Generating more faithful and consistent SOAP notes using attribute-specific parameters

Sanjana Ramprasad RAMPRASAD.SA@NORTHEASTERN.EDU
Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA

Elisa Ferracane ELISA@ABRIDGE.COM
Abridge AI Inc.
Pittsburgh, PA, USA

Sai P. Selvaraj PRABHAKARSAI@ABRIDGE.COM
Abridge AI Inc.
Pittsburgh, PA, USA

Abstract

The widespread adoption of SOAP notes for documenting diverse aspects of patient information in healthcare has been prevalent. However, the conventional process of manual note-taking is laborious and can distract healthcare providers from addressing patients' needs. Prior work by Krishna et al. (2021a) has introduced an end-to-end pipeline for generating SOAP notes, but model-generated notes are susceptible to inaccuracies, irrelevant and missing information. In this work, we assess the performance of large language models (GPT-3.5) for SOAP note generation, compare them with fine-tuned models using automated metrics, and propose a solution to improve the consistency and faithfulness of notes by incorporating attribute-specific information via SOAP section information. To achieve this, we integrate an extra layer of unique section-specific cross-attention parameters to existing encoder-decoder architectures. Our approach is evaluated using a comprehensive suite of automated metrics and expert human evaluators, demonstrating that it leads to more accurate, relevant, and faithful information.

1. Introduction

Electronic Health Records (EHRs) have become an indispensable tool for healthcare providers to document and monitor various facets of patient information efficiently. Despite its benefits, the vast amount of data contained within these records can pose difficulties for clinicians, such as information overload, which may result in adverse patient outcomes if the information recorded is incorrect, insufficient, or irrelevant. The SOAP (Subjective, Objective, Assessment, and Plan) method is widely adopted to mitigate these challenges and organize notes succinctly and articulately while effectively conveying vital patient information.

The SOAP framework (Podder et al., 2021) systematically documents patient information during clinical interactions. This framework encompasses four components: the

patient's self-reported symptoms and complaints, represented in the Subjective section; the clinician's observations, recorded in the Objective section; the clinician's diagnosis, documented in the Assessment section; and the treatment plan, outlined in the Plan section. The SOAP method is employed in clinical practice as a framework to structure and communicate patient information. However, manual note-taking utilizing the SOAP method can hinder clinicians during patient interactions, potentially compromising their ability to fully attend to the patient's needs. Developing an automated SOAP note generation system could be a potential solution for improving the standard of care provided by healthcare providers.

Subjective (Past Medical History)
DR: How's your breathing going?
PT: Breathing has been doing great
DR: Really? Good, even with the cold
PT: Yeah
DR: Good good, yeah that's -
PT: Oh numbers look borderline, but just round them up
DR: Yeah

Reference (Human written): Patient had shortness of breath

GPT 3.5(Zero Shot): During the conversation, the doctor asked the patient about their breathing and the patient responded that it has been great, even with a cold or other issues. The doctor expressed satisfaction with this response and the conversation briefly touched on other topics.

GPT 3.5(Few Shot) : No relevant past medical history related to breathing.

BART (baseline) : Breathing has been doing great

BART (Our modification) : Patient had breathing problem

Objective (Immunizations)
DR: Just be coming in fasted and do it that way, okay?
DR: So I would just take an iron. So, 5 refills and you go to ePrescribe. And you're going-
PT: [DEIDENTIFIED]. Oh Okay.
DR: Why don't I just send over there?
PT: Oh and they tell me this other one cannot be sent in.
DR: We know that. I already sent your prescription. All right, then.

Reference Summary (Human written): Prescribed iron tablet (already sent the prescripton, 5 refills)

GPT 3.5(Zero Shot): The doctor advises the patient to come in fasting for a test and prescribes five refills, suggesting ePrescribe. They discuss sending a prescription and mention that one cannot be sent in. The doctor confirms that they have already sent the prescription.

GPT 3.5(Few Shot) : Prescribed five refills to be sent through ePrescribe. Advised to come in fasting for a test. One prescription cannot be sent in, but already sent the others.

BART (baseline) : Eprescribe

BART (Our modification) : Iron (5 refills)

Figure 1: Comparison of summaries written by humans with those generated by GPT-3.5, BART baseline and our proposed model.

Previous work has investigated methods for summarizing transcripts into notes by using entire conversations as input into a summarization system (Enarvi et al., 2020; Krishna et al., 2021a). However, the most effective approach was found to be the summarization of localized dialogic exchanges or segments, which inherently serve as concise sources of evidence for the SOAP sections and subsections of interest (Krishna et al., 2021a). In line with that framing, work on generating SOAP notes from conversational data has adopted a two-stage strategy. The first stage involves the extraction of evidence snippets or utterances, where relevant and salient utterances that support the targeted section or sub-section are singled out (Schloss and Konam, 2020; Krishna et al., 2020). The second stage encompasses summarization, in which crucial information from the identified evidence utterances is condensed and integrated into the SOAP note. This step has typically included using

a standard sequence-to-sequence model conditioned on evidence utterances along with the section of interest by pre-pending to the input (Krishna et al., 2021a; Su et al., 2022).

While systems can achieve a high level of accuracy for the task of predicting evidence utterances (Krishna et al., 2021a), existing models employed for summarizing evidence utterances and conversations are subject to limitations such as hallucination, where the generated summaries are inconsistent or contain unsupported information in comparison with the source text (Cao et al., 2017; Kryściński et al., 2019; Maynez et al., 2020a; Nan et al., 2021a; Pagnoni et al., 2021; Tang et al., 2022). This problem is particularly worrisome in consequential domains such as medical documents (Wallace et al., 2021; Otmakhova et al., 2022). Generating inaccurate SOAP notes can have detrimental effects on patient care and may even result in legal repercussions (Seo et al., 2016). While there has been some work on improving the consistency of model outputs in generic news domains (Cao et al., 2017; Zhu et al., 2021; Nan et al., 2021b; Maynez et al., 2020b; Wan and Bansal, 2022), there has been limited research on evaluating and improving these systems in the biomedical domain (Wallace et al., 2021). Notably, no specific work has focused on improving the faithfulness and consistency of SOAP notes. In this work, we aim to generate summaries that refrain from introducing new facts (faithful) and that do not misconstrue existing information (consistent). Our model architecture changes are motivated by the following observations (highlighted in Figure 1).

- Different sections of a SOAP note encompass distinct types of information by focusing on different aspects of the conversation. For example, the Subjective section typically comprises symptoms and medication names, while the Objective section encompasses procedures, dosages, and frequencies.

- Each section summary articulates information in different writing styles. The subjective and objective sections are concise, one- or two-line summaries of evidence utterances, while the assessment and plan sections are more extensive and descriptive.

More recent work has highlighted the promising results of large language models (LLMs) in zero-shot and few-shot automatic summarization with news datasets (Goyal et al., 2022; Zhang et al., 2023), but their effectiveness in niche domains and datasets remains largely unexplored. In this study, we investigate the performance of OpenAI's GPT-3.5 model for summarizing SOAP notes using both zero-shot and few-shot settings. We further evaluate and compare the performance of LLMs specifically GPT-3.5 against fine-tuned models (BART) to determine their relative effectiveness in this task. Using zero-shot and few-shot techniques with prompts detailing the SOAP sections and subsections, we find that the generated summaries lack the stylistic coherence of SOAP notes when compared to human-written examples. We highlight examples in Figure 1. In all the examples displayed, GPT-3.5 tends to summarize all information in the snippet, rather than limiting itself to pertinent information for a given section. Since dialogue exchanges often include back and forth between multiple topics, this leads to surfacing irrelevant information for a respective section. Furthermore, the model fails to draw inferences based on dialogue exchanges. For instance, in the few-shot example for the Subjective section shown in Figure 1, the model is unable to infer information based on the conversation where the doctor and

Subsection	Data Distribution	Avg Summ Len	Avg Input Len
Allergies	2.31	9.60	90.58
Chief Complaint	8.81	6.31	86.42
Family Medical History	6.44	11.11	89.93
Medications	8.84	8.78	92.43
Past Medical History	8.81	6.43	86.57
Past Surgical History	8.78	8.17	88.84
Social History	7.94	8.44	84.89
Laboratory and Imaging Results	8.82	12.89	97.74
Immunizations	3.91	9.27	82.11
Assessment	8.84	56.71	254.82
Diagnostics and Appointments	8.82	11.20	100.58
Prescriptions and Therapeutics	8.80	12.34	118.87

Table 1: The statistics for subsections in Subjective, Objective and Assessment and Plan sections respectively. Data distribution indicates the percentage of subsection data in relation to total data. The average length of the summaries and input sequences is determined by utilizing the BART tokenizer.

patient discuss breathing issues without directly mentioning it. Our fine-tuned model, on the other hand, infers information based on the conversation and summarizes notes that conform to the appropriate section. While we present more empirical results later (section 6), our observation that LLM-generated summaries do not adhere to the SOAP note style leads us to use fine-tuning as a means of controllable summarization to include only relevant content for SOAP sections.

Generalizable Insights about Machine Learning in the Context of Healthcare

To our knowledge, there is no previous work that specifically focuses on improving the faithfulness and consistency of automated SOAP note generation. Thus our main contribution in this work is our model architecture which seamlessly integrates section-specific information leading to more accurate SOAP notes. In particular, we demonstrate the effectiveness of our approach through Figure 1. Our method not only improves the faithfulness and consistency of the generated SOAP notes, as confirmed by a suite of automated metrics, but also outperforms existing state-of-the-art approaches. These findings are also reinforced by human evaluation. In addition to our above contribution we also benchmark the performance of different categories of models for SOAP note generation from medical conversations. Furthermore, we introduce a new way of automatically evaluating SOAP note generation through the use of structured data.

2. Dataset

This study builds on the SOAP note dataset developed by Abridge AI[1], similar to what was introduced by Krishna et al. (2021a). The dataset comprises transcripts of English-language interactions between healthcare providers and patients, obtained through informed consent from the patients. To preserve patient privacy, all personal health information was removed through a meticulous de-identification process. Due to the confidential nature of the data, it is not publicly accessible.

Experienced annotators proficient in the SOAP note protocol were tasked with creating SOAP summaries from the conversational text and identifying respective evidence utterances. Our objective in this study was to generate summaries from evidence utterances; thus, we implemented a uniform sampling strategy across all sections and subsections to establish a well-balanced dataset that accurately reflects performance across all sections. As a result, each data point in our dataset comprises information on the section, subsection, evidence utterance, and summary.

The summaries inherently show variation in the type of information and the writing style in different sections of the SOAP notes. The Subjective section, which comprises the information reported by the patient, and the Objective section consisting of observations made by the clinician, were found to include verbatim text from the conversation and were 1-2 sentences long. The Assessment section synthesizes all available evidence and is typically more abstractive and lengthier. The Plan section typically outlines the prescribed medications and diagnostic procedures and is short, often just including a couple of words, such as medication names.

We present an overview of the dataset statistics in Table 1. An analysis of the distribution of various subsections in the dataset revealed a relatively even spread, except for the allergies subsection within the Subjective section and the immunization subsection within the Objective section, which was less frequently encountered in our transcripts. The BART (Lewis et al., 2019) tokenizer was utilized to determine the sequence lengths of the summaries and evidence utterances. Our findings indicated that while some subsections tend to have longer summaries, most sections contain fewer than 12 tokens, emphasizing the need for succinctness as a prevailing characteristic; this highlights the significance of capturing accurate information efficiently.

Our experiments used a dataset of 146,920 pairs of evidence utterances and corresponding SOAP summaries. To form our test set, we randomly sampled 10% of the data across all subsections, while the remaining data was used for training, with an additional 10% set aside as a validation set. We fine-tune the hyperparameters of the model based on the model's performance on the validation set.

3. Preliminaries

3.1. Adapter Module

The emergence of adapter modules (Houlsby et al., 2019; Wang et al., 2020; Pfeiffer et al., 2020) introduced a new mechanism of transfer learning for adapting pre-trained models without the need for fine-tuning all the weights. These modules incorporate a limited

1. https://www.abridge.com/

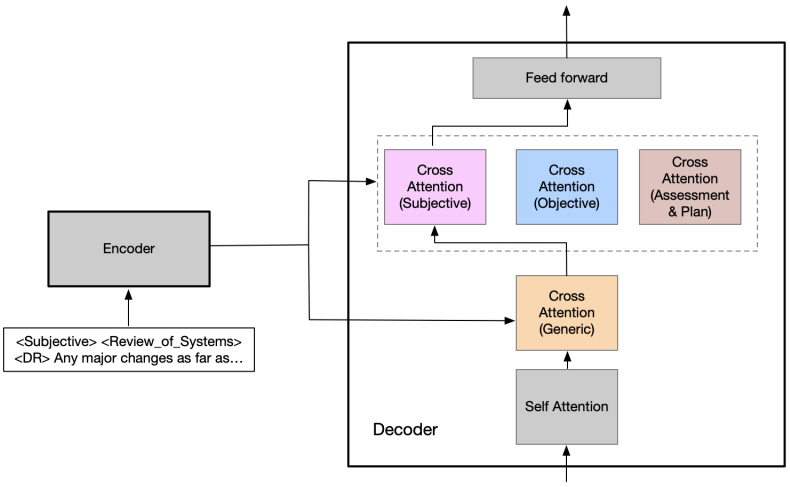

Figure 2: Overview of our architecture on an example evidence utterance to generate a note for the Subjective section.

number of new parameters (ϕ) that are learned in the context of a specific target task, while maintaining the pre-trained model parameters (θ) in a fixed state. This approach enables the adapters to learn task-specific representations. Adapter modules typically consist of a two-layer feed-forward neural network with a down-projection followed by an up-projection, creating a bottleneck structure. These networks are introduced at each transformer layer. Previous work (Houlsby et al., 2019; Bapna and Firat, 2019; Wang et al., 2020; Stickland and Murray, 2019) has investigated the impact of different adapter placements and layer normalizations.

4. Methods

In the following experiments, we employ BART (Lewis et al., 2019) as the transformer-based encoder-decoder architecture for fine-tuning(Vaswani et al., 2017), although the methods used can be applied to any similar architecture. We provide a concise overview of our methodology to 1) Establish a baseline of LLM generated summaries against fine-tuned baselines 2) incorporate section-specific parameters with token embeddings, 3) implement adapters for individual sections, and 4) introduce a novel approach for adding a distinct cross-attention layer for sections in sequence after the standard cross-attention layer.

4.1. Models

OpenAI's GPT-3.5: GPT-3 (Brown et al., 2020) is a state-of-the-art language model capable of performing on a wide range of tasks including summarization. In this work, we use a more recent variant gpt-3.5-turbo-0301 to query for zero shot and few-shot. In the zero shot setting we use the template as shown below.

```
Dialogue Snippet: [Dialogue Snippet]
```

```
Summarize the above dialogue snippet for SOAP
subsection [subsection] under the [section] section
Summary:
```

For the few shot setting we prepend in-context examples using the same template above. For each section/subsection dialogue snippet, we randomly select two examples from the training set and use as in-context examples.

BART: is a sequence-to-sequence model. The model consists of a bidirectional encoder and an autoregressive decoder and has been trained using a noising function that corrupts the input text, followed by training to reconstruct the original text (Lewis et al., 2019). As a baseline we use the vanilla bart model and provide section and subsection information through special tokens as part of the input. We pre-pend the special tokens before the dialogue snippets similar to the approach followed by Krishna et al. (2021a). Furthermore, we include speaker information using special speaker tokens. Eg. $\langle Subjective \rangle \langle Review\_of\_Systems \rangle \langle DR \rangle ...$

BART + section embeddings In order to improve the input representation of the BART model, we add section and subsection embeddings to its standard input and position embeddings. This approach draws upon previous research that leverages special embeddings to convey various types of information (Sundararaman et al., 2019). To obtain the embeddings, we define additional special tokens for each of the sections and subsections and compute the corresponding token embeddings, in addition to the position and input embedding. The objective of incorporating section and subsection embeddings is to impart explicit guidance on the type of information that should be reflected in the generated output.

BART + section adapters To improve the model's versatility in handling different sections of the SOAP note, we incorporate an adapter module for each section using the adapter configuration proposed by Bapna and Firat (2019). While conventional approaches have recommended freezing model parameters and only fine-tuning the adapters for optimal performance, our objective is to utilize the adapters to obtain separate section representations. Therefore, we adopt a fine-tuning strategy similar to Stickland and Murray (2019), where the entire architecture, including the adapter modules, is fine-tuned. The input to the model is the same as what we use for the vanilla BART model with special tokens indicating section, subsection and speaker information prepended to the evidence utterance. The section information contained within the input is then used to select the appropriate adapter module to employ.

4.2. Proposed Cross-Attention Layer

We propose a modified Transformer architecture (Vaswani et al., 2017) as shown in Figure 2 that adds unique, randomly initialized cross-attention parameters for each SOAP section as an extra sequential layer on top of the cross-attention layer in the original architecture (which we refer to as global cross-attention). The two layers allows the model to capture both section-specific information and shared information between sections.

Consider the decoder with input X and self attention layer o_{sa} where

$$o_{sa} = \text{softmax}\left(\frac{Q_{sa}{K_{sa}}^T}{\sqrt{d_k}}\right)V_{sa} \tag{1}$$

$$o_{sa} = \text{LayerNormalization}(X + o_{sa}) \tag{2}$$

The output of the self attention layer o_{sa} acts as the query to the cross attention layer which captures shared information. The encoder hidden states act as keys and values.

$$o^g_{ca} = \text{softmax}\left(\frac{Q_g {K_g}^T}{\sqrt{d_k}}\right) V_g \tag{3}$$

$$o^g_{ca} = \text{LayerNormalization}(o^g_{ca} + o_{sa}) \tag{4}$$

The main modification in our proposed architecture includes the incorporation of an additional layer of three distinct cross-attention parameters for each section (Subjective, Objective, Assessment and Plan) in a sequential manner to the global cross attention. Specifically, the output o^g_{ca} serves as the query, while the encoder hidden states act as the keys and values. The choice of cross-attention parameter in this layer is determined by the section of interest, which is provided as a prefix in the input. Eg. consider generating a summary for a dialogue snippet of the Subjective section.

$$o^{subj}_{ca} = \text{softmax}\left(\frac{Q_{subj} {K_{subj}}^T}{\sqrt{d_k}}\right) V_{subj} \tag{5}$$

$$o^{subj}_{ca} = \text{LayerNormalization}(o^{subj}_{ca} + o^g_{ca}) \tag{6}$$

The output from o_{ca} is then passed through the fully connected network followed by a language model head. We provide implementation and training details of the models in Appendix A

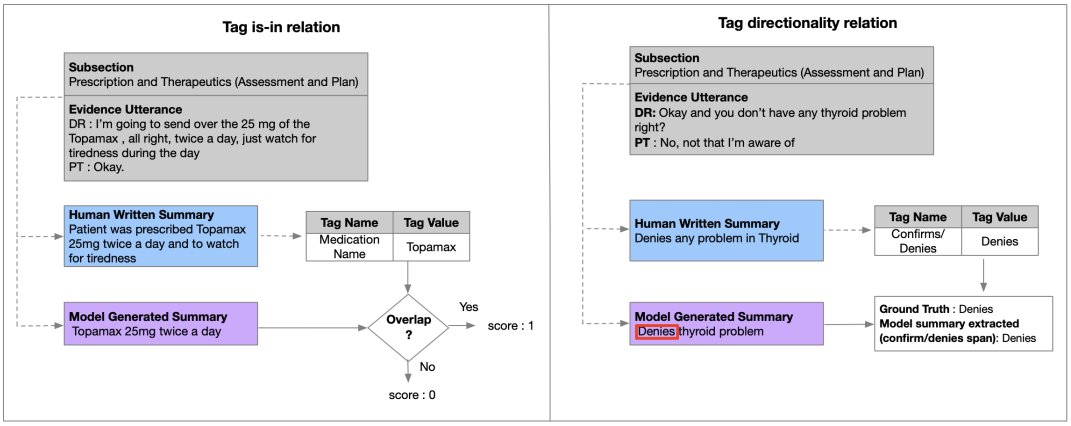

Figure 3: Example of how tags with "is-in" relation are scored given a medication name and summaries (left). Directionality evaluation with an example tag name and value is shown where spans that match pre-defined values (either confirms or denies here) is extracted from model generated summary and compared with ground truth tag values through precision/recall scores (right).

5. Evaluation and Results

5.1. Automatic Evaluation of Summaries

To measure the consistency and faithfulness improvement of generated SOAP summaries, we use a variety of standard and structured automated metrics. For all models we report scores on the test set created by sampling data from each section as described in Section 2.

5.1.1. Automated Standard Metrics

ROUGE : We report performance of our model generated summaries by comparing them to human-written reference summaries using the ROUGE-L metric. The ROUGE scores (Lin, 2004) are based on the exact word overlap and therefore provide insight into the informativeness of the generated summaries, but do not necessarily reflect their factual accuracy (Maynez et al., 2020b).

UMLS concept overlap : We use the UMLS concept overlap metric to evaluate the factual accuracy of generated medical summaries. Unified Medical Language System (Bodenreider, 2004) plays a crucial role in promoting interoperability in the biomedical domain by integrating and distributing a vast range of biomedical terminologies, classification systems, and coding systems from various sources. This ontology helps resolve differences in meaning and representation of biomedical concepts across different sources. To extract medical named entities in text and align them with the relevant biomedical concepts in UMLS, we utilize QuickUMLS (Soldaini and Goharian, 2016), a tool designed to distinguish and disambiguate entities. The use of QuickUMLS enables us to accurately map the named entities in the generated summaries to the corresponding concepts in UMLS and thereby evaluate their factual accuracy.

We evaluate summaries through the computation of precision and recall metrics where precision is computed as the fraction of concepts in the generated summary present in the reference summary, serving as an indicator of the factual accuracy of the generated summary compared to the reference. On the other hand, recall is utilized to assess the relevance of the information contained within the generated summary with regard to the targeted content. The dual determination of precision and recall was deemed essential, as it is imperative for the generated notes to not only be pertinent to the designated section/subsection but also to exhibit a high degree of accuracy.

Given concepts from the reference summary C_{ref} and the generated summary C_{gen} we calculate recall as $\frac{C_{\text{ref}} \cap C_{\text{gen}}}{C_{\text{ref}}}$ and precision as $\frac{C_{\text{ref}} \cap C_{\text{gen}}}{C_{\text{gen}}}$. We report F1 scores calculated using the precision and recall as described above.

5.1.2. Automated Structured Metrics

The reference metrics described above effectively capture the relevance and accuracy of the generated summaries compared to the oracle, based on the analysis of words and medical concepts. However, we delved into finer details to obtain a more comprehensive evaluation of factuality by conducting a structured annotation study. For each section, the evidence utterance and the corresponding human-written summary were analyzed and annotated using a pre-defined set of tags, providing a more nuanced evaluation of factuality as shown in Figure 3.

Tags with an "is-in" relation: In order to comprehensively evaluate the relevance of the generated summaries, we developed an extensive list of tags for each subsection, which define the various types of information that can surface in each subsection. The complete list of tags is included in Appendix B. During the annotation process, annotators were instructed to select relevant spans from the human-written summary and the corresponding evidence utterance and assign the appropriate tag. For each human-written summary and annotated tag, we compared the presence of the tag value in the generated summary. These tags, which encompass spans from the reference summary, were utilized to evaluate the "is-in" relationship between the generated summary and the reference.

Tag values are of freeform nature and encompass phrases or single medical words and dosages as presented in the human-written summary which makes this assessment a reliable indicator for omitted information. We assess the overlap between tag values and words that appear in the generated summary. If there is no overlap between the tag value and the generated summary, a score of 0 is assigned. The final score is computed as the fraction of tags found in the generated summaries out of all annotated tags.

Tags with a directionality relation: In several sections, ensuring the proper directionality of information pertaining to medications, symptoms, and side effects is a crucial concern. For instance, it is critical to differentiate between past and current medications or the initiation and discontinuation of a medication. Similar directionality considerations arise in sections such as the Review of Systems, where patients are queried about the presence or absence of symptoms. These types of tags, which reflect the directional aspect of the information, are referred to as "directionality tags" and are based on a limited set of possible values specific to the section being evaluated. The complete list of tags is included in Appendix C. Since medical summaries often include information about directionality, such as dosages, terms, or medical concepts, evaluating the accuracy of this aspect is essential to ensure the reliability of the generated summaries. For example, a summary that indicates starting a medication when the ground truth states stopping the medication can lead to critical errors if not reviewed by physicians. Traditional n-gram overlap metrics or concept coverage metrics do not accurately reflect such critical inaccuracies. To address these limitations, we include a section in our evaluation process that assesses the overlap between the pre-defined directionality tag value and words in the generated summary. If overlaps are detected, the predicted value is assigned as the overlapping span, if direction in the generated summary differs from the pre-defined or there is no mention of direction then value the predicted value is considered "incorrect." We report precision and recall scores for these evaluations to provide a comprehensive analysis of the accuracy of the generated summaries.

6. Results and Discussion

6.1. Fine-tuned vs LLM generated summaries

We first evaluate SOAP note summaries generated by large language models (GPT-3.5) and a fine-tuned BART model on a sample of 500 test set data points and present results in Table 2 . We train BART using the same approach as outlined in section 2. Prompt template details for large language models are provided in the section.

Model	Rouge-L	UMLS(F-1)	Direc(F-1)	Is-in
GPT 3.5 (Zero-Shot)	15.4	54.34	16.08	**86.42**
GPT 3.5 (Few-Shot)	26.15	55.36	18.61	78.96
BART (baseline)	**42.87**	59.81	29.93	68.88
BART + section CA	42.4	**60.63**	**31.66**	74.13

Table 2: Automated standard and structured metric comparison of GPT3.5, vanilla BART and BART + section CA on a sample of 500 data points.

GPT generated summaries show lower ROUGE-L scores compared to human-written reference summaries and the baseline method, possibly indicating stylistic differences. GPT summaries also perform poorly on F-1 scores for UMLS overlap and directionality compared to BART, but achieve a higher score for relationship "is-in" measure since it is a recall score and GPT sumamries tend to include all information found in evidence utterances. Overall, BART baseline outperforms GPT and leads to better performance on our proposed approach.

6.2. Comparison of different approaches to introduce section specific parameters

Since fine-tuned BART model outperforms zero-shot/few-shot GPT generated summaries, we use BART to incorporate section specific information. We present the results of various approaches employed for integrating section specific parameters with the BART model, as shown in Table 3.

BART-based models show similar performance based on ROUGE scores, but there is a decrease in UMLS precision when using section embeddings and a slight increase when including section-specific adapters and cross-attention parameters as evaluated by UMLS scores (Table 3). This suggests that the generated summaries are more accurate compared to baselines. Additionally, there is a more noticeable improvement in recall with the use of section-specific parameters, which implies that the summaries are more faithful to the style. Overall, our approach shows improvements in UMLS-F1 scores.

We also find that the integration of section specific information leads to a consistent and marked improvement in the performance of automated structured metrics for "is-in" relationships, albeit bearing some resemblance to UMLS recall. Notably, however, our findings indicate that the aforementioned improvement surpassed that of UMLS recall scores, possibly owing to the tags' concentration on pivotal concepts and also being an entirely recall-based measure. Moreover, our analysis demonstrates that directionality metrics were most affected by section embeddings, whereas adapter and cross-attention modules also contributed to a noticeable enhancement. In conclusion, our automated metric results reveal that the incorporation of section-specific parameters leads to superior performance compared to the baseline, with our approach emerging as the top-performing method in three out of four metrics evaluated.

Model	Rouge-L	UMLS(F-1)	Direc(F-1)	Is-In
BART	42.27	61.72	24.90	66.87
BART + section EMB	42.34	60.78	**28.80**	69.54
BART + section ADAPT	<u>42.86</u>	<u>62.17</u>	27.11	<u>71.87</u>
BART + section CA	**43.28**	**62.83**	<u>28.08</u>	**72.35**

Table 3: Performance comparison of all BART based models described in the methods section on automated standard and structured metrics on entire test set. Best performing scores are highlighted and second to best is underscored.

Model	Rouge-L	UMLS(F-1)	Direc(F-1)	Is-in
T5 pointer gen	30.60	51.2	19.98	58.85
T5 + section prefix	38.61	59.00	18.64	64.92
T5 + section CA (ours)	**39.00**	**59.00**	**22.29**	**65.83**

Table 4: Automated metric comparison with Krishna et al. (2021a)(T5 +section prefix) and Enarvi et al. (2020)(T5 pointer gen) who use a pretrained model with section information appended to the beginning of the text and a transformer-based pointer generator respectively. All experiments are based on T5-base.

6.3. Comparison with previous approaches

We finally compare our approach to previous approaches that use doctor-patient conversations and summarization techniques to generate medical notes, which are pertinent to our work. These include work done by Krishna et al. (2021a) and Enarvi et al. (2020), who utilize a pretrained T5 model with section information appended to the beginning of the text and a transformer-based pointer generator, respectively. To ensure fair comparison, we implement our approach using section specific cross attentions with T5 model as well as shown in Table 4 . Implementation details are same as in Appendix A

Overall, our proposed method yields improved performance compared to approaches presented in prior works. We see comparable performance with Krishna et al. (2021a) in terms of UMLS-based metrics, yet see improved results using Rouge and greater improvements through structured metrics.

6.4. Qualitative Evaluation

A qualitative comparison between the BART baseline model and our proposed section-specific cross attention model was performed to demonstrate the advantage of incorporating section-specific parameters as shown in Figure 4. The results reveal information loss in the baseline summary, as seen in Example 1 for the "Review of Systems" subsection and Example 3 for the "Laboratory and Imaging" subsection. The most critical information, such as details of the "eye" and "blood work" systems, are missed by the baseline in the respective sections. In Example 2, the baseline model fails to emphasize the crucial information in

Evidence Utterances	Section/ Subsection	Reference Summary	BART summary	BART+ section CA
DR: Any, any major changes at all as far as - PT: Hum-um. DR: The inflammation per the ophthalmologist? PT: No.	Subjective (Review of Systems)	Denies changes with eye inflammation	Denies any major changes	Denies inflammation per ophthalmologist
PT: My mom passed away at age [AGE] of massive heart attack. PT: Um, I lost my father in, uh, March, then I lost an aunt a few weeks after that and then my brother. DR: So, you think it's more stress related flare up or? PT: You mean my knees and stuff? DR: Your joints. DR: This could be the more osteo, too, but this is definitely, uh, rheumatoid.	Subjective (Chief Complaints)	Stress. Rheumatoid Arthritis.	Patient's mother passed away due to massive heart attack. Patient lost father in March then lost aunt a few weeks later	Patient has rheumatoid arthritis
PT: And I did my blood work yesterday. DR: Oh, and I got it. DR: Your white blood count's always low, your blood, your hematocrit is always low, but everything else is good. PT: Okay.	Objective (Laboratory and Imaging Results)	Blood work was done yesterday revealed low white blood count and low hematocrit but everything else as good	Patient had blood work yesterday which was stable	Blood work yesterday was stable. White blood count is low. Hematocrit is low but everything else is good
DR: Do you have other questions or concerns for me before I let the lab folks come get you? DR: I'll have you wait here. DR: And I'll send in the lab folks to check to see how things are going. PT: Well, I'm usually over this way every day anyway. PT: I got to go see my wife every day.	Assessment and Plan: (Diagnostics and Appointments)	Patient to do lab test today	Patient will be seen everyday	Lab work will be done today
PT: I keep on saying 19, though, every once in a while. PT: [DEIDENTIFIED], yeah. DR: And who is the president now? PT: Uh, Bush. PT: Um, Trump. DR: Who was the president before that? PT: Uh, I can't remember.	Assessment and Plan (Assessment)	Patient was asked to mention the name of the present president, which he fumbled answering. He was asked to name the earlier president which he could not remember.	Patient is still on 19mg of metformin every once in a while	Patient was not able to remember the name of the president.

Figure 4: Qualitative differences through the introduction of section specific parameters (cross attention).

the Chief Complaints subsection and instead focuses on irrelevant parts of the conversation. This tendency is also evident in Example 4 for the Diagnostics and Appointments subsection. Our qualitative analysis shows that by including section-specific information, the model is able to surface more relevant information.

In Example 5 of the assessment subsection, the generated summary exhibits both irrelevance and inconsistency. Our analysis shows that the baseline model focuses on inaccurate spans within the evidence clusters and incorporates unsupported information. Specifically, the model misinterprets "19" as a dosage and mentions "metformin" as a medication name, which is not mentioned in the evidence utterance. These limitations of the baseline model underline the significance of our proposed approach in enhancing the consistency and faithfulness of generated summaries for the assessment subsection.

6.5. Manual Evaluation of Summaries

A manual annotation process was carried out on a set of 60 summary triples, consisting of a human written summary, along with the baseline summary and our proposed section-specific cross attention model summary. To maintain a blind evaluation, the baseline and our proposed summary were shown in a randomized order. To evaluate the summaries,

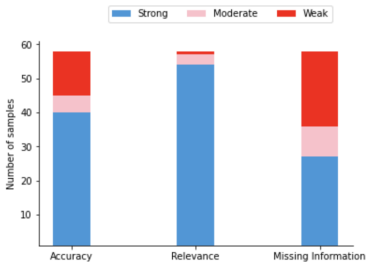

Figure 5: Manual Evaluation results on model generated summaries using BART model.

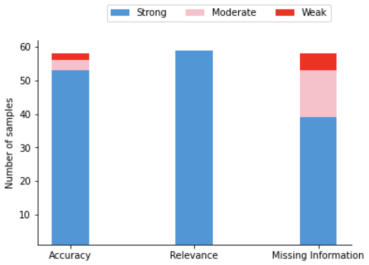

Figure 6: Manual Evaluation results on model generated summaries using our proposed model (BART + section CA).

the annotation questions included the SOAP section, subsection, evidence utterances, and human-written summary as reference, along with three questions for each system. For each system summary, the first question was to assess the summary's 'relevance' to the section and subsection, in order to determine if the information presented was appropriate. The second question was to evaluate the 'accuracy' of the summary by comparing it to the reference summary and the evidence utterances, and determining if the summary under consideration was precise or showed inconsistencies. The third question was to assess if system summary contained any 'missing information', compared to the reference summary. Through these three questions, the evaluation aimed to obtain a comprehensive assessment of the system, encompassing precision, relevance, and missing information. In order to perform the assessment, we procured the services of a medical expert with familiarity with SOAP notes, sourced from UpWork. Financial constraints limited us to securing the involvement of only one expert.

We highlight the evaluation results in Figures 5 and 6. Upon evaluating the generated summaries, it is evident that both models produce highly relevant summaries. However, the baseline model has a tendency to include some weakly irrelevant information. Our proposed model, which incorporates section-specific parameters, effectively addresses this issue. The primary distinctions between the two models is in terms of factual accuracy and the inclusion

of missing information. Our proposed model demonstrates a noticeable improvement in the accuracy of generated summaries, as well as a reduction in the occurrence of major errors (labelled as "weak"). Similarly, the performance in terms of missing information also shows an improvement with the inclusion of these parameters.

7. Conclusion

This paper proposes new techniques to improve the accuracy and fidelity of SOAP notes. We explore input concatenation, specialized embeddings, and adapters to incorporate section-specific information. Additionally, we introduce a novel approach that utilizes section-specific cross-attention parameters, yielding improved accuracy, relevance, and reduced missing information. Our study utilizes automated metrics and human evaluation to validate our approaches.

8. Limitations and Ethics

Limitation: Although our research has aimed to improve the faithfulness and consistency of SOAP notes, manual evaluation results show that automated notes may still be incomplete or inaccurate. Therefore, it is essential for medical professionals to carefully monitor generated notes to guarantee their accuracy and completeness. The human annotation evaluation was limited in sample size due to the high cost of expert annotators, who were paid 50$USD/hour and required 3-4 hours to annotate a batch size of 60 annotations. To achieve a more accurate assessment of new methods, it is necessary to conduct annotations on a larger sample size. Our methods show improvements on human annotated evidence utterances and assume the availability of accurate snippets, however at inference time these utterances will be predicted using a supervised model (Krishna et al., 2021a). Inaccuracies in predicted utterances for sections could further lead to inaccurate summaries.

We acknowledge that our implementation utilizing LLMs has provided a useful starting point in establishing a baseline for SOAP note generation. However, we recognize that the scope of our work is limited to this initial exploration and that further research is required to fully realize the potential of these models in this domain. Specifically, we recommend investigating different prompt techniques to improve the quality of generated summaries. We emphasize that extensive validation of generated summaries by healthcare professionals is necessary before practical implementation. This is because zero-/few-shot prompting of LLMs, without specific training on SOAP note data, may not capture the full nuances and context of clinical narratives. In this work, we also propose a method to automatically assess model-generated summaries using annotated structured tags from human-written references. We compare model-generated summaries using a naive approach that verifies the presence of a tag value in the generated summary. However, utilizing advanced techniques like automatic structured tag prediction (Krishna et al., 2021b) may yield a more precise metric.

Ethics: This paper used a dataset collected with full consent and all personal health information was de-identified. Examples of evidence utterances were rephrased to protect identification. The summarization model presented is intended as an aid for clinicians, however its output should always be reviewed by a medical professional. Finally, if a

classifier is employed for the task of identifying relevant utterances (when access to the ground truth is not available as in this study), then model practitioners should measure and monitor possible disparities in this type of classifier as described in Ferracane and Konam (2020).

References

Ankur Bapna and Orhan Firat. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1165. URL https://aclanthology.org/D19-1165.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

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

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. *CoRR*, abs/1711.04434, 2017. URL http://arxiv.org/abs/1711.04434.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpmc-1.4. URL https://aclanthology.org/2020.nlpmc-1.4.

Elisa Ferracane and Sandeep Konam. Towards fairness in classifying medical conversations into SOAP sections. *arXiv preprint arXiv:2012.07749*, 2020.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

Kundan Krishna, Amy Pavel, Benjamin Schloss, Jeffrey P. Bigham, and Zachary C. Lipton. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. *CoRR*, abs/2007.07151, 2020. URL https://arxiv.org/abs/2007.07151.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.384. URL https://aclanthology.org/2021.acl-long.384.

Kundan Krishna, Amy Pavel, Benjamin Schloss, Jeffrey P Bigham, and Zachary C Lipton. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, pages 155–169, 2021b.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020a.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. *CoRR*, abs/2005.00661, 2020b. URL https://arxiv.org/abs/2005.00661.

Feng Nan, Cícero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen R. McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. Improving factual consistency of abstractive summarization via question answering. *CoRR*, abs/2105.04623, 2021a. URL https://arxiv.org/abs/2105.04623.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. Improving factual consistency of abstractive summarization via question answering. *arXiv preprint arXiv:2105.04623*, 2021b.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.350. URL https://aclanthology.org/2022.acl-long.350.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*, 2021.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *CoRR*, abs/2005.00247, 2020. URL https://arxiv.org/abs/2005.00247.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. Soap notes. In *StatPearls [Internet]*. StatPearls Publishing, 2021.

Benjamin Schloss and Sandeep Konam. Towards an automated soap note: Classifying utterances from medical conversations. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 610–631. PMLR, 07–08 Aug 2020. URL https://proceedings.mlr.press/v126/schloss20a.html.

Ji-Hyun Seo, Hyun-Hee Kong, Sun-Ju Im, HyeRin Roh, Do-Kyong Kim, Hwa-ok Bae, and Young-Rim Oh. A pilot study on the evaluation of medical student documentation: assessment of soap notes. *Korean journal of medical education*, 28(2):237, 2016.

Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4, 2016.

Asa Cooper Stickland and Iain Murray. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/stickland19a.html.

Jing Su, Longxiang Zhang, Hamidreza Hassanzadeh, and Thomas Schaaf. Extract and abstract with bart for clinical notes from doctor-patient conversations. In *Interspeech*, 2022.

Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. Syntax-infused transformer and bert models for machine translation and natural language understanding. *arXiv preprint arXiv:1911.06156*, 2019.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605, 2021.

David Wan and Mohit Bansal. Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. *arXiv preprint arXiv:2205.07830*, 2022.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters. *CoRR*, abs/2002.01808, 2020. URL https://arxiv.org/abs/2002.01808.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking large language models for news summarization, 2023.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.58. URL https://aclanthology.org/2021.naacl-main.58.

Appendix A. Implementation and Training Details

We use the Huggingface Transformer library to implement all models. All the models were initialized to *bart-base* and fine-tuned for 3 epochs using a batch size of 8. We used the Adam optimizer and a learning rate of 3e-5.

During inference we use beam search with a beam size of 3.

Appendix B. Tags with an is-in relation

We describe a full list of tag names and descriptions used to evaluate the is-in relationship in Table 5

Appendix C. Tags with directionality relation

We describe the full set of tags with a pre-defined set of directionality values in Table 6

Subsection(s)	Tag Name
Past Medical History	Medical Condition
Past Surgical History	Procedure
Past Surgical History	Date
Past Surgical History	History of Present Illness
Family Medical History	Medical Condition
Medications	Medication Name
Medications	Dosage
Allergies	Allergen
Allergies	Symptoms
Immunizations	Immunizations
Laboratory and Imaging Results	Lab Tests
Laboratory and Imaging Results	Date
Laboratory and Imaging Results	History of Present Illness
Assessment	Health Problems
Diagnostics and Appointments	Health Problems
Diagnostics and Appointments	Follow up type
Prescription and Therapeutics	Health Problems
Prescription and Therapeutics	Medication Name
Prescription and Therapeutics	Dosage

Table 5: Tag names used in is-in relationship evaluation for subsections in Subjective, Objective, and Assessment and Plan sections respectively.

Subection(s)	Tag Name	Tag Values
Chief Complaint	Appointment Type	[annual checkup, follow-up]
Review of Systems	Patient Response	[confirms, denies]
Family Medical History	Patient Response	[confirms, denies]
Social History	Drug/Alcohol Use	[current user, previous user]
Medications	Usage	[current user, previous user]
Prescription and Therapeutics	Medication	[start, stop, continue, increase, decrease]

Table 6: Tag names used in directionality relationship evaluation for subsections in Subjective and Assessment and Plan sections respectively.