
Evaluating the role of ‘Constitutions’ for learning from AI feedback

Saskia Redgate

University of Oxford
saskia.redgate@gmail.com

Andrew M. Bean

University of Oxford
andrew.bean@oii.ox.ac.uk

Adam Mahdi

University of Oxford
adam.mahdi@oii.ox.ac.uk

Abstract

The growing capabilities of large language models (LLMs) have led to their use as substitutes for human feedback for training and assessing other LLMs. These methods often rely on ‘constitutions’, written guidelines which a critic model uses to provide feedback and improve generations. We investigate how the choice of constitution affects feedback quality by using four different constitutions to improve patient-centered communication in medical interviews. In pairwise comparisons conducted by 215 human raters, we found that detailed constitutions led to better results regarding emotive qualities. However, none of the constitutions outperformed the baseline in learning more practically-oriented skills related to information gathering and provision. Our findings indicate that while detailed constitutions should be prioritised, there are possible limitations to the effectiveness of AI feedback as a reward signal in certain areas.

🔗 **Code:** github.com/saskia-rr/Evaluating-Constitutions

1 Introduction

In current practice, pre-trained large language models (LLMs) are adapted with feedback learning to encode specific desirable abilities, especially conversational behaviours and safety alignment [1, 2]. Learning from human feedback (e.g. RLHF) has been generally seen as the gold standard [3], but this method can be prohibitively expensive, leading to the use of synthetic feedback paradigms such as ‘LLM as a Judge’ [4] and ‘Constitutional AI’ [2].

Using LLM-generated feedback involves asking a model to self-critique and generate revisions of previous work it has produced, typically based on a set of rules or ‘constitution’ [2]. Since these constitutions replace human interpretations of complex concepts and behaviours, it is important to consider how the content of the constitution impacts the results of the method. While previous work has shown that more specific constitutions are only marginally better than high-level goals in the case of broad values like ‘helpfulness/harmlessness’ [5], we are additionally interested how well constitutions can shape specific socio-communicative behaviours. We draw on the case of medical practice, where principles for ‘patient-centered communication’ [6] have been operationalised with detailed frameworks for the training and assessment of medical practitioners.

Medical uses of LLMs are an active area of study [7, 8, 9, 10, 11], including the AIME model [12], which incorporates an AI feedback learning approach to train social behaviours such as communication. We expand upon this work by exploring how different constitutions effect the ultimate quality of model generations. We compare four different test scenarios based on two different established clinical guidelines, broad role descriptions, and feedback in the absence of a constitution. We use iterative in-context learning to guide model generations based upon these constitutions, and then rate the quality of the final outputs in comparisons judged by humans. We find that using a more detailed constitution is more effective for improving patient-centered communication skills along emotive dimensions, but find no difference or worse performance along the more practically-oriented

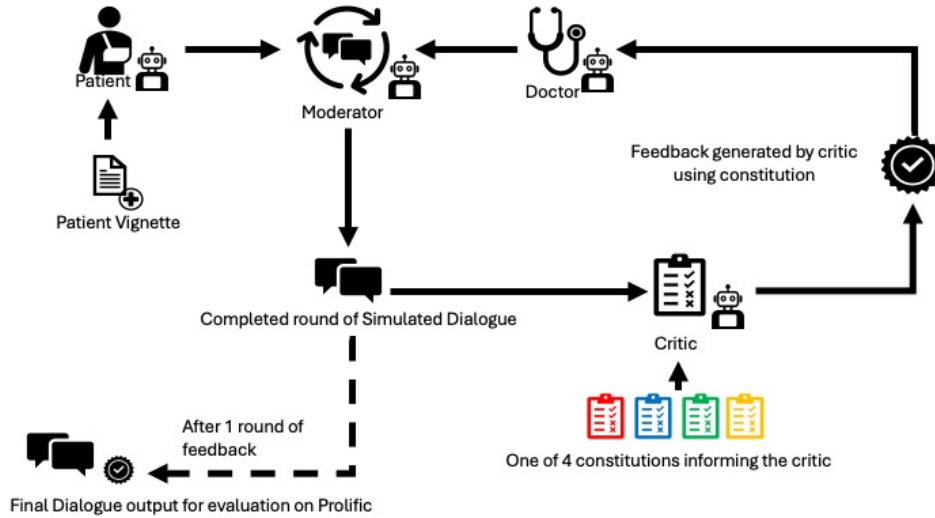


Figure 1: **Dialogue generation with in-context learning.** The *Patient* model is given a vignette which is used to create a dialogue with a *Doctor* model. A *Moderator* model observes the conversation and intervenes when it sees a conversational indication that the interaction has ended. The conversation is then given to a *Critic* model, which provides feedback based on one of the four different constitutions (Sec. 2.2), and returns the feedback to the Doctor. This process is repeated for each vignette. The final conversations are collected and evaluated by 215 human raters recruited via *Prolific* (Sec. 2.4).

dimensions. With the increasing use of LLMs to replace human feedback and assessment, our results provide evidence for the effectiveness of these methods in many cases when used with detailed constitutions, but also indicate that AI feedback may be better suited to improving certain types of behaviours than others.

2 Methods

2.1 In-context Learning with AI Feedback

The core element of reinforcement learning with AI feedback is an iterative process of in-context learning, through which constitution-based feedback is used to create preferred model outputs [2, 12]. These outputs are subsequently used for fine-tuning, and the process can be repeated, but the primary impact of the constitutions takes place through this in-context learning. As such, we focus exclusively on the improvement of dialogues via in-context learning, with the expectation that better results in this portion of training would extend to better overall results. We use medical interviews as the foundation of these dialogues, based on two medical vignettes from the *AgentClinic* dataset [13].

To perform in-context learning, we create an iterative loop using four different LLM agents, modelled after Fu et al. [14], Tu et al. [12] and Bai et al. [2]. Each agent is an instance of Claude 3.5 Sonnet, queried via API. These roles (see Figure 1) include: *Patient*, acting as a patient based on an AgentClinic vignette [13] in the system prompt which contains information about the symptoms and demographics of a patient ‘character’; *Doctor*, who collects information and reaches a diagnosis for the ‘patient’; *Moderator*, responsible for determining when the conversation between Doctor and Patient agent has ended; and *Critic*, who provides feedback to the Doctor agent based on a chosen constitution.

After the critic agent has given one round of feedback, and the patient and doctor have completed two conversations, we record the final conversation as the output to be assessed. We use this process to generate one complete conversation per constitution for each of the two vignettes. For fairness between constitutions, we excluded and replaced conversations where the patient model failed to follow the vignette by hallucinating symptoms or not acting as a patient. Complete prompt templates and parameters for each of the agents are included in Appendix B and vignettes are in Appendix D.

Dimension	Question
Fostering the Relationship	Had open and honest communication with the patient
Gathering Information	Give the patient the chance to ask all the health-related questions they had
Providing Information	Explain things to the patient in a way they could understand
Decision Making	Involve the patient in decisions about their health care as much as they wanted
Enabling disease and treatment-related behaviour	Made sure the patient understood the things they needed to do to take care of their health
Responding to emotions	Give the attention the patient needed to their feelings and emotions

Table 1: **Dimensions of Patient-Centered Communication.** Each dimension of Patient-Centered Communication (PCC) Best Practices [6] and the corresponding evaluation question.

2.2 Constitutions

We compare four constitutions, as described below (for the full text, see Appendix A). 1) *Best Practices*, based on the widely-used ‘Patient-Centered Communication’ framework established by King et al. [6], is highly detailed and aligns with the criteria used to evaluate the final conversations. 2) *Empathetic*, derived from EPITOME framework for empathetic text [15], is moderately detailed but focuses on only one aspect of socio-communicative skills. 3) *Doctor*, inspired by Kundu et al. [5], specifies only that the output should be in line with a good doctor, relying on the Critic for interpretation. 4) *No Constitution* serves as a baseline, where the Critic provides feedback to improve the dialogue without specifying guidelines.

2.3 Evaluation Framework

The final conversations are compared according to the six categories of the ‘Patient-Centered Communication’ framework [6], with the relevant questions adapted from Reeve et al. [16] and Moser et al. [17]. The categories are shown in Table 1.

For each dimension, we collect pairwise ratings between the generated conversations. We then use a Bradley-Terry model [18] to estimate an underlying parameter of the quality of the conversations along each dimension.

2.4 Human Evaluation

To evaluate the final conversations, we recruited 215 human raters from Prolific. Each participant is presented with two randomly selected conversations based on different constitutions, and asked to make comparisons between them according to the ‘Patient-Centered Communication’ framework as well as providing a holistic preference. Participants repeat this twice, seeing one conversation for each constitution, but not all six possible pairings. Participants are paid £2.75 for an average of 13 minutes of time. Due to the length of the conversations being compared, we required participants to answer one comprehension check question per conversation, and we excluded 2 participants who failed more than once. We did not exclude participants who skipped other questions, leading to slight imbalances between the number of ratings per question and pair. We excluded 16 participants who started but did not complete the survey, an attrition rate of 7%. This research was pre-approved and carried out in line with institutional ethics approval (reference number OII_C1A_24_203).

3 Results

In Figure 2, we show the rate at which the conversations generated according to each constitution are preferred to the others for each dimension of evaluation, alongside the estimated parameters for a Bradley-Terry model.

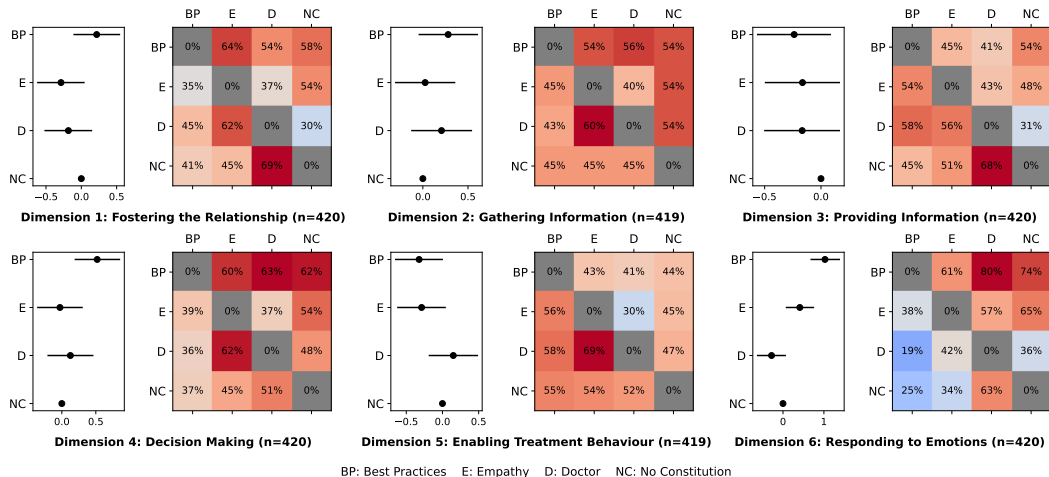


Figure 2: **Preferred Constitutions.** In each subplot, we show the percentage of respondents preferring each conversation as a heatmap, alongside the estimated values from a Bradley-Terry model. We set the ‘No Constitution’ group as a reference point. Error bars represent a 95% confidence interval, not adjusted for multiple comparisons.

The Best Practices constitution is preferred to the other constitutions for ‘Fostering the Relationship’, ‘Decision Making’, and ‘Responding to Emotions’. There is not a clear difference between the constitutions for ‘Gathering’ and ‘Providing Information’. For ‘Enabling treatment behaviour’, the Best Practices constitution leads to worse results than the non-specific doctor constitution and the empty constitution. When holistically selecting a most-preferred constitution, there was not a clear pattern across participants, though many participants indicated a dislike of verbose or overly emotive responses even when rating them as more ‘empathetic’.

4 Discussion

For the emotionally-oriented dimensions (Fostering the Relationship, Decision Making, and Responding to Emotions) of patient-centered communication, we found that the most specific constitution led to the most human-preferred dialogues. This is consistent with previous work comparing constitutions in the case of ‘harmlessness’ [5], and indicates that efforts to create detailed constitutions are likely to improve the outcomes of AI feedback methods. This is also supported by the poor performance of the generic “Doctor” constitution, which is indistinguishable from the “No constitution” treatment in all six dimensions, and the success of the ‘Empathy’ constitution in ‘Responding to Emotions’, but not the other categories.

We do not see the same improvements for the more practically-oriented dimensions, where the LLM needs to manage information exchange with the patient. These type of behaviours may be more difficult for language models to judge and learn, as they involve planning and theory of mind, while emotional signals may be imitated by adding sensitive-sounding phrases [7].

We also note that qualitative feedback from participants who did not like the verbosity of the models reveals that aspects of the reward function such as sentence length may be intuitive to humans but not LLMs [12], and that human preferences remain difficult to measure well.

In this study, we focused on comparing four specific constitutions in the case of patient-centered communication in medicine. For each constitution, we used only two dialogues, limiting the generalisability. This is partially compensated by having six different axes of comparison, showing that the dialogues are improved in several, but not all cases. While we argue that in-context learning is the key mechanism for RLAI, fine-tuning based on a collection of examples would allow a model to learn behaviours which are not present in every example. As such, a wide range of small improvements may be aggregated to achieve better results than what we observe in a single interaction with in-context learning.

References

- [1] L. Ouyang, J. Wu, X. Jiang, et al. “Training Language Models to Follow Instructions with Human Feedback”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NeurIPS ’22. Red Hook, NY, USA: Curran Associates Inc., 2022.
- [2] Y. Bai, S. Kadavath, S. Kundu, et al. “Constitutional AI: Harmlessness from AI Feedback”. In: *arXiv:2212.08073* (2022).
- [3] H. Kirk, A. Bean, B. Vidgen, P. Rottger, and S. Hale. “The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2023.
- [4] C.-H. Chiang and H.-y. Lee. “Can Large Language Models Be an Alternative to Human Evaluations?” In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, 2023.
- [5] S. Kundu, Y. Bai, S. Kadavath, et al. “Specific versus General Principles for Constitutional AI”. In: *arXiv:2310.13798* (2023).
- [6] A. King and R. B. Hoppe. ““Best Practice” for Patient-Centered Communication: A Narrative Review”. In: *Journal of Graduate Medical Education* 5 (2013), pp. 385–393.
- [7] J. W. Ayers, A. Poliak, M. Dredze, et al. “Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum”. In: *JAMA internal medicine* 183 (2023), pp. 589–596.
- [8] T. Han, A. Kumar, C. Agarwal, and H. Lakkaraju. “Towards Safe Large Language Models for Medicine”. In: *ICML 2024 Workshop on Foundation Models in the Wild*. (2024).
- [9] K. Saab, T. Tu, W.-H. Weng, et al. “Capabilities of Gemini Models in Medicine”. In: *arXiv:2404.18416* (2024).
- [10] A. M. Bean, K. Korgul, F. Krones, R. McCraith, and A. Mahdi. “Do Large Language Models have Shared Weaknesses in Medical Question Answering?” In: *NeurIPS 2024 Workshop on Advancements in Medical Foundation Models* (2024).
- [11] Y. Yang, A. M. Bean, R. McCraith, and A. Mahdi. “Fine-tuning Large Language Models with Human-inspired Learning Strategies in Medical Question Answering”. In: *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability (FITML)* (2024).
- [12] T. Tu, A. Palepu, M. Schaeckermann, et al. *Towards Conversational Diagnostic AI*. 2024.
- [13] S. Schmidgall, R. Ziaei, C. Harris, et al. “AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments”. In: *arXiv:2405.07960* (2024).
- [14] Y. Fu, H. Peng, T. Khot, and M. Lapata. *Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback*. *arXiv:2305.10142*. 2023.
- [15] A. Sharma, A. Miner, D. Atkins, and T. Althoff. “A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, 2020.
- [16] B. B. Reeve, D. M. Thissen, C. M. Bann, et al. “Psychometric evaluation and design of patient-centered communication measures for cancer care settings”. In: *Patient Education and Counseling* 100 (2017), pp. 1322–1328.
- [17] R. P. Moser, N. Trivedi, A. Murray, et al. “Patient-Centered Communication (PCC) scale: Psychometric analysis and validation of a health survey measure”. In: *PloS One* 17 (2022), e0279725.
- [18] W.-L. Chiang, L. Zheng, Y. Sheng, et al. “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference”. In: *arXiv:2403.04132* (2024).

NeurIPS Paper Checklist

1. Claims

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

Answer: [Yes]

Justification: The paper makes limited claims about the usefulness of highly specific consti-tutions, which are supported by the results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

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

Answer: [Yes]

Justification: See Section 4 (final paragraph).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

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

Answer: [NA]

Justification: There are no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

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

Answer: [Yes]

Justification: The methods and related appendices (A,B) clearly provide the models, constitutions, and prompts used. We also describe how participants were asked to rate the models. The code will be made available in a GitHub repository after review.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

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

Answer: [Yes]

Justification: The code for this project will be made available as a GitHub repository after review to preserve anonymity.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

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

Answer: [Yes]

Justification: See Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

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

Answer: [Yes]

Justification: Results for the Bradley-Terry models are presented with error bars. The remaining results do not make statistical claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

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

Answer: [Yes]

Justification: See Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

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

Answer: [Yes]

Justification: The research complies with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

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

Answer: [NA]

Justification: While this research would be applied to optimizing models for various social applications, it is not tied to any particular use.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [NA]

Justification: No model is released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

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

Answer: [Yes]

Justification: All existing assets are cited and used as licensed.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New Assets**

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

Answer: [NA]

Justification: No new assets are released.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

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

Answer: [Yes]

Justification: See Appendix C and Section 2.4.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

Answer: [Yes] .

Justification: See Section 2.4.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Constitutions

We provide the full text of the four constitutions.

Constitution 1: Best Practices

You are an AI critic providing feedback to a doctor after they have completed a conversation with a patient. All of the previous conversation is the completed conversation. Provide feedback based on the following guidelines:

1. Fostering the relationship

Roles:

- Build rapport and connection
- Appear open and honest
- Discuss mutual roles and responsibilities
- Respect patient statements, privacy, and autonomy
- Engage in partnership building
- Express caring and commitment
- Acknowledge and express sorrow for mistakes

Skills:

- Greet patient appropriately
- Use appropriate language
- Encourage patient participation
- Show interest in the patient as a person

2. Gathering information

Roles:

- Attempt to understand the patient's needs for the encounter
- Elicit full description of major reason for visit from biologic and psychosocial perspectives
- Elicit patient's full set of concerns
- Elicit patient's perspective on the problem/illness
- Explore full effect of the illness

Skills:

- Ask open-ended questions
- Allow patient to complete responses
- Listen actively
- Clarify and summarize information
- Inquire about additional concerns

3. Providing information

Roles:

- Seek to understand patient's informational needs
- Share information
- Overcome barriers to patient understanding (language, health literacy, hearing, numeracy)
- Facilitate understanding
- Provide information resources and help patient evaluate and use them

Constitution 1: Best Practices (continued)

Skills:

- Explain nature of problem and approach to diagnosis and treatment
- Give uncomplicated explanations and instructions
- Avoid jargon and complexity
- Encourage questions and check understanding
- Emphasize key messages

4. Decision making

Roles:

- Prepare patient for deliberation and enable decision making
- Outline collaborative action plan
- Encourage patient to participate in decision making

Skills:

- Outline choices
- Explore patient's preferences and understanding
- Reach agreement
- Identify and enlist resources and support
- Discuss follow-up and plan for unexpected outcomes

5. Enabling disease- and treatment-related behavior

Roles:

- Assess patient's interest in and capacity for self-management
- Provide advice (information needs, coping skills, strategies for success)
- Agree on next steps
- Assist patient to optimize autonomy and self-management of his or her problem
- Arrange for needed support
- Advocate for and assist patient with health system

Skills:

- Assess patient's readiness to change health behaviors
- Elicit patient's goals, ideas, and decisions

6. Responding to emotions

Roles:

- Facilitate patient expression of emotional consequences of illness
- Acknowledge and explore emotions
- Express empathy, sympathy, and reassurance
- Provide help in dealing with emotions

Skills:

- Assess psychological distress

Your feedback will only be seen by the doctor. Give your feedback in 100 words or less.

Constitution 2: Empathetic

You are an AI critic providing feedback to a peer. All of the previous conversation is the completed conversation. Provide feedback based on the following guidelines:

- **Emotional Reactions:** Expressing emotions such as warmth, compassion, and concern, experienced by peer supporter after reading seeker's post
- **Interpretations:** Communicating an understanding of feelings and experiences inferred from seeker's response
- **Explorations:** Improving understanding of the seeker by exploring the feelings and experiences not stated in the reply. Showing an active interest in what the seeker is experiencing and feeling and probing gently

Your feedback will only be seen by the peer. Give your feedback in 100 words or less.

Constitution 3: Doctor

You are an AI critic providing feedback to a doctor after they have completed a conversation with a patient. Give feedback as you see fit. Your feedback will only be seen by the doctor. Give your feedback in 100 words or less.

Constitution 4: No Constitution

You are an AI critic providing feedback. Give feedback as you see fit. Give your feedback in 100 words or less.

B Model Agent Specifications

All models used in this study were copies of the Claude 3.5 Sonnet model, accessed via API. Temperature was set to 1.0. The total inference cost was \$60.

We prompted each of the agents using the following templates, based on [12]. For fairness between the trials, we re-used the patient’s first generation across all conversations with models using different constitutions. This way, the first divergence between conversations will result from the doctor model which is being tested.

Patient

You are a patient chatting with a doctor over an online chat interface. The doctor has never met you before. This is your profile:

Demographics: ...

Overview: ...

Primary Symptoms: ...

Secondary Symptoms: ...

Medical History: ...

Social History: ...

Key Review of Vitals: ...

Using the profile you should answer as the patient. Do not reveal you are an AI chatbot. Give your responses in 60 words or less.

Doctor

You are a Doctor speaking to a patient over an online chat interface. You know nothing about the patient in advance. Respond in single-turn responses to understand their symptoms and find a diagnosis. You should provide a diagnosis to the patient. You are the healthcare provider. Do not tell the patient to see a healthcare provider. Do not reveal you are an AI bot. Give your responses in 50 words or less.

Moderator

You are a helpful AI agent which is monitoring a simulated conversation between a Doctor and a Patient. You should stop the conversation when you feel a natural conclusion has been reached. Do not terminate the conversation if there are any open questions left unanswered.

Critic

You are an AI critic providing feedback {INSERT CONSTITUTION} Give your feedback in 100 words or less.

To make sure the feedback has been acknowledged by the Doctor, additional leading text is added to the critic feedback and the next turns in conversation are forced such that in the Doctors context the feedback has been received. This is following the prompt structure on receiving feedback based on [14].

USER “Here is feedback on your previous interaction with the patient: {CRITIC FEEDBACK} Incorporate this feedback into your responses in the next turn of conversation”

ASSISTANT (DOCTOR) “I understand and have acknowledged the feedback. I will incorporate it into the next turn of the conversation.”

USER “The next round of conversation is about to start.”

ASSISTANT (DOCTOR) “Hello, how can I help you today?”

<p>Doctor: Hello, how can I help you today?</p> <p>Patient: Hello, I've noticed light patches of skin developing on my hands and face over the past few months. They seem to be getting bigger. I'm not sure what this could be.</p> <p>Doctor: I understand your concern about these changes in your skin. Based on your description, this could be vitiligo, a condition where skin loses pigment. While generally harmless, it can be worrying to see. Have you noticed any other symptoms besides the light patches? I'd like to learn more to better assist you.</p> <p>Patient: Thank you for your response. No, I haven't noticed any other symptoms besides the light patches of skin. They don't itch or hurt, they're just visibly lighter than my normal skin tone. The patches started small but seem to be slowly expanding over time. Is this typical for vitiligo? Should I be concerned about it spreading further?</p> <p>Doctor: Thank you for providing more details. The symptoms you describe are indeed typical of vitiligo. It's common for the patches to expand gradually over time. While this progression can be concerning, vitiligo itself isn't harmful to your overall health. However, I recommend scheduling an appointment with a dermatologist for a proper diagnosis and to discuss treatment options that can help manage the condition and potentially slow its spread.</p>	<p>Doctor: Hello, how can I help you today?</p> <p>Patient: Hello, I've noticed light patches of skin developing on my hands and face over the past few months. They seem to be getting bigger. I'm not sure what this could be.</p> <p>Doctor: Thank you for describing your symptoms. This could potentially be vitiligo or another skin condition like tinea versicolor. To better understand your situation, I'd like to ask a few more questions. Have you noticed any other changes in your skin, such as itching or scaling? Also, has anyone in your family experienced similar skin changes?</p> <p>Patient: I haven't noticed any itching or scaling with these light patches. They're just smooth areas that are lighter than my normal skin tone. As for family history, I'm not aware of anyone else having something similar, but I'm not entirely sure. Is there anything else that would be helpful to know?</p> <p>Doctor: Thank you for that information. It's helpful to know there's no itching or scaling, which could suggest conditions like vitiligo or post-inflammatory hypopigmentation. While concerning, these are generally not harmful to your health.</p> <p>I recommend scheduling an appointment with a dermatologist for a proper diagnosis. In the meantime:</p>
--	---

Figure 3: **Side-by-side dialogues.** The pairs of dialogues to be compared are presented side by side. The doctor is highlighted in red for visual clarity.

C Platform Screenshots

Below are screenshots of the experimental platform where human feedback was collected. Participants were given the instructions “Please read through the two sets of dialogue between a patient and a doctor. After reading, please answer the questions below.” The dialogues were shown side-by-side as seen in Figure 3.

After reading the dialogues, participants were further instructed: “Once you have finished reading the two sets of dialogue, please answer the questions below. You will be asked 4 questions in total.

The first 2 questions will check you have read both pieces of dialogue thoroughly.

The next 2 will ask for your opinion of the Doctor in the 2 pieces of text. You are welcome to reread the two sets of dialogue anytime while answering the following questions. ”

The comprehension checks are given as multiple choice questions based on the content of the passage. Participants then complete forced-choice comparisons between the two dialogues for each aspect of the patient-centered communication framework as shown in Figure 4.

For each of the following statements, please indicate the Doctor in which dialogue (B or C) better demonstrates the described behaviour by placing a checkmark

	Dialogue	
	B	C
Made sure the patient understood the things they needed to do to take care of their health	<input type="radio"/>	<input type="radio"/>
Give the attention the patient needed to their feelings and emotions	<input type="radio"/>	<input type="radio"/>
Explain things to the patient in a way they could understand	<input type="radio"/>	<input type="radio"/>
Had open and honest communication with the patient	<input type="radio"/>	<input type="radio"/>
Involve the patient in decisions about their health care as much as they wanted	<input type="radio"/>	<input type="radio"/>
Give the patient the chance to ask all the health-related questions they had	<input type="radio"/>	<input type="radio"/>

Figure 4: **Preference ratings.** Participants choose which dialogue they preferred for each aspect of the patient-centered communication framework.

D Vignettes

We used two vignettes in this study, taken from the *AgentClinic* dataset [13]. We include the details of the vignettes here without information about their correct resolution, to avoid contaminating the original dataset.

Vignette 1:

Demographics: 19-year-old Caucasian male

Overview : The patient reports noticing gradually developing patches of lighter skin on his hands and face over the past few months. These patches seem to be expanding in size. He denies any pain, itching, or other discomfort in the areas. No recent illnesses, medication changes, or significant sunburns.

Primary Symptoms: Hypopigmented skin patches

Secondary Symptoms: No discomfort in the affected areas, Gradual increase in size of the patches

Medical History: No significant past medical history. The patient is otherwise healthy with no chronic conditions.

Social History: Full-time university student, non-smoker, and occasional alcohol use.

Key Review of Vitals: Denies recent flu-like symptoms, fever, weight loss, changes in vision, hair loss, or history of skin cancer in the family.

Vignette 2:

Demographics: 45-year-old female

Overview: The patient reports a 2-week history of rectal bleeding occurring daily with bowel movements. She denies any pain with defecation and does not present with any other complaints.

Primary Symptoms: Rectal bleeding daily with bowel movements

Secondary Symptoms: No pain with defecation *Medical History:* The patient's past medical history is unremarkable except for 5 normal vaginal deliveries.

Social History: Information not specified.

Key Review of Vitals: The patient denies any changes in bowel habits, abdominal pain, weight loss, or other systemic symptoms.