Physician Edits to AI-Drafted Patient Messages and Their Impact on Clinical Workload

Lily Poursoltan[1,2]; Jie Cao[2]; Wei Chen[3]; Eileen Kim[4]; Tejpreet K. Nakai[4]; Aaron Boussina[2,5]; Andrew Chua[2]; Jeffrey Pan[2]; Marlene Millen[4]; Ming Tai-Seale[2,5,6]; Christopher A. Longhurst[2,5]; Kevin Zhu[1]; Karandeep Singh[2,5]

[1] UC San Diego, Rady School of Management; [2]UC San Diego Health, Jacobs Center for Health Innovation; [3] University of Connecticut; [4] UC San Diego Health, General Internal Medicine; [5] UC San Diego, Biomedical Informatics; [6] UC San Diego, Department of Family Medicine

Summary

- **Why this matters.** EHR-integrated LLMs draft patient-message replies at scale, but in practice the "human-in-the-loop" step can **add time and cognitive load** — so we map *which edits* actually drive that burden.
- **Study design.** Real-world deployment at UC San Diego Health: **7,913** AI-assisted replies across **994** physicians; **15-category** edit taxonomy (LLM + expert validation); mixed-effects models controlling for message complexity, clinician habits.
- **Results.** All **15 edit types increase (↑) response time; high-burden edits** (radiology, referrals, side effects, diagnosis) **> +45% ↑**; frequent **low-load edits** (scheduling, lifestyle) drive the **largest system-level burden**.
- **Key Takeaways. AI-assisted drafts ≠ free lunch** → every edit adds cognitive burden**; 15-category taxonomy** → first empirical map of **human-in-the-loop AI edits**; **Beyond accuracy** → cognitive burden is evaluated and hidden costs shown.
- **Managerial implications.** optimize high-burden tasks, automate/delegate low-burden, build trust via phased adoption.

Problem

Messaging surge → burnout

- Patient-portal messages jumped **≈157%** with telemedicine
- After-hours *"pajama time"* is a major burnout driver (**> 60%** physicians affected)

AI drafts aren't a free lunch.

- LLM-drafted replies are now embedded in EHRs
- Early real-world studies show **mixed or no time savings** — in some cases, total **review/edit time** actually **increases**

Unknown drivers of effort.

- No clear **map** of which **edits** drive **cognitive burden**
- Without this, **prompt & workflow** design remain **blind**

Research Questions

- **RQ1 — What gets changed (and how much)?**
  Which categories of clinician edits appear when physicians review AI-drafted patient messages in the EHR, and how extensive are those edits?
- **RQ2 — What does each edit cost in time (cognitive load)?**
  By how much does each edit category increase response/editing time (proxy for cognitive burden) after adjusting for message complexity and temporal factors?

Study Design

Real-world deployment

- UC San Diego Health, **Epic EHR-integrated LLM** (Apr 2024–2025)
- Clinicians: start from **AI draft replies → review/edit → send**

Cohort & Data

- **7,913** AI-assisted replies across **994** physicians, diverse specialties
- **Data:** Patient message, prompts, AI drafts, final response, timestamps (outliers > 97th removed); handled system timeouts / interruptions

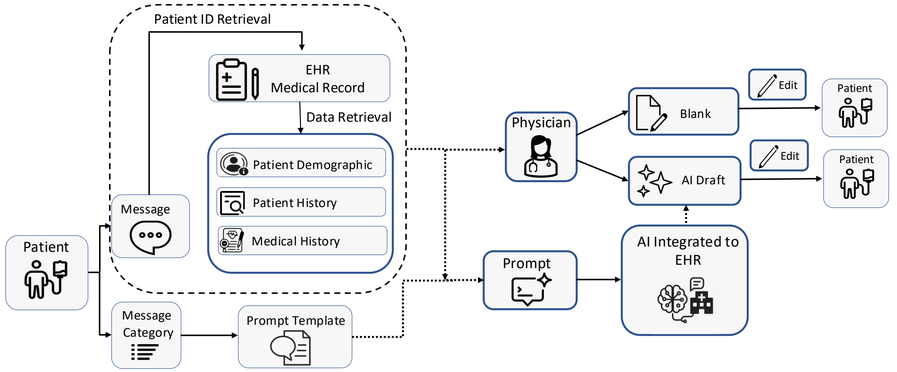
Edit taxonomy & validation

- **15-category taxonomy** (Llama + iterative expert review)
- **Edit extent: 0–3 scale** (identical → major change)
- **Validation: κ ≈ 0.82** (substantial–almost perfect agreement)

Analysis

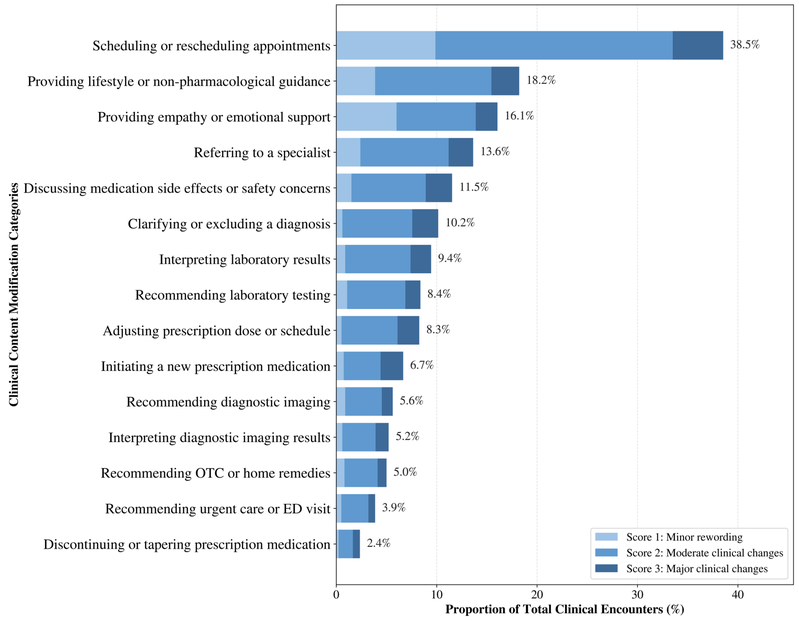
- **DV: Editing time (reply init → send)** = proxy for cognitive burden
- **Model: Mixed-effects** with physician random intercepts
- **Controls: message complexity** (readability, linguistics, terminology + SOTA embedding models → PCA), **temporal factors**

Human-in-the-Loop AI Workflow



Patient message → EHR → LLM draft → physician edits → final message; we analyze edits and their time cost

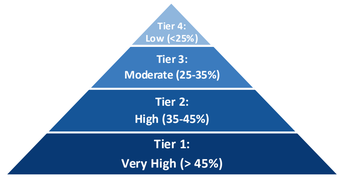
Human-in-the-Loop AI Edits: What & How Much



- **Scheduling dominates volume. Scheduling/Rescheduling = 38%** of all edits (largest); **Discontinue/Taper meds = 2.4%** (smallest).
- **Next most common. Lifestyle/Non-pharmacologic advice = 18.2%; Diagnosis clarification/rule-out = 10.2%.**
- **Edit size is usually "moderate."** Most edits change meaning but **not** care plan; **major** edits cluster in **Scheduling (7.5%)** and **Diagnosis clarification (4.6%)**.

Human AI Cognitive Load

	Rank	Category	Effect size (% Increase in Edit Time)	95% CI	p -value
Tier 1	1	Interpreting Radiology Result	45.9	[33.2, 59.8]	<0.001
	2	Discussing Medication Side Effects	45.6	[36.6, 55.3]	<0.001
	3	Offering a Referral to a Specialist	45.5	[37.2, 54.3]	<0.001
	4	Clarifying or Ruling Out Diagnosis	45.3	[37.5, 53.6]	<0.001
Tier 2	5	Lifestyle, Diet, Non-Pharmacological Advice	43.9	[36.5, 51.7]	<0.001
	6	Interpreting Laboratory Result	38.8	[29.4, 48.8]	<0.001
	7	Starting Prescription Medication	37.4	[28.0, 47.5]	<0.001
	8	Adjusting Medication Dose	37	[27.6, 47.0]	<0.001
	9	Recommending a Radiology Test	35.4	[24.6, 47.2]	<0.001
	10	Recommending a Laboratory Test	35.4	[26.3, 45.1]	<0.001
	11	Stopping/Tapering Medication	35.1	[18.9, 53.6]	<0.001
Tier 3	12	Urgent Care/Emergency Recommendation	34.4	[22.1, 48.1]	<0.001
	13	Recommending OTC/Home Remedies	28.1	[16.5, 40.8]	<0.001
Tier 4	14	Scheduling/Rescheduling Appointment	23.3	[18.2, 28.7]	<0.001
	15	Empathy/Reassurance/Emotional Support	19.8	[12.2, 27.8]	<0.001



- **All 15 edit categories increased editing time** → every type of edit adds measurable cognitive burden
- **Highest-burden tasks:** interpretive/diagnostic edits (e.g., radiology & side effects **+46%**, referrals **+45%)**
- **Lowest-burden tasks:** supportive/administrative edits (e.g., empathy **+20%**, scheduling **+23%)**, but high frequency means **cumulative workload is substantial**

Contributions & Implications

Key Contributions

- **Developed 15-category Taxonomy.** human–AI edits in clinical messaging
- **Expanding Evaluation Metrics. Cognitive burden** as a core measure, revealing **hidden costs** of human-in-the-loop AI
- **Extending Cognitive Load Theory. Systematic variability** in real-world contexts, Burden differs by **task type**, not just complexity

Managerial Implications

- **AI System Design.** Optimize high-burden tasks; build **feedback loops**
- **Workforce Management.** Automate/delegate low-burden tasks; monitor **real-time load**
- **Adoption Strategy.** Start with **low-burden** → build trust → expand to **high-burden**

Responsible AI requires managing workload + trust — not just accuracy