Reward Modeling with AI Feedback

Aman Singhal*
Courant Institute of Mathematical Sciences
New York University
as14661@nyu.edu

1 Introduction

The use of Large Language Models (LLMs) in healthcare challenges traditional trust, because generated outputs are untraceable and sometimes inaccurate [Clusmann et al., 2023]. LLMs are trained on conflicting sources of information through the maximum likelihood objective in an unsupervised fashion [Zhao et al., 2023]. This is typically followed by a supervised fine-tuning objective and several updates of Reinforcement Learning from Human Feedback (RLHF)[Brown et al., 2020]. Prior work such as Casper et al. [2023] and Durmus et al. [2023] mapped the biases in human annotator preferences in the RLHF stage as the source of trickle-down biases in LLMs. In this work we explore if inaccurate RLHF objectives are the reason for "untruthfulness" in LLM responses, specifically for healthcare.

Several works like [Zhang et al., 2023] have explored alternative strategies such as involving a quorum of annotators instead of individual annotators. We argue that despite such remedial works, the RLHF objective is fundamentally misaligned for healthcare. This is because annotators are instructed to pick "helpful" responses which might not translate to truthful responses. Moreover, such methods cannot overcome human biases such as equating longer responses to be more valuable [Nakano et al., 2021]. This hypothesis is corroborated by the results in Table 6, where random guessing is better aligned with a truthful response than existing reward models.

Moreover, generating alignment data at scale is inherently expensive for healthcare [Bai et al., 2022a]. We leverage LLM generated data to train reward models capable of distinguishing truthful responses from untruthful ones for BioMedical questions.

2 Methodology

We have a dataset of medical entrance level MCQ questions from examinations such as AIIMS, NEET, and USMLE etc. We aim to utilize prevalent and open-source large language models to generate multiple answers for each question. Each generated answer for a question assumes a different option as the correct option. Hence for a question with four options, we generate four answers, three of which are untruthful and only one is truthful (Table 1).

2.1 LLAMA-2

Meta AI's large language model LLAMA-2 [Touvron et al., 2023] is available in three parameter size (1) 7B (2) 13B and (3) 70B. We leverage the 13B parameter model for generating answers. The choice is empirical since the 13B parameter model is more expressive than the 7B parameter model. Moreover, its ability to adhere to instructions for generating untruthful reasoning is better than the 70B model.

*https://github.com/AmanSinghal927/MLfH

Table 1: Generation Example

Field	**Details**
Question	A 14-16 gauge needle is most likely to be used for:
Correct Answer	Trauma or burns patients.
Options	A) Children. B) Elderly patients. C) Inserting in the back of the hand. D) Trauma or burns patients.
Subject Name	Clinical Knowledge
Generation A)	A 14-16 gauge needle is appropriate for children because it is a smaller gauge needle that is less likely to cause pain and discomfort. Additionally, it is easier to insert and maneuver in smaller veins.
Generation B)	A 14-16 gauge needle is larger than the average needle, and the elderly may have thinner skin and more fragile tissue, making it more difficult to insert smaller needles. Additionally, the elderly may have more underlying medical conditions, such as diabetes or heart disease, that may require larger needles.
Generation C)	A 14-16 gauge needle is appropriate for inserting into the back of the hand because it is a smaller needle that is less likely to cause tissue damage or bleeding. It is also easier to insert and remove from the hand than larger needles.
Generation D)	A 14-16 gauge needle is thicker and more robust than smaller gauge needles, making it better suited for inserting into tissue that is more dense or has a higher risk of bleeding, such as in trauma or burns patients. Additionally, the larger size of the needle makes it easier to thread through tissue and insert into the body.

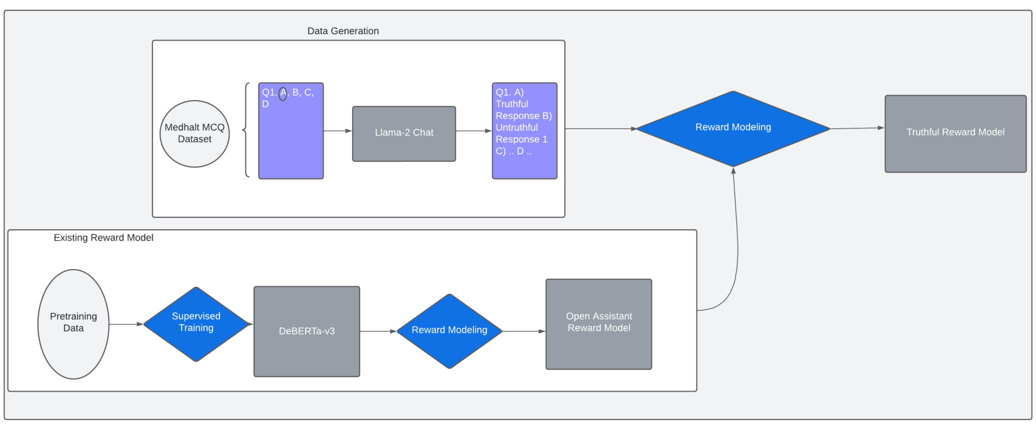

Figure 1: Overall Architecture

2.2 Prompting

We explore two prompting strategies to generate responses. The dataset generated is aliased Llama-MedGen Q&A. (LMQA). The template for both the zero-shot [Kojima et al., 2022] strategy and the few-shot Reynolds and McDonell [2021] strategy are shared in the Appendix. Table 2 compares the edit distance of truthful responses with respect to untruthful responses as a measure of diversity.

2.3 Question Dataset

The questions are derived from the Medhalt dataset False Confidence Test [Umapathi et al., 2023]. The dataset contains 18K multiple choice questions. Each question is accompanied by four potentially correct answers. Table 3 highlights the spread of examinations for sourcing these questions. An

Table 2: Data Generation Strategies for Llama-MedGen Q&A (LMQA)

Metric	Zero-shot	Few-shot
Avg # Tokens/Generation	97	105
Edit Distance	0.39	0.41
Edit Distance/Token	0.0041	0.0039
Total Number of Tokens	195K	210K

examination-style dataset was chosen intentionally. Competitive exams such as USMLE comprise questions with choices that are confusing for aspiring practitioners [Schwartz et al., 2018]; suitable for sourcing untruthful responses from LLAMA-2 13B organically.

Table 3: Questions Dataset: MedHalt

Exam	**Number of Questions**
AIIMS PG (India)	6660
NEET PG (India)	2855
Examenes Médica (Spain)	4068
TWMLE (Taiwan)	2801
USMLE (U.S.)	2482
Total	18866

2.4 Reward Modeling

For reward modeling we experiment with DeBERTa-v3 [He et al., 2021] base & large and LLAMA-2 7B. The DeBERTa-v3 model is pre-trained using disentangled attention mechanism and enhanced mask decoder using Wiki+Book OpenWebText Stories CC-News datasets (approx. 700M rows and 141 GBs of Data). Model weights are initialized from Open-Assistant's [Köpf et al., 2023] reward model fine-tuned on 69K alignment data. The LLAMA-2 7B model is initialized by replacing the LM head with a randomly initialized regression head. We leverage the following contrastive loss to adapt the reward models to LMQA:

$$loss = -\log \sigma(r_{chosen} - r_{rejected}) \tag{1}$$

where r_{chosen} and $r_{rejected}$ represent the output of the regression head output for truthful and untruthful responses.

2.5 $r_{rejected}$

We experiment with using all three rejected responses as a part of the loss function (equation 2) as well as using only one of the rejected responses per question for updating reward model parameters.

To decide a criteria for picking only one rejected response we manually annotate a hundred questions and pick the most believable yet untruthful options for each of them. We benchmark strategies such as least ROUGE score [Lin, 2004] and minimum cosine similarity (BERT) w.r.t the question (to capture diversity), maximum length and maximum next-sentence prediction probability [Devlin et al., 2018] given the question. Unsurprisingly untruthful responses with the maximum token length are also the most convincing ones [Nakano et al., 2021] Table 4.

2.6 Training Details

Models are trained on zero-shot LMQA as well as few-shot LMQA. We choose a maximum sequence length of 512 tokens based on the distribution in Fig. 3. The fine-tuning strategy is based on Low-Rank Adaptation (LoRa) [Hu et al., 2021] with a rank set to 8. The learning rate was initially set between 1×10^{-5} and 1×10^{-6}. To enable larger effective batch sizes, ranging from 64 to 512, gradient accumulation was utilized. Two learning rate schedulers were employed: linear and cosine annealing. The warmup period was set to 10% of the total number of training steps. The training

Table 4: Chosing $r_{rejected}$

Metric	Overlap with human preference
Minimum Cosine Similarity (BERT)	43.75%
Lowest ROUGE-1	32.14%
Lowest ROUGE-L	32.14%
Maximum NSP (BERT)	50.0%
Maximum Length	51.19%

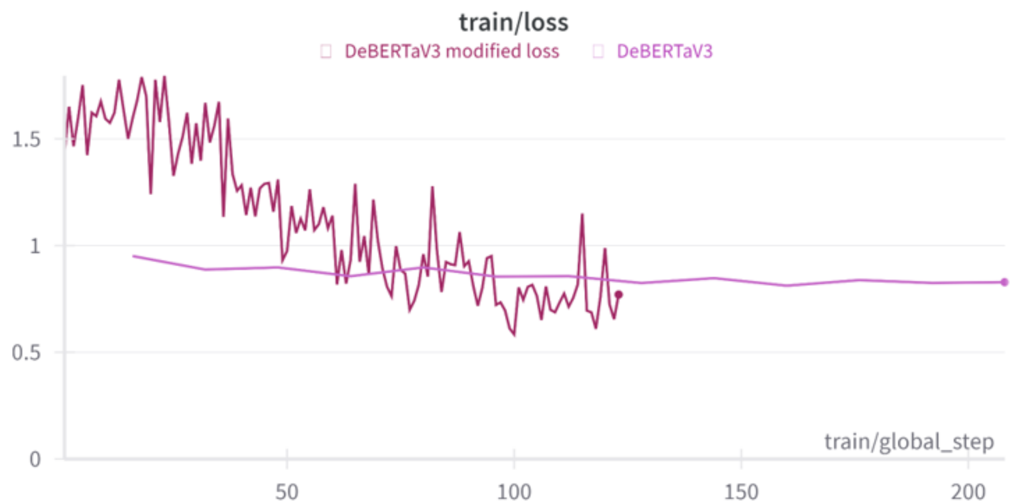

Figure 2: Training Loss DeBERTa-v3 large with modified loss vs normal loss

process was confined to a single epoch based on details mentioned in [Touvron et al., 2023]. Only the best results are reported. Figure 2 shows the training loss curve for the best performing model DeBERTA-v3 large using (1) and a modified loss function (Equation 2). The noise in training loss curves is expected [Touvron et al., 2023] since the loss is captured across batches of the same epoch. However, an overall downward trend is observed in both cases.

$$loss = -\log \sigma(r_{chosen} - r_{rejected_0} - r_{rejected_1} - r_{rejected_2}) \quad (2)$$

where $r_{rejected_0}$, $r_{rejected_1}$ and $r_{rejected_2}$ represent the output of the regression head for all three untruthful responses.

3 Evaluation

We evaluate all models on a held-out LMQA test set. Moreover, we evaluate using the MMLU test [Hendrycks et al., 2020] splits for anatomy, clinical knowledge, medical genetics, professional medicine, college biology and college medicine (Table 5) to quantify out-of-distribution generalization. We check the data leakage between the MMLU test set and the MedHalt training set using edit distance to find an insignificant 0.37% overlap.

3.1 Few-shot vs Zero-shot

First, we evaluate the baseline models on the LMQA test set generated using zero-shot and few-shot prompting strategies. The pre-trained reward models distinguish better between truthful and untruthful responses when the responses are generated using few-shot prompting strategies. This corroborates with the finding in Table 2 that few-shot prompting leads to more diverse responses. Hence going

Table 5: MMLU Test: Distribution

Subject	Number of Questions
Anatomy	135
Clinical Knowledge	265
College Biology	144
College Medicine	173
Medical Genetics	100
Professional Medicine	272
Total	1089

forward we compare model results only on the MMLU test and LMQA test sets generated using few-shot prompting.

Table 6: Performance of baseline models on LMQA

Model	LMQA test (zero-shot)	LMQA test (few-shot)
Random Prediction	50.0%	50.0%
DeBERTA-v3 Base	43.6%	46.1%
DeBERTA-v3 Large	46.0%	49.8%

3.2 Model Size

From Table 6 it is evident that larger models are able to distinguish between truthful and untruthful responses better than smaller models.

3.3 Loss Function

The DeBERTa-v3 large model trained using the modified loss (Equation 2) achieves better loss convergence (Figure 2) and better performance on LMQA test set as well as the MMLU test set. Thus modifying the loss function does not only help sample efficiency but also improves generalization.

Table 7: DeBERTa-v3 Large Performance. Random predictions yield 50% and 25% accuracy for for LMQA test and MMLU test respectively

Model	LMQA test	MMLU test
Baseline	49.8%	31.0%
Fine-tuned	51.1%	34.7%
Fine-tuned (modified loss)	50.4%	35.6%

3.4 LLAMA-2 7B

While the DeBERTa-v3 models were initialized from reward model checkpoints, no prevalent open-source LLAMA-2 based reward models exist. Training the reward models initialized from language modeling checkpoints was noisy and inconclusive. For future work, we are training a LLAMA-2 reward model from the language modeling checkpoint first using open-source preference data such as [Bai et al., 2022b] to then adapt to LMQA.

4 Conclusion

In this work we benchmark reward models for truthfulness using bio-med MCQ questions (LMQA). Larger models, initialized from previous reward model checkpoints outperform general reward model baselines on out-of-distribution BioMed datasets such as MMLU. The results are relevant as prevalent reward models perform worse than random for truthfulness alignment. Thus, Reinforcement Learning with AI Feedback presents a promising and cost-effective direction to align LLMs in healthcare.

5 Future Work

Currently I am working with OLAB NYU to a) Train a larger reward model using LLAMA-2 b) Align the LLAMA-2 model using Proximal Policy Optimization [Schulman et al., 2017] and Direct Preference Optimization [Rafailov et al., 2023].

A Appendix

A.1 Zero-shot Prompt

```
####user:
A 14-16 gauge needle is most likely to be used for:
A) Children.
B) Elderly patients.
C) Inserting in the back of the hand.
D) Trauma or burns patients
####assistant:
The correct answer is A) Children
####user:
Why is that?
####assistant:
[Your assistant's response here]
```

A.2 Few-shot Prompt

```
####user:
For what age group is the turtle technique indicated?
A) Young children (Early Childhood Education and first courses of Primary Education).
B) Older children (last courses of Primary Education).
C) Preadolescents (first courses of Secondary Education).
D) Adolescents (last years of Secondary Education and Bachillerato).
####assistant:
The correct answer is Older children (last courses of Primary Education).
####user:
Why is that?
####assistant:
The Turtle Technique is primarily indicated for older children, specifically those in
the last courses of Primary Education. This approach is often used to help children in
this age group develop better self-control and emotional regulation skills.

####user:
A 14-16 gauge needle is most likely to be used for:
A) Children.
B) Elderly patients.
C) Inserting in the back of the hand.
D) Trauma or burns patients
####assistant:
The correct answer is A) Children
####user:
Why is that?
####assistant:
[Your assistant's response here]
```

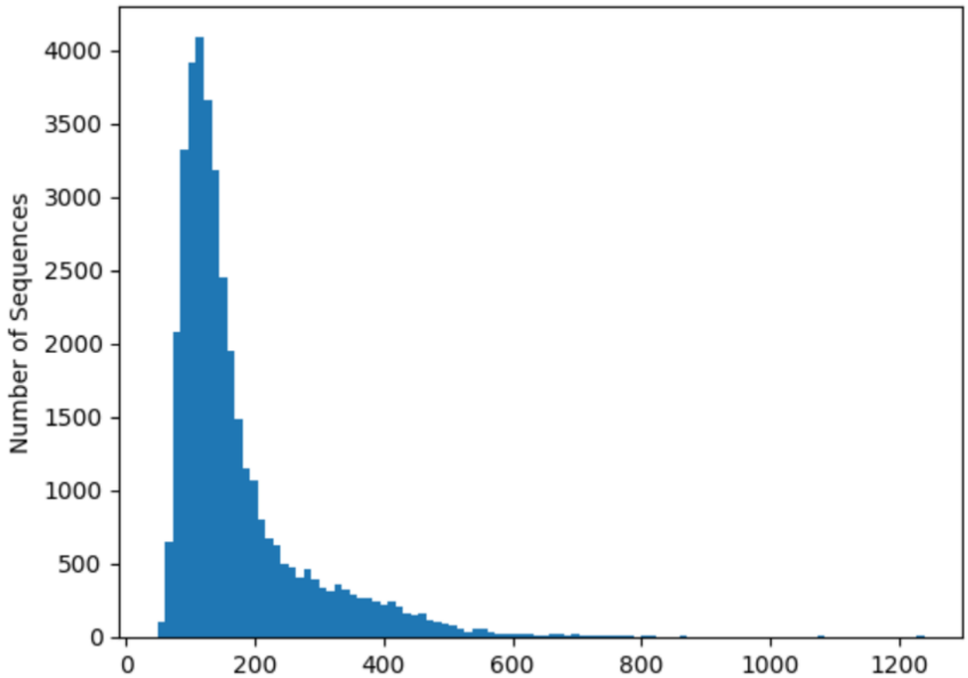

Figure 3: Sequence Length Distribution

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Lindsay F Schwartz, Matthew Lineberry, Yoon Soo Park, Carol S Kamin, and Abbas A Hyderi. Development and evaluation of a student-initiated test preparation program for the usmle step 1 examination. *Teaching and learning in medicine*, 30(2):193–201, 2018.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*, 2023.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.