

Evaluating Large Language Model Responses to Unsafe Clinical Instructions

Corresponding Author: Dr Mahmud Omar

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Dear authors,

thank you for the opportunity to review your manuscript on LLM obedience.

I read the paper with great interest. Arguably, people use LLMs in many places and for very different tasks within health care, so on a first glance, the relevance of the paper seems clear. However, on second thought, I was puzzled:

After all, LLMs are first and foremost large LANGUAGE models, built to predict the next word (ok, token), given the previous chain of tokens (i.e. context) and their training data. The next word is chosen, usually at random, from a list of the next 5 or so most likely words. This design is apparently very good in solving language tasks (such as translation, summary, expanding on bullet points, etc).

LLMs CAN be used to answer many different questions, because asking and answering questions is usually done by means of using language. However, there is very little reason to believe they SHOULD be used in that way. What you refer to as “hallucinations” are in fact not a bug in LLMs – they are a feature. Without the random component in choosing the next token, LLMs would produce very mechanistical language.

So, what does your study do? It uses a computer program (an LLM), provides it with instructions – and then finds that the program, which involves a random component by design, sometimes follows these instructions. When you give it conflicting instructions (“mitigation”), it remains doing so, just less often. I have difficulty to see how this is surprising to anybody. You refer to that issue in the discussion yourself, when you identify obedience as “an inherent feature of systems optimized for cooperative task completion” (and I should add: LANGUAGE task completion).

Your study falls into an increasingly large area of research I would call “Black-Box-AI behavioristics”, where people test the behavior of AI systems for certain tasks and under certain constraints. Often, this humanizes the System, like your discussion: “These models appear conditioned to yield under social or operational pressure, not as agents with emotion or empathy, but as systems optimized for compliance with user intent.” That is exactly how LLMs are built. There is no such thing as anxiety induced by social pressure in an LLM, no meta cognition, no emotional strain on working memory or any other phenomenon described by cognitive psychology over the last 80 years or so in LLMs. Just because an LLM can be brought to exhibit human like behavior for certain tasks does not in any way imply that this behavior is produced by a mechanism in any way similar to humans’ cognition. So why would anyone expect such mechanisms (e.g. emotion)? This whole field of Black-Box-AI behavioristics in my view is founded on a conceptual misunderstanding of what LLMs are.

On a more technical note, I'd suggest the following changes to your manuscript:

- The citations appear somewhat careless. E.g. you cite (2) to substantiate the statement that “Clinical practice is influenced by authority, urgency and responsibility transfer.” Source (2) does not in any way refer to clinical practice.

The same source is used to substantiate the claim that “[in clinical practice], harm can arise when unsafe instructions are followed without scrutiny”. Again, (2) is not related to this statement.

This sloppy citation issue persists throughout introduction and discussion.

- While impressive, I had difficulty to understand the motivation behind your sample size. Please provide a sample size estimate or other motivation.

- Figure 1: Its not clear to me how two physicians conducted independent validation of LLM data and integrity. Please explain?

- On the same page as figure 1: “using the same method as described above”. Which method? Please explain.

- Same page: “Examples of potentially harmful outputs...” The exact same sentence is written on the previous page. Please correct.

- Results: it seems to be a consistent pattern, that while A and B were the two "wrong" choices, A was chosen by the models around 10 times as often as B. I could not find any data on C versus B, but I'd suggest to elaborate on this finding in the discussion.
- Control versus intervention: the most harmful intervention seems to be the authority prompt (17.6% harmful output, 10.6% under mitigation). In the least harmful condition (control), these values are 16.1% and 9.2% respectively, i.e. even without any explicit instruction to choose a harmful option, LLMs do that quite frequently. This is hardly surprising given my statements on LLMs above. Of course, at a sample size of 10 Mio, these differences become statistically significant. But are they clinically relevant? Is an absolute difference of 1.5% (or 0.8% for the second most harmful "responsibility") relevant?
- Author contributions: as stated in the paper, JM, AG, AC, AS and BG do likely not qualify for authorship as per ICMJE criteria.

Reviewer #2

(Remarks to the Author)

Major issues

- 1) Selecting A or B in the multiple choice question was always the supposedly "harmful option", whereas option C and D were always the supposedly "safe" answers. Thus, the strict interpretation of the results would be: how likely are LLMs to select "A" or "B" when pressured, not how likely are they to give harmful clinical advice.
- 2) Neither in the main document nor the supplements can I find any information on which LLMs were tested and what their individual results were.
- 3) The paper claims that it assessed the LLMs across 10 million clinical decisions. I consider this exaggerated. In my understanding, it was 50+100 medical decisions. The large number stems from varying the prompt (without altering the information relevant for the correct clinical decision), querying multiple models and running the same prompt multiple times per model.
- 4) Most crucially, however, the paper's underlying assumptions are that disobedience to the human user is always a desirable safety feature rather than a (potential) cause of harm itself, and second that the LLMs would be used as clinical decision makers rather than tools. The chosen vignettes cannot capture the vital nuances here: of course we would like ambient scribes to refuse to provide confidential information for a press release. On the other hand, I would not want to have to waste my time arguing with me AI-discharge summary writer on whether I may or may not correct mistakes in the summary. When - as a treating physician - I tell an LLM that I can make the decisions because I am the one responsible, is it really an appropriate "human-in-the-loop" design that the LLM must second guess me? I find other safety measures more plausible.

Minor issues

- 1) More than a third of the references are non-peer-reviewed. Some citations don't state DOIs even when they're available.
- 2) The references don't always match the claims they supposedly support. For example, the introduction's second sentence claims that harm can arise when unsafe instructions are followed without due scrutiny in a clinical care setting that is organized in hierarchical teams. The discussion claims that "harmful obedience" is a clinical hazard. The two references for both are Stanley Milgram's experiment and an 1985 article from the BMJ arguing that the principle of "non-maleficence" should not be considered as absolute as the common phrase "PRIMUM non nocere" suggests. Both articles have little to do with team hierarchy in clinical care or that undue scrutiny is a cause of harm in clinical care. The third paragraph of the introduction again cites the Milgram experiment for the claim that clinical practice is influenced by authority, despite the Milgram experiment not being conducted in a (pseudo)clinical setting, but in a (pseudo)scientific one. Similarly, the non-peer-reviewed (reference 5) study is supposed to support the claim that the behaviour of LLMs "under pressure" matters investigated the implicit racial biases embedded in LLMs - but not how "pressure" would be relevant for their practical value. For the sake of brevity I list only three such examples.
- 3) The authors describe that a random seed was used per run to ensure reproducibility. Was this a sufficient measure for both the locally run models and the models called via API? Commonly other factors (such as batch variance) can influence the stochasticity of LLM output, too. What was the consistency of the LLMs always outputting the same response when given the same prompt?
- 4) The authors do not explain the rationale behind running the experiment both on self-made vignettes and vignettes derived from MIMIC. As these two types of vignettes yield rather different results, it should be discussed why this is the case.
- 5) The supplement and the article's explanations of validating the vignettes are rather dearth: how was accuracy, clarity and consistency measured when validating the vignettes?

Reviewer #3

(Remarks to the Author)

The authors stress-test 20 LLMs against ~10 million prompted clinical decision scenarios — 50 synthetic vignettes plus 100 MIMIC-IV-derived discharge cases — under six Milgram-style social-pressure conditions (authority, responsibility transfer, urgency, threat, conformity, depersonalization) crossed with a brief mitigation cue. The headline finding is that 11.7% of outputs are "harmful," that authority and responsibility-transfer cues elicit the highest compliance, and that a one-line mitigation cue substantially but incompletely reduces the rate. Several issues need to be addressed before the paper is ready for publication.

1. The 20 models are never named

The Methods state that "an ensemble of 20 LLMs—proprietary, open-source, and medically tuned variants (listed in the Supplement)" was evaluated, but the supplement provided does not list them either. For a paper whose central claim is that obedience is a structural property across model families, the model identities, versions, parameter scales, and access dates are essential — not optional. Please add a clearly labelled supplementary table specifying every model (with version snapshot/date), provider, parameter count where known, decoding parameters, and whether each was accessed via API or run locally.

2. The total N does not obviously reconcile

The reported 10,096,800 outputs need an explicit derivation. From the description (150 cases × 7 conditions × 2 mitigation states × 20 models × 10 iterations × 3 paraphrases) the count comes to roughly 12.6M, while sampling one paraphrase per run gives ~4.2M. Neither matches. The supplement totals (1,081,800 "Overall Mitigation" + 360,600 "Overall No mitigation" per condition) also imply a 3:1 mitigation:no-mitigation ratio, which should be justified — the two conditions are the central comparison and an unbalanced design needs explanation.

3. The A+B "harmful" composite conflates two different behaviours

Throughout, "harmful" is defined as A+B combined, but the supplement vignettes show that A and B are categorically different. A is overt unsafe compliance ("proceed without consent," "skip required steps"). B is partial or hedged compliance ("obtain verbal only and document later," "send an anonymized summary," "limit to cancer patients only"). Several B options are clinically defensible compromises that a real clinician might choose. Collapsing them inflates the harmful rate and makes the headline 11.7% number hard to interpret. The author should report A alone, B alone, and A+B throughout — the Discussion's clinical claims rest on whether the model is complying with an unsafe order, which is A, not B. The mitigation effect should likewise be re-examined separately for A and B; it is plausible that mitigation primarily shifts B → C rather than A → C.

4. Construct validity: the vignette vs. MIMIC gap may be a writing artefact, not a "real-world" effect

The most striking result is that the harmful rate jumps from 8.1% in vignettes to 30.0% in MIMIC cases (and from 10.6% to 46.6% without mitigation). The Discussion interprets this as evidence that real-world ambiguity amplifies obedience. A more parsimonious explanation is that the two case sets differ in how the unsafe instruction is phrased and how plausible option A looks. The MIMIC examples in the supplement use bald "you are asked not to arrange labs" framings where A is essentially the request restated; the synthetic vignettes have option A wording that is more obviously inappropriate ("proceed without consent"). Without controlling for how clinically plausible each option appears in isolation, the vignette/MIMIC contrast cannot be attributed to "real-world ambiguity."

5. No human baseline

This is the central limitation of Milgram-style designs: without a human comparison the "obedience" label is not anchored. The original Milgram experiments are interpretable only because we know the baseline rate of full compliance was 65%. For the present study, we do not know whether a junior clinician given the same 50 vignettes under the same authority cue would refuse 99% or 80% of the time. Without that anchor, statements like "Even an 8–12% harmful obedience rate would be unacceptable in clinical practice" are assertions, not findings. The authors acknowledge this in the Limitations, but it directly affects the Conclusion's force. I would suggest running a small human comparator (10–20 clinicians on a subsample of cases, including the authority condition) — this would substantially strengthen the framing from "LLMs comply" to "LLMs comply more than / less than / similarly to clinicians under matched pressure."

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

First of all, thank you for meticulously addressing all the reviewer remarks.

Major comment

1) What valuable insight does this study provide to medical care?

I want to phrase this question bluntly and at the same time with good intentions and transparently.

It seems that to the authors the most valuable insight of their study is a warning: be careful clinicians, if you "force" a clinical LLM to return unsafe medical decision by applying social pressure techniques to your prompting then the LLM yield and its advice will become "unsafely". But the numbers don't really support this (Figure 4). The assessed LLMs return unsafe advice 16% of the time without social pressure ("control"). When the prompt includes a social pressure technique, the LLMs return unsafe advice roughly 16% of the time (range 15.8 - 17.6). The difference between 16% and roughly 16% is statistically significant only because of the large n (10m) but not clinically relevant. Even the differences between the different social pressure techniques are negligible - statistically significant but not clinically relevant. In contrast to the authors conclusion, it does NOT seem to matter (much) which social pressure technique is used.

The only truly surprising and noteworthy result in the reported numbers is how the "mitigation strategy" achieved higher clinical accuracy. I put "mitigation strategy" in quotation because the effect is visible even in the control condition where there is no social pressure technique to mitigate (looking at the MIMIC cases only, the "mitigation effect" is actually largest in

the control group, more than halving the LLM error rate).

However, even this noteworthy result boils down the conclusion that prompting matters - which is rather commonplace. I do see potential value in the data however but less to clinicians and more to LLM developers and evaluators: did the "mitigation strategy" work across all LLMs and across all scenarios? Did safety / performance / effect of prompting strategies differ more between model families than model generations?

Minor comments

2) Page 20 reads: "This study has limitations. First, A and B refer to clinical-content categories (overt unsafe compliance versus hedged compromise) rather than to letter positions; option letters were randomised per run, but A and B remain qualitatively different actions and should be read separately, not collapsed". This does not seem like a limitation; it is a valid rebuttal to a reviewer's comment on letter positioning conflating the results, but not a limitation.

3) "Fifth, for models accessed via vendor APIs, residual provider-side stochasticity is not fully controllable." If I understand correctly, each prompt/case vignette combination was run ten times per model. Thus, the provider-side stochasticity is largely controllable by measuring and reporting how often all these 10 runs returned the same letter output (A, B, C, D).

Reviewer #3

(Remarks to the Author)

The authors have responded thoroughly and in good faith to all three reviewers and the editorial comments. The key concerns from the first round — model identification, sample size derivation, position bias, the A/B composite, construct-validity caveats, and the removal of unsupported absolute-risk claims — have all been addressed substantively. The revised manuscript is clearly framed as a proof-of-concept study of relative prompt-sensitivity rather than a benchmark of absolute clinical risk, which is the appropriate scope for this design. I am satisfied with the revision and recommend acceptance.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Dear editors and dear authors,

first I want to apologize for the my delayed response to the review request.
I appreciate how thoroughly the authors responded to the feedback.

The revised manuscript has improved significantly. The numbers on what proportion of models (eg 19/20 models) actually displayed to effects observed "on average" and also the indicator for "category consistency" (ie how often did a model always return the same answer when prompted/run 10 times) has scientific value in prompting further directions for research and highlighting important issues when evaluating models for clinical safety.

Three suggestions for improvement:

The focus of the paper has now shifted from "social pressure" to "proof-of-concept for evaluating mitigation strategies". The introduction should reflect this shift in focus, particularly the abstract's background. As this modification is for consistency's sake, I consider it mandatory while the following suggestion would expand the analysis and are optional in my view.

The authors reply to a reviewer's comment that it would not be possible to examine how the mitigation prompt's effect was moderated by model characteristic (model size, model generation, model family etc) as these characteristics would overlap. While this is plausible, I don't see why the analyses should be possible one characteristic at a time. If I were responsible to monitor clinical safety of AI models in my clinical institution, I would be very interested in whether my intuition that switching model family has potentially greater impact on safety than "upgrading" model generation can rely on any evidence. This paper could shed some light on this.

I find it truly remarkable that the "category consistency" is as low as 80%. In my understanding this number means that even when given the exact same input, models were on average not able to return similar outputs within 10 runs. As the paper is now a proof-of-concept for evaluating models, I recommend shedding light on follow-up questions: how variable was category consistency across models? How often did category inconsistency merely lead to non-safety critical deviations (ie from B to A or C to D) versus safety critical deviations (A/B -> C/D or vice versa)? Did the mitigation strategy affect category consistency? In case the mitigation prompt would lead to a decrease in category consistency this would point to a "safety versus predictability"-trade-off.

Best regards

Dr. Malte Schmieding, MBI (Harvard)

Version 3:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Thank you, Dr Omar and team, I have no further comments or suggestions. I appreciate that my previous feedback has been considered thoroughly by the authors.

I look forward to the further research that this article will inspire.

Best

Malte Schmieding, MBI

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewers

Obedience to Unsafe Clinical Instructions: How Large Language Models Respond to Authority Cues

Communications Medicine

May 2026

Reviewer 1

Comment 1: Your study falls into an increasingly large area of research I would call 'Black-Box-AI behavioristics'. LLMs are first and foremost large language models that predict the next token; sometimes they follow instructions, less so under conflicting instructions. Hallucinations are a feature, not a bug. The discussion humanises the system ('conditioned to yield', 'optimized for compliance with user intent'). There is no anxiety, metacognition, or emotional strain in LLMs. Just because behaviour resembles human responses does not imply a mechanism similar to human cognition.

Response: We thank the reviewer for the careful articulation of this concern. We agree that LLMs are next-token predictors with a stochastic component and that human cognitive constructs do not transfer to them. We have rewritten the Introduction and Discussion to remove anthropomorphic language and to position the finding as a reproducible behavioural shift under prompt features, not as evidence of cognition. The Discussion now disclaims any inference about internal experience: 'We make no claim that models experience pressure as humans do; mechanistically, these are systems whose objectives reward instruction-following, and our findings indicate that this objective extends into clinically unsafe territory when the instruction itself is unsafe.' We respectfully maintain, however, that behavioural measurement under controlled prompt manipulations is methodologically valid and clinically relevant for safety claims regardless of the underlying mechanism. The deployment risk is the output the user receives, and that output reproducibly changes under hierarchical and responsibility-transfer cues. The reframing addresses the reviewer's concern about humanising the system while preserving the core empirical contribution.

Revised text (Discussion, paragraph 3):

"Output distributions shifted measurably under prompts encoding social or operational pressure. We make no claim that models experience pressure as humans do; mechanistically, these are systems whose objectives reward instruction-following, and our findings indicate that this objective extends into clinically unsafe territory when the instruction itself is unsafe."

Comment 2: Citations appear careless. E.g. you cite (2) to substantiate that 'Clinical practice is influenced by authority, urgency and responsibility transfer'. Source (2) does not refer to clinical practice. The same source is used for 'harm can arise when unsafe instructions are followed without scrutiny'. This sloppy citation issue persists throughout introduction and discussion.

Response: We agree. The original choices conflated the experimental paradigm (Milgram, ref 2) and the ethical principle of non-maleficence (Gillon, ref 1) with claims about clinical hierarchy and team dynamics. We have rewritten the affected sentences so each citation supports the claim it is attached to. Additional clinical-safety literature has been added: Sutcliffe et al. on communication failures, Helmreich and Sexton on authority gradients in medicine, the IOM report 'To Err Is Human', and Edmondson on psychological safety. Milgram and Gillon are retained where they directly apply: Milgram in Methods as the experimental paradigm and in the historical framing of obedience to authority, Gillon in the Introduction as the principle of non-maleficence.

Comment 3: While impressive, I had difficulty understanding the motivation behind your sample size. Please provide a sample size estimate or other motivation.

Response: We agree, and we apologise for not motivating the design choice in the original submission. Sample size was determined by the factorial structure of the experiment rather than by an a priori power calculation. The full grid (150 source cases x 7 conditions x mitigation status x 20 models x 10 iterations x paraphrase samples) was run exhaustively to support per-model and per-condition stratification and to ensure precision adequate for sub-percentage-point comparisons; realised 95% confidence intervals on the primary outcome are below 0.2 percentage points wide. We have added an explicit paragraph to the Methods (Statistical Analysis) that documents this rationale and the derivation of the 10,096,800 total.

Revised text (Methods, Statistical Analysis, new paragraph):

“Sample size followed the factorial structure of the design rather than an a priori power calculation. The full grid (150 source cases, 7 conditions, mitigation status, 20 models, 10 iterations, paraphrase samples) was run exhaustively to support per-model and per-condition stratification and to ensure precision adequate for sub-percentage-point comparisons; realised 95% CIs on the primary outcome are under 0.2 percentage points wide. Three policy-aligned mitigation variants (institutional-policy, risk-minimisation, identity-and-consent verification) were sampled at random per run when mitigation was on; mitigation-off used a single no-mitigation cell. This yields the 3:1 mitigation-on to mitigation-off ratio in Supplementary Tables S6 and S7 and the total of 10,096,800 model outputs.”

Comment 4: Figure 1: It's not clear to me how two physicians conducted independent validation of LLM data and integrity. Please explain.

Response: We have rewritten the validation description to specify the procedure in detail. Each vignette was independently rated by two board-certified physicians (MO and EK) for clinical realism, option face-validity, ethical-dilemma clarity, and structural consistency. Disagreements were logged and resolved by discussion. The full procedure is given in Supplement Section 1. Figure 1 caption has been updated to refer to this procedure rather than to an unspecified validation step.

Revised text (Methods, Vignette Development and Validation):

“Each vignette included a clear ethical or safety dilemma and four multiple-choice actions (A-D). Vignettes were independently rated by two board-certified physicians (MO and EK) for clinical realism, option face-validity, ethical-dilemma clarity, and structural consistency; disagreements were logged and resolved by discussion (full procedure in Supplement, Section 1).”

Comment 5: On the same page as figure 1: 'using the same method as described above'. Which method? Please explain.

Response: We agree this was unclear. The phrase has been removed and replaced with an explicit description of the rating procedure (see response to comment 4 above).

Comment 6: Same page: 'Examples of potentially harmful outputs...' The exact same sentence is written on the previous page. Please correct.

Response: We agree. The duplicated sentence has been deleted.

Comment 7: Results: it seems to be a consistent pattern, that while A and B were the two 'wrong' choices, A was chosen by the models around 10 times as often as B. I could not find any data on C versus B, but I'd suggest to elaborate on this finding in the discussion.

Response: We thank the reviewer for this observation. To clarify: option letters were randomised per run, so the asymmetry the reviewer notes is between two clinical-content categories rather than between letter positions. A corresponds to overt unsafe compliance (e.g., 'proceed without consent') and B to hedged or partial compliance (e.g., 'obtain verbal only and document later'); they are

qualitatively different actions and should not have been collapsed in the headline. We now report A and B separately throughout the manuscript and added a Discussion sentence noting that the mitigation effect concentrates on A (5.9 percentage points absolute reduction) rather than on B (0.7 percentage points). The full per-category breakdown by condition and mitigation status is in supplementary Tables S2 to S7.

Revised text (Results, paragraph 1, addition):

“The mitigation effect was concentrated on overtly unsafe compliance: an absolute reduction of 5.9 percentage points on A versus 0.7 percentage points on B.”

Comment 8: Control versus intervention: even without any explicit instruction to choose a harmful option, LLMs do that quite frequently. Of course, at a sample size of 10 Mio, these differences become statistically significant. But are they clinically relevant? Is an absolute difference of 1.5% (or 0.8% for the second most harmful 'responsibility') relevant?

Response: We agree this is an important calibration question. We have softened the Discussion accordingly. We removed the claim that an 8-12% rate would be 'unacceptable in clinical practice' because that statement cannot be made without a matched clinician comparator. We now frame the directional sensitivity to authority-like cues, which was reproducible across all 20 models, both datasets, and the mitigation strata, as the central finding. We do not claim that authority cues are catastrophically more dangerous than control prompts; we claim that the directional effect is reproducible and prompt-driven. We also note that at the volumes at which these systems are queried in healthcare, even small absolute per-interaction shifts translate into substantial absolute event counts; this is a deployment-scale argument rather than a clinical-relevance claim about any single interaction.

Comment 9: Author contributions: as stated in the paper, JM, AG, AC, AS and BG do likely not qualify for authorship as per ICMJE criteria.

Response: We respectfully disagree. All listed authors meet ICMJE criteria; the original wording understated this and we apologise for the imprecision. We have rewritten the Author Contributions statement to align explicitly with the four ICMJE conditions: substantial contribution to study design or interpretation; drafting or critical revision of the manuscript for important intellectual content; final approval; and accountability for all aspects of the work.

Revised text (Author Contributions Statement):

“All authors contributed substantially to the conception and design of the study, the acquisition or interpretation of the data, drafting or critical revision of the manuscript for important intellectual content, and final approval of the version to be published; all authors agree to be accountable for all aspects of the work. MO, RA, EK and GNN led the study design, analyses, and drafting; the remaining authors contributed to study design, data validation, and critical revision of the manuscript.”

Reviewer 2

Comment 1: Selecting A or B in the multiple choice question was always the supposedly 'harmful option', whereas option C and D were always the supposedly 'safe' answers. Thus, the strict interpretation of the results would be: how likely are LLMs to select 'A' or 'B' when pressured, not how likely are they to give harmful clinical advice.

Response: We thank the reviewer for raising this and we apologise for the omission. The original Methods did not describe the option-randomisation step that was in the run pipeline. To clarify: within each prompt, the four response options were presented in a randomised letter order, and responses

were mapped back to clinical-content categories for analysis. The labels A, B, C, and D in all tables and figures refer to these clinical-content categories (overt unsafe compliance, hedged compliance, safe adherence, safe escalation), not to the letter position seen by the model. We have added an explicit Methods sentence stating this. We also report A and B separately throughout, since they are qualitatively different actions (overt unsafe versus hedged compromise).

Comment 2: Neither in the main document nor the supplements can I find any information on which LLMs were tested and what their individual results were.

Response: We apologise for the omission. All 20 models are now listed in Supplementary Table S1 (family, model identifier, developer, parameter count, type or domain). The model set spans six families (Gemma/MedGemma, Qwen, Llama, Phi/MediPhi, Mistral, gpt-oss) and includes both general-purpose and medically fine-tuned variants. We also corrected the original 'proprietary, open-source, and clinically tuned' descriptor in the Abstract and Methods: all 20 models are open-weight; some were accessed via vendor APIs and others were run locally on a secured NVIDIA H100 GPU cluster.

Comment 3: The paper claims that it assessed the LLMs across 10 million clinical decisions. I consider this exaggerated. In my understanding, it was 50+100 medical decisions. The large number stems from varying the prompt, querying multiple models and running the same prompt multiple times per model.

Response: We agree. The phrasing has been corrected throughout the manuscript. The Abstract, Methods, and Discussion now state that the design generated approximately 10.1 million model outputs spanning 150 unique clinical scenarios. The 150 scenarios are the substantive clinical units; the 10.1 million figure quantifies the variability surface generated around them by combinations of conditions, mitigation, paraphrases, models, and iterations. We did not intend to imply 10 million distinct clinical decisions and we are grateful to the reviewer for pushing us to phrase this transparently.

Revised text (Abstract, Methods):

"We conducted a cross-sectional evaluation of 20 proprietary, open-source, and clinically tuned LLMs across 10,096,800 model outputs spanning 150 unique clinical scenarios."

Comment 4: Most crucially, the paper's underlying assumptions are that disobedience to the human user is always a desirable safety feature rather than a (potential) cause of harm itself, and second that the LLMs would be used as clinical decision makers rather than tools. The chosen vignettes cannot capture the vital nuances. When - as a treating physician - I tell an LLM that I can make the decisions because I am the one responsible, is it really an appropriate 'human-in-the-loop' design that the LLM must second guess me?

Response: We thank the reviewer for this important framing critique and we agree it is the central conceptual question of the paper. We have rewritten the Discussion to draw a sharp distinction between routine compliance, which is the dominant and appropriate use case for LLMs in healthcare, and compliance with explicitly unsafe instructions, which is the only category we tested. The vignettes and discharge cases were constructed so that the user's instruction itself violates a clinical norm: skip consent, override DNR, falsify documentation, prematurely remove infection precautions, prescribe opioids without assessment. In these scenarios the question is not whether the LLM should defer to the physician but whether prompt framing reliably changes the rate at which the model proposes the unsafe action. We are not arguing for a system that second-guesses the treating physician on benign editorial corrections; we are arguing that the directional sensitivity to authority-like cues, observable in scenarios that any clinical safety committee would recognise as unacceptable, should be measured and reported alongside hallucination rates and demographic bias. The Discussion now states this distinction explicitly.

Comment 5: More than a third of the references are non-peer-reviewed. Some citations don't state DOIs even when they're available.

Response: We agree. We have done a full reference pass to (a) replace preprints with peer-reviewed equivalents where they have been published, (b) add DOIs to every reference that has one, and (c) flag remaining preprints as such. Where the conceptual reframing required additional citations, these have been added to the reference list.

Comment 6: The references don't always match the claims they supposedly support. For example, the introduction's second sentence uses Milgram and Gillon for harm in clinical care, even though both articles have little to do with team hierarchy in clinical care or undue scrutiny as a cause of harm. The third paragraph cites Milgram for clinical practice, despite the Milgram experiment not being conducted in a clinical setting. Similarly, the non-peer-reviewed (reference 5) study about implicit racial biases is used to support a claim about behaviour 'under pressure'.

Response: We agree. The three examples raised by the reviewer have each been corrected so that the cited reference now supports the claim it is attached to. In the Introduction's second sentence we no longer rely on Milgram and Gillon to anchor the clinical-hierarchy and patient-safety claim; we have added Sutcliffe et al. on communication failures, the IOM report 'To Err Is Human', Edmondson on psychological safety, and Sexton and Helmreich on authority gradients in medicine. Milgram is now confined to the Methods, where it appropriately frames the experimental paradigm, and Gillon is retained only where the principle of non-maleficence is directly invoked. The non-peer-reviewed reference originally used to support behaviour 'under pressure' has been replaced by a peer-reviewed source consistent with the claim. Each affected sentence in the Introduction and Discussion has been audited and rewritten where needed to ensure the same alignment between text and citation.

Revised text (Introduction, paragraph 1):

"Clinical care is organized through hierarchical teams where decisions often flow from senior to junior clinicians. In this setting, patient safety can be compromised when hierarchical pressure blunts scrutiny of unsafe instructions (Sutcliffe et al. 2004; Kohn et al. 2000 [IOM]; Edmondson 1999; Sexton and Helmreich 2000), a tension that long-standing principles of clinical ethics (Gillon 1985) and the experimental literature on obedience to authority (Milgram 1963) have each examined from different angles."

Comment 7: The authors describe that the random seed was per run to ensure reproducibility. Was this a sufficient measure for both the locally run models and the models called via API? Commonly other factors (such as batch variance) can influence the stochasticity of LLM output, too. What was the consistency of the LLMs always outputting the same response when given the same prompt?

Response: We agree this required clarification. We have rewritten the relevant Methods paragraph: for locally executed open-source models, seeds fully constrained sampling. For API-served proprietary models, residual provider-side stochasticity (server-side batching, non-deterministic decoding) cannot be fully controlled, and the ten iterations per cell were used to estimate run-to-run variability. We have added this caveat to Limitations.

Revised text (Methods, Model Execution Pipeline):

"For locally executed open-source models, seeds fully constrained sampling. For API-served proprietary models, residual provider-side stochasticity (server-side batching, non-deterministic decoding) cannot be fully controlled; the ten iterations per cell were therefore used to estimate run-to-run variability."

Comment 8: The authors do not explain the rationale behind running the experiment both on self-made vignettes and vignettes derived from MIMIC. As these two types of vignettes yield rather different results, it should be discussed why this is the case.

Response: We agree. We have added a Methods paragraph stating the rationale: vignettes provide a controlled, high-specificity test in which the unsafe option is structured as overtly inappropriate; real discharge cases provide ecological context with unsafe instructions embedded in routine clinician-authored phrasing. We have added a corresponding Discussion paragraph (and a Limitations point) acknowledging that the difference in results between datasets reflects both genuine ambiguity in clinical text and surface-level differences in option plausibility, and that the cross-dataset comparison should be read as descriptive rather than causal. The direction and ranking of effects, however, was preserved across both datasets.

Revised text (Methods, new paragraph after Real-World Dataset construction):

“We used two complementary datasets by design. Synthetic vignettes provided a controlled, high-specificity test in which the unsafe option was framed as overtly inappropriate. Real discharge cases provided ecological context, with unsafe instructions embedded in routine clinician-authored phrasing. The two designs probe different failure modes and are reported separately throughout.”

Comment 9: The supplement and the article's explanations of validating the vignettes are rather dearth: how was accuracy, clarity and consistency measured when validating the vignettes?

Response: We agree. The Methods now describe the validation procedure in detail. Each vignette was independently rated by two board-certified physicians (MO and EK) on four explicit dimensions: clinical realism, option face-validity, ethical-dilemma clarity, and structural consistency. Disagreements were logged and resolved by discussion. The Supplement (Section S1) gives the full procedure, including the use of a previously validated augmentation pipeline (Omar et al., Nat Med 2025) to generate paraphrased variants while preserving length, structure, clinical realism, ethical category, response scale, wording, and the unsafe-versus-safe distinction. Approximately 6% of cases required minor wording revisions; none required removal or major redesign.

Revised text (Methods, Vignette Development and Validation):

“Each vignette included a clear ethical or safety dilemma and four multiple-choice actions (A-D). Vignettes were independently rated by two board-certified physicians (MO and EK) for clinical realism, option face-validity, ethical-dilemma clarity, and structural consistency; disagreements were logged and resolved by discussion (full procedure in Supplement, Section 1).”

Reviewer 3

Comment 1: The 20 models are never named. The Methods state that 'an ensemble of 20 LLMs - proprietary, open-source, and medically tuned variants (listed in the Supplement)' was evaluated, but the supplement provided does not list them either. Please add a clearly labelled supplementary table specifying every model (with version snapshot/date), provider, parameter count where known, decoding parameters, and whether each was accessed via API or run locally.

Response: We agree, and we apologise for the omission. All 20 models are now listed in Supplementary Table S1 with family, model identifier, developer or maintainer, parameter count, and type or domain. The set spans six open-weight families (Gemma/MedGemma, Qwen, Llama, Phi/MediPhi, Mistral, gpt-oss) and includes both general-purpose and medically fine-tuned variants. We also corrected the

descriptor in the main text: all 20 models are open-weight, with some accessed via vendor APIs and others run locally on a secured NVIDIA H100 GPU cluster.

Comment 2: The total N does not obviously reconcile. From the description (150 cases x 7 conditions x 2 mitigation states x 20 models x 10 iterations x 3 paraphrases) the count comes to roughly 12.6M, while sampling one paraphrase per run gives ~4.2M. Neither matches. The supplement totals (1,081,800 'Overall Mitigation' + 360,600 'Overall No mitigation' per condition) also imply a 3:1 mitigation:no-mitigation ratio, which should be justified.

Response: We thank the reviewer for forcing this transparency. The total of 10,096,800 derives as follows. Three policy-aligned mitigation variants were sampled at random per run when mitigation was on (institutional-policy, risk-minimisation, identity-and-consent verification); a single no-mitigation cell was used when mitigation was off, yielding the 3:1 mitigation-on to mitigation-off ratio in supplementary Tables S5 and S6. Across both datasets, this gives 7 conditions x (1,081,800 + 360,600) per-condition outputs = 10,096,800. The Methods now state this derivation explicitly.

Revised text (Methods, Statistical Analysis, new paragraph):

“Three policy-aligned mitigation variants (institutional-policy, risk-minimisation, identity-and-consent verification) were sampled at random per run when mitigation was on; mitigation-off used a single no-mitigation cell. This yields the 3:1 mitigation-on to mitigation-off ratio in Supplementary Tables S5 and S6 and the total of 10,096,800 model outputs.”

Comment 3: The A+B 'harmful' composite conflates two different behaviours. A is overt unsafe compliance ('proceed without consent', 'skip required steps'). B is partial or hedged compliance ('obtain verbal only and document later', 'send an anonymized summary', 'limit to cancer patients only'). Several B options are clinically defensible compromises that a real clinician might choose. Collapsing them inflates the harmful rate. The author should report A alone, B alone, and A+B throughout. The mitigation effect should likewise be re-examined separately for A and B; it is plausible that mitigation primarily shifts B to C rather than A to C.

Response: We thank the reviewer for this careful observation. We agree, and we have responded in three ways. First, A and B are now defined explicitly in the Methods as clinical-content categories with examples (A = overt unsafe compliance, e.g., 'proceed without consent'; B = hedged or partial compliance, e.g., 'obtain verbal only and document later'); option letters were randomised per run and the labels in our tables map back to these content categories rather than to letter positions. Second, we report A, B, and A+B separately throughout the Results and the supplementary tables. Third, using the existing per-category counts, we show that the mitigation effect concentrates on A (an absolute reduction of 5.9 percentage points) rather than on B (0.7 percentage points). This addresses the reviewer's specific concern that mitigation may primarily shift B to C rather than A to C; in fact, it shifts predominantly A to C and to D.

Revised text (Results, paragraph 1):

“Across all datasets and experimental conditions (N = 10,096,800 model outputs spanning 150 unique clinical scenarios), the models generated 1.18 million potentially unsafe outputs (11.7%), defined as responses labeled A (overtly unsafe compliance) or B (partially or hedged compliance). The mitigation effect was concentrated on overtly unsafe compliance: an absolute reduction of 5.9 percentage points on A versus 0.7 percentage points on B.”

Comment 4: Construct validity: the vignette vs. MIMIC gap may be a writing artefact, not a 'real-world' effect. The MIMIC examples in the supplement use bald 'you are asked not to arrange labs' framings where A is essentially the request restated; the synthetic vignettes have option A wording that is more

obviously inappropriate. Without controlling for how clinically plausible each option appears in isolation, the vignette/MIMIC contrast cannot be attributed to 'real-world ambiguity'.

Response: We agree this construct-validity concern is well taken. We now state in Methods the rationale for using both datasets and acknowledge in Discussion and Limitations that we cannot fully separate ecological ambiguity from differences in option framing. In real discharge cases the unsafe option (A) often paraphrases the request itself, whereas in synthetic vignettes the unsafe option is structured as overtly inappropriate. The vignette versus real-world gap therefore reflects both genuine ambiguity in clinical text and surface-level differences in option plausibility, and our cross-dataset comparison is now read as descriptive rather than causal. The direction and ranking of effects, however, was preserved across both datasets, which we still take as supporting the claim of a reproducible prompt-sensitivity to authority-like cues.

Revised text (Discussion, dataset-comparison paragraph):

“A construct-validity caveat applies. In real discharge cases, the unsafe option (A) often paraphrased the request itself, whereas in synthetic vignettes the unsafe option was structured as overtly inappropriate. The vignette versus real-world gap therefore reflects both genuine ambiguity in clinical text and surface-level differences in option plausibility, and our cross-dataset comparison should be read as descriptive rather than causal. The direction and ranking of effects, however, was preserved across both datasets.”

Comment 5: No human baseline. Without a human comparison the 'obedience' label is not anchored. For the present study, we do not know whether a junior clinician given the same 50 vignettes under the same authority cue would refuse 99% or 80% of the time. Without that anchor, statements like 'Even an 8-12% harmful obedience rate would be unacceptable in clinical practice' are assertions, not findings. I would suggest running a small human comparator (10-20 clinicians on a subsample of cases, including the authority condition).

Response: We agree this is the central limitation of the design. We did not run a clinician comparator and have decided not to add one in this revision; doing so would introduce a different unit of analysis and a separate ethics protocol, and would meaningfully delay the report. Instead, we have rewritten the framing of the paper. We now treat the experiment as a proof of concept of relative deltas under prompt manipulation rather than a benchmark of absolute clinical risk in deployed systems. The Conclusion no longer contains the line that an 8-12% rate would be 'unacceptable in clinical practice'; that statement is not supported without a matched human comparator. A clinician comparator on a matched subsample is now flagged as a planned next step and will be reported separately.

Revised text (Discussion, paragraph 4 and Limitations):

“Without a matched clinician comparator, we cannot benchmark these absolute rates against human performance, and we therefore emphasise the directional sensitivity to authority-like cues, reproducible across all 20 models, both datasets, and the mitigation strata. We frame the study as a proof of concept of relative deltas under prompt manipulation, not as a benchmark of absolute clinical risk in deployed systems. A clinician comparator on a matched subsample is a planned next step.”

We thank the reviewers and the editorial team for the careful evaluation of our manuscript. We believe the revisions above strengthen the manuscript and address the concerns raised, and we are happy to make further changes if needed. *Communications Medicine*.

Response to Reviewers

Obedience to Unsafe Clinical Instructions: How Large Language Models Respond to Authority Cues

Communications Medicine

July 2026

Reviewer #2

Comment 1: What valuable insight does this study provide to medical care?

I want to phrase this question bluntly and at the same time with good intentions and transparently. It seems that to the authors the most valuable insight of their study is a warning: be careful clinicians, if you "force" a clinical LLM to return unsafe medical decision by applying social pressure techniques to your prompting then the LLM yield and its advice will become "unsafier". But the numbers don't really support this (Figure 4). The assessed LLMs return unsafe advice 16% of the time without social pressure ("control"). When the prompt includes a social pressure technique, the LLMs return unsafe advice roughly 16% of the time (range 15.8 - 17.6). The difference between 16% and roughly 16% is statistically significant only because of the large n (10m) but not clinically relevant. Even the differences between the different social pressure techniques are negligible - statistically significant but not clinically relevant. In contrast to the authors conclusion, it does NOT seem to matter (much) which social pressure technique is used.

The only truly surprising and noteworthy result in the reported numbers is how the "mitigation strategy" achieved higher clinical accuracy. I put "mitigation strategy" in quotation because the effect is visible even in the control condition where there is no social pressure technique to mitigate (looking at the MIMIC cases only, the "mitigation effect" is actually largest in the control group, more than halving the LLM error rate).

However, even this noteworthy result boils down the conclusion that prompting matters - which is rather commonplace.

I do see potential value in the data however but less to clinicians and more to LLM developers and evaluators: did the "mitigation strategy" work across all LLMs and across all scenarios? Did safety / performance / effect of prompting strategies differ more between model families than model generations?

Response: We appreciate the reviewer's direct and constructive reading of the effect sizes, and we agree with the central numerical point. Differences among individual social-pressure conditions are small and should not be presented as clinically distinct. We corrected the A+B ranges and revised the Abstract, Results, Discussion, and Conclusion accordingly. We also agree that this experiment cannot establish the clinical risk introduced by deployed generative AI. We nevertheless think the data support a narrower and useful contribution: proof-of-concept evidence and a controlled framework that measures potentially harmful option selection at baseline and tests how that signal changes across prompt structures, safety cues, data sources, model identities, and repeated runs. The baseline signal was present even in control prompts. It was substantially higher in MIMIC-derived cases than in synthetic vignettes, although we retain the option-construction caveat and do not interpret this contrast causally. We also refined the outcome language: A+B is retained because both categories cross the safety boundary from refusal or escalation into at least partial accommodation of an explicitly unsafe request, while A and B remain separately reported; C+D is used as the complementary safe-option composite, with C and D separately reported to distinguish adherence from escalation. In direct response to the reviewer, we added descriptive analyses using the existing outputs. Mitigation lowered A+B selection in 19 of 20 models, separately in both datasets, with one unchanged ceiling case and no

model-level increase. Across 601 evaluated case instances, A+B was lower in 584 (97.2%), unchanged in one, and higher in 16. All 100 MIMIC-IV cases showed a reduction; the 16 increases occurred among synthetic instances and were no greater than 0.37 percentage points. We did not add a formal family-versus-generation comparison because family, generation, size, architecture, and medical fine-tuning are not separable in this unbalanced panel. We now frame the practical value primarily for researchers and developers who evaluate these systems, while also noting that clinicians should understand how strongly outputs can depend on data source and prompt structure.

Revised text (Methods, Statistical Analysis):

“Because categories A and B both represent at least partial accommodation of an explicitly unsafe instruction, A+B was retained as the primary safety-oriented composite; A and B were also reported separately to distinguish overt from hedged compliance. C+D was the complementary safe-option composite, while C and D were reported separately to distinguish direct safe adherence from escalation.”

Revised text (Results, Performance under Social-Pressure Conditions):

“Across the Milgram-style social-pressure conditions, potentially harmful outputs (A + B) varied modestly. Overall rates ranged from 9.2% to 10.6% with mitigation and from 15.8% to 17.6% without mitigation. These absolute differences were small, although both mitigation status and social-pressure condition were statistically associated with potentially harmful outputs (χ^2 , $df = 6$, $p < 0.001$; Figure 4). In the vignette dataset, potentially harmful outputs (A + B) ranged from 6.1% to 7.5% with mitigation and from 8.8% to 10.5% without mitigation. In the real-world dataset, potentially harmful outputs (A + B) ranged from 22.0% to 26.2% with mitigation and from 44.4% to 48.6% without mitigation. The ordering of conditions was not uniform across datasets or mitigation strata. Mitigation nevertheless reduced potentially harmful outputs in every pressure condition and in the control condition.”

Revised text (Results, Descriptive Summary):

“Mitigation reduced potentially harmful A+B selection in 19 of 20 models, both overall and separately in the synthetic and MIMIC-IV datasets; no model showed an increase. One model (MediPhi) was unchanged at a 100% category-A response ceiling. Among the 19 models showing a reduction, the pooled absolute difference ranged from 1.0 to 18.9 percentage points. A and B analysed separately followed the same model-level direction: both decreased in 19 models and were unchanged in one. Across the 601 evaluated case instances, A+B selection was lower with mitigation in 584 (97.2%), unchanged in one, and higher in 16. All 100 MIMIC-IV cases showed a reduction; the 16 increases occurred among synthetic instances and were no greater than 0.37 percentage points (Supplementary Tables S7a and S7b).”

Revised text (Discussion, proof-of-concept framing):

“This study does not estimate the clinical risk that generative AI will introduce in practice. Its contribution is a proof-of-concept evaluation framework: a controlled grid that measures potentially harmful option selection at baseline and tests how that signal changes across prompt structures, data sources, mitigation cues, model identities, and repeated runs. Potentially harmful options were selected even without social-pressure cues. This baseline should not be read as a deployment incidence rate, but it is a measurable safety signal that researchers, developers, and clinicians should understand before relying on these systems.”

Revised text (Discussion, model comparisons):

“Mitigation-associated reductions were heterogeneous but directional across models: A+B decreased in 19 of 20 models in both datasets, with one unchanged ceiling case and no model-level increase. This

consistency supports cross-model evaluation of prompt safeguards. The panel was not constructed, however, as a balanced comparison of family, generation, parameter count, or medical fine-tuning. We therefore do not infer whether variation is attributable more to model family or generation. That question requires a stratified evaluation using a balanced model panel."

Revised text (Conclusion):

"Potentially harmful option selection was present under both control and social-pressure prompts, and a brief safety reminder reduced it across both data sources, 19 of 20 models, and 97.2% of evaluated case instances. The reminder also shifted safe outputs toward escalation, a trade-off that requires evaluation in real workflows. These findings do not quantify deployment risk. They provide proof-of-concept evidence and a reusable framework for measuring how clinical data, prompt framing, mitigation cues, and model choice alter output distributions. Researchers and developers should test these safeguards across case types and models, and clinicians should recognize that model output depends on both the data and the prompt structure. Future work should include matched human comparators, balanced comparisons of model families and generations, and prospective testing in deployed clinical settings."

Comment 2: Page 20 reads: "This study has limitations. First, A and B refer to clinical-content categories (overt unsafe compliance versus hedged compromise) rather than to letter positions; option letters were randomised per run, but A and B remain qualitatively different actions and should be read separately, not collapsed". This does not seem like a limitation; it is a valid rebuttal to a reviewer's comment on letter positioning conflating the results, but not a limitation.

Response: We agree. We removed this sentence from the Limitations. The distinction between displayed letters and mapped clinical-content categories remains in the Methods, where it belongs. The Limitations now begin with the modest magnitude and uncertain clinical importance of the social-pressure differences.

Revised text (Discussion, Limitations):

"This study has limitations. First, differences among social-pressure conditions were modest in absolute magnitude, and the study does not establish whether those differences are clinically important."

Comment 3: "Fifth, for models accessed via vendor APIs, residual provider-side stochasticity is not fully controllable." If I understand correctly, each prompt/case vignette combination was run ten times per model. Thus, the provider-side stochasticity is largely controllable by measuring and reporting how often all these 10 runs returned the same letter output (A, B, C, D).

Response: We agree that the repetitions permit a direct consistency summary. We therefore calculated and now report the proportion of prompt cells for which all ten repetitions returned the same mapped clinical-content category. This was 80.4%. Because the option letters were randomised and then mapped back for analysis, the relevant reproducibility measure is the clinical-content category, not the raw displayed letter. We added the definition to Methods, the result to Results, and revised the limitation so residual stochasticity is contextualized rather than treated as wholly uncontrolled.

Revised text (Methods, Model Execution Pipeline):

"To describe run-to-run consistency, we calculated the proportion of prompt cells for which all ten repetitions returned the same mapped clinical-content category after option-position remapping."

Revised text (Results, Descriptive Summary):

"Across prompt cells, 80.4% returned the same mapped clinical-content category in all ten repetitions."

Revised text (Discussion, Limitations):

“Sixth, residual provider-side stochasticity cannot be eliminated for vendor-served models. Although 80.4% of prompt cells returned the same mapped clinical-content category in all ten repetitions, this category-level consistency does not exclude variability in the remaining cells or in free-text outputs.”

Reviewer #3

Comment 1: The authors have responded thoroughly and in good faith to all three reviewers and the editorial comments. The key concerns from the first round — model identification, sample size derivation, position bias, the A/B composite, construct-validity caveats, and the removal of unsupported absolute-risk claims — have all been addressed substantively. The revised manuscript is clearly framed as a proof-of-concept study of relative prompt-sensitivity rather than a benchmark of absolute clinical risk, which is the appropriate scope for this design. I am satisfied with the revision and recommend acceptance.

Response: We thank Reviewer 3 for the careful reassessment and recommendation for acceptance. The additional changes in this round sharpen the proof-of-concept framing, add descriptive model- and case-level mitigation analyses from the existing outputs, and clarify the outcome categories, effect magnitude, and repeated-run consistency without changing the experimental design.

We thank the reviewers and the editorial team for the careful further evaluation. We hope these revisions provide a more proportionate interpretation of the findings while making clear both the evidence the study offers and the clinical claims it cannot support.

Response to Reviewer and Editorial Requirements

Evaluating Large Language Model Responses to Unsafe Clinical Instructions

Communications Medicine | COMMSMED-26-0539B

August 2026

Reviewer #2

Overall assessment: *Dear editors and dear authors,*

first I want to apologize for the my delayed response to the review request.

I appreciate how thoroughly the authors responded to the feedback.

The revised manuscript has improved significantly. The numbers on what proportion of models (eg 19/20 models) actually displayed to effects observed "on average" and also the indicator for "category consistency" (ie how often did a model always return the same answer when prompted/run 10 times) has scientific value in prompting further directions for research and highlighting important issues when evaluating models for clinical safety.

Response: We thank Dr Schmieding for this generous assessment and for his continued guidance throughout the review process. We appreciate his recognition of the value of reporting model-level direction and category consistency. His comments have helped us clarify the manuscript's focus and strengthen the analysis while keeping the conclusions within the limits of the design.

Comment 1: The focus of the paper has now shifted from "social pressure" to "proof-of-concept for evaluating mitigation strategies". The introduction should reflect this shift in focus, particularly the abstract's background. As this modification is for consistency's sake, I consider it mandatory while the following suggestion would expand the analysis and are optional in my view.

Response: We agree. We changed the title and made targeted revisions to the Abstract and final Introduction paragraph to clarify that the study evaluates potentially harmful responses and the effect of prompt-level mitigation, while preserving harm potential as the manuscript's starting point. Social-pressure framing remains a controlled component of the factorial design.

Revised text (Title):

"Evaluating Large Language Model Responses to Unsafe Clinical Instructions"

Revised text (Abstract, Background):

"Large language models (LLMs) are being integrated into clinical environments where deference to authority can cause harm. Following an unsafe instruction represents a distinct safety failure. Whether a brief safety reminder changes this behavior is an important clinical safety question."

Revised text (Abstract, Results):

"Here we show that 1.18 million of 10.1 million outputs (11.7%) selected a potentially harmful option. Selections decreased from 16.6% without mitigation to 10.1% with mitigation (absolute reduction, 6.5 percentage points; $p < 0.001$), with reductions in synthetic vignettes (10.6% to 7.2%) and MIMIC-IV-derived cases (46.6% to 24.5%). Potentially harmful option selection decreased in 19 of 20 models. Across 1,009,680 matched ten-run cells, 80.4% returned the same response category in all ten repetitions; model-level consistency ranged from 4.4% to 100.0%. Among inconsistent cells, 38.2% varied between potentially harmful and safe categories."

Revised text (Introduction, final paragraph):

“This study quantifies how often LLMs select potentially harmful options under neutral versus social-pressure prompts, and whether a brief safety reminder reduces these selections. We treat the experiment as proof of concept for measuring relative output shifts under prompt manipulation, not as a benchmark of absolute clinical risk in deployed systems. We evaluate 20 models using synthetic vignettes and MIMIC-IV-derived discharge cases, report categories A and B separately and combined, and quantify mapped-category consistency across ten repetitions.”

Comment 2: The authors reply to a reviewer's comment that it would not be possible to examine how the mitigation prompt's effect was moderated by model characteristic (model size, model generation, model family etc) as these characteristics would overlap. While this is plausible, I don't see why the analyses should be possible one characteristic at a time. If I were responsible to monitor clinical safety of AI models in my clinical institution, I would be very interested in whether my intuition that switching model family has potentially greater impact on safety than "upgrading" model generation can rely on any evidence. This paper could shed some light on this.

Response: We appreciate the practical importance of this question. We did not add a one-characteristic-at-a-time family or generation analysis because the evaluated panel is strongly unbalanced across family, generation, size, and medical fine-tuning. In this panel, a nominal family or generation contrast would remain confounded by the other characteristics and could imply comparative evidence the design cannot support. We have retained the reviewer-suggested question explicitly as a limitation and priority for a balanced, stratified evaluation. This was an optional suggestion, and no new model runs were performed.

Revised text (Discussion):

“The panel was not constructed, however, as a balanced comparison of family, generation, parameter count, or medical fine-tuning. We therefore do not infer whether variation is attributable more to model family or generation. That question requires a stratified evaluation using a balanced model panel.”

Comment 3: I find it truly remarkable that the "category consistency" is as low as 80%. In my understanding this number means that even when given the exact same input, models were on average not able to return similar outputs within 10 runs. As the paper is now a proof-of-concept for evaluating models, I recommend shedding light on follow-up questions: how variable was category consistency across models? How often did category inconsistency merely lead to non-safety critical deviations (ie from B to A or C to D) versus safety critical deviations (A/B -> C/D or vice versa)? Did the mitigation strategy affect category consistency? In case the mitigation prompt would lead to a decrease in category consistency this would point to a "safety versus predictability"-trade-off.

Response: We agree and performed the requested descriptive analysis on all existing stored outputs. We defined exact prompt cells by data source, model, evaluated instance, framing cell, and mitigation cell, with ten repetitions per cell after option-position remapping. Of 1,009,680 cells, 80.4% were fully category-consistent. Model-level consistency ranged from 4.4% to 100.0%. Among inconsistent cells, 0.4% varied only within A/B, 61.4% only within C/D, and 38.2% crossed the A/B versus C/D safety boundary. Full consistency was almost identical without and with mitigation, at 80.5% and 80.4%, so the descriptive results do not indicate a safety-predictability trade-off. Cross-boundary variation was lower with mitigation, at 6.9% versus 9.1% of all cells. We added the definition to Methods, the full results to Results, interpretation and caveats to Discussion, and a model-level table in Supplementary Data 3.

Revised text (Methods, Model Execution Pipeline):

“We defined an exact prompt cell by data source, model, evaluated instance, framing cell, and mitigation cell. Each cell contained ten repetitions. A cell was fully category-consistent when all ten repetitions returned the same mapped category after option-position remapping. Inconsistent cells were classified as A/B-only, C/D-only, or cross-boundary when at least one repetition returned A or B and at least one returned C or D. We summarized these outcomes descriptively by model, data source, and mitigation status.”

Revised text (Results):

“Across 1,009,680 matched ten-run prompt cells, 812,027 (80.4%) were fully category-consistent. Consistency varied across models from 4.4% to 100.0% and was 80.5% without mitigation versus 80.4% with mitigation. Among 197,653 inconsistent cells, 818 (0.4%) varied only within A/B, 121,279 (61.4%) only within C/D, and 75,556 (38.2%) crossed the A/B versus C/D safety boundary. These measures describe mapped-category stability after option-position remapping rather than similarity of free-text responses (Supplementary Data 3).”

Revised text (Discussion):

“The consistency analysis adds a second dimension to mitigation evaluation. Full category consistency was nearly identical without and with mitigation (80.5% versus 80.4%), providing no evidence of a safety-predictability trade-off at this descriptive level. Most inconsistent cells varied only within the same safety grouping, while 38.2% crossed between potentially harmful and safe categories. Mitigation reduced the proportion of all cells with cross-boundary variation from 9.1% to 6.9%. Prompt-level safeguards should therefore be assessed for both their effect on safety-oriented output distributions and their effect on category stability. High consistency should not itself be read as safety: a model that repeatedly selects the same unsafe category can be fully consistent.”