Editing with AI: How Doctors Refine LLM-Generated Answers to Patient Queries

Rahul Sharma[*]
JazzX AI , Bangalore, India

Pragnya Ramjee[*]
Stanford University, USA

Kaushik Murali
Sankara Eye Hospital, Bangalore, India
`kaushik@sankaraeye.com`

Mohit Jain
Microsoft Research, Bangalore, India
`mohja@microsoft.com`

Abstract

Patients frequently seek information during their medical journeys, but the rising volume of digital patient messages has strained healthcare systems. Large language models (LLMs) offer promise in generating draft responses for clinicians, yet how physicians refine these drafts remains underexplored. We present a mixed-methods study with nine ophthalmologists answering 144 cataract surgery questions across three conditions: writing from scratch, directly editing LLM drafts, and instruction-based indirect editing. Our quantitative and qualitative analyses reveal that while LLM outputs were generally accurate, occasional errors and automation bias revealed the need for human oversight. Contextualization—adapting generic answers to local practices and patient expectations—emerged as a dominant form of editing. Editing workflows revealed trade-offs: indirect editing reduced effort but introduced errors, while direct editing ensured precision but with higher workload. We conclude with design and policy implications for building safe, scalable LLM-assisted clinical communication systems.

1 Introduction

Patients seek information throughout their medical journeys [9]. Effectively addressing these information needs is central to patient-centered care [30], supporting informed decision making [9] and improving health outcomes like treatment adherence and emotional well-being [24]. Digital communication platforms, such as chat groups, instant messengers, and web portals, have emerged as powerful tools for meeting these needs, by enabling healthcare providers to deliver timely, accurate responses to patient concerns [20, 28, 29]. However, the increasing ubiquity of these platforms, accelerated by the COVID-19 pandemic and a surge in telemedicine, has strained healthcare systems with unprecedented volumes of interactions [12, 18].

The advent of generative AI has raised the possibility that large language models (LLMs) could help manage this burden by assisting physicians in patient communication [20, 22, 29]. Early studies evaluating the quality of LLM responses to medical questions found physicians frequently rate these answers as 'safe', and sometimes even prefer them to peer-written responses [23]. However, LLMs can still output inaccurate, outdated or inappropriate information [3, 5], and patients report lower trust in invalidated LLM responses [20]. Thus, doctor oversight remains indispensable.

One promising approach is using LLMs to generate *draft* responses to patient queries for clinicians to refine, rather than replace them [4, 20]. Initial deployments report clear benefits such as reduced

[*] Work done while employed at Microsoft Research, Bangalore, India.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

workload on doctors and improved response quality [7, 25].However, challenges remain, including physicians over-relying on LLM output instead of exercising their own clinical judgment [5, 7, 27] and producing verbose answers [25], potentially due to the labor required in correcting and shortening LLM responses.Thus, understanding how clinicians refine LLM drafts, and how different co-authoring strategies align with their workflows to maximize usability and effectiveness, is essential [1, 5, 10].

While the existing body of work has examined LLM response quality and physician acceptance, the specific approaches for refining LLM drafts remain understudied. We address this gap with a mixed-methods study comparing three answer generation approaches: *writing from scratch* (doctor writes answer, no LLM involved), *direct editing* (manually correcting LLM-generated drafts), and *instruction-based editing* (providing instructions to the LLM for correction). We evaluate accuracy, completeness, and safety of the answers, along with doctor's efficiency and preferences, to answer these research questions: How do different LLM co-authoring approaches affect physician efficiency and response quality when answering patient queries? What are physicians' perceptions of usability and workflow compatibility across different LLM co-authoring approaches?

To ground our investigation in a concrete clinical context, we focus on cataract surgery, the most common ophthalmic procedure worldwide and the second highest surgical procedure globally [17]. Doctor-patient communication becomes critical for surgical procedures, as patients seek reassurance about surgical risks and detailed information about post-operative care [9]. High patient query volumes have prompted development of LLM-based communication systems, including standardized patient education materials [26] and doctor-in-the-loop chatbots [20, 22].

We worked with 9 ophthalmologists who answered 144 cataract surgery questions across the three conditions, then participated in focus group discussions and evaluated answers generated by their peers. Doctors found LLM-generated answers to be generally accurate and complete, but occasional factual errors and the risk of automation bias underscored the need for human oversight. A central form of editing was contextualization—adapting otherwise generic answers to local practices, terminology, and patient expectations. Editing workflows revealed trade-offs: instruction-based editing reduced effort compared to direct text editing, yet introduced occasional technical errors and ambiguities. Finally, while doctors valued the polished language of LLM drafts, they often had to simplify and reframe answers into shorter, clearer, and more conversational forms suitable for patients. Together, these results extend prior work by shifting attention from LLM response quality alone to the process of refinement, illuminating design and policy directions for safe, scalable integration of LLMs into patient communication.

2 Methods

We conducted a mixed-method controlled user study during June-July 2025 in collaboration with Sankara Eye Hospital, a leading tertiary eye care and teaching hospital in India. The study took place at its Jaipur and Hyderabad branches. Approval was obtained from the Scientific and Ethics Committees of both sites. In line with hospital policy, participants received no financial compensation.

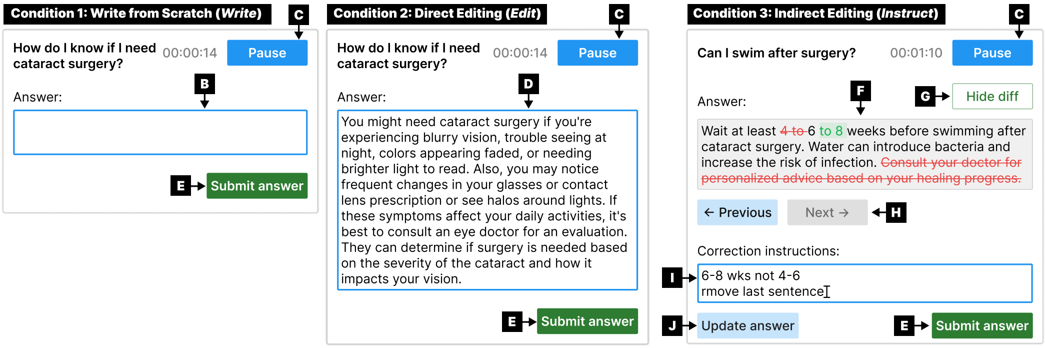
Participants Nine practicing ophthalmologists (4 female) from Sankara Eye Hospital (4 Hyderabad, 5 Jaipur) participated in the study. All performed at least 10 cataract surgeries per week. Their mean age was 37.1±5.5 years, with 11.6±5.1 years of professional experience.

Interface Design We designed a custom web-based application powered by GPT-4o (Figure 1) to compare the three conditions.

Condition 1 (*Writing from Scratch*) aka *Write*: Doctors answered questions in a blank text box (Figure 1B, without the pre-generated LLM answer).

Condition 2 (*Direct Editing*) aka *Edit*: Doctors revised an LLM-generated draft in an editable text box (Figure 1D).

Condition 3 (*Indirect Editing*) aka *Instruct*: Doctors received an LLM draft in a non-editable text box (Figure 1F) and indirectly edited it by providing instructions in a separate text box (Figure 1I). The system generated revised responses with changes highlighted (green for additions, red strikethrough for deletions), along with an optional toggle button (Figure 1G) to hide/show differences. Navigation



Condition 1: Write from Scratch (*Write*)

How do I know if I need cataract surgery?

Answer:

Condition 2: Direct Editing (*Edit*)

How do I know if I need cataract surgery?

Answer:

You might need cataract surgery if you're experiencing blurry vision, trouble seeing at night, colors appearing faded, or needing brighter light to read. Also, you may notice frequent changes in your glasses or contact lens prescription or see halos around lights. If these symptoms affect your daily activities, it's best to consult an eye doctor for an evaluation. They can determine if surgery is needed based on the severity of the cataract and how it impacts your vision.

Condition 3: Indirect Editing (*Instruct*)

Can I swim after surgery?

Answer:

Wait at least ~~4 to~~ 6 to 8 weeks before swimming after cataract surgery. Water can introduce bacteria and increase the risk of infection. ~~Consult your doctor for personalized advice based on your healing progress.~~

Correction instructions:

6-8 wks not 4-6
rmove last sentence

Figure 1: Web-based study interface. **(Left)** Condition 1: *Writing from Scratch* (Write), where doctors write answers in text box *B*. **(Middle)** Condition 2: *Direct Editing* (Edit), where doctors edit pre-generated LLM answers (in the textbox *D*). **(Right)** Condition 3: *Indirect Editing* (Instruct), where doctors provide instructions for revision (*I*), with changes visually highlighted (*F*).

arrows (Figure 1H) enabled participant to review past versions of the answer and selecting one for further editing or submission.

All conditions included a timer (Figure 1C), which doctors could pause only for interruptions. The order of conditions was randomized across participants for counterbalancing. All interactions were automatically logged.

Procedure After a pilot study, we conducted the main study in three phases (Figure 2):

Phase 0: Pilot Study. We conducted four training sessions with 9 doctors to familiarize them with the Phase 1 web interface (Figure 1). Each session, held on Microsoft Teams and facilitated by two researchers, began with a demonstration of the three conditions using two example questions. Doctors then practiced with five sample questions per condition. We collected demographic information via a web form at the start of each session. Feedback from these pilots informed minor refinements to the interface. Each session lasted ∼60 minutes.

Phase 1: Answer Generation. We conducted this phase over two sessions with nine doctors, organized into three triplets. Each doctor generated 48 answers (16 per condition). The question dataset comprised of 144 common cataract surgery questions sourced from a large-scale patient chatbot deployment in India [22]. In each session, triplets received the same 24 questions, distributed across conditions so that each question was answered once in each condition across the three triplets. This design enabled inter-condition comparison of response quality and efficiency. In total, doctors generated 432 answers. Afterward, they completed NASA-TLX ratings (5-point Likert scale) [11] and indicated their most/least preferred, most efficient, safest, and most workflow-aligned condition.

Phase 2: Focus Group Discussions (FGDs). We conducted three FGD sessions in English via Microsoft Teams, with each doctor participating in one session. Two researchers facilitated: one moderated, while the other took notes and prompted follow-ups. To aid recall, screenshots of the different conditions were shared during the discussion. Topics included doctors' experiences with each condition, perceived advantages and drawbacks, trust and accuracy of responses, and perspectives on the future of LLM-assisted patient communication. Each session lasted ∼75 minutes and was audio-recorded, with transcripts prepared by a researcher immediately afterward.

Phase 3: Answer Evaluation. Each doctor rated 48 questions (24 from each session), reviewing one answer from each condition per question. Ratings covered three dimensions: accuracy, completeness, and likelihood of harm—using the question: '*Please rate the [dimension] of this answer (1 = very low, 2 = low, 3 = medium, 4 = high, 5 = very high).*' Doctors could optionally provide written justifications. To minimize bias, all answers were blinded to condition and randomized in order. Each doctor evaluated answers generated by another triplet to avoid self-assessment. Each answer received two independent ratings, which we averaged for analysis.

Data Collection and Analysis We adopted a mixed-method approach combining quantitative and qualitative analyses. Data sources included interaction logs (time, answer length, edit distance),

Table 1: Demographic details of study participants (n=9).

ID	City	Age	Gender	Prof exp (years)	Surgeries/week	Participation
D1	Hyderabad	34	Female	7	10-20	FGD3
D2	Hyderabad	36	Male	10	40+	FGD1
D3	Hyderabad	34	Male	15	40+	FGD1
D4	Jaipur	43	Male	16	20-30	FGD2
D5	Jaipur	43	Male	17	20-30	FGD2
D6	Jaipur	45	Male	15	10-20	FGD2
D7	Jaipur	30	Female	4	20-30	FGD2
D8	Jaipur	31	Female	5	20-30	FGD2
D9	Hyderabad	38	Female	15	40+	FGD3

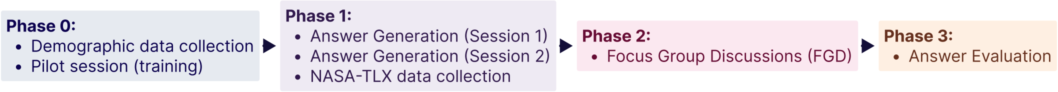

Phase 0:
• Demographic data collection
• Pilot session (training)
Phase 1:
• Answer Generation (Session 1)
• Answer Generation (Session 2)
• NASA-TLX data collection
Phase 2:
• Focus Group Discussions (FGD)
Phase 3:
• Answer Evaluation

Figure 2: Phases of our mixed-methods evaluation study.

NASA-TLX ratings, preference rankings, FGD transcripts and notes, and answer ratings on accuracy, completeness, and harmfulness. Quantitative data (logs and ratings) were analyzed using descriptive statistics, t-tests, and linear mixed-effects models (with condition as a fixed effect and random intercepts for questions and participants). Qualitative data (FGD transcripts and notes) were analyzed through inductive thematic analysis [6], initially coded by one author and iteratively discussed with three co-authors to derive broader themes. Edits in *Edit* and *Instruct* were also categorized and coded.

3 Findings

3.1 Answer Quality and Reliability

Accuracy, Completeness, and Safety Drawings on our interactions with doctors and prior research [20], we focus on three key qualities of doctor-generated answers: accuracy, completeness, and non-harmfulness. We analyzed 432 answers (3 per question for 144 questions in Phase 1) using ratings from two independent doctors on the three metrics in Phase 3. Ratings were averaged across the two doctors for analysis. Linear mixed-model analysis revealed significant differences across conditions (accuracy: $F(2, 278) = 22.4$, $p < .005$; completeness: $F(2, 278) = 76.7$, $p < .005$; non-harmfulness: $F(2, 278) = 23.8$, $p < .005$). Post-hoc comparisons showed that both *Edit* and *Instruct* significantly outperformed *Write* on all three metrics. While *Instruct* responses were rated slightly higher than *Edit*, these differences were not statistically significant for accuracy ($p = 0.07$) or completeness ($p = 0.7$). For non-harmfulness, however, *Instruct* was rated significantly less harmful than *Edit* ($p < .05$). Among the 432 answers, harmful ratings were relatively rare: 14 in *Write*, 7 in *Edit*, and 5 in *Instruct*. This suggests that most answers were high-quality and safe for end-users.

Here are examples of low accuracy and high harmfulness (Ans1), and low completeness (Ans2).
Q1: Should I take complete rest post-surgery?
Ans1: Complete bed rest is not recommended. You can resume your daily activities.
Q2: Can the patient take their regular BP medicine on the surgery day?
Ans2: Follow your doctor's instructions carefully to ensure a safe and successful procedure.

Consistency and Standardization Across Answers We observed a philosophical divide among doctors in how answers should be framed. Some preferred answers tailored to each patient's demographic and context, while others valued standardized responses across patients and doctors. The first group favored *Write* as it allowed them to "*frame answers in their own language*," whereas the latter leaned toward *Edit* or *Instruct* for their consistency. With an LLM providing a base draft, most doctors made minimal edits, resulting in more homogeneous answers across patients.

> "*Each doctor in Write will write the answer differently, and also... differently for each patient, as in, if the patient is educated or not... old or young. E.g., I would know a rural patient swims in borewells vs a patient from a city swims in a pool. With that, the answer to a query about swimming becomes location-specific. Manual answers are tailored and very subjective... AI-generated is more homogeneous across geography, across demography.*" – D2.

While some appreciated this uniformity, others criticized LLM-generated answers as "*generic*" or "*lacking depth*," contrasting with the expectation that doctors be "*very specific*." From a hospital's perspective, however, standardization was seen as beneficial for ensuring uniform care: "*If you ask 'when to come after the surgery', some surgeons may say next day, some may say next week... The hospital would want a standard treatment guideline, and expect all doctors to say 'one week'.*" – D5.

Automation Bias as an Emerging Risk A consequence of this standardization was the emergence of automation bias. Doctors reported that they often skipped correcting LLM answers if they were "*not completely wrong*," even when the advice diverged from their own practice. Over time, this tendency risked anchoring doctors to the LLM-generated phrasing and subtly shifting their judgment.

> "*I edited only where there was strong disagreement. The remaining were okay, so I let it be... E.g., it said 'take a head bath after 1 week' while I would say '2 weeks'... I didn't edit this as it's not completely wrong.*" – D3.

Such omissions increased standardization but reduced subjectivity, creating a risk of clinicians internalizing the AI's defaults. As D4 reflected: "*These (LLM-generated) answers can manipulate our thinking. It changes our mindset.*"

Perceived Verbosity vs. Useful Detail Answer length varied sharply across conditions. A linear mixed-effects analysis revealed significant differences ($F(2, 278) = 536.9$, $p < .001$): *Write* produced much shorter answers (89.6±74.5 characters/answer) than both *Edit* (335.1±90.1) and *Instruct* (351.6±103.8), with both pairwise comparisons showing p<.001. No difference was found between *Edit* and *Instruct* (p=0.7). Comparing the two edit conditions, *Edit* and *Instruct*, we found no statistical differences in insertion, deletion, substitution, or overall edit distance.

Doctors' editing behavior highlighted how length shaped their experience. In *Edit*, only 34 of 144 answers (23.6%) were edited. Edits typically shortened answers, reducing their length from 355.7±68.9 to 264.1±127.3 characters. Edits included deleting unnecessary details (e.g., D2 removed: "Also, medications used during surgery might cause drowsiness.") or condensing verbose phrasing. Additions of new content were rare (14 answers) compared to deletions (39) and substitutions (27).

In *Instruct*, editing unfolded differently. Of the 144 answers, 51 (35.4%) were edited, requiring 134 instructions in total (2.6±1.4 per answer). Two answers required as many as seven iterative comments. Doctors noted the difficulty of condensing LLM-generated answers through instructions:

> "*Every AI-generated answer usually is lengthy... and has multiple points. For each point, I need to do multiple modifications... I can't give instructions at one go. I have to keep modifying.*" – D8

Taken together, these results show that while LLM-supported conditions (*Edit*, *Instruct*) generated more detailed and comprehensive answers than *Write*, doctors often worked to shorten or streamline them. This tension suggests that answer length, though positively associated with completeness, could burden clinicians with extra review effort or risk overwhelming patients with information.

3.2 Editing Workflows and Doctor Effort

Cognitive and Physical Demands NASA-TLX ratings revealed clear differences across conditions: *Instruct* was rated as the least physically and mentally demanding, while *Write* was rated the highest on both dimensions. Doctors stated that *Write* felt "*cumbersome*" (D9) because it required extensive typing and recall to compose answers from scratch. In contrast, *Edit* and *Instruct* shifted effort toward reviewing and editing existing drafts. Editing in *Edit* often involved small textual adjustments, while *Instruct* enabled concise, instruction-based edits: "*I provided very short, crisp instructions.*" (D5).

The nature of mental effort varied by condition. In *Write*, doctors worried about completeness: "*I always thought* 'Am I forgetting something?'*... E.g., for the Q on 'do's and don'ts for cataract surgery', that's a long list. It is hard to get everything.*" (D9). In *Edit*, effort centered on scanning long drafts for issues: "*There is a maze of information in the AI answers... It's not trivial to read 8 lines and decide what to add, where to add it, and what to delete.*" (D5). In *Instruct*, cognitive load came from sequencing effective instructions and verifying results: "*I need to think what is there and not, what instruction needs to be given... and whether that instruction worked or not.*" (D1). Some doctors found this cognitive model appealing—"*computer making the edit is better, as my thoughts fit well with it*" (D2)—while others pointed the learning curve and stochasticity involved in *Instruct*: "*What command to give, how to make it work... there is also the uncertainty of how my instruction will be interpreted*" (D9).

Editing efficiency reflected these trade-offs. A linear mixed-effects analysis revealed significant differences across conditions ($F(2, 278) = 41.3$, $p < .001$). *Write* (52.0±44.5s per answer) took significantly longer than both *Edit* (27.9±39.9s) and *Instruct* (46.4±54.3s) (p<.001), and *Instruct* was also slower than *Edit* (p<.05). This pattern is partly explained by the fact that only a fraction of answers were edited in *Edit* (23.6%) and *Instruct* (35.4%). However, when restricting analysis to edited answers, *Instruct* (85.3±60.4s) was significantly slower than *Write* (52.0±44.4s), while *Edit* (50.2±37.2s) did not differ from either. Focus group discussions suggested this was partly due to the novelty of *Instruct*, which required time to formulate instructions, review edits, and manage uncertainty about corrections. Although doctors noted a small wait time for the LLM to process instructions in *Instruct*, it was not statistically significant.

Types of Edits Manual analysis of the 84 corrections in *Edit* and *Instruct* revealed four categories: corrections to incorrect or ambiguous content (27), deletions of redundant details (63), additions of missing or context-specific information (43), and rewording or restructuring for clarity (9). Each reflected distinct limitations of the LLM drafts and clinicians' priorities in tailoring them for patients.

Corrections. Several answers required factual correction, often because the LLM produced misleading or overly generalized statements. For example, to the question "Do they put a needle in your eye during cataract surgery?", the LLM-generated answer was "No, a needle is not put into your eye during cataract surgery. The surgery is...", which D4 corrected in *Edit* to "Yes, needles are used in a standard way for some steps in cataract surgery. The surgery is...". For the same question, D5 in *Instruct* simply instructed the system to "remove the first sentence". Similarly, D6 flagged that the LLM's advice on fasting before surgery—"no food for six hours"—was inaccurate, since cataract surgery is typically performed under local anesthesia, requiring only two hours of fasting, with a light breakfast permitted. Beyond factual inaccuracies, doctors also corrected vague or hedged responses. As D5 put it, "*Patients want concrete, clear answers. AI is overly cautious and ambivalent. For a question like 'Can I swim?', the answer should be definitive... 'no, not for the first month.'*"

Deletions. All doctors shortened answers by removing verbose or repetitive content. In total, 63 deletions were recorded, including 34 single-sentence removals, 20 two-sentence deletions, and 9 cases of deleting more than two sentences. Doctors felt that long, elaborate answers risked confusing patients, who typically preferred concise advice. As D4 emphasized, "*Shorter, to-the-point answers are definitely better.*" D5 elaborated: "*With AI, a short answer becomes a long answer. E.g., 'Can I swim after surgery?' I would simply say 'no'. But AI will say the same 'no' in a more appropriate, but very lengthy manner... In face-to-face conversations, I would have given a short 'no' as answer.*" Doctors also ensured that the final answer contained as much "*direct information*" as possible. Redundant closing statements such as "consult your doctor" were almost universally removed.

Additions and Rewording. Doctors also added clarifying details that the LLM drafts had omitted, ensuring responses were complete and clinically useful. In total, 43 such additions were recorded, often involving patient-relevant advice. For example, D9 added practical lifestyle advice such as "Avoid sunlight and dusty environments" as post-surgery precautions, which the LLM had omitted. Rewording was less frequent (9 cases), but served to improve readability and presentation. Doctors sometimes requested alternative formats such as bullet points, or relied on *Instruct*'s ability to automatically adjust grammar and flow.

Across categories, edits reflected a shared goal: make responses direct, clear, and patient-appropriate.

Contextualization Beyond these structural edits, doctors frequently engaged in contextualization. Although the study was conducted in India, the prompt used to generate the initial LLM answers was not tailored to this context. As a result, the answers in *Edit* and *Instruct* were often generic and universal in tone, lacking important cultural, clinical, and geographic grounding. Doctors frequently intervened to localize these answers so they would be relevant and actionable for their patients. Typical edits included substituting technical terms (e.g., replacing "*ECCE*" with the more commonly used "*SICS*" surgery type), correcting currency references from USD to INR when discussing costs, and adding India-specific lifestyle guidance such as "avoid sunlight and dusty environments."

Contextualization spanned *all* edit types—corrections, additions, deletions, and rewording. Clinicians emphasized that local adaptation was essential for patient comprehension and trust, and that answers without such contextualization risked feeling disconnected from the realities of Indian patients. At the same time, a few admitted that they sometimes skipped minor contextual corrections when the AI's response appeared "*good enough*," echoing the broader risk of automation bias.

Editing Strategies and Challenges When faced with major disagreements with the LLM-generated drafts, doctors adopted very different strategies in *Edit* and *Instruct*. In *Edit*, the common approach was to delete the entire draft and effectively revert to *Write*, composing a fresh answer from scratch. In contrast, in *Instruct*, doctors often either provided instructions to regenerate the response or rewrote the full answer within the instruction itself. As D9 explained: "*If it is a very big answer, and I don't agree with it (in Instruct), I need to write out the full answer as instructions.*" In *Instruct*, we observed two distinct strategies: issuing multiple short sequential instructions each fixing a single issue, or combining several fixes into one longer instruction. For example, one doctor combined several fixes into a single instruction: "Remove line 1. Add 'It is advisable to avoid cooking in the first week after cataract surgery.' Line 2: 'However, if required, cooking can be done, but it's important to be careful.' Remove line 3. Remove line 6."

With respect to smaller edits, doctors reported two main challenges in *Edit*: difficulty integrating new content seamlessly into the draft, and the need to repeat similar edits across multiple sections. For integration, doctors noted that adding a new sentence often disrupted the flow and required careful matching of grammar and style. As D2 explained:

> "*I don't like Edit... A newly added sentence needs to be of the same grammatical standard as the AI answer. It is hard to figure out where to fit this new sentence. It takes time and thinking.*" – D2.

In contrast, the same participant found *Instruct* far easier: "*Half the job is done. Adding anything here is much easier. It takes care of grammar, where to add the new points... framing, everything!*" – D2. A second issue was redundancy. Because similar phrasing could recur multiple times, doctors had to manually correct each instance. D5 described: "*In Edit, I might need to edit the first part of a sentence, but similar content may recur again in the next sentence, which I have to edit again. Such inconsistencies go away in Instruct automatically.*" In our manual review of *Edit* outputs, we found 9 answers (across 6 doctors) containing informal phrasing, punctuation errors, and grammatical mistakes–problems that were not present in *Instruct* due to its automated re-framing of edits.

Smartphone-Based Editing Constraints Although doctors were instructed to use a laptop/desktop for Phase 1–since the interface was not optimized for small screens–three doctors completed one session on their smartphones. As D5 explained: "*The phone is always with me, I prefer that. I rarely sit at a computer... In a real-world scenario, there is a higher probability that I use my phone for such tasks.*" All three doctors unanimously preferred *Instruct* over *Edit* on smartphones, stating that direct text edits were cumbersome. They noted that tasks such as correcting a spelling required "*moving the cursor and clicking before editing*" (D3), and that deleting long text meant "*holding down the back button for a long time*" (D4)—both difficult on a phone.

3.3 Role and Limitations of LLMs

LLM-Generated Answers Being Lengthy, Inaccurate, and Formal Doctors generally described the LLM-generated answers in *Edit* and *Instruct* as "*well-written*", "*complete*", and "*accurate*". As D2 remarked, "*AI didn't make any blunders. I edited a few timelines... but these are specific to each doctor,*" while D6 noted, "*Most of the (LLM) answers were correct. Only a few minor edits were required.*" At the same time, doctors consistently highlighted three limitations. First, answers were often unnecessarily long and verbose: what could be conveyed in one or two sentences was typically expanded into a multi-paragraph explanation, which doctors felt risked confusing patients. Second, the content occasionally contained factual inaccuracies or generic statements not aligned with standard cataract practice in India, e.g., instructions about fasting appropriate for general anesthesia but not for local anesthesia. Such errors required careful correction. Finally, the tone of the answers was described as overly formal and academic, often framed in textbook-like language. While doctors appreciated the grammatical polish, they emphasized that such formality could feel impersonal and disconnected from patient communication. In their edits, they sought to make responses more direct, conversational, and reassuring. As D8 summarized for *Edit*: "*The AI answer requires only minimal modifications, but I still change it to sound more natural for the patient.*"

Technical Errors in *Instruct* Manual analysis and focus group discussions revealed five instances where *Instruct* failed to execute editing instructions exactly as doctors intended. E.g., D2 noted that asking the system to "remove the third sentence" instead deleted the second, while D1 reported that "remove from eating onwards" removed two additional sentences. In such cases, doctors adopted two strategies: re-issuing the instruction more explicitly, or using recovery options such as the "Previous" button or typing "undo" as the next instruction. These worked because the system maintained conversation history. D5 recalled: "*I asked it to remove the second and fourth sentences. However, it only removed the second. So I had to ask again to remove the third, and it worked.*" These safeguards, along with the ability to see highlighted changes, helped doctors catch and correct errors quickly, though they remained cautious about how their instructions would be interpreted.

4 Discussion

Keeping Human Experts-in-the-Loop Our findings highlight the need to maintain a strong safeguard by keeping doctors firmly in the verification loop for all LLM-generated answers. While LLMs often produced drafts that were accurate and complete, occasional inaccuracies, omissions, or overly generic advice highlighted the risks of unchecked automation, consistent with prior work [3, 5, 20]. In the current medicolegal landscape, accountability for medical guidance ultimately rests with licensed practitioners, not AI systems. This makes human review indispensable, both for patient safety and for ensuring legitimacy of the information provided.

At the same time, there is a practical challenge of scalability. If every patient poses numerous questions, doctors cannot feasibly review all responses without significant burden. This opens a design and workforce question: who should validate LLM outputs at scale? One possibility is task-shifting, where trained non-doctor professionals (e.g., nurse educators, patient counselors) provide frontline validation of LLM answers, escalating only high-risk or ambiguous cases to physicians. Such tiered review mechanisms would align oversight intensity with the stakes of the advice, preserving safety without overwhelming experts. Their feasibility is evident in prior ethnographic work, where nurses in resource-constrained settings served as the first line of response to patient queries in digital chat groups [28], albeit without LLMs involved.

Policy frameworks will also need to clarify accountability and medicolegal responsibility for LLM-generated patient information. Clear guardrails on liability can both protect patients and provide confidence for clinicians to engage with these systems.

Balancing Standardization and Subjectivity A central tension in our findings lies between the standardization offered by LLMs and the subjective nuances contributed by individual doctors. On the positive side, AI-mediated standardization reduces variability in patient education—ensuring that every patient receives consistent and reliable core information. This consistency may reduce confusion and inequities in care that can arise from idiosyncratic differences across clinicians. Yet this consistency comes at a cost. The flip side of standardization is a loss of personalization. Patients may miss out on the individualized communication styles and nuanced emphases that human doctors naturally bring to their explanations. Over time, over-standardization risks making patient education less empathetic, less culturally attuned, and less responsive to individual concerns.

A further risk is overreliance. Because most LLM-generated answers were found to be "good enough," there is a temptation for clinicians to skip detailed review–a risk highlighted in prior work as well [5, 7, 27]. This echoes broader concerns in medical AI around automation bias: clinicians may gradually outsource judgment to AI systems, potentially eroding their clinical acumen and vigilance. Guardrails, such as requiring justification of edits, introducing periodic audits, or leveraging crowdsourced expert edits [19] may help mitigate this drift while preserving the efficiency of LLM-generated drafts. Design implications flow directly from these tensions [14]. Interfaces should foreground contextualization—making it easier for doctors to tailor generic LLM drafts to local cultural, clinical, and geographic realities. Systems should also embed transparency mechanisms, such as highlighting changes, to sustain clinician engagement and reduce the risk of blind acceptance. Finally, to preserve sustainability, tools should help minimize redundant edits and enable efficient review workflows, so that the efficiency gains of AI do not come at the cost of clinician burnout.

Design Implications *Qualities of Effective Answers.* Our findings highlight that useful answers share core qualities: they should be short, direct, and free of unnecessary complexity. Patients expect clarity and commitment rather than hedging or vague phrasing. For example, effective responses clearly state what is and is not allowed, often with timelines, instead of offering conditional or overly cautious language. This aligns with prior work on patient-centered communication [8, 16], which emphasizes plain language and actionable guidance. Embedding these principles into LLM prompting and fine-tuning could improve default outputs and reduce the need for rewording by doctors.

Improving LLM-Generated Outputs. While LLMs produced fluent drafts, they often lacked contextual grounding. Recommendations sometimes referred to procedures or practices not followed in India, used inappropriate currency or units, or generated verbose multi-paragraph explanations. To reduce such mismatches, prompts should explicitly incorporate geographic and institutional context (similar to [19]), and default outputs should be optimized for brevity and clarity. A design strategy could be to generate short, two-sentence responses by default, with optional expansion for additional detail. Such changes would reduce editing burden and help ensure that answers remain both relevant and comprehensible to patients.

Supporting Hybrid Editing Workflows. Editing strategies differed across conditions: *Edit* was efficient for small corrections but cumbersome for major rewrites, while *Instruct* handled grammar and flow well but was less effective for deletions or large changes. To accommodate this diversity, future systems should support hybrid workflows. For example, *Write* could be paired with LLM-based grammar refinement to produce polished drafts from doctor-authored text. *Edit* could integrate options for deleting large sections or starting fresh, while *Instruct* should include a transparent undo button to recover from misinterpretations. Designing for fluid transitions between modes can reduce cognitive load and support seamless editing, consistent with principles of mixed-initiative interaction [2, 13] and recent work on AI-assisted writing [21].

Voice Input for Mobile Use. Because many clinicians accessed the system via smartphones, typing was described as slow and screen-cluttering. Voice input emerged as a natural modality, particularly for *Write* and *Instruct*, where answers are short or instructions conversational. However, voice is less compatible with cursor-based direct editing in *Edit*. A promising direction is to combine voice dictation for drafting with lightweight correction tools for transcription errors. This approach builds on prior HCI findings [19, 20] that multimodal input, especially voice, can reduce physical effort and improve efficiency in mobile, resource-constrained settings.

Personalizing to Patient Values. Finally, answers should adapt not only to clinical accuracy but also to patient needs and preferences. Older patients may benefit from concise, directive responses, while more educated or information-seeking patients may prefer detailed explanations. Similarly, some patients require reassurance and empathy, while others prioritize brevity. This calls for adaptive answer generation that adjusts length, tone, and level of detail based on patient demographics and values. Prior work [8, 15] in health communication demonstrates that such tailoring improves comprehension, trust, and adherence, suggesting a valuable direction for LLM-assisted systems.

Limitations Although we made specific attempts to reduce the skew in our participant cohort, we used convenience sampling in our recruitment, which may introduce selection bias and affect the representativeness of the sample. Further, we worked with a population in urban India, and views of chatbots and their advice may be different in rural or periurban contexts. Finally, while doctors were central evaluators, patients–the ultimate consumers of these answers–were not included, and their perspectives remain an important direction for future work.

Conclusion This paper examined how doctors answered patient queries using three approaches: writing from scratch, directly editing LLM drafts, and instruction-based indirect editing. While LLMs generated accurate and polished responses, doctors' edits highlighted essential needs: contextualizing content to local practices, safeguarding against factual errors, and reframing verbose responses into clear, patient-centered communication. Our findings emphasize the importance of keeping human experts in the loop, not only for safety, but also to sustain personalization and empathy in patient education. This work contributes empirical evidence on human–AI co-authoring in high-stakes settings and identifies design opportunities. Ultimately, safe and scalable adoption of LLMs in healthcare will depend on balancing standardization with subjectivity, and automation with meaningful human oversight.

References

[1] Majid Afshar, Yanjun Gao, Graham Wills, Jason Wang, Matthew M Churpek, Christa J Westenberger, David T Kunstman, Joel E Gordon, Cherodeep Goswami, Frank J Liao, and Brian Patterson. Prompt engineering with a large language model to assist providers in responding to patient inquiries: a real-time implementation in the electronic health record. *JAMIA Open*, 7(3):ooae080, August 2024.

[2] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Mag.*, 35(4):105–120, December 2014.

[3] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. Ai chatbots not yet ready for clinical use. *Frontiers in Digital Health*, 5:60, 2023.

[4] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and Davey M Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*, 183(6):589–596, June 2023.

[5] Joshua M. Biro, Jessica L. Handley, J. Malcolm McCurry, Adam Visconti, Jeffrey Weinfeld, J. Gregory Trafton, and Raj M. Ratwani. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *npj Digital Medicine*, 8(1):222, Apr 2025.

[6] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[7] Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6(6):e379–e381, Jun 2024.

[8] Zhanming Chen, Alisha Ghaju, May Hang, Juan Fernando Maestre, and Ji Youn Shin. Designing health technologies for immigrant communities: Exploring healthcare providers' communication strategies with patients. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[9] Martina A Clarke, Joi L Moore, Linsey M Steege, Richelle J Koopman, Jeffery L Belden, Shannon M Canfield, Susan E Meadows, Susan G Elliott, and Min Soon Kim. Health information needs, sources, and barriers of primary care patients to achieve patient-centered care: A literature review. *Health Informatics Journal*, 22(4):992–1016, 2016. PMID: 26377952.

[10] Patricia Garcia, Stephen P. Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, Carlene Lugtu, Matthew Rojo, Steven Lin, Tait Shanafelt, Michael A. Pfeffer, and Christopher Sharp. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Network Open*, 7(3):e243201–e243201, 03 2024.

[11] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, USA, 1988.

[12] A Jay Holmgren, N Lance Downing, Mitchell Tang, Christopher Sharp, Christopher Longhurst, and Robert S Huckman. Assessing the impact of the covid-19 pandemic on clinician ambulatory electronic health record use. *Journal of the American Medical Informatics Association*, 29(3):453–460, 12 2021.

[13] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 159–166, New York, NY, USA, 1999. Association for Computing Machinery.

[14] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 895–906, New York, NY, USA, 2018. ACM.

[15] Rutuja Joshi, Yu-Jou Lee, and Klaus Bengler. User preferences in conversational ai for healthcare: Insights from an interview study. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, CUI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[16] Dingdong Liu, Yujing Zhang, Bolin Zhao, Shuai Ma, Chuhan Shi, and Xiaojuan Ma. Scaffolded turns and logical conversations: Designing humanized llm-powered conversational agents for hospital admission interviews. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[17] Charles NJ McGhee, Jie Zhang, and Dipika V Patel. A perspective of contemporary cataract surgery: the most common surgical procedure in the world. *Journal of the Royal Society of New Zealand*, 50(2):245–262, 2020.

[18] Frederick North, Kristine E Luhman, Eric A Mallmann, Toby J Mallmann, Sidna M Tulledge-Scheitel, Emily J North, and Jennifer L Pecina. A retrospective analysis of Provider-to-Patient secure messages: How much are they increasing, who is doing the work, and is the work happening after hours? *JMIR Med Inform*, 8(7):e16521, July 2020.

[19] Pragnya Ramjee, Mehak Chhokar, Bhuvan Sachdeva, Mahendra Meena, Hamid Abdullah, Aditya Vashistha, Ruchit Nagar, and Mohit Jain. Ashabot: An llm-powered chatbot to support the informational needs of community health workers. In *Proc of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. ACM.

[20] Pragnya Ramjee, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. Cataractbot: An llm-powered expert-in-the-loop chatbot for cataract patients. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(2), June 2025.

[21] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. Abscribe: Rapid exploration & organization of multiple writing variations in human-ai co-writing tasks using large language models. In *Proc of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. ACM.

[22] Bhuvan Sachdeva, Pragnya Ramjee, Rahul Sharma, Mithun Thulasidas, Geeta Fulari, Kaushik Murali, and Mohit Jain. Utility of an llm-powered experts-in-the-loop chatbot for pre- and post-operative care of cataract surgery patients. *European Journal of Ophthalmology*, 2025.

[23] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, Mar 2025.

[24] M A Stewart. Effective physician-patient communication and health outcomes: a review. *CMAJ*, 152(9):1423–1433, May 1995.

[25] Ming Tai-Seale, Sally L. Baxter, Florin Vaida, Amanda Walker, Amy M. Sitapati, Chad Osborne, Joseph Diaz, Nimit Desai, Sophie Webb, Gregory Polston, Teresa Helsten, Erin Gross, Jessica Thackaberry, Ammar Mandvi, Dustin Lillie, Steve Li, Geneen Gin, Suraj Achar, Heather Hofflich, Christopher Sharp, Marlene Millen, and Christopher A. Longhurst. Ai-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Network Open*, 7(4):e246565–e246565, 04 2024.

[26] Polly Thompson, Richard Thornton, and Conor M. Ramsden. Assessing chatbots ability to produce leaflets on cataract surgery: Bing ai, chatgpt 3.5, chatgpt 4o, chatsonic, google bard, perplexity, and pi. *Journal of Cataract & Refractive Surgery*, 51(5):371–375, 2025.

[27] Arpita Wadhwa, Aditya Vashistha, and Mohit Jain. Designing with culture: How social norms shape trust and preference in health chatbots, 2025.

[28] Ding Wang, Santosh D. Kale, and Jacki O'Neill. Please call the specialism: Using wechat to support patient care in china. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. ACM.

[29] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(2), May 2024.

[30] Jördis M Zill, Isabelle Scholl, Martin Härter, and Jörg Dirmaier. Which dimensions of patient-centeredness matter?-results of a web-based expert delphi survey. *PloS one*, 10(11):e0141978, 2015.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction reflect the scope and contributions of our research.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Section 4.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not produce theory.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see Section 2.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper does not include experiments requiring code.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: Our experiment doesn't involve training an AI model and hence no data split or hyperparameter tuning reported.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are accompanied by statistical significance tests and associated explanations.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Our experiment doesn't involve training an AI model.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere completely to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Section 4.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our experiment doesn't involve training an AI model.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see Section 2.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our experiment doesn't involve new AI model/data assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Please see Section 2 and Figure 1.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: IRB approval was obtained, and risk are highlighted and mitigated, please see Section 2.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Please see Section 2.