Synthetic Imitation Edit Feedback for Factual Alignment in Clinical Summarization

Prakamya Mishra[*1], **Zonghai Yao**[*1]
Shuwei Chen[2], **Beining Wang**[2], **Rohan Mittal** [1], **Hong Yu**[1,3]
University of Massachusetts, Amherst[1], Fudan University[2]
University of Massachusetts, Lowell[3]
{prakamyamish, zonghaiyao}@umass.edu

Abstract

Large Language Models (LLMs) like the GPT and LLaMA families have demonstrated exceptional capabilities in capturing and condensing critical contextual information and achieving state-of-the-art performance in the summarization task. However, community concerns about these models' hallucination issues continue to rise. LLMs sometimes generate factually hallucinated summaries, which can be extremely harmful in the clinical domain NLP tasks (e.g., clinical note summarization), where factually incorrect statements can lead to critically erroneous diagnoses. Fine-tuning LLMs using human feedback has shown the promise of aligning LLMs to be factually consistent during generation, but such training procedure requires high-quality human-annotated data, which can be extremely expensive to get in the clinical domain. In this work, we propose a new pipeline using ChatGPT[2] instead of human experts to generate high-quality feedback data for improving factual consistency in the clinical note summarization task. We focus specifically on edit feedback because recent work discusses the shortcomings of human alignment via preference feedback in complex situations (such as clinical NLP tasks that require extensive expert knowledge), as well as some advantages of collecting edit feedback from domain experts. In addition, although GPT has reached the expert level in many clinical NLP tasks (e.g., USMLE QA), there is not much previous work discussing whether GPT can generate expert-level edit feedback for LMs in the clinical note summarization task. We hope to fill this gap. Finally, our evaluations demonstrate the potential use of GPT edits in human alignment, especially from a factuality perspective.[3]

1 Introduction

Large language models (LLMs) such as GPT-3 [8], GPT-4 [37], PaLM [14], LLaMA [52], and LLaMA-2 [52] have significantly advanced generative AI, outperforming smaller models like T5 [42] and GPT-2 [40] in both language understanding and natural language generation capabilities, as demonstrated in recent studies [9, 44, 14, 35, 37, 56]. However, these LLMs still have an inherent tendency to generate hallucinations producing factually inconsistent outputs, hurting their reliability in being helpful, truthful, and harmless [27, 37, 63, 36]. Recently, alignment of these LLMs to generate factually consistent summaries has been an active area of research, especially in the generation tasks [45, 11, 29, 25], where human alignment with different optimization techniques like RLHF

---

[*]indicates equal contribution

[2]We used ChatGPT using Azure OpenAI Service: https://azure.microsoft.com/en-us/products/ai-services/openai-service/

[3]Our code and dataset will be released at https://github.com/seasonyao/LearnFromHumanEdit
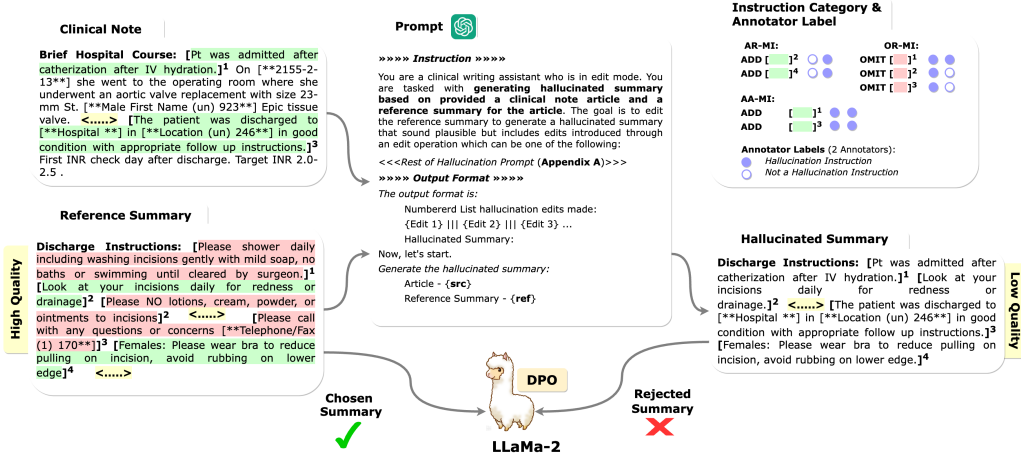
Figure 1: Illustration of the proposed synthetic imitation edit data generation and training pipeline. Given a clinical note and the corresponding reference summary (factually high-quality summary), we prompt ChatGPT to generate a hallucinated summary (factually low-quality summary) by generating edit instructions to the **ADD** hallucinated sentences and **OMIT** critically important sentences from the reference summary & the clinical note. The high & low-quality summary pairs are used to train LLaMA-2 using DPO, where the reference high-quality summary is the chosen one and the hallucinated low-quality summary is the rejected one. In the figure, the sentences with green background are the edits to be included in the hallucinated summary, whereas the sentences with red background are not. The annotated instruction category and hallucination labels corresponding to the ChatGPT generated edit instructions are also shown.

[38, 68, 47], RLAIF [32], RRHF [62], RAFT [51] using PPO [18], DPO [41] and SALT [59] have demonstrated their effectiveness in aligning these hallucinated models. Despite the effectiveness of these alignment methods, they require a significant amount of human-annotated data to illustrate human preference during training, which can be scarce in clinical domains where such expert-level annotations can be expensive to get. Synthetically generated data is used in such domains where there is a lack of preference-based human-annotated data [33, 61, 16].

In this work, we focus on the clinical note summarization task in the clinical domain. Clinical notes consist of a patient's history throughout their visit to a hospital, which includes nursing & physician notes, ECG reports, radiology reports, and etc. The clinical note summarization task is the task to generate a discharge summary from a clinical note that can later be used for the correct diagnosis of the patient based on the visit history present in the clinical note. Using LLMs to generate such discharge summaries could drastically improve the efficiency in the diagnosis of a patient's history. However, due to the hallucinations generated by LLMs, their generated outputs are not reliable, resulting in the need for alignment training using preference-based data.

Human feedback for summarization can come in different forms. Comparison-based feedback, Scalar feedback, and Label feedback are more common and well-studied types for NLP tasks [48, 67, 66, 5, 55, 22]. Edit feedback and Language feedback are two more informative but harder-processed types [12]. Recent works [12, 26, 59] discussed some limitations of current common feedback types and the advantages of adding Edit or Language feedback for better human alignment, especially for tasks that need expert domain knowledge. However, it is challenging to collect real-world doctor edit feedback due to privacy protection and strict data regulations like HIPAA [43, 4]. Generating a synthetic imitation edit feedback dataset is one potential solution. Manually constructing a large amount of such synthetic data using domain experts is time and effort-consuming [30, 2]. An alternative is to build the synthetic dataset by leveraging large language models such as ChatGPT [21, 33, 15].

In this work, we mainly focus on previously less-studied edit feedback data to better align LMs to generate factually correct clinical note summaries. Although GPT has reached the expert level in many clinical NLP tasks (e.g., USMLE QA) [31, 24], there is not much previous work discussing whether GPT can generate expert-level edit feedback for LMs in the clinical NLP tasks. Specifically, we propose to use high-quality synthetically LLMs-generated edit feedback for fine-tuning LLMs

and LMs using the recent SOTA alignment methods DPO [41] for improving factuality in the model-generated summaries.

In this work, we demonstrate how high-quality synthetic edit-feedback can be used to improve the factual alignment of the LLMs in the clinical domain for the clinical note summarization task. We propose a new pipeline to generate synthetic preference-based data. We use ChatGPT to add factual hallucinations to generate factuality-based low-quality summaries given the original factuality-based high-quality reference summaries and the corresponding clinical notes. We then treat the high-quality summaries as the DPO-chosen ones and the low-quality summaries as the DPO-rejected ones used for the synthetic edit feedback. Our experiments demonstrate the effectiveness of the synthetic edit-feedback for improving factuality in the model-generated summaries of the LLaMA-2 7B (4.26%↑ Rouge-1, 4.67%↑ Factuality) and GPT2 (5.97%↑ Rouge-1, 8%↑ Factuality) model when trained using DPO compared to simple fine-tuning (SFT). Outputs from models trained using our approach also received a higher preference for factuality by human annotators.

2 Related Work

Recent research has yielded promising findings regarding the application of Large Language Models (LLMs) for data augmentation, particularly in tasks like summarization, translation, and code generation [33, 15, 65, 17, 60]. [23] and [19] have investigated the efficacy and precision of LLMs in the context of data annotation. Their findings have demonstrated promising results, showcasing the potential of LLMs to perform on par with or even surpass human Crowd-Workers in terms of accuracy. [7] employed LLMs to generate positive sample pairs for the training of downstream models. Additionally, within the biomedical domain, recent research endeavors have explored the capabilities of LLMs in clinical text mining and doctor-patient conversation tasks, aiming to tackle issues related to suboptimal performance and privacy concerns [50, 1, 53, 54, 34].

On the other hand, [46] points out that standard Supervised Fine-tuning (SFT) makes important errors (such as hallucinations) and unimportant errors (such as minor grammatical errors) have the same impact on the final loss, which leaves the model still unable to consistently produce output of human-determined high quality (such as factuality). Recent work has shown the promise of learning with human feedback paradigms to produce human-determined high-quality text [6, 67, 48, 3, 20, 64, 62]. In the clinical domain, a lack of medical knowledge often results in LMs and LLMs generating a large number of factual errors [39, 49, 57, 58]. In this paper, we will focus on using feedback learning to improve factuality.

Recent works discussed some limitations of current common feedback types and the advantages of adding edit or language feedback for better human alignment, especially for tasks that need expert domain knowledge [12, 59]. [12] discussed the pros and cons of five different feedback types: comparison-based feedback, scalar feedback, label feedback, edit feedback, and language feedback, and suggested some more exploration for less-studied types like edit feedback and language feedback. [59] showed edit feedback is a natural way to collect feedback from doctors in workflows where they may be working off of an AI-generated summary in their workflow. This makes exploring edit feedback in the clinical domain attractive: when experts don't have time to spend extra time providing a large amount of feedback data to the research community, how can we use the data from their daily work to help human alignment? However, due to privacy reasons, the content of the doctor's daily work cannot be released for research usage. In the work, we hope to explore how to use ChatGPT, which has reached an expert level in the medical license exam [31, 24], to generate a large number of synthetic imitation edits. And then, we used it to train smaller LLMs such as Llama2 [52].

3 Problem Statement

Given an available dataset $D : \{X, Y_+\}$ of C clinical notes $X : \{x^1, x^2, ...x^C\}$, their corresponding ground truth reference discharge summaries $Y_+ : \{y_+^1, y_+^2, ...y_+^C\}$, and a reference model π_{ref}, the aim of the clinical note summarization task T is to train the model $\pi_{ref}(y_+^i|x^i)$. Here the ith clinical note $x^i : \{x_1^i, x_2^i, ...x_n^i\}$ consists of n tokens (j^{th} token represented by x_j^i) and the ith reference summary $y_+^i : \{y_{+,1}^i, y_{+,2}^i, ...y_{+,m}^i\}$ consists of m tokens (j^{th} token represented by $y_{+,j}^i$ & $m << n$). The standard way to fine-tune π^{ref} on T is to simply fine-tune π^{ref} using the cross-entropy loss

$\ell_c e(y_+^i, \pi^{ref}(x^i))$ over the original training dataset D. The model fine-tuned using this approach is represented by π_{sft}.

Aligning π_{ref} using DPO requires the need for preference-based data $D_{pref} : \{X, Y_+, Y_-\}$, where Y_+ is a set of preferred summaries, and Y_- are the dispreferred ones ($Y_- : \{y_-^1, y_-^2, ...y_-^k\}$). Such a preference is usually gathered through human annotation or is generated synthetically. As previously explained, not only gathering human annotations is expensive in the clinical domain, but even generating synthetic data using standard approaches like corruption [13] can be challenging.

So in this work, we (1) Propose a new pipeline to generate high-quality synthetic edit data Y_- from Y_+ to imitate D_{pref}, where Y_- acts as the dispreferred summary; (2) Use the synthetically generated preference data $D_{pref} : \{X, Y_+, Y_-\}$ to fine-tune π_{ref} using DPO to align π_θ to generate factually consistent outputs, where π_θ is the model that is being trained. In the following subsections, we describe the synthetic edit feedback generation pipeline and the training procedure in detail.

3.1   Synthetic Imitation Edit Feedback

For summarization alignment using DPO, the model learns from the preference data D_{pref} by learning to increase the likelihood of the preferred summaries Y_+ and to decrease the likelihood of the dispreferred summaries Y_-. Usually, Y_+ is easy to get from the ground truth labels (reference summaries) in D, but on the other hand, Y_- is usually hard to get. Synthetic data generating using any corruption function $f_c(Y_+)$ to generate Y_- from Y_+ have been explored to generate the corrupted summaries, which can be used to act as dispreferred summaries. However leveraging a corruption-based synthetic preference data generation approach for factuality alignment can be challenging and misaligned to the final factuality alignment objective, especially in the clinical domain where some phrases (clinical instructions) in the discharge summaries are extremely important for the correct diagnosis of the patient. So using incorrectly (corrupting incorrect phrases in the summaries of Y_+) corrupted Y_+ to act as Y_- for aligning the model using DPO can lead to poor factuality alignment.

In order to generate correct high-quality synthetic edit data for factuality alignment, we propose to use off-the-shelf LLMs like ChatGPT to act as a corruption function f_h that can be used to generate imitation edit data Y_- by adding hallucination to Y_+. In order to do so, we prompt ChatGPT to generate a hallucinated summary $f_h : \{x^i, y_+^i\} \rightarrow y_-^i$ given a clinical note x^i and the corresponding reference summary y_+^i, as shown in Figure 1. The detailed prompt is attached in **Appendix A** Table 3. For f_h, we use the clinical note x^i and the corresponding reference summary y_+^i to generate a hallucinated summary y_-^i that sounds plausible but includes edits introduced through the edit operations listed below:

- **ADD Operation**: Intentionally add medico-legally essential words from the article required for accurate diagnosis and treatment documentation.

- **OMIT Operation**: Intentionally omit medico-legally essential words in the reference summary required for accurate diagnosis and treatment documentation.

For generating the hallucinated summaries y_-^i's in Y_-, we prompt ChatGPT (f_h) to first list a set of I edit instructions $E^i : \{e_1^i, e_2^i, ...e_I^i\}$, where each instruction consists of either an ADD or OMIT operation on contents in the clinical note X^i or the corresponding reference summary Y_+^j, Then the prompt uses the generated instructions in E^i to generate the hallucinated summary y_-^i. Since we prompt ChatGPT to add/omit medico-legally unimportant/important information respectively, resulting in a decrease in the factual consistency of the content in the hallucinate summary y_-^i, so we treat y_-^i as the dispreferred summary and the y_+^i as the preferred one for generating the preference data D_{pref}. Since edit operations (ADD/OMIT) used in f_h generate Y_- that are of low quality in terms of clinical factuality, so in line with the final objective of clinical factuality alignment, we treat Y_- as a set of dispreferred summaries and the reference summaries Y_+ as a set of preferred ones (since they are the ground truth and are of higher quality with respect to factuality). Examples of the generated hallucination edit instructions E^i along with the hallucinated summaries Y_- are attached in the **Appendix D**.

3.2 Preference-based Training

$$\ell_{dpo}(\pi_\theta; \pi_{ref}) = - E_{(x^i, y_+^i, y_-^i) \sim D_{pref}} \left(\log \sigma \left[\beta \log \frac{\pi_\theta(y_+^i | x^i)}{\pi_{ref}(y_+^i | x^i)} - \beta \log \frac{\pi_\theta(y_-^i | x^i)}{\pi_{ref}(y_-^i | x^i)} \right] \right) \quad (1)$$

For aligning π_{ref} using DPO ($\pi_{ref} \rightarrow \pi_\theta$), we train the model by optimizing the loss function ℓ_{dpo} shown in equation 1, where given the preference data $D_{pref} : \{X, Y_+, Y_-\}$ consisting of a set of clinical notes x^i, preferred summaries Y_+^i (ground truth reference summaries), and the dispreferred summaries Y_-^i (hallucinated summaries), the model learns to increase the likelihood of the preferred summaries Y_+^i and to decrease the likelihood of the dispreferred summaries Y_-^i. In the equation, π_{ref} is the base model and π_θ is the model being trained to have improved alignment and β is used to scale the weight on how incorrect the model should treat the dispreferred summary y_-^i relative to the preferred summary y_+^i. The higher the β beta, the less the divergence from the initial policy π_{ref}.
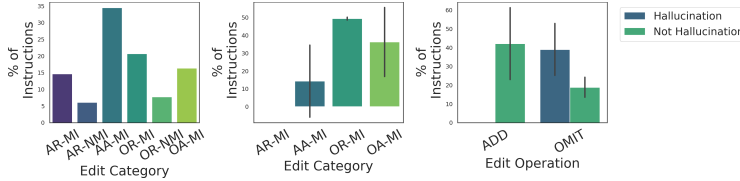

Figure 2: **First:** % of edits made in the human evaluation samples for each edit type listed in Table 1. **Second:** % of edits made by the ADD/OMIT edits mentioned in the generated instructions that resulted in hallucinations according to the annotators. **Third:** % of the ADD/OMIT edits for generating hallucinated/non-hallucinated edits summaries according to the annotators. Annotation for all the instructions from our annotators had a mean Kappa score k of 0.38 (**Appendix D**).

4 Results

In this section, we first evaluate the quality of the synthetic edit data generated by our pipeline, where we evaluate the correctness of hallucination edits and the diversity of edits made by our pipeline. Then we conduct experiments for external evaluation on the downstream summarization tasks. Following previous work [10], we used the discharge instructions from the MIMIC-III database [28] in our experiments for clinical note summarization, consisting of 25k/3k/3k train/valid/test respective clinical note and reference summaries. MIMIC-III is a publicly available repository of de-identified health records of over 40,000 patients collected from the Beth Israel Deaconess Medical Center in Massachusetts. Due to resource limitations, we restricted the train/valid/test set to 5k/128/128.

4.1 Synthetic Edit Feedback Evaluation

To quantify the quality of the hallucination instructions generated by our proposed pipeline, we use two domain expert medical students to annotate the generated instructions and hallucinated summaries with (1) Edit type (Listed in Table 1), (2) Edit Instruction hallucination label (0=Hallucination instruction, 1=Not and hallucination instruction), and (3) Comment to justify their reasoning for the annotated hallucination label. Detailed human evaluation guidelines are described in **Appendix C**.

Table 1: Hallucination Edit Types

Instruction Abbrivation	Description
AR-MI	Add from Reference Summary (AR) | Mentioned in Instruction (MI)
AR-NMI	Add from Reference Summary (AR) | Not Mentioned in Instruction (MI)
AA-MI	Add from Article | Mentioned in Instruction (MI)
OR-MI	Omit from Reference Summary (OR) | Mentioned in Instruction (MI)
OR-NMI	Omit from Reference Summary (OR) | Not Mentioned in Instruction (MI)
OA-MI	Omit from Reference Summary (OR) | Not Mentioned in Instruction (MI)

Using the annotations from the human evaluators, we focus on the following aspects to quantify the correctness and diversity of the imitation edit instructions and generated hallucinated summaries:

Types of edits in the hallucinated summary From our evaluation we observed that there were majorly 7 types of edit made from our pipeline as shown in Table 1. These edits are categorized mainly based on (1) the operation used for the edit (ADD/OMIT), (2) whether the edit was instructed in the instruction generated by our pipeline , and (3) whether the edit was made using the contents from the reference summary or the article (clinical note). Figure 2 (First) illustrates the percentage of edits made in the human evaluation samples for each edit type (Table 1). We observe that the majority

of hallucination edits made by our pipeline were mentioned in the instructions and were either to add content from the article (AA-MI 34%) or omit content from the reference summary (OR-MI 21%). A small proportion of the edits were not mentioned in the generated instruction (AR-NMI 6% & OR-NMI 8%), which we believe was because these edits instructions improve the factuality in the edited summary because of the addition of critical content from the reference summary as well as excluding useless content from the clinical note article. Since we constrained the pipeline to make the edited summary of approximately the same length as the reference summary, so we observed that there was some useful content that was still added from the reference summary (AR-MI 15%) and some useless content that was omitted from the article (OA-MI 16%). In conclusion, the edited summary consists of factually important and unimportant content, a high-quality synthetic edit data.

Imitation edit instructions for hallucination Focusing on the edits mentioned in the generated instructions, we observed that the majority of edits that led to hallucinations in the edited summary as annotated by the annotators by omitting useful content from the reference summary (OR-MI 49%), as shown in Figure 2 (Second). The other hallucinations were made by either adding/omitting useless/useful content from the article with approximately equal proportions (AA-MI 14% & OA-MI 36%). The annotations also showed that none of the hallucination edits in our pipeline were made by adding content from the reference summary (AR-MI 0%).

Contribution of ADD/OMIT instructions for hallucination We also investigated the use of ADD vs OMIT operations for generating hallucination edits, as shown in Figure 2 (Third). According to the annotators, both OMIT operations contribute more wrt ADD operation towards generating hallucination edits (ADD 0% & OMIT 39%), and a majority of ADD operations are responsible for generating edits that do not result in hallucinations (ADD 42% & OMIT 18%).

From these figures, we conclude that according to the annotators, the majority of edit instructions generated by our pipeline are considered to produce hallucinations in the edited summary, and not only is our pipeline able to generate a variety of edits (listed in Table 1), but also the majority of edit instructions generated by our pipeline result in hallucinations in the edited summary. Additionally, from Figure 3,

Figure 3: Word cloud of the comments from the annotators illustrating the reason why they consider an edit instruction to generate hallucination

we illustrate the main comments on why the domain expert annotators believed an edit instruction is responsible for generating hallucination. Their response reasons further validate the quality of our edits for generating hallucinations as the majority of the time they hallucinate content that is usually an *"Important/Required instruction/artifact"* for the discharge summaries to be factually consistent.

4.2 External Evaluation

To evaluate the effectiveness of our proposed edits for improving factuality in the model-generated outputs, we compare the summarization performance of a model trained using a simple fine-tuning (STF) approach optimizing on $\ell_c e$ vs the preference-based DPO training approach that optimizes ℓ_{dpo} and uses our proposed synthetic imitation edit pipeline for generating

Table 2: External Evaluation Results

	ROUGE			Factuality		
	R1	R2	RL	G-Eval	UMLS-F1	Human Rank↓
GPT2-SFT	30.80	13.84	23.05	5.085	20.76	3.078
GPT2-DPO	32.64	**15.81**	**24.51**	5.492	**22.91**	2.578
LLaMA2-SFT	31.92	13.25	22.57	6.351	20.16	2.242
LLaMA2-DPO	**33.28**	14.56	24.41	**6.648**	22.05	**2.085**

preference data. In simple SFT, the input to the model is the clinical note and the objective is to increase the likelihood of the reference summary, whereas in the case of DPO, the input is the same but the objective is to increase the likelihood of the reference summary while decreasing the likelihood of the hallucinated summary. We experiment with GPT-2 and LLaMA-2 and evaluate the quality of the trained models for summarizatio using ROUGE score (R1,R2,RL), and for factuality using G-Eval & UMLS-F1. G-Eval evaluates factual alignment using GPT-4 chain-of-thought to assess the factuality when prompted to generate a factuality score (**Appendix C**), whereas UMLS-F1 calculates the F1-score between the UMLS CUI (Unified Medical Language System Controlled Unclassified Information) codes present in the reference summary and the generated summary. For both G-Eval and UMLS-F1, the higher the score/F1, the higher the factuality in the generated output.

From Table 2, we observe that not only does the use of preference-based DPO training using our proposed synthetic data result in significant improvement in the ROUGE scores, but also there is consistent improvement over the factuality metrics indicating the effectiveness of our proposed pipeline. We also asked the annotators to compare the factuality of the summaries generated by both the models using SFT & DPO training, resulting in the highest win-rate by the summaries generated using LLaMA2-DPO and the least by GPT2-SFT (both the DPO are preferred over the SFT counterparts).

5   Conclusion

In this paper, our study demonstrates the efficacy of utilizing synthetic edit feedback to enhance factual alignment in LLMs for clinical note summarization. We introduce a novel pipeline for generating synthetic preference-based data. Additionally, human annotators consistently favored the factuality of summaries generated through our approach.

6   Limitations

In this paper, we focus only on the task of factuality alignment in clinical summarization, and adapting the proposed method to other domains is yet to be explored. Since we only had access to medical students as annotators, we relied on them for the factuality ranking metrics (Human Rank) inorder to quantify the preference results. Although they are qualified to read and annotate clinal notes and their corresponding discharge summaries, using more qualified domain experts as annotators would further increase the statistical significance of our results, which we leave to future work. We further leave it to future work to address concerns about fairness, generalizability to other domains/languages, and potential biases inherent in LLMs.

References

[1] Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, 2023.

[2] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2283–2294, 2023.

[3] Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*, 2023.

[4] George J Annas. Hipaa regulations: a new era of medical-record privacy? *New England Journal of Medicine*, 348:1486, 2003.

[5] Erdem Bıyık, Malayandi Palan, Nicholas C Landolfi, Dylan P Losey, and Dorsa Sadigh. Asking easy questions: A user-friendly approach to active reward learning. *arXiv preprint arXiv:1910.04365*, 2019.

[6] Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*, 2019.

[7] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*, 2022.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott

Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[10] Pengshan Cai, Fei Liu, Adarsha Bajracharya, Joe Sills, Alok Kapoor, Weisong Liu, Dan Berlowitz, David Levy, Richeek Pradhan, and Hong Yu. Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6234–6247, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[11] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[12] 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.

[13] Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. Purr: Efficiently editing language model hallucinations by denoising language model corruptions, 2023.

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[15] Haixing Dai, Zheng Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007, 2023.

[16] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. Auggpt: Leveraging chatgpt for text data augmentation, 2023.

[17] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*, 2022.

[18] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. https://github.com/openai/baselines, 2017.

[19] Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*, 2022.

[20] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

[21] Gunther Eysenbach et al. The role of chatgpt, generative language models, and artificial intelligence in medical education: A conversation with chatgpt and a call for papers. *JMIR Medical Education*, 9(1):e46885, 2023.

[22] Rachel Freedman, Rohin Shah, and Anca Dragan. Choice set misspecification in reward inference. *arXiv preprint arXiv:2101.07691*, 2021.

[23] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

[24] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312, 2023.

[25] Navita Goyal, Ani Nenkova, and Hal Daumé III. Factual or contextual? disentangling error types in entity description generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8322–8340, Toronto, Canada, July 2023. Association for Computational Linguistics.

[26] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2023.

[27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023.

[28] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[29] Daniel Kang and Tatsunori B. Hashimoto. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online, July 2020. Association for Computational Linguistics.

[30] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.

[31] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

[32] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback, 2023.

[33] Zongjie Li, Chaozheng Wang, Pingchuan Ma, Chaowei Liu, Shuai Wang, Daoyuan Wu, and Cuiyun Gao. On the feasibility of specialized ability extracting for large language code models, 2023.

[34] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, and Xiang Li. Differentiate chatgpt-generated and human-written medical texts, 2023.

[35] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023.

[36] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[37] OpenAI. Gpt-4 technical report, 2023.

[38] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[39] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] 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, 2023.

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[43] Thomas C Rindfleisch. Privacy, information technology, and health care. *Communications of the ACM*, 40(8):92–100, 1997.

[44] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207, 2021.

[45] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR, 23–29 Jul 2023.

[46] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[47] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020.

[48] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[49] Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*, 2021.

[50] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.

[51] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiao-qing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[53] Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, Zhichao Yang, and Hong Yu. Umass_bionlp at mediqa-chat 2023: Can llms generate high-quality synthetic note-oriented doctor-patient conversations? *arXiv preprint arXiv:2306.16931*, 2023.

[54] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. Notechat: A dataset of synthetic doctor-patient conversations conditioned on clinical notes, 2023.

[55] Nils Wilde, Erdem Bıyık, Dorsa Sadigh, and Stephen L Smith. Learning reward functions from scale feedback. *arXiv preprint arXiv:2110.00284*, 2021.

[56] Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Beining Wang, Dan Berlowitz, and Hong Yu. Performance of multimodal gpt-4v on usmle with image: Potential for imaging diagnostic support with explanations. *medRxiv*, 2023.

[57] Zonghai Yao, Yi Cao, Zhichao Yang, Vijeta Deshpande, and Hong Yu. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. *arXiv preprint arXiv:2209.07859*, 2022.

[58] Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. Context variance evaluation of pre-trained language models for prompt-based biomedical knowledge probing. *arXiv preprint arXiv:2211.10265*, 2022.

[59] Zonghai Yao, Benjamin J Schloss, and Sai P Selvaraj. Improving summarization with human edits. *arXiv preprint arXiv:2310.05857*, 2023.

[60] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*, 2021.

[61] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[62] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears, 2023.

[63] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball, 2023.

[64] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

[65] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

[66] Banghua Zhu, Jiantao Jiao, and Michael I Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

[67] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[68] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.

A Hallucination Edit Prompt

Table 3: Edit Prompt

»»»» Instruction »»»»
You are a clinical writing assistant who is in edit mode. You are tasked with generating hallucinated summary based on provided a clinical note article and a reference summary for the article. The goal is to edit the reference summary to generate a hallucinated summary that sounds plausible but includes edits introduced through an edit operation which can be one of the following:

Add Operation: Intentionally add medico-legally essential words from the article not required for accurate diagnosis and treatment documentation.

Omit Operation: Intentionally omit medico-legally essential words in the reference summary required for accurate diagnosis and treatment documentation.

For these operations focus on words that, if missing or incorrect in the hallucinated summary, could lead to wrong diagnoses and treatments in the future. Maintain coherence while excluding essential terms. The hallucinated summary should be concise and contain no more than FIVE EXTRA WORDS compared to the reference summary and should have an equal number of Add/Omit operations.

Steps for generating the hallucinated summary:

Step 1: List the proposed edit operations to introduce hallucination on the reference summary.

Step 2: Use the proposed edit operations to edit the reference summary.

»»»» Output Format »»»»
The output format is:
Numbererd List hallucination edits made:
{Edit 1}, {Edit 2}, {Edit 3} ...
Hallucinated Summary:

»»»» Follow the above Instructions, Hallucination Method and Output Format »»»»
Now, let's start.
Generate the hallucinated summary:
Article - {src}

B Human Evaluation Annotation Guidelines

Table 4: Human Evaluation Guideline

Op.	Label	Description
Hallucination Instruction Identification Guideline for Add/Omit Operations		
ADD	0	Including medico-legally phrases from the Article/Reference Summary that are required for accurate diagnosis and treatment documentation.
ADD	1	Including medico-legally phrases from the Article/Reference Summary that are not required for accurate diagnosis and treatment documentation.
OMIT	0	Not Including medico-legally phrases from the Article/Reference Summary that are not required for accurate diagnosis and treatment documentation.
OMIT	1	Not Including medico-legally phrases from the Article/Reference Summary that are required for accurate diagnosis andtreatment documentation.

For the human evaluation, we provided the annotators with a set of clinical note articles (article), reference summaries with the corresponding discharge instructions with reference to the article, and a list of edit instructions (edit instructions). The edit instructions consisted of two operations (Add & Omit operations) using which a new summary called edited summary can be generated.

The two operations are described below:

1. **Add Operation**: Intentionally including medico-legally phrases in the edited summary from the article or reference summary that are not required for accurate diagnosis and treatment documentation.

2. **Omit Operation**: Intentionally not including medico-legally phrases in the edited summary from the article or reference summary that are required for accurate diagnosis and treatment documentation.

Both the above operations in the edit instruction can be used to generate hallucinations in the edited summary, where hallucinations are the phrases that are either (1) not present in the edited summary that is crucial for accurate diagnosis and treatment documentation, or (2) present in the edited summary that are not crucial for accurate diagnosis and treatment documentation. Edit instruction that leads to hallucination is called a hallucination instruction.

The condition for an edit instruction with either Add or Omit operation is a hallucination instruction is listed in Table 4. In the table above, the hallucination label is used to label if an instruction leads to hallucination in the edited summary or not (0=Hallucination instruction, 1=Not and hallucination instruction).

Given the article, reference summary, and edit instructions generated by our pipeline, we asked the annotators to annotate each instruction with its:

1. Hallucination Label: 0 if the instruction is a hallucination instruction or 1 if not.

2. Comment: Justification for the hallucination label.

Further, we also asked them to annotate the edit type listed int Tabel 1 (Section 4.1) for each edit in the generated hallucinated summary.

C G-Eval Factuality Metric Prompt

Table 5: G-Eval Factuality Prompt

You will be given one discharge summary written for a Clinical Note.
Your task is to rate the summary on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Factual Consistency (1-10): Is the summary has missing or incorrect facts that are not supported by the source text and could lead to wrong diagnoses and treatments?

Evaluation Steps:

1. Read the clinical note carefully and identify the main topic and key points.

2. Read the discharge summary and compare it to the clinical notee. Check if the summary covers the main topic and key points of the clinical note, and Is the summary has missing or incorrect facts that are not supported by the source text and could lead to wrong diagnoses and treatments?

3. Assign a score for Factual Consistency on a scale of 1 to 10, where 1 is the lowest and 10 is the highest based on the Evaluation Criteria.

Clinical Note Text:
{Document}
Reference Discharge Summary:
{Reference Summary }
System Output Discharge Summary:
{System Output Summary}

Return the scores as dictionary objects, adhering to the following structure:
{"Factual Consistency": ...}
Please provide your response solely in the dictionary format without including any additional text.

D   Human Evaluation Examples
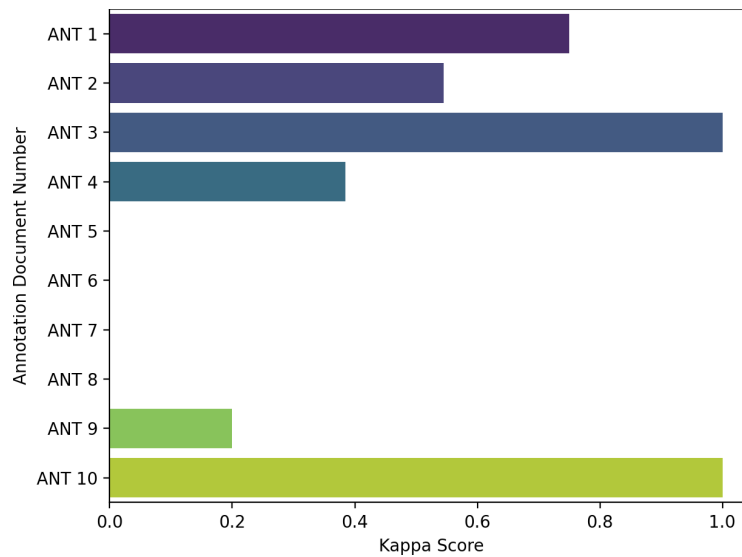
Figure 4: Kappa score for all annotation documents (ANN #) used in the human evaluation

For our human evaluation, we also calculated the Kappa score for inter-annotator agreement for hallucination labels in each document annotated by our annotators. We observed a mean Kappa score of 0.38, and the Kappa score plot for each of the documents is shown in Figure 4. We observe that for the majority of the documents, the annotators were in high agreement with each other for the hallucination label, but in some documents in which the ADD/OMIT operation was being done using the contents of the article instead of the reference, there was a high degree of disagreement for deciding the hallucination label. We provide two examples for our human annotation in Table 6 & Table 7.

Table 6: Human Annotation No. 1

Clinical Note

Brief Hospital Course: Pt was given 3 units of PRBCs and 4L NS for resuscitation in the ED and admitted to the MICU where he was NPO, having serial HCT q4hr, a 4th unit of PRBCs, [**Hospital1 **] IV PPI, and close monitoring with 2 large bore IVs. Pt got an EGD which showed 2 ulcers, 1x1am and 1x2cm near the anastomosis from his prior gastric bypass. The larger ulcer had an adherent clot underneath which there was a visible vessel and more clot. BICAP was applied to cauterize. There was no bleeding noted at the conclusion of the EGD. KUB and CXR post-procedure revealed no free air in the abdomen. Pt's HCT went from 21.8 on admission to 26 after his transfusions. Pt remained NPO with maintenance IVF, electrolyte replacement, and serial HCT until HD# 3. Pt had no further bleeding and was hemodynamically stable and was transferred to a monitored floor bed on HD#3. The pt began taking sips and then clears by mouth on HD#3. He was able to take in a Stage III diet and oral PPI medication on HD#4 and his HCT was stable at 27.6 and 27.4. The pt was sent home on a PPI and instructed to f/u with his surgeon within a week or to call sooner if her experienced any further bleeding.

Reference Summary

Discharge Instructions: Please call you doctor if you experience any of the following: - fever >101 - blood with bowel movements or blood in vomit - dizziness or lightheadedness - persistent nausea and vomiting - inability to eat or drink - severe abdominal pain - any other questions or concerns Please take your medications as prescribed. Please also take a chewable multivitamin, like Flintstones, daily. Please follow up with Dr. [**Last Name (STitle) **]. Please remain on Stage III diet until you follow up with your doctor. Do not self-advance your diet. Also, do not chew gum or drink out of a straw.

Edit Instructions

Numbered List hallucination edits made:

1. Add Operation: "Please call your doctor if you experience any of the following:"
2. Omit Operation: "- any other questions or concerns"
3. Add Operation: "Please take your medications as prescribed."
4. Omit Operation: "Please also take a chewable multivitamin, like Flintstones, daily."
5. Add Operation: "Please follow up with Dr. [**Last Name (STitle) **]."
6. Omit Operation: "Please remain on Stage III diet until you follow up with your doctor."
7. Add Operation: "Do not self-advance your diet."
8. Omit Operation: "Also, do not chew gum or drink out of a straw."'

Hallucinated Summary

Discharge Instructions: Please call your doctor if you experience any of the following: - fever >101 - blood with bowel movements or blood in vomit - dizziness or lightheadedness - persistent nausea and vomiting - inability to eat or drink - severe abdominal pain. Please take your medications as prescribed. Please follow up with Dr. [**Last Name (STitle) **]. Do not self-advance your diet.

Human Annotation #1

Instruction 1:
Hallucination Label: 0
Comment: Important instruction for any potential emergencies or progression. Required.
Instruction 2:
Hallucination Label: 1
Comment: Important instruction for any potential emergencies or progression. Required.
Instruction 3:

Hallucination Label: 0
Comment: Required.
Instruction 4:
Hallucination Label: 1
Comment: Multivitamin provides essential components required to produce new blood cells. Required.
Instruction 5:
Hallucination Label: 0
Comment: Tell the patient how to contact. Required.
Instruction 6:
Hallucination Label: 1
Comment: Important note for aftercare.
Instruction 7:
Hallucination Label: 0
Comment: Emphasis on instruction 6.
Instruction 8:
Hallucination Label: 1
Comment: Chewing gum may stimulate gastric acid to secrete. Drinking through a straw may increase the pressure in the cavity of upper GI tract and consequently trigger rebleeding, which may be controversial. Overall, it is better not to omit these two suggestions.

Human Annotation #2
Instruction 1:
Hallucination Label:1
Comment: more detail needed
Instruction 2:
Hallucination Label:1
Comment:very simply
Instruction 3:
Hallucination Label:1
Comment: Useful doctor's advice
Instruction 4:
Hallucination Label:0
Comment: Useful doctor's advice
Instruction 5:
Hallucination Label:1
Comment: It adds more detail
Instruction 6:
Hallucination Label:0
Comment: Useful doctor's advice
Instruction 7:
Hallucination Label:1
Comment: Useful doctor's advice
Instruction 8:
Hallucination Label:0
Comment: Useful doctor's advice

Table 7: Human Annotation No. 2

Clinical Note

Brief Hospital Course: Pt was admitted after catherization after IV hydration. On [**2155-2-13**] she went to the operating room where she underwent an aortic valve replacement with size 23-mm St. [**Male First Name (un) 923**] Epic tissue valve. See operative note for full details. Overall the patient tolerated the procedure well and post-operatively was transferred to the CVICU in stable condition for recovery and invasive monitoring. POD 1 found the patient extubated, alert and oriented and breathing comfortably. The patient was neurologically intact and hemodynamically stable on no inotropic or vasopressor support. Low dose beta blocker was initiated but then stopped due to hypotension with a SBP in the 80's. The patient was gently diuresed toward the preoperative weight. The patient was transferred to the telemetry floor for further recovery. Chest tubes were split due to high drainage and mediatinal chest tubes were removed with subsequent removal of left pleural chest tube once drainage had decreased. She was anticoagulated with Coumadin for chronic atrial fibrillation. The patient was evaluated by the physical therapy service for assistance with strength and mobility. By the time of discharge on POD #7 the patient was ambulating with assistance, the wound was healing well and pain was controlled with Tylenol only due to somnolence with Ultram. The patient was discharged to [**Hospital **] in [**Location (un) 246**] in good condition with appropriate follow up instructions.First INR check day after discharge. Target INR 2.0-2.5 .

Reference Summary

Discharge Instructions: Please shower daily including washing incisions gently with mild soap, no baths or swimming until cleared by surgeon. Look at your incisions daily for redness or drainage Please NO lotions, cream, powder, or ointments to incisions Each morning you should weigh yourself and then in the evening take your temperature, these should be written down on the chart No driving for approximately one month and while taking narcotics, will be discussed at follow up appointment with surgeon when you will be able to drive No lifting more than 10 pounds for 10 weeks Please call with any questions or concerns [**Telephone/Fax (1) 170**] Females: Please wear bra to reduce pulling on incision, avoid rubbing on lower edge **Please call cardiac surgery office with any questions or concerns [**Telephone/Fax (1) 170**]. Answering service will contact on call person during off hours**

Edit Instructions

Numbered List hallucination edits made:

1. Add: Pt was admitted after catherization after IV hydration.

2. Omit: Please shower daily including washing incisions gently with mild soap, no baths or swimming until cleared by surgeon.

3. Add: Look at your incisions daily for redness or drainage.

4. Omit: Please NO lotions, cream, powder, or ointments to incisions.

5. Add: Each morning you should weigh yourself and then in the evening take your temperature, these should be written down on the chart.

6. Omit: No driving for approximately one month and while taking narcotics, will be discussed at follow up appointment with surgeon when you will be able to drive.

7. Add: The patient was evaluated by the physical therapy service for assistance with strength and mobility.

8. Omit: No lifting more than 10 pounds for 10 weeks.

9. Add: The patient was discharged to [**Hospital **] in [**Location (un) 246**] in good condition with appropriate follow up instructions.

10. Omit: Please call with any questions or concerns [**Telephone/Fax (1) 170**].

11. Add: Females: Please wear bra to reduce pulling on incision, avoid rubbing on lower edge.

12. Omit: **Please call cardiac surgery office with any questions or concerns [**Telephone/Fax (1) 170**]. Answering service will contact on call person during off hours**.

Hallucinated Summary

Discharge Instructions: Pt was admitted after catherization after IV hydration. Look at your incisions daily for redness or drainage. Each morning you should weigh yourself and then in the evening take your temperature, these should be written down on the chart. The patient was evaluated by the physical therapy service for assistance with strength and mobility. The patient was discharged to [**Hospital **] in [**Location (un) 246**] in good condition with appropriate follow up instructions. Females: Please wear bra to reduce pulling on incision, avoid rubbing on lower edge.

Human Annotation #1

Instruction 1:
Hallucination Label: 1
Comment: It should not be included in a discharge instruction. But it's necessary as a part of the clinical note.
Instruction 2:
Hallucination Label: 1
Comment: Important aftercare instruction. Helps prevent infection.
Instruction 3:
Hallucination Label: 0
Comment: Important aftercare instruction. Redness or drainage may indicate infection.
Instruction 4:
Hallucination Label: 1
Comment: Important aftercare instruction, decreasing the risk of infection or non-healing wound.
Instruction 5:
Hallucination Label: 0
Comment: Necessary and required to keep a trace of the recovery course. Weight may indicate volume status (hypovolemia, hypervolemia, or euvolemia) which helps judge heart function. Elevated temperature is associated with infection.
Instruction 6:
Hallucination Label: 1

Comment: Important. For traffic safety. Because narcotic medications may sedate the patient, disturbing judgment and agility.
Instruction 7:
Hallucination Label: 1
Comment: Unrelated. Similar to Instruction 1.
Instruction 8:
Hallucination Label: 1
Comment: Required. For better wound healing and overall recovery.
Instruction 9:
Hallucination Label: 1
Comment: Unrelated. Similar to Instruction 1.
Instruction 10:
Hallucination Label: 1
Comment: Tell the patient how to contact. Required.
Instruction 11:
Hallucination Label: 0
Comment: Required aftercare. For better incision healing.
Instruction 12:
Hallucination Label: 1
Comment: Tell the patient how to contact. Required.

Human Annotation #2

Instruction 1:
Hallucination Label:0
Comment: makes summary more accurate
Instruction 2:
Hallucination Label:0
Comment:should be more accurate
Instruction 3:
Hallucination Label:0
Comment: Increase doctor's orders and reduce postoperative complications
Instruction 4:
Hallucination Label:1
Comment:should be more details
Instruction 5:
Hallucination Label:0
Comment:makes summary more accurate
Instruction 6:
Hallucination Label:1
Comment:should be more accurate
Instruction 7:
Hallucination Label:0
Comment:adds more details
Instruction 8:
Hallucination Label:1
Comment: Medical advice should be increased
Instruction 9:
Hallucination Label:0
Comment:Adds more details
Instruction 10:
Hallucination Label:1
Comment: patients need to know how to contact the hospital
Instruction 11:
Hallucination Label:0
Comment:adds more details
Instruction 12:
Hallucination Label:1
Comment: patients need to know how to contact the hospital