When Experts and Students Disagree: Divergent Perceptions of Bias Mitigation Strategies In a Stroke Prognosis Support Chatbot

Daniel Hansen
ddbh19@student.aau.dk

Arlonsompoon P. Lind
alind19@student.aau.dk

ABSTRACT

This study investigates the effectiveness of cognitive bias mitigation strategies in an AI-powered CDSS, focusing on reducing premature closure bias among clinicians during prognosis. We designed and evaluated two experimental conditions: one employing a single strategy "hear the story first," which prompts users to review patient data before receiving AI recommendations, and another condition which adds with "consider the opposite," which encourages reflection on alternative prognoses. Our mixed methods evaluation involved 10 medialogy students and 3 practising clinicians. Quantitative results from the TLX and CUQ revealed that students perceived the single-mitigation chatbot as more usable (CUQ score: 69.1 vs. 59.8), with no significant differences in workload. Qualitative feedback showed that students often overlooked mitigation prompts, while clinicians strongly favoured "consider the opposite" for its role in fostering critical reflection and transparency. Notably, clinicians dismissed "hear the story first" as redundant, highlighting a divergence between expert and non expert user needs. The findings underscore the importance of tailoring bias mitigation strategies to the target audience: passive prompts may go unnoticed by non experts, whereas clinicians value active challenges to their reasoning. The study also demonstrates the risks of over reliance on non expert feedback during design, as clinician insights fundamentally reshaped our understanding of effective AI support. Future work should explore standalone implementations of "consider the opposite".

INTRODUCTION

For several decades Conversational agents (CA's) meant to improve healthcare have been developed. "Eliza"[22] was one of the first examples of this, built to respond roughly as a psychotherapist would, and other early natural language processing have been developed for healthcare use since then [5, 20].

In recent years, advancements in artificial intelligence, natural language processing and deep learning, have created renewed interest in natural language user interfaces, such as chatbots. These advancements are not without their challenges, including the issues of hallucinations and inherent biases in communication [3]. In healthcare, AI-powered chatbots are used to address patients' medical inquiries and assessing symptoms based on the patient's own descriptions. These technologies enhance the accessibility of healthcare services, particularly for individuals in remote areas, or during times with physician shortages [1].

The integrations of AI in clinical decision support systems have potential to enhance medical decision making and improve patient outcomes. Despite this, clinicians often resist AI-driven recommendations due to concerns about trust, autonomy, and transparency[18], as well as the aforementioned challenges concerning AI-hallucinations.

In this paper we will design a chatbot meant to support clinicians during prognosis of stroke patients, and evaluate how the use of bias mitigation strategies [2] could be used to address the concerns of autonomy that cause resistance to trusting AI-driven recommendations, while encouraging deeper thought into prognostic decision making.

RELATED WORK

Chatbots have evolved significantly, from early rule-based systems like ELIZA to today's advanced AI-powered models capable of generating human like responses. While early chatbots relied on predefined rules and scripted dialogue trees, modern systems harness deep learning and large-scale language models to enable more dynamic and context-aware interactions. These AI-driven chatbots support users in various ways, often by generating content that humans then review and refine for quality, or by reducing cognitive load through the automation of time-consuming tasks. Such tasks are seen across multiple domains, like education, business and healthcare. Healthcare chatbots following the same idea as Eliza are also still in use[8, 16], by engaging users in cognitive behavioural therapy exercises, for stress and anxiety. [12, 3]. The adoption of AI in clinical decision making has the potential to improve healthcare outcomes, and are often seen as objective and reliable. However, the black box nature of these systems remain a barrier, which prevents clinicians from understanding and trusting their recommendations.[18, 2]

According to Panigutti et al.[9], cognitive bias can be defined as *"the class of effects through which an individual's preexisting beliefs, expectations, motives and situational, context influence the collection, perception, and interpretation of information"*. More simply put, cognitive bias is how previous experience, personality and situational context affects decision making. Ko and Glusac[9], outline several different types of cognitive bias, among them is "Premature Closure Bias", defined as closing the decision making process too soon.

Medical decision making is inherently uncertain and cognitive biases contribute significantly to prognostic errors, affecting patient care at all stages. These biases are mental shortcuts that can lead to errors in information gathering and diagnosis.

Despite the widespread recognition of these biases, reducing them remains a challenge[15]. The cognitive error of premature closure in clinical decision making occurs when a clinician settles on an initial diagnosis without sufficiently considering other plausible alternatives, this phenomenon is a common cause of delayed diagnosis and misdiagnosis[11].

Research indicates that the risk of premature closure can increase with years of experience among clinicians. A study by Eva et al.[6] found that more experienced physicians tend to weigh their first impressions more heavily, which can lead to a higher likelihood of prematurely closing their diagnostic search[6]. This tendency highlights the importance of understanding cognitive biases in clinical practice and suggests that educational strategies should be tailored to address these experiences specific cognitive tendencies.
One method to mitigate the risk of premature closure is the use metacognitive strategies, strategies to monitor ones own thinking, such as a checklist, which encourages the clinician to reflect on their thought process and consider alternative diagnoses[4].

A study by Bach et al.[2] propose and explore several bias mitigation strategies in a AI support system, and highlight previous work that show the importance of mitigating biases in decision support tools. They present:

- **Hear the story first**: Encouraging clinicians to formulate their own diagnosis before seeing AI recommendations.
- **Decision justification**: Prompting clinicians to explain their reasoning before receiving AI generated outputs.
- **Consider the opposite**: Encouraging users to reflect on alternative explanations when their assessment differs from the AI's suggestion.

Participants in bach et al. [2] with a prototype in a simulated workflow and provided feedback through structured interviews. The findings revealed varied levels of trust in AI decision support. Some clinicians appreciated the system's ability to highlight abnormalities, while others expressed concern about false positives and negatives. Green labeled images (indicating no abnormalities) often led to quicker decision making, but also introduced overconfidence clinicians sometimes deferred to AI even when aware of its limitations.

- Diagnosis-before-AI encouraged independent reasoning but reduced efficiency.
- Justifying decisions promoted reflection but was sometimes seen as redundant.
- Reconsideration prompts (when AI and clinician diagnoses conflicted) helped identify overlooked issues but risked positioning the AI as a fact-checking authority.

Bach et al.[2] emphasize a trade off between enhancing diagnostic accuracy and maintaining workflow efficiency. Bias mitigation strategies encouraged more thoughtful decision making but also added cognitive load, raising concerns about their practicality in time pressured environments. Clinicians recognized the potential benefits of these strategies but questioned their feasibility during routine practice.

DESIGN

The design of this prototype centers on selecting appropriate mitigation strategies to address premature closure bias. Based on the list of mitigation methods presented by Bach et al.[2], we excluded Decision Justification as it requires additional time and effort from the user, which according to Bach et al. is exactly what clinicians want to avoid.

To evaluate the effectiveness of cognitive bias mitigation strategies in AI-assisted decision making, we implemented two experimental conditions. The first condition applies a single mitigation method, while the second combines two methods. This setup allows us to investigate not only the individual impact of a single strategy, but also whether a combination of strategies produces a greater or different effect on Mental workload.Comparing a single strategy condition with a dual strategy condition allows us to evaluate whether combining them provides an additive benefit

Hear the story first

We selected "hear the story first" as the single mitigation strategy because it aligns with prior research indicating that prompting users to consider contextual information before forming a judgment can reduce anchoring and premature closure biases[2]. This approach encourages users to engage more deeply with the narrative before being exposed to AI generated suggestions, potentially improving prognostic reasoning, see Figure 1.

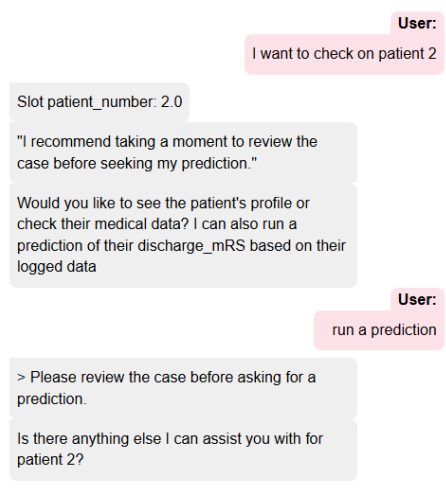

Figure 1: Screenshot of chatbot telling user to review the data first

Consider the opposite

The second condition additionally employs the "consider the opposite" strategy. This approach explicitly prompts users to reflect on and challenge their assumptions, encouraging them to actively consider alternative possibilities, see Figure 2. Research has shown that this method can reduce confirmation bias by disrupting intuitive reasoning patterns and fostering more deliberate analysis[19]. We hypothesize that incorporating this strategy will promote reflective thinking and help users avoid premature closure bias.

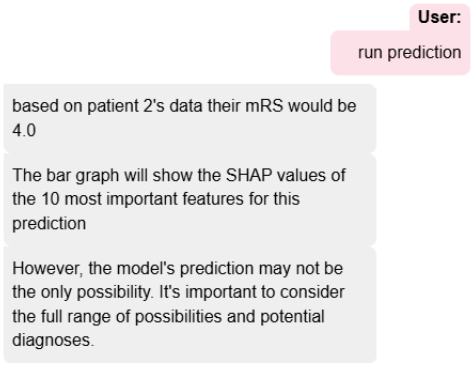

Figure 2: Screenshot of chatbot reminding the user to consider the full scale of possibilities

Hypothesis

We wish to evaluate on the use of multiple bias mitigation strategies, to promote deeper thought into decision making, without over reliance on the chatbots output. We therefore present the following hypothesis:

- **Null Hypothesis:** Increasing the number of bias mitigation strategies implemented in the chatbot does not affect the amount of thought users put into their answers, reflected in a workload that is not significantly different.
- **Alternative Hypothesis:** Increasing the number of bias mitigation strategies implemented in the chatbot leads users to put more thought into their answers, reflected in a higher workload.

Prediction model - LightGBM Accuracy

Label	Precision	Recall	F1-score	Support
0	0.68	0.73	0.71	365
1	0.54	0.59	0.56	447
2	0.38	0.29	0.33	283
3	0.50	0.49	0.50	216
4	0.81	0.83	0.82	310
5	0.58	0.54	0.56	160
6	0.81	0.84	0.83	122
Accuracy	0.61 (1903 samples)			
Macro avg	0.61	0.62	0.61	1903
Weighted avg	0.60	0.61	0.61	1903

Table 1: Classification Report from LightGBM model. Label = discharge_mrs classes

The LightGBM model achieved an overall accuracy of 61% on 1,903 samples, with performance varying across the seven discharge_mrs classes. Classes 4 and 6 showed the strongest results, with F1 scores of 0.82 and 0.83 respectively, reflecting high precision and recall. In contrast, class 2 was the most challenging to predict accurately, evidenced by the lowest F1 score of 0.33. The macro-averaged metrics hovered around 0.61, indicating consistent but moderate performance across all classes. Cross validation accuracy was higher at 73%

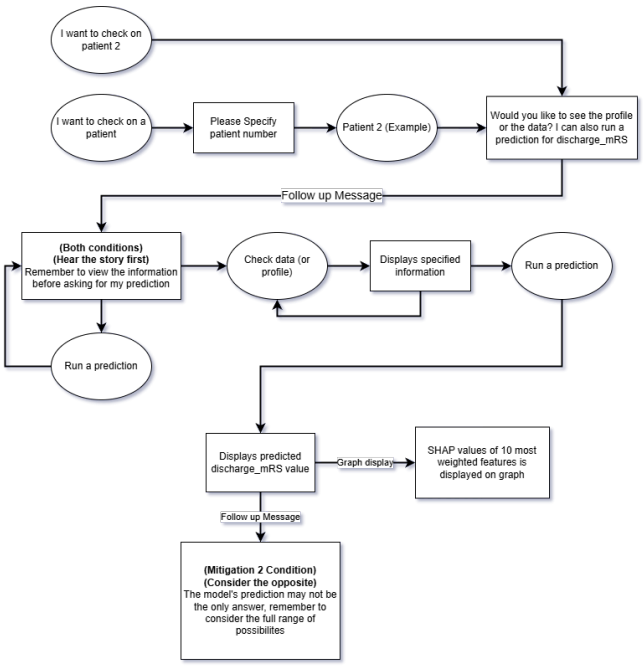

Figure 3: Abstracted flowchart of conversation with chatbot, showing both mitigation strategies

EVALUATION PROCEDURE

After a test participant signed the consent form, screen and audio recording was started and the user was given a piece of paper explaining the test scenario and their objective, see Figure 4.

Every participant completed both evaluation conditions, with every other participant testing the conditions in reverse order to counterbalance order effects.

During testing, if participants had trouble getting the bot to perform an action, facilitators would wait a few attempts before stepping in to help.

Usually facilitator interference would be limited to giving advice on how to reword user input to make the bot respond correctly, but in the event of a complete malfunction, a facilitator would take control of the computer to reset things, and then write messages to get the participant back to where they were in the task prior to the malfunction. If the bot gave incorrect responses, such as inaccurate explanations of feature names, a facilitator would step in a give a correct explanation.

After each condition, the participant filled out a NASA Task Load Index (TLX)[7] and Chatbot Usability Questionnaire (CUQ)[10] according to their experience with the condition. At the end of the evaluation a short interview was conducted about the participants experience with the prototype and how the two conditions compared in their opinion.

You are a clinician who needs to evaluate a patient who has suffered a stroke. Based on their previous history and the data that has been logged about the patient since the onset of the stroke.

The goal is for you to read and understand the information about three patients and choose whether you think the stroke they have suffered is **mild**, **moderate** or **severe**

Just assign one of those three options to each of the patients.

If there are any terms you want explained you can try to ask the bot to explain them to you.
In the event that the bot gives an incorrect explanation, an evaluator will step in and give a correct explanation.

The bot can provide you with information about the patients if you ask about it.
To start off the test, write that you want to check on a patient.

Figure 4: Text description of evaluation scenario given to participant

The evaluation was performed on 10 Medialogy students, in addition our supervisor evaluated on 3 practising neurologist at the European Stroke Organisation Conference (ESOC) in Helsinki, who gave their feedback on the prototype and answered the same questionaries and interview questions.

RESULTS

CUQ Likert-scale responses were converted to numeric values (1–5) for analysis. Normality of each TLX subscale and CUQ score was assessed using the Shapiro-Wilk test. The Difference in scores between conditions were also tested for normality. If difference scores were normally distributed ($p > 0.05$), a paired t-test was used [14, 17]. If not ($p <= 0.05$), the Wilcoxon signed-rank test was applied as a non-parametric alternative[23, 17]

Quantitative Results

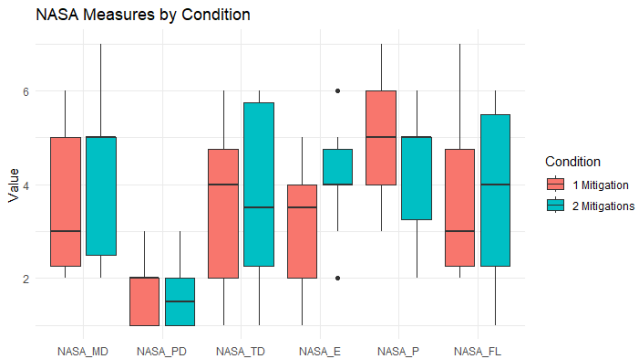

Figure 5: Boxplots for NASA TLX subscales for both conditions

Figure 5 illustrates the distribution of student participant responses across TLX subscales under two conditions: 1 Mitigation (red) and 2 Mitigations (blue). While the overall distributions are similar across conditions, there are slight variations, such as increased Temporal Demand (NASA_TD), Frustration (NASA_FL) and Effort (E) in the 2 Mitigations condition. However, no NASA subscale showed statistically significant differences between conditions as seen in table 2.

Figure 6: Histogram of CUQ scores for Mitigation 1 and Mitigation 2.

The over all CUQ score for mitigation 1, across the 10 student participants was 69.1 (SD = 8.1) indicating a moderately high usability for condition 1. And the CUQ score mean for mitigation 2 was 59.8 (SD = 9.7) indicating a moderate usability. Most participants reported higher usability scores for the chatbot in the Mitigation 1 condition. This trend suggests that the chatbot employing only the first mitigation strategy was perceived as more usable than the version implementing both mitigation strategies. These differences are reflected in the statistical tests, which found significant differences in CUQ scores favouring Mitigation 1 also seen in table 2

Q	1/2	1F/2F	1F/1S	2F/2S	1S/2S
MD	0.2785	1.0000	0.3375	0.8662	0.3262
PD	0.7728	1.0000	0.7489	1.0000	1.0000
TD	0.4679	0.4766	0.7780	0.7632	0.7780
E	0.0947	0.2355	0.5415	0.1419	0.3046
P	0.1727	0.4676	0.8501	1.0000	0.2943
FL	1.0000	0.2355	0.4130	0.0516	0.3046
CUQ	**0.0170**	0.2983	0.3372	0.9594	**0.0252**

Table 2: Paired t-test or Wilcoxon Signed-rank test p-values for TLX Subscales and CUQ scores across conditions. 1 = Condition 1, 2 = Condition 2, F = First, S = Second.

Table 2 presents the p-values from paired t-tests or Wilcoxon signed-rank tests comparing TLX subscales and CUQ scores across different condition sequences. Significant differences were observed only in two comparisons, both related to CUQ scores. Participants rated mitigation 1 as more usable (**p = 0.017**), meaning they found the chatbot more usable when it had only one mitigation strategy. Second, a significant difference was found in the 1S/2S sequence (**p = 0.025**), suggesting participants preferred experiencing Mitigation 1 in the second session rather than Mitigation 2.

Student NASA TLX & CUQ scores

Participant	CUQ 1	CUQ 2	TLX 1	TLX 2
P1	78.1	67.2	40	48
P2	67.2	59.4	38	29
P3	54.7	42.2	38	55
P4	76.6	54.7	42	67
P5	67.2	65.6	54	64
P6	76.6	71.9	38	26
P7	76.6	53.1	35	62
P8	73.4	64.1	42	48
P9	57.8	70.3	57	43
P10	73.4	50.0	44	60
Mean ± SD	69.1 ± 8.1	59.8 ± 9.7	44 ± 7.7	50 ± 14.2

Table 3: Combined Student CUQ and TLX for Mitigations 1 and 2.

Table 3 shows CUQ and TLX scores for 10 student participants across both conditions. CUQ scores range from 42.2 to 78.8. On average, CUQ scores were slightly higher in mitigation 1 (mean = 69.1, SD = 8.1) compared to mitigation 2 (mean = 59.8, SD = 9.7). The TLX values represent overall workload as percentages, calculated by summing the six unweighted subscale ratings (each from 1 to 7), dividing by the maximum possible score (42), and multiplying by 100. TLX scores in Mitigation 1 ranged from 35 to 57, while in Mitigation 2 they ranged from 26 to 67. The average workload was lower in Mitigation 1 (Mean = 44, SD = 7.7) than in Mitigation 2 (Mean = 50, SD = 14.2), suggesting that participants perceived the second mitigation condition as more mentally demanding or effortful.

Clinicians NASA TLX & CUQ scores

Participant	CUQ 1	CUQ 2	TLX 1	TLX 2
C1	53.1	65.6	30.8	35.0
C2	50.0	48.4	34.2	31.7
C3	50.0	65.6	27.5	31.7
Mean ± SD	51.0 ± 1.8	59.9 ± 9.9	30.8 ± 3.4	32.8 ± 1.9

Table 4: Combined clinician CUQ and TLX Scores for Mitigations 1 and 2.

Table 4 presents the CUQ and TLX scores for each clinician across the two mitigation conditions. CUQ scores increased for two of the three when using the chatbot with mitigation 1, indicating a possible improvement in perceived usability. Similarly, TLX scores which reflect perceived workload, slightly increased on average, suggesting a marginal increase in perceived effort when adding the 'consider the opposite' mitigation method. No statistical tests were performed due to the small sample size (n = 3).

Participant messages and time elapsed per task

C	M	MMean	TE	TE Mean
1/2	348	17.4	03:58:25	00:11:55
F	198	19.8	02:31:54	00:15:11
S	150	15.0	01:26:31	00:08:39
1F	91	18.2	01:14:15	00:14:51
1S	70	14.0	00:43:23	00:08:41
2F	107	21.4	01:23:52	00:16:46
2S	80	16.0	00:43:08	00:08:38

Table 5: Summary of messages and elapsed time across conditions. M = Messages, TE = Time Elapsen. 1 = Condition 1, 2 = Condition 2, F = First, S = Second.

Participants in the Mitigation 1 & 2 condition exchanged a total of 348 messages (mean = 17.4), with an average interaction time of 11 minutes and 55 seconds, indicating moderate engagement under the dual mitigation strategy. In the first round, interaction was higher (mean messages = 19.8; mean time = 15:11) compared to the second round (mean = 15.0; mean time = 8:39), suggesting a decline in engagement over time.

The double mitigation first condition had the highest average message count (mean = 21.4) and longest interaction time (16:46), while the single mitigation second condition had the lowest (mean = 14.0; time = 8:41). These results suggest that both the type and timing of mitigation strategies influenced interaction levels, with earlier exposure to dual strategies driving more sustained engagement.

Qualitative Results

Qualitative data was gathered through interviews at the end of the evaluation, the interviews were transcribed and thematically analysed from the audio recordings of the evaluation.

Thematic Analysis of students

When asked if they noticed the difference between the two conditions 6 out of 10 participants stated they did not notice a difference.

Two participants believed they had identified a difference, but the aspects they mentioned were actually the same to both conditions.

- *"Felt exactly the same."*(P1)
- *"I couldn't spot what the difference was between the two."*(P6)

1 participant, correctly identified and described the difference.

- *"*[Mitigation 2] *egged me on to make my own conclusions, when I asked it to make a prediction."*(P5)

1 participant described mitigation 2 as proactive and robotic, reasoning that the bot gave him more instruction, this seemed to be about the extra "consider the opposite" message, implying this participant did notice the difference, and interpreted it as one bot being more proactive than the other.

In contrast another participant, described mitigation 2 as more human-like.

- *"*[I preferred] *the one that wasn't proactive, I don't know, it felt more robotic"*
(P8, when asked about their preferred condition).

- *"*[Mitigation 2] *was more descriptive and in a way more human-like in how it explained things."*
(P10, when asked about the difference between the conditions)

7 out of 10 participants listed mitigation 1 as their preferred version, though one of the seven added that he imagined mitigation 2 would be better for a medical professional.

- *"*[Mitigation 1] *because I am inexperienced in the field, but I think [Mitigation 2] would be better if I was a clinician"*(P5)

When asked if the bot's prediction made them rethink their initial prognosis of the patient, 9 participants said yes. 6 of the 9 also mentioned that they hadn't formed a prognosis of the patient before getting the prediction.

- *"My choice was completely taken from it's prediction"*(P3)

- *"I trusted it blindly" - "I didn't really form my own opinion"*(P4)

- *"Before the robot gave me a prediction, I didn't really have an idea of how* [the patients] *were doing"*(P7)

4 of these 9 reasoned that they thought the bot knew better than them.

- *"It is outside my field."*(P1)

- *"It is hard for me to say, when I do not have a medical background."*(P9)

Thematic Analysis of clinicians

When asked about their opinions about the two mitigation strategies, all 3 clinicians preferred "Consider the opposite" over "hear the story first".
All 3 explained that they would already have access to the patient's information from other sources (e.g. scans, lab results and clinical notes, according to Clinician 2), so being reminded by the bot to check the information was not useful.

- *"Usually I collect the data myself"- -"So when it asked me to stop and review again, it felt like… not needed."*
(Clinician 1)

- *"When I use a tool like this, I've already seen the scans, lab results, the clinical notes. I don't need the bot to tell me to look again."* (Clinician 2)

- *"'Hear the story first' wasn't particularly helpful for me, just because I already make it a habit to review the full picture before deciding anything."* (Clinician 3)

All three clinicians liked that the bot made them stop and double check themselves, even though none of them changed their answers based on the prediction. Clinicians 2 and 3 both said that the extra check made them more confident in their choice.

- *"It felt more like the bot was helping me think better, not just guiding me."* (Clinician 2)

- *"It's easy to get anchored on a score too early, especially when we're tired or busy. Having the bot say, "Hold on are you sure?" feels more like a colleague giving you a quick tap on the shoulder."* (Clinician 3)

Clinician 1 explained that consider the opposite helped, the bot feel more transparent.

- *"I think it helps make the bot feel more open, more… transparent? Like it's not just saying "here is the answer," but asking you to check. I also think that these bots are not perfect so it is good that it can show it's own insecurity"*
(Clinician 1)

Clinicians 1 and 2, believed a system like this could help junior clinicians.

- *"This kind of help* [consider the opposite]*—it shows what to ask yourself. Like a reminder to look again. I think for residents* [junior clinicians]*, this can be very useful."*
(Clinician 1)

- *"I think for junior doctors, this would be very good. They don't always know what to challenge in their own decision-making."* (Clinician 2)

DISCUSSION

Participant Perception and Preference

Even though most participants could not identify the difference between the two conditions, when asked about which one they preferred, the majority listed mitigation 1. This also fits with the quantitative data, where mitigation 1 scored higher in usability with 9 out of 10 participants, see Figure 6.

3 participants who started with the mitigation 1 preferred it over mitigation 2, and 4 of those who started with mitigation 2 also listed mitigation 1 as their favourite.

This preference is further supported by the CUQ item scores, where mitigation 1 scored consistently higher on statements related to ease of navigation, system clarity, and overall intuitiveness. Additionally, TLX responses showed that participants rated mitigation 1 as less mentally demanding and less frustrating, reinforcing the perception of smoother interaction with fewer cognitive barriers.

It should be mentioned that familiarity with the bot might have had an effect on how difficult the second round felt and thus affected which condition was considered the favourite.

Impact and Visibility of Mitigation strategies

Given that six participants admitted to not actually forming a prognosis before asking for a prediction, it seems that the "hear the story first" mitigation strategy often went unnoticed, despite generating a prompt every time a patient is selected.

These findings suggest that passive mitigation strategies, even when repeated, can be insufficient if not perceived as essential to the task. Despite the intention to foster deeper engagement, mitigation 2 was not salient enough to consistently alter participants' behaviour or decision making. This is reflected in CUQ items related to perceived usefulness and relevance of responses, which were not notably higher in mitigation 2.

Experience Effects and Learning Curve

The additional data of messages sent and time per patient also seem to be affected by experience with the bot, as both fall quite drastically in the second round, regardless of condition order.

Participants always required some time to get familiar with the bot during their first round. The extra mitigation strategy in mitigation 2 does not seem to have made this learning period faster. The average amount of time for the first round was 1 minute and 55 seconds longer for those who started with mitigation 2 than for those who started with mitigation 1. This could support the idea that the extra mitigation strategy made them put more thought into what they were doing, making the prognosis take more time. However it seems this effect only lasts for their initial experience with the bot

These results align with TLX ratings showing that mental demand and effort scores were higher during the first round, particularly for mitigation 2. This suggests that the additional prompts may have increased complexity or cognitive load without providing a meaningful usability benefit for the students.

Knowledge Transfer Across Rounds

Regardless of condition order, by the second round, most questions about explaining terms or the meaning of data values had already been asked in the first round. With 4 out of 10 participants not asking any additional questions—beyond viewing patient information and predicting `discharge_mRS`, this indicates a strong transfer of task knowledge and system familiarity across rounds.

This plateau in interaction suggests that the perceived need for chatbot assistance diminished over time, possibly due to users internalizing task structure or recognizing the limits of the system's support. Interestingly, this effect occurred regardless of the mitigation strategy, reinforcing that user adaptation may outweigh mitigation-based influence over time.

General public vs Experts

The answers we got from the clinicians differed from the Medialogy students.

7 students preferred mitigation 1, the one without "consider the opposite", while all 3 clinicians found "hear the story first" to not be useful, but liked how they were encouraged to double check themselves by consider the opposite. This is due to the clinician already knowing everything related to the stroke data and history about their respective patients. 1 clinician specifying that the prediction from the trained model added confidence to their choice. Potentially adding a new role, a validation role, to Li et al. [12] as this were not any of the proposed chatbot roles.

Clinician 3 specifically mentioned 'Consider the opposite' was a great method to mitigate them being tired or busy, or wanting to choose treatment early, which is entirely premature closure. Opposed to this, are the student participants, who did not like the 2 strategy mitigation method, likely due to wanting to complete their task quickly.

6 students also did not notice the difference between the two conditions and 2 students described differences that weren't there.
Differing from the Clinicians who all noticed the difference immediately while they were performing the task.
The students tended to "filter" the chatbot's responses as they read them, for example when they asked for a prediction, they generally did not read the follow-up text after seeing the answer they were looking for.
The clinicians however read the responses thoroughly, which helped them notice the difference immediately.

As mentioned earlier, it is likely the students were trying the finish the task quickly, where the clinicians had a stronger interest in really testing the prototype, because results from the test could contribute to software solutions in their field. A survey by Longo et al.[13] hypothesises with their novel definition of mental workload, that when a user has an interest in the task before them they put in more effort when performing it. The difference between the students and clinicians' data in our evaluation would support this hypothesis.

These differences in their approach to the evaluation, led to different results being recorded from their evaluation.
Had we not had the data and feedback from the ESOC clinicians, it is likely we would have concluded "consider the opposite" to possibly be the less appropriate strategy of the ones we tested, since the condition where it was excluded was generally better received.
This was challenged when the clinicians showed positive interest the "Consider the opposite" strategy and a generally negative interest in "Hear the story first".
A study by Wang et al.[21] found that there were clear differences between the evaluation results they got from university student participants, and crowdsourced participants from Amazon mTurk.
When building for expert users, it seems that testing on crowd workers or users from outside the target group can give results that do not fit the actual context where the data would be used to build further. Our study seems to support this, showing that even college students who also possess high expertise, can not reliably substitute experts of the field the prototype is supposed to be used in, when testing.

NASA TLX

Due to a design oversight, the TLX subscales were administered using a 1–7 Likert scale instead of the standard 0–100 scale or the weighted version typically employed in workload assessment. To enable comparability and maintain interpretability, we calculated total scores across the six subscales and expressed them as a percentage of the maximum possible score (i.e., 42), effectively rescaling responses to a 0–100 range. While this approach preserves relative differences between participants and conditions, it does not capture the full granularity of the TLX methodology. As such, results should be interpreted with caution, and future studies should employ the validated TLX format to ensure methodological rigour.

Machine Learning and Prediction Accuracy

The machine learning model used to predict `discharge_mRS` served as the foundation for the chatbot's decision support functionality. Trained on structured patient data, the model aimed to provide reliable outcome predictions to guide user decision-making. Although the primary focus of the study was on the impact of mitigation strategies rather than predictive performance, the model's integration allowed participants to compare their own prognosis with algorithmic suggestions in real time. Importantly, no participants reported the predictions as obviously inaccurate, suggesting a baseline level of trust or plausibility in the model's outputs. However, qualitative feedback and participant behavior indicated that some users accepted predictions without critical reflection, particularly in the double mitigation strategy, underscoring the importance of interface design in promoting appropriate model reliance. The LightGBM model never misclassified classes 0–3 as any of 4–6, demonstrating a clear distinction between mild and severe outcomes. This separation minimizes the risk of critical misclassification, which is vital for accurate patient prognosis and informed clinical decision-making. While we experimented with several algorithms—including XGBoost, CatBoost, logistic regression, SVM, random forest, and naive Bayes, we ultimately selected LightGBM due to computational constraints on our machines, as training other ensemble models required significantly longer times. We also explored stacking CatBoost, LightGBM, and XGBoost, but this approach increased training time by approximately threefold.

FUTURE WORK & LIMITATIONS

Since 6 out 10 participants stated they had not formed an opinion about the patients prior to asking for a prediction, a future study should look into the use of the mitigation strategy "Decision justification", it was unused in this study due to requiring another input from the user, but that extra input might be needed to force users to form an opinion of their own before asking the chatbot. However based on comments from the Clinicians, encouraging them to form their own opinion is not helpful, and requiring them to input their own opinion into the bot would be a waste of time[2]. So it is likely that even if Decision justification works better for the general public, clinicians might not find it helpful. Based on the clinicians preference for consider the opposite, another test should be performed where consider the opposite is the strategy used in both conditions. Since our quantitative data seems to support the hypothesis that adding more mitigation strategies does not make the system more usable, there would be value in seeing how the clinicians preferred strategy could perform alone.

It should be noted that the clinicians that tested our prototype at ESOC do not necessarily represent the average clinician, as all 3 of them have more experience in data processing and analysis and testing new systems.

During the development of the machine learning model, we experimented with automatic hyperparameter tuning using Optuna. However, contrary to expectations, this optimization process consistently reduced model accuracy by 1–4% across several algorithms. This suggests that the search space or evaluation criteria may not have been well-aligned with the underlying data characteristics, and highlights the need for more tailored tuning strategies in future work.

Additionally, there is room for improvement in feature engineering. While efforts were made to remove administrative and non clinical variables such as hospital identifiers, it is possible that other irrelevant or misleading features remained. Conversely, some features we excluded may have held clinical significance, especially given our limited domain expertise. Collaborating with medical professionals in future iterations could help identify more meaningful predictors and eliminate confounding variables.

Future work should aim to improve the model's predictive performance through a combination of better feature selection, more advanced preprocessing, and the use of more sophisticated models or ensemble methods. Reevaluating the hyperparameter optimization process with refined constraints could also yield better results.

CONCLUSION

This study explored the impact of different mitigation strategies in a clinical decision support chatbot, comparing a single mitigation strategy, "hear the story first", with a double strategy that added "consider the opposite." Usability data, workload assessments and Thematic analysis, revealed that the single mitigation condition was generally preferred by student participants, who found it easier to navigate and less cognitively demanding.

In contrast, clinicians valued the double mitigation strategy, though not necessarily due to multiple strategies, but because of the "consider the opposite" prompts inclusion. They felt the strategy supported critical reflection and helped guard against premature closure. The students However, were not consistently influenced by the double mitigation strategy, the added mitigation prompt often went unnoticed by student users.

Given these outcomes, we fail to reject the null hypothesis: increasing the number of mitigation strategies did not significantly lead users to put more thought into their answers. These findings suggest that simply adding more prompts is insufficient without ensuring they are perceived as meaningful and relevant, particularly for student users. However, the clinicians, did see 'Consider the opposite' as positive in their chatbot interaction.

Despite our small sample size, our study provided useful data, not just in the use of mitigation strategies to counter premature closure bias, but also in testing on users who do fit intended target group.

The expert user feedback challenged many of the assumptions made based on our initial tests, highlighting the importance of matching the evaluation participants with the intended target group of the prototype.

REFERENCES

[1] Matthew R. Allen, Sophie Webb, Ammar Mandvi, Marshall Frieden, Ming Tai-Seale, and Gene Kallenberg. 2024. Navigating the doctor-patient-AI relationship - a mixed-methods study of physician attitudes toward

artificial intelligence in primary care. *BMC family practice* 25, 1 (2024), 42–12.

[2] Anne Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. 2023. "If I Had All the Time in the World" : Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. (2023).

[3] Avyay Casheekar, Archit Lahiri, Kanishk Rath, Kaushik Sanjay Prabhakar, and Kathiravan Srinivasan. 2024. A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer science review* 52 (2024), 100632–.

[4] Keng Sheng Chew, Jeroen J. G. van Merrienboer, and Steven J. Durning. 2019. Perception of the usability and implementation of a metacognitive mnemonic to check cognitive errors in clinical setting. *BMC medical education* 19, 1 (2019), 18–18.

[5] K.M. Colby, A.P. Goldstein, and L. Krasner. 2013. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. Elsevier Science. https://books.google.dk/books?id=M-RFBQAAQBAJ

[6] Kevin W. Eva and John P. W. Cunnington. 2006. The difficulty with experience: Does practice increase susceptibility to premature closure? *The Journal of continuing education in the health professions* 26, 3 (2006), 192–198.

[7] Sandra G. Hart and Lowell E. Staveland. 1988. *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*. Technical Report NASA-TR-101. NASA Ames Research Center. https://ntrs.nasa.gov/citations/20000012465

[8] Marcos Iglesias, Chaitali Sinha, Ramakant Vempati, Sarah Elizabeth Grace, Madhavi Roy, William C. Chapman, and Monica Lynn Rinaldi. 2023. Evaluating a Digital Mental Health Intervention (Wysa) for Workers' Compensation Claimants: Pilot Feasibility Study. *Journal of occupational and environmental medicine* 65, 2 (2023), e93–e99.

[9] Christine J. Ko and Earl J. Glusac. 2023. Cognitive bias in pathology, as exemplified in dermatopathology. *Human pathology* 140 (2023), 267–275.

[10] Katarzyna Kuligowska, Rana Bawaneh, Firas Ismail, Yasser Taher, and Ibrahim Al-Mashaqbeh. 2020. Chatbot Usability Questionnaire (CUQ): A tool for measuring chatbot usability. *International Journal of Human–Computer Interaction* 36, 12 (2020), 1146–1161. DOI: http://dx.doi.org/10.1080/10447318.2020.1733512

[11] Bharat Kumar, Balavenkatesh Kanna, and Suresh Kumar. 2011. The pitfalls of premature closure: clinical decision-making in a case of aortic dissection. *BMJ case reports* 2011, oct03 1 (2011), bcr0820bcr0820114594–0820114594.

[12] Haotian Li, Yun Wang, and Huamin Qu. 2023. Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration. (2023).

[13] Luca Longo, Christopher D. Wickens, Gabriella Hancock, and P. A. Hancock. 2022. Human Mental Workload: A Survey and a Novel Inclusive Definition. *Frontiers in psychology* 13 (2022), 883321–883321.

[14] John H. McDonald. 2014. *Handbook of Biological Statistics* (3rd ed.). Sparky House Publishing, Baltimore, Maryland. http://www.biostathandbook.com/

[15] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, Dino Pedreschi, Simone Barbosa, Caroline Appert, Cliff Lampe, David A. Shamma, Koji Yatani, Steven Drucker, and Julie Williamson. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–9.

[16] Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. *Journal of medical Internet research* 23, 3 (2021), e24850–e24850.

[17] Samuel S. Shapiro and Martin B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611. DOI: http://dx.doi.org/10.2307/2333709

[18] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, Adam Perer, Tesh Goyal, Anicia Peters, Stefanie Mueller, Kaisa Väänänen, Per Ola Kristensson, Albrecht Schmidt, Julie R. Williamson, and Max L. Wilson. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18.

[19] Suzan van Brussel, M.C.L. Timmermans, Peter Verkoeijen, and Fred Paas. 2020. Consider the Opposite: Effects of Elaborative Feedback and Correct Answer Feedback on Reducing Confirmation Bias – A Pre-registered Study. *Contemporary Educational Psychology* 60 (2020), 101844. DOI: http://dx.doi.org/10.1016/j.cedpsych.2020.101844

[20] Richard S. Wallace. 2009. *The Anatomy of A.L.I.C.E.* Springer Netherlands, Dordrecht, 181–210. DOI: http://dx.doi.org/10.1007/978-1-4020-6710-5_13

[21] Ben Wang, Jiqun Liu, Jamshed Karimnazarov, Nicolas Thompson, Paul Clough, Morgan Harvey, and Frank Hopfgartner. 2024. Task Supportive and Personalized Human-Large Language Model Interaction: A User Study. In *Proceedings of the 2024 Conference on*

Human Information Interaction and Retrieval. ACM, New York, NY, USA, 370–375.

[22] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. DOI: http://dx.doi.org/10.1145/365153.365168

[23] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. DOI:http://dx.doi.org/10.2307/3001968

APPENDIX A - GITHUB-RASA

https://github.com/LuckyDDBH/RasaProCalm

APPENDIX B - GITHUB-WEBSTORM

https://github.com/LuckyDDBH/chartjsdash/tree/DanielSommeBranch

APPENDIX C - WORKSHEET

Worksheet can be found in the folder Appendix/Med10Worksheet.pdf

APPENDIX D - AV PRODUCTION

https://youtu.be/S8_ALDcYr5A

APPENDIX E - RAW DATA

The raw data can be found in the folder Appendix/data/