SYNFAC-EDIT: Synthetic Imitation Edit Feedback for Factual Alignment in Clinical Summarization

Prakamya Mishra [*†1,3], **Zonghai Yao** [*1]
Parth Vashisht[1], **Feiyun Ouyang**[3], **Beining Wang**[2], **Vidhi Dhaval Mody**[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) such as GPT & Llama have demonstrated significant achievements in summarization tasks but struggle with factual inaccuracies, a critical issue in clinical NLP applications where errors could lead to serious consequences. To counter the high costs and limited availability of expert-annotated data for factual alignment, this study introduces an innovative pipeline that utilizes >100B parameter GPT variants like GPT-3.5 & GPT-4 to act as synthetic experts to generate high-quality synthetics feedback aimed at enhancing factual consistency in clinical note summarization. Our research primarily focuses on edit feedback generated by these synthetic feedback experts without additional human annotations, mirroring and optimizing the practical scenario in which medical professionals refine AI system outputs. Although such 100B+ parameter GPT variants have proven to demonstrate expertise in various clinical NLP tasks, such as the Medical Licensing Examination, there is scant research on their capacity to act as synthetic feedback experts and deliver expert-level edit feedback for improving the generation quality of weaker (<10B parameter) LLMs like GPT-2 (1.5B) & Llama 2 (7B) in clinical domain. So in this work, we leverage 100B+ GPT variants to act as synthetic feedback experts offering expert-level edit feedback, that is used to reduce hallucinations and align weaker (<10B parameter) LLMs with medical facts using two distinct alignment algorithms (DPO & SALT), endeavoring to narrow the divide between AI-generated content and factual accuracy. This highlights the substantial potential of LLM-based synthetic edits in enhancing the alignment of clinical factuality [1].

1 Introduction

The advent of generative artificial intelligence (AI) has been markedly accelerated by the development of large language models (LLMs) such as GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), Llama (Touvron et al., 2023). These models have demonstrated superior capabilities in natural language understanding and natural language generation, outperforming language model predecessors (LMs) like T5 (Raffel et al., 2020) and GPT-2 (Radford et al., 2019) in a variety of linguistic tasks. Despite these advancements, LLMs confront significant challenges, primarily their propensity for generating hallucinations—fabricated information not grounded in source text—and producing factually inconsistent outputs (Ji et al., 2023b; Zhang et al., 2023a; Maynez et al., 2020). Such limitations critically undermine the models' reliability, particularly critical in clinical NLP applications, where inaccuracies could result in serious misdiagnoses.

The NLP community has discussed many reasons for the hallucination problem, including some limitations stemming from traditional supervised fine-tuning (SFT). SFT fails to differentiate between significant errors, such as hallucinations, and minor inaccuracies, like grammatical mistakes, treating all errors equally in their loss calculations. Moreover, SFT applies a uniform loss weighting across all data, regardless of its type, quality, or complexity, potentially diluting the training signal for more critical learning objectives. In response to these limitations, recent research has explored learning paradigms incorporating human feedback, such as RLHF (Ouyang et al., 2022; Ziegler et al., 2020; Stiennon et al., 2020b), RLAIF (Lee et al., 2023), RRHF (Yuan

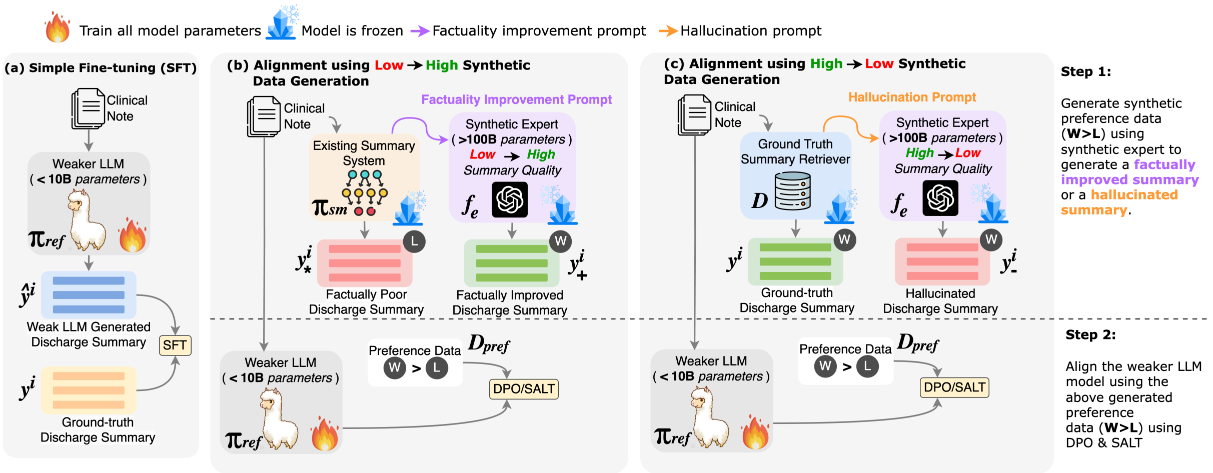
[*] indicates equal contribution

[†] Presently in AMD AI

[1] Dataset is released here: https://huggingface.co/datasets/bio-nlp-umass/SYNFAC-EDIT.

† To appear in proceedings of the Main Conference on Empirical Methods in Natural Language Processing (EMNLP) 2024

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20061–20083
November 12-16, 2024 ©2024 Association for Computational Linguistics



Train all model parameters
Model is frozen
Factuality improvement prompt
Hallucination prompt
(a) Simple Fine-tuning (SFT)
Clinical Note
Weaker LLM (< 10B parameters)
π_{ref}
$\hat{y}^i$
Weak LLM Generated Discharge Summary
SFT
y^i
Ground-truth Discharge Summary
(b) Alignment using Low → High Synthetic Data Generation
Clinical Note
Existing Summary System
π_{sm}
Factuality Improvement Prompt
Synthetic Expert (>100B parameters)
Low → High
Summary Quality
f_e
L
W
y^i_*
Factually Poor Discharge Summary
y^i_+
Factually Improved Discharge Summary
Weaker LLM (< 10B parameters)
π_{ref}
Preference Data
W > L
D_{pref}
DPO/SALT
(c) Alignment using High → Low Synthetic Data Generation
Clinical Note
Ground Truth Summary Retriever
D
Hallucination Prompt
Synthetic Expert (>100B parameters)
High → Low
Summary Quality
f_e
W
W
y^i
Ground-truth Discharge Summary
y^i_-
Hallucinated Discharge Summary
Weaker LLM (< 10B parameters)
π_{ref}
Preference Data
W > L
D_{pref}
DPO/SALT
Step 1:
Generate synthetic preference data (W>L) using synthetic expert to generate a factually improved summary or a hallucinated summary.
Step 2:
Align the weaker LLM model using the above generated preference data (W>L) using DPO & SALT

Figure 1: **(a):** The illustration of a standard simple fine-tuning pipeline. **(b & c):** The illustration of our proposed synthetic edit feedback generation & alignment training pipeline. In **Step 1** of our synthetic edit feedback generation pipeline we generate preference data in two directions: (1) **Low**→**High [b]**, where we generate a factually improved summary given an unaligned existing summary system generated summary (Section 3.2); (2) **High**→**Low [c]**, where we generate a hallucinated summary given a clinical not article and a ground truth reference summary (Section 3.1). In **Step 2** we align the Weaker LLM model using the Step 1 generated preference data using two alignment algorithms namely DPO & SALT (Section 3.3).

et al., 2023), and RAFT (Teed and Deng, 2020). Techniques including PPO (Dhariwal et al., 2017), DPO (Rafailov et al., 2023), and SALT (Yao et al., 2023) have demonstrated effectiveness in aligning these hallucination-prone models. However, these alignment methods require substantial amounts of human-annotated data to illustrate human preferences during training, which can be challenging to obtain in clinical domains (Li et al., 2023; Yoo et al., 2021; Dai et al., 2023b).

In this work, we mainly focus on previously less-studied edit feedback data to better align LMs (`GPT-2 (1.5B)`) & LLMs (`Llama-2 (7B)`) to generate factually correct clinical note summaries. Recent works (Casper et al., 2023; Ji et al., 2023a; Yao et al., 2023) discussed some limitations of current common feedback types (comparison or rating feedback) and the advantages of adding edit feedback for better human alignment. Human edits are a more natural way to collect feedback from clinicians as they fix AI-generated text for their workflow to improve generation (Yao et al., 2023). Collecting other forms of feedback that are not directly tied to the clinician's workflow will not scale as much, this is especially true in domains requiring expert domain knowledge and with nuanced user goals. Considering the cost, time, and availability of the experts, it is important to collect edit feedback from the expert's daily workflow. However, it is challenging to collect real-world clinician's edit feedback due to privacy protection and strict data regulations like HIPAA (Annas, 2003).

Generating a synthetic imitation edit feedback dataset by leveraging large (>`100B` parameters) `GPT` variants like `GPT-3.5` & `GPT-4` (Eysenbach et al., 2023; Li et al., 2023; Dai et al., 2023a) is one potential solution [2]. Such dataset can then be used for alignment training. Although these large `GPT` variants have reached expert-level performance in many clinical NLP tasks (e.g., Medical Licensing Examination) (Kung et al., 2023; Gilson et al., 2023; Yang et al., 2023), there is not much previous work discussing whether they can generate expert-level edit feedback for LMs and LLMs in the clinical NLP tasks. Therefore, we propose to use these large GPT variants as synthetic experts for generating high-quality edit feedback for fine-tuning LLMs and LMs using the recent SOTA alignment methods like DPO (Rafailov et al., 2023) & SALT (Yao et al., 2023) for improving factuality in the clinical domain for the clinical note summarization task. Specifically, we propose a new pipeline to generate synthetic preference-based data in two directions for alignment training: 1) **High**→**Low:** where we use these large GPT variants as synthetic experts to add factual hallucinations to generate factuality-based low-quality dispreferred

[2] We used Azure OpenAI Service, which is HIPAA-regulated: https://azure.microsoft.com/en-us/products/ai-services/openai-service/

summaries given the original factuality-based high-quality preferred ground-truth summaries & the corresponding clinical notes. 2) **Low→High:** where we use these large GPT variants as synthetic experts to add factual information to generate factuality-based high-quality preferred summaries given the factuality-based low-quality dispreferred unaligned model generated summaries and the corresponding clinical notes. We then treat the high-quality summaries as the preferred ones and the low-quality summaries as the dispreferred ones used in our synthetic preference data pairs.

Our experiments demonstrate the efficacy of utilizing synthetic edit feedback to enhance the factual accuracy of model-generated summaries. Specifically, for `Llama2 (7B)`, we observed a 2.44% ↑ in `ROUGEL` and a 1.35% ↑ in factuality using the DPO. Similarly, SALT resulted in a 2.47% ↑ in `ROUGEL` and a 2.04% ↑ in factuality. For `GPT-2`, DPO led to a 3.04% ↑ in `ROUGEL` and a 2.93% ↑ in factuality, while SALT yielded a 4.04% ↑ in `ROUGEL` and a 4.64% ↑ in factuality. Moreover, our top-performing model garnered a 78% preference rate for factuality among human evaluators, highlighting its superior performance.

2 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 i^{th} 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 i^{th} 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 over the original training dataset D, as shown in Figure 1(a).

Aligning π_{ref} using alignment training requires the need for preference-based data $D_{pref} : \{X, Y_w, Y_l\}$, where Y_w is a set of preferred summaries, and Y_l are the dispreferred ones. Such preference-based data 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 (Chen et al., 2023) can be challenging. So in this work, we **(1)** propose a new pipeline to generate high-quality synthetic preference data $D_{pref} : \{X, Y_w, Y_l\}$; **(2)** use D_{pref} to align π_{ref} to generate factually consistent outputs using alignment methods like DPO training & SALT loss. In the following subsections, we describe the synthetic edit-based preference data generation pipeline and the edit feedback-based alignment training method in detail.

Algorithm 1: Synthetic preference data generation (**High→Low** & **Low→High**).

```
Dataset: $D$:$\{X, Y\}$
Clinical Note Articles: $X$:$\{x^1, ..., x^c\}$
Reference Summaries: $Y$:$\{y^1, ..., y^c\}$
Small LM: $\pi_{sm}$
Function get_$D_{pref}^{High\to Low}$($D$,$f_e^{High\to Low}$):
    $D_{pref} : \{\}$                                   ▷ Preference Data
    for $i = 1$ to $c$ do
        $y_-^i, E^i \leftarrow f_e^{High\to Low}(x^i, y^i)$
        $y_w^i \leftarrow y^i$                          ▷ Preferred
        $y_l^i \leftarrow y_-^i$                        ▷ dispreferred
        $D_{pref} \leftarrow D_{pref} + \{x^i, y_w^i, y_l^i\}$
    end
    return $D_{pref}$
Function get_$D_{pref}^{Low\to High}$($D$,$f_e^{Low\to High}$,$\pi_{sm}$):
    $D_{pref} : \{\}$                                   ▷ Preference Data
    for $i = 1$ to $c$ do
        $y_*^i \leftarrow \pi_{small}(x^i)$  ▷ $\pi_{sm}$ Unaligned Output
        $y_+^i, E^i \leftarrow f_e^{Low\to High}(x^i, y_*^i)$
        $y_w^i \leftarrow y_+^i$                        ▷ Preferred
        $y_l^i \leftarrow y_*^i$                        ▷ dispreferred
        $D_{pref} \leftarrow D_{pref} + \{x^i, y_w^i, y_l^i\}$
    end
    return $D_{pref}$
```

3 Synthetic Imitation Edit Feedback

For summarization alignment, the model learns from the preference data pairs (Y_w,Y_l)) in D_{pref} by learning to increase the likelihood of Y_w and to decrease the likelihood of Y_l. Usually, Y_w is easy to get from the ground truth labels (reference summaries) in D, but on the other hand, Y_l is usually not readily available. Heuristic-based data augmentation functions have been previously explored to tackle this problem in low-resource settings (Kryściński et al., 2019). In this work, we propose to use LLMs as synthetic experts (specifically acting as edit functions (f_e) imitating as domain experts) to synthetically generate $D_{pref} : \{X, Y_w, Y_l\}$ in two directions, i.e., (1) **High→Low** or (2) **Low→High**. The synthetic preference data generation procedure for the above-mentioned two directions is explained in detail in the following two sections.

3.1 High→Low Synthetic Preference Data Generation

For the **High** $\rightarrow$ **Low**, to generate high-quality synthetic edit-based preference data for factuality alignment, we use off-the-shelf LLMs like GPT-3.5 & GPT-4 to act as synthetic domain experts specifically acting as edit function $f_e^{High\rightarrow Low}$ to mirror edits made by actual domain experts, to generate imitation edit data Y_{-} by adding hallucination to Y (as shown in get_$D_{pref}^{High\rightarrow Low}$() synthetic edit data generation function in Algorithm 1). For the **High** $\rightarrow$ **Low** preference data, since Y is the original ground truth in D, we treat it as the Y_w, whereas Y_{-} is treated as Y_l as it is the hallucinated summary w.r.t Y. In $f_e^{High\rightarrow Low} : \{x^i, y^i\} \rightarrow y_{-}^i$, we prompt synthetic experts to generate a hallucinated summary given a clinical note x^i and the corresponding reference summary y^i, as shown in Figure 1(c). The prompt for $f_e^{High\rightarrow Low}$ is designed to generate y_{-}^i using edits introduced through the edit operations listed in Table 6 of Appendix A, and the resulting y_{-}^i sounds plausible but includes hallucinated information that is not required for accurate diagnosis and treatment documentation of x^i. The detailed prompt is attached in Table 7 of Appendix A.

3.2 Low→High Synthetic Preference Data Generation

Generation using smaller LMs like GPT-2 (even after fine-tuning) has been observed to generate hallucinations and factually incorrect text (Zhang et al., 2023b). By leveraging this phenomenon and treating the summaries generated from smaller LMs (represented by Y_*) as factually unaligned, we propose an alternative direction (**Low** $\rightarrow$ **High**) to generate high-quality synthetic edit-based preference data for factuality alignment where we use off-the-shelf LLMs (GPT-3.5 & GPT-4) to act as synthetic domain experts specifically acting as edit function $f_e^{Low\rightarrow High}$ to generate imitation edit data Y_{+} by improving factuality in Y_* (as shown in get_$D_{pref}^{Low\rightarrow High}$ synthetic edit data generation function in Algorithm 1). Here for **Low** $\rightarrow$ **High** preference data, since $Y*$ is generated from an unaligned smaller model (susceptible to hallucinations and poor generation), we treat it as Y_l, whereas Y_{+} is treated as Y_w as it is the factually improved summary w.r.t Y_*. In $f_e^{Low\rightarrow High} : \{x^i, y_*^i\} \rightarrow y_{+}^i$, we prompt synthetic experts to generate a factually improved summary given a clinical note x^i and the corresponding smaller model generated summary y_*^i, as shown in Figure 1(b). The prompt for $f_e^{Low\rightarrow High}$ is designed to generate y_{+}^i using edits introduced through the edit operations listed in **Table 6** of Appendix A, and the resulting y_{+}^i is factually consistent and includes information that is required for accurate diagnosis and treatment documentation of x^i. The detailed prompt is attached in Table 8 of Appendix A.

Similar to chain-of-thought phenomenon (Chu et al., 2023), for both $f_e^{High\rightarrow Low} : \{x^i, y^i\} \rightarrow y_{-}^i$ (**High** $\rightarrow$ **Low**) & $f_e^{Low\rightarrow High} : \{x^i, y_*^i\} \rightarrow y_{+}^i$ (**Low** $\rightarrow$ **High**), we prompt the synthetic experts to first generate 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 (Table 6 of Appendix A; we also provide justification for only using ADD & OMIT operations for our edits in Appendix A) to be done on the contents X^i or Y^i/Y_*^i. Then the prompt further leverages E^i to generate y_{-}^i/y_{+}^i. In summary, we prompt the synthetic expert to ADD or OMIT medico-legally unimportant/important factual information resulting in a decrease/increase in the factual consistency of the contents of y_{-}^i/y_{+}^i respectively. Examples of the generated edit instructions E^i along with the edited summaries y_{-}^i/y_{+}^i are attached in the Appendix D.

3.3 Factual Alignment with Edit Feedback

After collecting preference data for imitating edit feedback, we naturally obtain a pair of summaries (low-quality dispreferred and high-quality preferred). For SFT, since it aims to maximize the probability of the model generating certain token distributions, it is evident that only high-quality summaries can be used for optimization. In contrast, preference training utilizes both low and high-quality summaries from the dataset to align the model toward the desired direction based on their differences. Specifically, in this paper, we employ two alignment algorithms, DPO and SALT, to align π_{ref} based on the differences in factuality levels between the high and low-quality summaries.

3.3.1 DPO Training

For aligning π_{ref} using DPO ($\pi_{ref} \rightarrow \pi_\theta$), we train the model by optimizing the loss function ℓ_{dpo} shown in Algorithm 2, where given the preference data $D_{pref} : \{X, Y_w, Y_l\}$ consisting of a set of clinical notes x^i, preferred summaries Y_w^i, and the dispreferred summaries Y_l^i, the model learns to

Algorithm 2: DPO & SALT loss functions for alignment training.

```
Preference Dataset: $D_{pref}$:$\{X, Y_w, Y_l\}$, where $Y_w$ & $Y_l$ are generated using get_$D_{pref}^{Low \rightarrow High}$ or get_$D_{pref}^{High \rightarrow Low}$
  from Algorithm 1.
Reference Model: $\pi_{ref}$
Aligned Model: $\pi_\theta$
$\ell_{dpo}(\pi_\theta; \pi_{ref}) = - E_{(x^i, y_w^i, y_l^i) \sim D_{\text{pref}}} \left( \log \sigma \left[ \beta \log \frac{\pi_\theta(y_w^i|x^i)}{\pi_{ref}(y_w^i|x^i)} - \beta \log \frac{\pi_\theta(y_l^i|x^i)}{\pi_{ref}(y_l^i|x^i)} \right] \right)$    ▷ DPO Loss
$\ell_{\text{salt}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x^i, y_w^i, y_l^i) \sim D_{\text{pref}}} (\alpha_1 \Omega_1 + \alpha_2 \Omega_2 - \alpha_3 \Omega_3)$    ▷ SALT Loss
$\Omega_1 = \sum_{a \in A} \log \pi_\theta(a|x^i)$    ▷ For aligned tokens
$\Omega_2 = \sum_{u_w \in U_{y_w^i}} \log \pi_\theta(u_w|x^i)$    ▷ For unaligned tokens in $y_w^i$
$\Omega_3 = \sum_{u_l \in U_{y_l^i}} \log(1 - \pi_\theta(u_l|x^i))$    ▷ For unaligned tokens in $y_l^i$
```

increase the likelihood of Y_w^i and to decrease the likelihood of the Y_l^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 y_l^i relative to y_w^i. The higher the β beta, the less the divergence from π_{ref}.

3.3.2 SALT Training

For situations where the differences between the preferred and dispreferred summaries are minimal, the DPO training method may not be optimal, as it tends to reward many tokens in y_w^i while penalizing the same tokens in y_l^i. To address this, we explore the SALT (Yao et al., 2023) training method in Algorithm 2, specifically designed for feedback involving minor edits. SALT training requires the sequence alignment of y_w^i and y_l^i, identifying:

- Ω_1: The set of tokens in y_w^i and y_l^i that are aligned through sequence alignment, indicating similarity or identical parts between the preferred and dispreferred summaries.
- Ω_2: The set of tokens in the preferred summary y_w^i that cannot be matched with any tokens in the dispreferred summary y_l^i, representing unique or important aspects of the preferred summary.
- Ω_3: The set of tokens in the dispreferred summary y_l^i that cannot be matched with any tokens in the preferred summary y_w^i, representing aspects to be avoided or corrected.

Then we can calculate SALT loss as shown in Algorithm 2. Here, the SALT loss is calculated to enhance model alignment by optimizing the following objectives: promoting the likelihood of tokens in Ω_1 and Ω_2 while discouraging the likelihood of tokens in Ω_3. The model, denoted as π_θ, aims to align more closely with the refined preferences represented by the preferred summaries. Weights α_1, α_2, and α_3 adjust the significance of aligned tokens, unique preferred summary tokens, and unique dispreferred summary tokens, respectively, in the loss function. This method allows for a nuanced adjustment of the model's predictions, ensuring that minor but critical edits are appropriately incorporated into the training process.

4 Results

4.1 Experimental Setup

In the next section, we evaluate the quality of our synthetic edit-based preference data (edit instructions E^i as well as the edited summaries $y_{\_}^i$ & y_+^i) generated using our synthetic experts in our pipeline for both the directions: **`High`** $\rightarrow$ **`Low`** & **`Low`** $\rightarrow$ **`High`**. For our pipeline, we experimented with both (`GPT-3.5` & `GPT-4`) as our synthetic experts. We leverage human annotations from domain expert human annotators on a small sample[3] of our pipeline-generated edits (called human evaluation sample set D_{eval}) to quantify these results. We also conduct experiments for external evaluation of our synthetic edit-based preference data on the downstream clinical note summarization task. Following previous works (Cai et al., 2022; Adams et al., 2022) in this domain, we used their cleaned discharge instruction dataset which is based on `MIMIC-III` database (Johnson et al., 2016) (instead of `MIMIC-IV`) in our experiments for clinical note summarization. This dataset consists of 25k/3k/3k train/valid/test respective clinical notes and reference summaries. Due to resource limitations, we restricted the downstream task train/valid/test set to 5k/128/128 samples, whereas for **`Low`**→**`High`**, we used the held-out training set of 20k samples to fine-tune π_{sm}.

In Section 4.3, we further evaluate the effective-

[3] For our human evaluation, we used 10 samples from our synthetic data using three domain experts: 1 doctor & 2 medical students. Human evaluation guideline: Appendix B.

Data Generation Setting	Synthetic Expert	% **ADD** Instructions (Out of Total)	% **OMIT** Instructions (Out of Total)	% of Total **Hallucination** Instructions (**ADD & OMIT**)	% **Hallucination** **ADD** Instructions (Out of Total)	% **Hallucination** **OMIT** Instructions (Out of Total)	% **Hallucination** **ADD** Instructions (Out of ADD Ins.)	% **Hallucination** **OMIT** Instructions (Out of OMIT Ins.)
High → Low	`GPT-4`	52.00	48.00	$54.33_{23.35}$	$18.50_{16.12}$	$34.00_{7.94}$	$35.00_{30.41}$	$71.67_{16.07}$
	`GPT-3.5`	64.16	35.84	$33.06_{15.49}$	$14.48_{13.30}$	$28.11_{4.20}$	$18.89_{18.36}$	$70.67_{11.14}$
				% of Total **Factuality** Instructions (**ADD & OMIT**)	% **Factuality** **ADD** Instructions (Out of Total)	% **Factuality** **OMIT** Instructions (Out of Total)	% **Factuality** **ADD** Instructions (Out of ADD Ins.)	% **Factuality** **OMIT** Instructions (Out of OMIT Ins.)
Low → High	`GPT-4`	54.33	45.67	$42.59_{17.14}$	$28.67_{15.74}$	$4.39_{1.29}$	$51.67_{28.87}$	$11.67_{2.89}$
	`GPT-3.5`	55.00	45.00	$28.64_{3.11}$	$24.50_{2.36}$	$4.14_{1.47}$	$47.22_{4.81}$	$7.50_{2.50}$

Table 1: **[Columns 3 & 4]:** Statistics for % of ADD & OMIT instructions present in $E^i \in D_{eval}$. **[Column 5]:** Statistics for % of both ADD & OMIT instructions present in $E^i \in D_{eval}$ that were annotated as hallucinating (up)/factuality aid (down) instruction. **[Column 6 & 7]:** Statistics for % of only ADD or OMIT instructions respectively out of total (ADD + OMIT) instructions present in $E^i \in D_{eval}$ that were annotated as hallucinating (up)/factuality aid (down) instruction. **[Column 8 & 9]:** Statistics for % of only ADD or OMIT instructions respectively out of respective ADD/OMIT instructions present in $E^i \in D_{eval}$ that were annotated as hallucinating (up)/factuality aid (down) instruction.

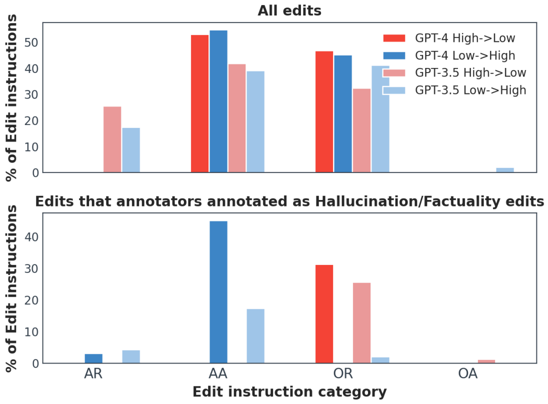


Figure 2: **Top:** % of edits made in the D_{eval} for each edit type listed in Table 12 of Appendix C. **Bottom:** % of edits that resulted in hallucinations/factuality aid according to the annotators. Plot legend format: `<Synthetic Expert> <Data Generation Setting>`

ness of our generated synthetic preference data for improving factuality in the weaker LLM-generated outputs. For showcasing extrapolation of our approach to different model parameter scales, in our experiments, we use `GPT-2 (1.5B)` & `Llama-2 (7B)` models (hyperparameter details in Appendix F). Although `GPT-2` is an old model and now there exist larger and better models like Llama variants, it is still important to showcase the applicability of our approach on smaller and weaker models as they are very frequently used on device deployed models, discussed more in Section 7. We compare the summarization performance of the models trained using a simple STF approach vs the preference-based DPO/SALT training approach. Human evaluation examples are listed in Appendix E.

4.2 Synthetic Edit Feedback Evaluation

To quantify the quality of the edit instructions generated by our proposed pipeline, we used domain expert annotators to annotate the generated edit instructions and edited summaries in D_{eval}. In our proposed pipeline, both $f_e^{High \rightarrow Low}$ & $f_e^{Low \rightarrow High}$, first generates a set of edit instruction E^i which are then used to generate y_-^i or y_+^i respectively. To quantify the quality of y_-^i/y_+^i, we first used human evaluation to evaluate the quality of the corresponding E^i by annotating whether a generated instruction e_j^i in E^i is useful or not in generating hallucinations (in the case of **`High`** → **`Low`**)/factuality improvements (in the case of **`Low`** → **`High`**). Refer to Table 1 for human evaluation results on generated edit instructions. The numbers in Table 1 over the last five columns report the mean and standard deviation of the hallucination/factuality statistics over annotations by all our annotators, whereas columns 3 & 4 are the statistics derived directly from $E^i \in D_{eval}$. For **`High`**→**`Low`** setting, we ask the annotators to annotate whether the edit instruction is hallucinating instructions or not, whereas for **`Low`**→**`High`** setting, we ask the annotators to annotate if the edit instructions are factuality aid instructions or not. Here are the results from our human evaluation of synthetic edit data generation:

`GPT-4` is much better at following prompt instructions We designed our prompt to use equal number of ADD & OMIT hallucination (**`High`**→**`Low`**)/factuality aid (**`Low`**→**`High`**) edit instructions, but we observe that `GPT-3.5` is poor at following these instructions relative to `GPT-4`, which in both $f_e^{High \rightarrow Low}$ & $f_e^{Low \rightarrow High}$ generates approximately equal number of ADD & OMIT instructions. From Table 1 we also observe that `GPT-4` is better at generating a higher percentage of desired edit instructions (last two columns of Table

Edit Data Setting	High→ Low		Low→ High	
Prompt Model	GPT-4	GPT-3.5	GPT-4	GPT-3.5
Preference %	$6.67_{0.05}$	$6.67_{0.12}$	$\mathbf{26.67_{0.15}}$	$\mathbf{33.33_{0.15}}$

Table 2: Mean & std. statistics for annotated preference percentage of (1) **Column [2 & 3]:** y^i_- over y^i, and (1) **Column [4 & 5]:** y^i_+ over y^i_*. ↓ pref. % in **High→Low** & ↑ pref. % in **Low→High** is good.

1) which according to our annotators actually leads to desired edits. Using GPT-4 for edits also resulted in a higher percentage of the total number of desired edits generated relative to GPT-3.5 (**Column 5**), suggesting better prompt following tendencies.

OMIT instructions leads to hallucination whereas ADD instructions lead to factuality improvements We observe that in the case of **High→Low**, majority of actual hallucination edits are generated using OMIT instructions (**Column 9 [Up]**), whereas in the case of **Low→High**, majority of actual factuality edits are generated using ADD instructions (**Column 8 [Down]**). While the percentage of ADD & OMIT edits are relatively higher in the case of **Low→High** & **High→Low** respectively, in the case of **Low→High** it is relatively much more skewed towards the ADD instructions leading to highly skewed edit type distribution in E^i from $f_e^{Low\to High}$. Counter-intuitively, this suggests that for our synthetic experts, it is relatively difficult to identify and remove factually incorrect information compared to factually correct information. To quantify the type of edits made in our pipeline, we prompted (C) GPT4 to categorize edit instructions into one of the edit-type categories listed in Table 12 of Appendix C. From the results in Figure 2 for **High→Low** setting, the majority of desired hallucination edits are generated by either omitting useful information from the reference summary/article, whereas in the case of **Low→High**, majority of desired factuality improvement edits are generated adding useful information from the article/unaligned model generated summary. This further validates our findings.

High→Low synthetic data results in better quality preference data compared to Low→High From Table 2, we observe that for **High→Low** the annotators had a very low preference in terms of factuality (<10%,) towards the edited summary, validating our desired output. On the other hand, for **Low→High** a relatively higher preference towards the edited summary was observed. Although this is in line with the desired outcome, in case of **Low→High**, despite a significant relative increase in preference rate towards the edited summaries (hypothesized to be factually improved), the annotators still preferred the low-quality π_{sm} generated summary on average, suggesting a relatively poor desired preference data.

GPT-4 exhibits tendencies to generate higher granularity of synthetic edits relative to GPT-3.5 We further analyzed the edits by quantifying the granularity of edits using ROUGE scores. In Table 3, we calculated ROUGE-1/2/L/Lsumm between the y^i & y^i_- in the case of **High→Low** preference data, and between the y^i_* & y^i_+ in the case of **Low→High** to measure the token level difference between the pair of summaries. From our results we observe that for both directions, GPT-4 had significantly higher ROUGE scores relative to GPT-3.5, suggesting GPT-4 edits are of high granularity not only at the token level but also for longer spans of tokens.

4.3 External Evaluation

To evaluate the effectiveness of our proposed edits for improving factuality in the model-generated outputs, we compare the summarization performance of the model trained using a simple STF approach vs the preference-based DPO/SALT training approach, where we use either **High→Low** or **Low→High** edit pipeline for generating preference data. In SFT, the model takes a clinical note as input and aims to generate a summary that matches the reference as closely as possible. In preference training, using the DPO approach, the goal remains similar but with a focus on favoring a high-quality summary over a low-quality one. For SALT, we first target GPT-verified tokens (Ω_1), GPT-preferred tokens (Ω_2), and GPT-dispreferred tokens (Ω_3) using sequence alignment, and then the objective is to increase the likelihood of Ω_1 and Ω_2 (can be different weights) while decreasing the likelihood of Ω_3. We experiment with GPT-2 and Llama2 and evaluate the quality of the trained models for summarization using ROUGEL and for factuality using G-Eval and UMLS-F1. G-Eval evaluates factual alignment using the GPT-4 chain-of-thought to assess the factuality when prompted to generate a factuality score (Appendix D), whereas UMLS-F1 calculates the F1-score between the UMLS medical terms present in the reference summary and the generated summary. For both G-Eval and UMLS-F1, the higher the score, the higher the factuality in the generated output. We also had two medical students

Data Generation Setting	Synthetic Expert	R-1	R-2	R-L	RLsum
High–>Low	GPT-4	**0.86**	**0.82**	**0.84**	**0.85**
	GPT-3.5	0.55	0.47	0.52	0.52
Low–>High	GPT-4	**0.73**	**0.68**	**0.71**	**0.71**
	GPT-3.5	0.55	0.50	0.50	0.53

Table 3: Edit granularity analysis using ROUGE between (1) **Row [2 & 3]:** y^i_- & y^i, and **Row [4 & 5]:** y^i_+ & y^i_*. Here, ↑ ROUGE scores signifies ↑ granularity edits.

review 50 summaries for factual accuracy, specifically looking for missing or incorrect information that could lead to errors in medical treatment after discharge. Each method listed in Table 4 and 5 was evaluated in a head-to-head comparison against its corresponding SFT baseline, with the stipulation that no ties were allowed in the assessment.

As demonstrated in Table 4 and 5, we observed for both GPT-2 and Llama2, preference training with GPT-4 edits surpassed GPT-3.5 edits in performance. Specifically, SALT using GPT-4 edits excelled in all metrics, while SALT with GPT-3.5 edits occasionally fell short compared to DPO with GPT-3.5 edits. This aligns with the conclusions drawn in Section 4.2, suggesting that SALT benefits from higher granularity edit feedback, enabling better alignment outcomes. GPT-3.5 edits, often involving extensive sentence-level modifications, reduce the accuracy of sequence alignment in SALT, introducing noise that degrades the alignment's effectiveness. In addition, GPT-2's performance improved with **High**→**Low** over **Low**→**High** in ROUGEL scores, without noticeable differences in factuality and human assessments. We only reported **Low**→**High** outcomes for GPT-2 in Table 5 because Llama2, when trained with GPT-edits from the **Low**→**High** pipeline, lagged behind the SFT baseline significantly (Appendix Table 15). This discrepancy could be attributed to the **Low**→**High** data being generated by the smaller LM in our experiments, making the corrections from GPT-edits on low-quality data too simplistic for LLMs like Llama 2 to achieve factual alignment. Conversely, the **High**→**Low** data, crafted by leveraging GPT based on the reference summary, is model-agnostic, enabling more effective utilization for both GPT-2 and Llama 2.

5 Related Work

Recent studies have demonstrated the efficacy of LLMs in enhancing data augmentation processes (Li et al., 2023; Dai et al., 2023a; Zhou et al., 2022; Dai et al., 2022). Investigations into the precision and effectiveness of LLMs for data annotation have revealed their potential to match or even exceed the accuracy of human annotators, as reported by Gilardi et al., 2023 and Ding et al., 2022. Moreover, the use of LLMs for generating positive sample pairs, crucial for training downstream models, has been explored with promising outcomes (Bonifacio et al., 2022). Within the biomedical field, LLMs are increasingly utilized for tasks including clinical text mining, question answering, summarization, medical documentation, and other clinical generation tasks for enhancing data augmentation processes. These efforts aim to address challenges such as suboptimal performance, adherence to instructions, and privacy concerns, showcasing the broad applicability of LLMs in this domain (Tang et al., 2023; Tran et al., 2023; Sarker et al., 2023; Wang et al., 2023; Liao et al., 2023).

On the other hand, Stiennon et al., 2020a notes that standard sequence-to-sequence training (SFT) incorrectly weighs significant errors (like hallucinations) and minor mistakes (such as grammatical inaccuracies) equally, impacting the ability to consistently generate text of high quality as determined by human standards, such as factuality. Recent studies highlight the potential of learning with human feedback paradigms to produce text that meets these high-quality standards (Ziegler et al.,

	ROUGEL	UMLS-F1	G-Eval	Human H2H (vs X_sft)
	GPT-3.5 Synthetic Dataset			
GPT2-sft	34.51	31.83	3.96	-
GPT2-dpo	34.70	32.82	4.24	62% win
GPT2-salt	34.34	32.03	4.04	44% win
Llama2-sft	38.03	35.47	6.48	-
Llama2-dpo	38.15	36.30	6.61	60% win
Llama2-salt	38.32	36.85	6.54	56% win
	GPT-4 Synthetic Dataset			
GPT2-sft	34.51	31.83	3.96	-
GPT2-dpo	37.35	34.76	4.42	78% win
GPT2-salt	38.55	36.47	4.60	72% win
Llama2-sft	38.03	35.47	6.48	-
Llama2-dpo	40.47	36.82	6.71	60% win
Llama2-salt	40.50	37.51	6.82	74% win

Table 4: External Evaluation **High**→**Low**. First column represents <Weaker LLM>-<Training Algorigthm>

	ROUGEL	UMLS-F1	G-Eval	Human H2H (vs X_sft)
	GPT-3.5 Synthetic Dataset			
SFT	34.51	31.83	3.96	-
DPO	34.34	32.94	4.44	66% win
SALT	34.55	33.33	4.12	53% win
	GPT-4 Synthetic Dataset			
SFT	34.51	31.83	3.96	-
DPO	34.30	33.64	4.34	72% win
SALT	36.95	34.35	4.48	74% win

Table 5: External evaluation **Low**→**High** results (GPT2).

2019; Stiennon et al., 2020c; Akyürek et al., 2023; Dong et al., 2023; Zhao et al., 2023; Yuan et al., 2023). In the clinical realm, the risk of factual errors by large language models (LLMs) due to gaps in medical knowledge is significant, potentially leading to severe consequences like misdiagnoses (Petroni et al., 2019; Sung et al., 2021; Yao et al., 2022a,b). Various feedback mechanisms—such as comparison-based, scalar, label, edit, and language feedback—have been explored, with calls for further research into less conventional methods like edit and language feedback (Casper et al., 2023). The use of edit feedback in clinical settings, where doctors review AI-generated summaries, presents a practical method for acquiring expert feedback without compromising privacy (Yao et al., 2023). The proposal to create a synthetic dataset of imitation edit feedback using models such as `GPT-3.5` and `GPT-4` offers a promising solution to address privacy concerns (Mishra et al., 2023). Despite the proficiency of `GPT` models in numerous clinical NLP tasks, including passing the Medical Licensing Examination (Kung et al., 2023; Gilson et al., 2023; Yang et al., 2023), their ability to generate expert-level edit feedback for clinical applications has not been thoroughly investigated. This paper aims to fill this research gap by evaluating the capability of these `GPT` variants to effectively generate such feedback.

6 Conclusion

This study leverages synthetic edit feedback to improve factual accuracy in clinical summarization using DPO and SALT techniques. Our approach demonstrates the effectiveness of `GPT`-generated edits in enhancing the reliability of clinical NLP applications.

7 Limitations and Ethical Considerations

This study offers valuable insights but also comes with several limitations that we would like to highlight:

- **Domain Specificity:** Our research exclusively focuses on the task of factuality alignment in clinical summarization. The adaptation of the proposed method to other domains remains unexplored. This suggests that our approach may need further validation and adjustments before being applied to different fields.
- **Expertise of Annotators:** We relied on 1 doctor and 2 medical students as annotators for human evaluation and preference results. While they are qualified to read and annotate clinical notes and their corresponding discharge summaries, employing more qualified domain experts as annotators would enhance the statistical significance of our results. We leave this to future work, along with addressing concerns about fairness, generalizability to other domains/languages, and potential biases inherent in LLMs.

Privacy Implications Privacy protection is crucial when dealing with clinical text and patient data. Even though we utilized a de-identified public dataset (such as MIMIC-III), generating and using synthetic data in practical applications must strictly adhere to data protection laws and ethical standards to prevent misuse of patient information.

Bias Considerations LLMs may inherently contain or amplify biases present in the training data. When generating edit feedback and synthetic data using LLMs, these biases must be carefully considered to avoid propagating inaccurate or biased information in clinical decision-support tools.

Broader Impacts Our research aims to enhance the factual accuracy of clinical summarizations through synthetic edit feedback, potentially positively impacting the reliability of healthcare decision-support systems and reducing patient risk. However, technology usage should be approached cautiously to ensure that technological errors do not endanger patient safety.

Experimentation with more capable LLMs We focused our experimentation on only `GPT-2 (1.5B)` & `Llama-2 (7B)` as our weaker LLMs and `GPT-3.5` & `GPT-4` as our synthetic experts, but in future we would also like to explore those capabilities of our model using more recent, capable and domain-specific models like BioGPT and BioLlama, we will leave that to future work.

In summary, while our study demonstrates the potential of using LLMs to improve the factual accuracy of clinical summaries, practical applications must consider domain adaptability, annotator expertise, privacy, bias, and broader ethical societal implications. Future work should focus on addressing these limitations and ethical considerations to ensure the safe, fair, and effective use of technology.

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

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

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pengshan Cai, Fei Liu, Adarsha Bajracharya, Joe Sills, Alok Kapoor, Weisong Liu, Dan Berlowitz, David Levy, Richeek Pradhan, and Hong Yu. 2022. 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. International Committee on Computational Linguistics.

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

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

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

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. 2023a. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007.

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. 2023b. Auggpt: Leveraging chatgpt for text data augmentation.

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

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

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

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

Gunther Eysenbach et al. 2023. 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.

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

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. 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.

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. 2023a. Ai alignment: A comprehensive survey.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea

Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

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. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

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

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

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

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

Prakamya Mishra, Zonghai Yao, Shuwei Chen, Beining Wang, Rohan Mittal, and Hong Yu. 2023. Synthetic imitation edit feedback for factual alignment in clinical summarization. *arXiv preprint arXiv:2310.20033*.

OpenAI. 2023. Gpt-4 technical report.

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. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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

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

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

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

Shouvon Sarker, Lijun Qian, and Xishuang Dong. 2023. Medical data augmentation via chatgpt: A case study on medication identification and medication event classification. *arXiv preprint arXiv:2306.07297*.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020b. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

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

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

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

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

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, Xiaoqing 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. 2023. Llama 2: Open foundation and fine-tuned chat models.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2023. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *arXiv preprint arXiv:2310.19975*.

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

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

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

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

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

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. 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. Association for Computational Linguistics.

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

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

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

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

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

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

A Edit Prompts

<table>
<tr><th colspan="2">Hallucination Edit Operations</th></tr>
<tr><th>Edit Operation</th><th>Description</th></tr>
<tr><td><b>ADD</b></td><td>Intentionally <b>including</b> medico-legally phrases in the edited summary from the article or reference summary that <b>are not required</b> for accurate diagnosis and treatment documentation.</td></tr>
<tr><td><b>OMIT</b></td><td>Intentionally <b>not including</b> medico-legally phrases in the edited summary from the article or reference summary that <b>are required</b> for accurate diagnosis and treatment documentation.</td></tr>
<tr><th colspan="2">Factuality Edit Operations</th></tr>
<tr><th>Edit Operation</th><th>Description</th></tr>
<tr><td><b>ADD</b></td><td>Intentionally <b>including</b> medico-legally phrases in the edited summary from the article or reference summary that <b>are required</b> for accurate diagnosis and treatment documentation.</td></tr>
<tr><td><b>OMIT</b></td><td>Intentionally <b>not including</b> medico-legally phrases in the edited summary from the article or reference summary that <b>are not required</b> for accurate diagnosis and treatment documentation.</td></tr>
</table>

Table 6: Description of edit operations used in our prompts for generation synthetic data in **Hight**→**Low** (hallucination edit operations) & **Low**→**High** (factuality edit operations)

We used ADD & OMIT operations as the only to operations available to the synthetic experts to edit and generate new summaries. The definitions for these ADD & OMIT operations in both **High**→**Low** & **Low**→**High** settings are given in Table 6. The detailed prompts for generating synthetic edit data in both directions, **High**→**Low** & **Low**→**High** used for generating hallucinations and factuality improvements are provided in Table 7 & 8 respectively.

We only include these ADD & OMIT operations in our prompts because ADD/OMIT is the natural way in which medical professionals create/correct this data. When human experts try to correct AI summary, they can modify or delete a span of tokens, insert a new span of tokens, or not change anything to a span of tokens, all these conditions can be decomposed into ADD & OMIT combination. Here's how different editing situations can be decomposed into ADD and OMIT combinations:

1. Modify a Span of Tokens:

   Decomposition: First, OMIT the span of tokens that need modification. Then, ADD the new span of tokens with the corrected information.

   Example: "The patient has a high fever" to "The patient has a mild fever", first OMIT "high", then ADD "mild" in its place.

2. Delete a Span of Tokens:

   Decomposition: OMIT the span of tokens without adding any new content.

   Example: To remove "due to viral infection" from "The patient has a mild fever due to viral infection", simply OMIT "due to viral infection".

3. Insert a New Span of Tokens:

   Decomposition: ADD the new span of tokens at the specific location without omitting any existing content.

   Example: To add "and coughing" to "The patient has a mild fever", ADD "and coughing" at the end of the sentence.

4. No Change to a Span of Tokens:

   Decomposition: Neither ADD nor OMIT actions are performed on the span of tokens.

   Example: If "The patient has a mild fever" is accurate and requires no modification, then no action is taken.

B Human Evaluation Annotation Guidelines

For the human evaluation, we provided the annotators with a set of clinical note articles (article), reference or unaligned model-generated summaries with the corresponding, and a list of edit instructions (edit instructions) generated by prompting `GPT-4` & `GPT-3.5`. The edit instructions consisted of two operations (Add & Omit operations) using which a new summary called edited summary is generated.

The two operations in the case of **`High`**→**`Low`** 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.

The two operations in the case of **`Low`**→**`High`** are described below:

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

Both the above operations in the edit instruction in **`High`**→**`Low`** & **`Low`**→**`High`** can be used to generate hallucinations & factuality improvements in the edited summary respectively. Here 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. Similarly, factuality improvements in the edited summary are either by (1) the addition of phrases that are not present in the original summary but are crucial for accurate diagnosis and treatment documentation or (2) the omission of phrases that are present in the original summary by are not crucial for accurate diagnosis and treatment documentation. Edit instruction that leads to hallucinations is referred to as hallucination instruction, whereas edit instructions

Table 7: Hallucination 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}
Reference Summary - {ref}

Table 8: Factuality Prompt

»»»» Instruction »»»»
You are a writing assistant who is in edit mode. You are tasked with generating edited summary based on provided a clinical note article and a model generated summary for the article. The goal is to edit the model generated summary to generate an edited summary that is factually consistent with respect to the article and contains edits introduced through an edit operation which can be one of the following:

Add Operation: Intentionally add medico-legally essential words from the article to the edited summary required for accurate diagnosis and treatment documentation. Only add a single sentence in a single edit.
Omit Operation: Intentionally omit medico-legally non-essential words from the model generated summary to the edited summary not required for accurate diagnosis and treatment documentation.

For these operations focus on words that, if present or correct in the edited summary, could lead to the right diagnoses and treatments in the future. Maintain coherence while including essential terms. The edited summary should be concise and contain no more than FIVE EXTRA WORDS compared to the model generated summary and should have an equal number of Add & Omit operations.

Steps for generating the edited summary:
Step 1: List the proposed edit operations to improve factually consistent in the model generated summary.
Step 2: Use the proposed edit operations to edit the model generated summary.
»»»» Output Format »»»»
The output format is:
Numbered List factuality edits made:
{Edit 1}, {Edit 2}, {Edit 3} ...
Edited Summary:
»»»» Follow the above Instructions, Factuality Improvement Method and Output Format »»»»
Now, let's start.
Generate the edited summary:
Article - {src}
Model Generated Summary - {ref}

Hallucination Instruction Identification Guideline for Add/Omit Operations		
Op.	Label	Description
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.

Table 9: Human annotation instructions for annotating whether an ADD/OMIT instruction is a hallucination instruction or not.

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

Table 10: Human annotation instructions for annotating whether an ADD/OMIT instruction is a factuality instruction or not.

	Edit Instruction Agreement	
	Edit Data Setting	
Expert	High to Low (Annotators agree to Hallucination)	Low to High (Annotators agree to Factuality)
GPT-4	0.13	0.30
GPT-3.5	**0.73**	**0.59**
	Prefrence Agreement	
	Edit Data Setting	
Expert	High to Low (Annotators agree to Hallucination)	Low to High (Annotators agree to Factuality)
GPT-4	0.73	**0.37**
GPT-3.5	**0.74**	0.33

Table 11: Inter-annotator agreement (mean kappa score between annotations) for hallucination/factuality edits annotation.

that lead to factuality improvements are referred to as factuality instructions.

The conditions for an edit instruction with either ADD or OMIT operation is a hallucination instruction is listed in Table 9, and the conditions for an edit instruction with either ADD or OMIT operation is a factulaity instruction is listed in Table 10. In these tables, the hallucination label is used to label if an instruction leads to hallucination in the edited summary or not (0=Hallucination instruction, 1=Not a hallucination instruction), and the factuality label is used to label if an instruction leads to factuality improvements in the edited summary or not (1=Factuality instruction, 0=Not a factuality instruction).

Given the article, reference/unaligned model-generated summary, and edit instructions generated by our pipeline, we asked the annotators to annotate each instruction with its hallucination/factuality label along with a justification comment for the annotation.

For preference analysis, we also asked the annotator to give a preference label to each of the reference - edited summary pair in the case of **High**→**Low**, and unaligned model generated - edited summary pair in the case of **Low**→**High**. A preference label was given for each of these summary pairs as follows:

High→**Low**:

- `Preference Label: 0` if the annotator would prefer the reference summary over the edited summary as the discharge instructions for the corresponding article.

- `Preference Label: 1` if the annotator would prefer the edited Summary over the reference summary as the discharge instructions for the corresponding article.

Low→**High**:

- `Preference Label: 0` if the annotator would prefer the unaligned model-generated summary over the edited Summary as the discharge instructions for the corresponding article.

- `Preference Label: 1` if the annotator would prefer the editted Summary over the unaligned model-generated summary as the discharge instructions for the corresponding article.

We also report the Kappa scores for both edit-instruction and preference label annotation used in our human evaluation, to further validate the quality of edits generated by our pipeline. Table 11 reports the mean kappa score between all the annotations (1) of the edit instructions from our human evaluation samples, where each annotator gives a hallucination/factuality label (0 or 1) to an edit instruction to annotate if the edit instructions leads to a hallucination/factuality edit; (2) of the preference label (0 or 1) given between the edit summaries and ground truth summaries by our annotators. We observe an overall high agreement between the annotators, where edit instructions were observed to have an higher agreement when GPT-3.5 was used as the expert for both the type of edits, whereas in the case of preference labels overall **High**→**Low** edit had an higher agreement for both the experts.

Table 12: Hallucination Edit Types

Instruction Abbrivation	Description
AR	Add from Reference Summary
AA	Add from Article
OR	Omit from Reference Summary
OA	Omit from Article

C Edit Instruction Categories

From our evaluation, we observed that there were majorly 4 types of edits made from our pipeline as shown in Table 12. These edits are categorized mainly based on (1) the operation used for the edit (ADD/OMIT), and (2) whether the edit was made using the contents from the reference summary (or unaligned model generated summary) or the article (clinical note).

In order to analyze the type of edits, we promoted GPT-4 to categorize each generated instruction into one of the edit type categories listed in Table 12. The prompt used is shown in Table 14.

For Figure 2, we first used the above prompt to categorize each generated instruction into one of the edit-type categories a identify the percentage-wise contribution of each edit type over all the generated instructions (top plot in Figure 2). Then we used the human annotation labels to calculate the percentage of instructions of each type responsible for generating the desired (hallucination/factuality improvement) outcome (bottom plot in Figure 2).

D G-Eval Factuality Metric Prompt

The prompt for our G-Eval evaluation is given in Table 13.

E Human Evaluation Examples

Examples of our human annotation are given in Table 16 & Table 17.

F External Evaluation Experimental setting

In this paper, We trained GPT 2 and Llama 2 on the summarization dataset with 3 epochs (batch size of 8). For GPT 2, the experiments take about 2 hours. For Llama 2, the experiments take about 20 hours. We did all the experiments with 4 NVIDIA Tesla RTX8000 GPU - 48 GB memory, with Adam optimizer – betas=(0.9,0.999), epsilon=1e-08, learning rate=1e-04. In all our summary generation, we used a beam size of 4, no-repeat-ngram-size=2, and minimum length and maximum length of sentences were set as (10, 256). We used five different random seeds to sample training data for all our experiments, and the scores reported in the tables are the average of these random seeds.

Table 13: G-Eval 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.

Table 14: Edit type categorization prompt

You would be given an Article, a reference summary and an edited summary along with some edit instructions.
An article is a clinical note and the reference summary is a summarisation of the clinical note.
Article: Article
Reference Summary: Ref_sum
Edited Summary: edit_sum
Instructions: ins_truc
The "edited summary" is generated from the "Article" and the "reference summary" using the Instructions.
Instruction Types: AR, AA, OR, OA
Task:
Your task is to act as an expert evaluator and categorize each instruction to exactly one of the above 4 instruction types.
Refer to the evaluation guide and the rules:
If there are no instructions. then the counts for each category would be zero.
Evaluation Guide:
Step 1: Classify each instruction to either an Add instruction or an Omit Instruction.
Step 2: Focus at the information mentioned in the instruction and find out, which paragraph among the reference summary and the article contains this information.
Step 3: If an exact match is not found, find which paragraph among the two contains a contextually similar content.
Step 4: Categorize the instructions using the rules.
Step 5: Output the count of each type of instruction in the
format: AR: <number>, AA: <number>, OR: <number>, OA: <number>
Rules:
- An instruction is to be categorized as AR if the instruction type is "Add" and if similar information can be found in the reference summary or if the instruction clearly tries to add some information before or after an instruction present in the reference summary.
- An instruction is to be categorized as AA if the instruction type is "Add" and if similar information can be found in the article.
- An instruction is to be categorized as OR if the instruction type is "Omit" and if similar information can be found in the reference summary.
- An instruction is to be categorized as OA if the instruction type is "Omit" and if similar information can be found in the article.

	ROUGEL	UMLS-F1
	GPT-3.5 Synthetic Dataset	
SFT	38.03	35.47
DPO	36.48	34.51
SALT	34.87	32.84
	GPT-4 Synthetic Dataset	
SFT	38.03	35.47
DPO	34.92	32.95
SALT	34.70	32.04

Table 15: Llama2-L2H

Table 16: Human Annotation for **High**→**Low**

Clinical Note
Brief Hospital Course: Patient was found to have blood loss anemia (HCT 40s –> 22) for which she was recuscitated in ED, received 4Units PRBC, 4 FFP, 2 mg vit K, 14 mg morphine, 1 mg ativan, 1600 cc NS. A foley was placed which revealed frank blood. She also had ARF (BUN/Cr 21/1.8). Abdomenal U/S, CT and cystoscopy revealed prevesicular and intrabladder hematomas in the context of anticoagulation along with blood loss into right lower extremity. The hematomas were evacuated (urology and vascular [**Doctor First Name **]), and an intra-op right ureteral stent was placed for her obstructive uropathy w/ right hydronephrosis. Pt developed post-surgical ileus that resolved in two days. RLE U/S was negative for IVC thrombus or LE venous compression. Intial ARF was likely pre-renal secondary to major volume loss and resolved with fluid resusitation. The ureter stent was removed once her Cr normalized. In transitioning her back to coumadin via heparin, she developed HIT (Plt 200s —–> 23 over 2 days). Given her high risk for thrombosis (PAF, Obese, immobile) she was initially treated with Argatroban (Per heme recs, Dr. [**Last Name (STitle) **]. However, she became jaundice w/ elevated bilirubin. It was reasoned that this could possibly be secondary to Argatroban liver toxicity in addition to low grade hemolysis (resorbing residual hematoma). A RUQ U/S failed to show any changes in her liver, biliary tree, or gallbladder. She was ultimately switched to Lepirudin for anticoagulation, bridged to coumadin when platellet levels reached 100,000, and obtained a theurapeutic INR before discharge.
Reference Summary
Discharge Instructions: 1. If you develop chest pain, shortness of breath, palpitations, lightheadedness, abdominal pain develop seek immediate medical attention. 2. Take all medications as listed below. A number of medications have been changed so please review the list carefully. Check with Dr. [**Last Name (STitle) 5310**] concerning when to restart your aspirin. 3. Follow up as directed below. 4. You will need an INR checked in 2 days with instructions on adjusting your coumadin dose by Dr. [**Last Name (STitle) 5310**] 5. You have been provided with a note excusing you from work for the next month. Please contact your PCP for instructions on when to return after that."
GPT-4 Generated Edit Instructions
Numbered List hallucination edits made: 1. Add Operation: Add "blood loss anemia" to the summary. 2. Omit Operation: Omit "shortness of breath" from the summary. 3. Add Operation: Add "HIT" to the summary. 4. Omit Operation: Omit "palpitations" from the summary. 5. Add Operation: Add "Argatroban liver toxicity" to the summary.
GPT-4 Generated Edited Summary
Discharge Instructions: 1. If you develop chest pain, lightheadedness, abdominal pain or symptoms of blood loss anemia and HIT, seek immediate medical attention.

2. Take all medications as listed below. A number of medications have been changed so please review the list carefully. Check with Dr. [**Last Name (STitle) 5310**] concerning when to restart your aspirin and in case of symptoms of Argatroban liver toxicity.
3. Follow up as directed below.
4. You will need an INR checked in 2 days with instructions on adjusting your coumadin dose by Dr. [**Last Name (STitle) 5310**]
5. You have been provided with a note excusing you from work for the next month. Please contact your PCP for instructions on when to return after that."

Human Annotation for GPT-4 Edited Summary (by a doctor)

Instruction 1:
Hallucination Label: 0
Comment: a description
Instruction 2:
Hallucination Label: 1
Comment: patient may ignore this
Instruction 3:
Hallucination Label: 0
Comment: nothing changes
Instruction 4:
Hallucination Label: 1
Comment: patient may ignore this
Instruction 5:
Hallucination Label: 0
Comment: nothing changes

GPT-3.5 Generated Edit Instructions

Numbered List hallucination edits made:

1. Add Operation: "Please note that" before "If you develop chest pain"
2. Omit Operation: Omit "shortness of breath"
3. Add Operation: "and" before "palpitations"
4. Omit Operation: Omit "lightheadedness"
5. Add Operation: "and" before "abdominal pain"
6. Add Operation: "Please" before "review the list carefully"
7. Add Operation: "Please" before "Check with Dr. [**Last Name (STitle) 5310**]"
8. Add Operation: "Please" before "Follow up as directed below"
9. Add Operation: "Please" before "You will need an INR checked in 2 days"

GPT-3.5 Generated Edited Summary

Discharge Instructions: Please note that if you develop chest pain and palpitations and abdominal pain develop seek immediate medical attention. Take all medications as listed below. Please review the list carefully. Please Check with Dr. [**Last Name (STitle) 5310**] concerning when to restart your aspirin. Please Follow up as directed below. Please You will need an INR checked in 2 days with instructions on adjusting your coumadin.

Human Annotation for GPT-3.5 Edited Summary (by a doctor)
Instruction 1: Hallucination Label: 0 Comment: Instruction 2: Hallucination Label: 1 Comment: this is important Instruction 3: Hallucination Label: 0 Comment: Instruction 4: Hallucination Label: 1 Comment: import disease Instruction 5: Hallucination Label: 0 Comment: Instruction 6: Hallucination Label: 0 Comment: Instruction 7: Hallucination Label: 0 Comment: Instruction 8: Hallucination Label: 0 Comment: Instruction 9: Hallucination Label: 0 Comment:

Table 17: Human Annotation for **Low**→**High**

Clinical Note
Brief Hospital Course: As mentioned in the HPI, pt was initially seen by cardiac surgery following his cardiac cath. Patient was eventually consented for surgery and on [**2155-2-13**] he was brought to the operating room where he underwent a coronary artery bypass graft x 3 and aortic valve replacement. Please see op note for surgical details. Patient tolerated the procedure well and was transferred to the CSRU in stable condition receiving Neo-Synephrine, Dobutamine, and Levophed drips. Later on op day sedation was weaned and patient awoke neurologically intact. He was then weaned from mechanical ventilation and extubated. He was weaned from all Inotropes/Pressors on post op day one and was then transferred to the cardiac step-down unit. B Blockers and Diuretics were initiated and patient was gently diuresed towards his pre-op weight. Chest tubes and Foley catheter were removed on post op day two. And epicardial pacing wires were removed on post op day three. Pt was followed by physical therapy during his entire post op course for strength and mobility. Patient had a relatively uncomplicated post op course and was at level 5 by post op day five. His exam and labs were stable on post op day five and he was discharged home with VNA services and the appropriate follow-up appointments.
Unaligned Model Generated Summary

Discharge Instructions: [**Month (only) 116 **] shower, no baths or swimming Monitor wounds for infection redness, drainage, or increased pain Report any fever greater than 101 Report any weight gain of greater than 2 pounds in 24 hours or 5 pounds in a week No creams, lotions, powders, or ointments to incisions No driving for approximately one month No lifting more than 10 pounds for 10 weeks Please call with any questions or concerns.

GPT-4 Generated Edit Instructions

Numbered List hallucination edits made:

1. Add Operation: Add "Patient underwent a coronary artery bypass graft x 3 and aortic valve replacement." from the article to the model generated summary.
2. Omit Operation: Omit "No creams, lotions, powders, or ointments to incisions" from the model generated summary.

GPT-4 Generated Edited Summary

Discharge Instructions:
Patient underwent a coronary artery bypass graft x 3 and aortic valve replacement. Shower, no baths or swimming. Monitor wounds for infection - redness, drainage, or increased pain. Report any fever greater than 101. Report any weight gain of greater than 2 pounds in 24 hours or 5 pounds in a week. No driving for approximately one month. No lifting more than 10 pounds for 10 weeks. Please call with any questions or concerns.

Human Annotation for GPT-4 Edited Summary (by a doctor)

Instruction 1:
Factuality Label: 1
Comment: useful
Instruction 2:
Hallucination Label: 0
Comment: This is not useful

GPT-3.5 Generated Edit Instructions

Numbered List hallucination edits made:

1. Add Operation: "Patient underwent a coronary artery bypass graft x 3 and aortic valve replacement."
2. Omit Operation: "No creams, lotions, powders, or ointments to incisions."
3. Add Operation: "Patient was transferred to the cardiac step-down unit."
4. Omit Operation: "No driving for approximately one month."
5. Add Operation: "Patient was discharged home with VNA services and the appropriate follow-up appointments."

GPT-3.5 Generated Edited Summary

Discharge Instructions: [**Month (only) 116**] shower, no baths or swimming. Monitor wounds for infection - redness, drainage, or increased pain. Report any fever greater than 101. Report any weight gain of greater than 2 pounds in 24 hours or 5 pounds in a week. No lifting more than 10 pounds for 10 weeks. Please call with any questions or concerns. Patient underwent a coronary artery bypass graft x 3 and aortic valve replacement. Patient was transferred to the cardiac step-down unit. Patient was discharged home with VNA services and the appropriate follow-up appointments.

Human Annotation for GPT-3.5 Edited Summary (by a doctor)

Instruction 1:
Hallucination Label: 1
Comment: Knowing what surgeries were performed is important to patient.
Instruction 2:
Hallucination Label: 0
Comment: This is common sense like notification, but I think this is also importatnt.
Instruction 3:
Hallucination Label: 0
Comment: This explanation sounds not helpful in discharge note.
Instruction 4:
Hallucination Label: 0
Comment: Be careful with all kinds of risks.
Instruction 5:
Hallucination Label: 1
Comment: A notification for future plan