *J Biomed Inform*. 2017;74:59–70.

Ren Y, Wu D, Khurana A, Mastorakos G, Fu S, Zong N, et al. Classification of patient portal messages with BERT-based language models. In: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI). Piscataway (NJ): IEEE; 2023. p. 176–82.

Ren Y, Wu Y, Fan JW, Khurana A, Fu S, Wu D, et al. Automatic uncovering of patient primary concerns in portal messages using a fusion framework of pretrained language models. *J Am Med Inform Assoc*. 2024;31(8):1714–24.

Davoudi A, Lee NS, Luong T, Delaney T, Asch E, Chaiyachati K, et al. Identifying medication-related intents from a bidirectional text messaging platform for hypertension management using an unsupervised learning approach: Retrospective observational pilot study. *J Med Internet Res*. 2022;24(6):e36151.

English E, Laughlin J, Sippel J, DeCamp M, Lin CT. Utility of artificial intelligence–generative draft replies to patient messages. *JAMA Netw Open*. 2024;7(10):e2438573.

Kim J, Chen ML, Rezaei SJ, Liang AS, Seav SM, Onyeka S, et al. Perspectives on artificial intelligence–generated responses to patient messages. *JAMA Netw Open*. 2024;7(10):e2438535.

Cavalier JS, Goldstein BA, Ravitsky V, Bélisle-Pipon JC, Bedoya A, Maddocks J, et al. Ethics in patient preferences for artificial intelligence–drafted responses to electronic messages. *JAMA Netw Open*. 2025;8(3):e250449.

Liang AS, Vedak S, Dussaq A, Yao DH, Villarreal JA, Thomas S, et al. Artificial intelligence-generated draft replies to patient messages in pediatrics. *JAMIA Open*. 2025;8(6):ooaf159.

Small WR, Wiesenfeld BM, Brandfield-Harvey B, Jonassen Z, Mandal S, Stevens ER, et al. Large language model–based responses to patients' in-basket messages. *JAMA Netw Open*. 2024;7(7):e2422399.

Mandal S, Wiesenfeld BM, Szerencsy AC, Small WR, Major V, Richardson S, et al. Utilization of generative AI-drafted responses for managing patient-provider communication. *NPJ Digit Med*. 2025;8(1):591.

TaftiAhmad P. Probing patient messages enhanced by natural language processing: A top-down message corpus analysis. *Health Data Sci*. 2021.

Chen J, Lalor J, Liu W, Druhl E, Granillo E, Vimalananda VG, et al. Detecting hypoglycemia incidents reported in patients' secure messages: Using cost-sensitive learning and oversampling to reduce data imbalance. *J Med Internet Res*. 2019;21(3):e11990.

Hu D, Guo Y, Zhou Y, Flores L, Zheng K. A systematic review of early evidence on generative AI for drafting responses to patient messages. *NPJ Health Syst*. 2025;2(1):27.