No perspective, no perception!! Perspective-aware Healthcare Answer Summarization

Gauri Naik[1], **Sharad Chandakacherla**[2], **Shweta Yadav**[2], **Md. Shad Akhtar**[1]
[1]IIIT Delhi, [2]University of Illinois at Chicago
{gaurin, shad.akhtar}@iiitd.ac.in, {schand65, shwetay}@uic.edu.in

Abstract

Healthcare Community Question Answering (CQA) forums offer an accessible platform for individuals seeking information on various healthcare-related topics. People find such platforms suitable for self-disclosure, seeking medical opinions, finding simplified explanations for their medical conditions, and answering others' questions. However, answers on these forums are typically diverse and prone to off-topic discussions. It can be challenging for readers to sift through numerous answers and extract meaningful insights, making answer summarization a crucial task for CQA forums. While several efforts have been made to summarize the community answers, most of them are limited to the open domain and overlook the different perspectives offered by these answers. To address this problem, this paper proposes a novel task of perspective-specific answer summarization. We identify various perspectives, within healthcare-related responses and frame a perspective-driven abstractive summary covering all responses. To achieve this, we annotate 3167 CQA threads with 6193 perspective-aware summaries in our PUMA dataset. Further, we propose PLASMA, a prompt-driven controllable summarization model. To encapsulate the perspective-specific conditions, we design an energy-controlled loss function for the optimization. We also leverage the prefix tuner to learn the intricacies of the health-care perspective summarization. Our evaluation against five baselines suggests the superior performance of PLASMA by a margin of ~$1.5 - 21\%$ improvement. We supplement our experiments with ablation and qualitative analysis.

1 Introduction

In this digital age, community question-answering (CQA) platforms like Quora, Reddit, and Yahoo! Answers have significantly transformed the way we exchange information. These platforms enable users from around the globe to share knowledge, experiences, and opinions, fostering a unique collaborative environment. Among the various topics discussed on these platforms, medical advising and interactions have gained notable popularity, such as Reddit's `r/AskDocs`. Users seek help by posting their questions and peers respond to them. However, the diverse nature of responses, as well as their overwhelming number, make finding reliable medical insights a challenging and non-trivial task. Summarizing the responses (or answers) in a concise and meaningful way offers a tangible solution. Moreover, these responses offer a wide range of perspectives such as *personal experiences*, *factual information*, *advice*, etc., that an end user finds relevant (Fabbri et al., 2022; Chan et al., 2012; Chowdhury and Chakraborty, 2019). For example, Figure 1 depicts an instance of a CQA thread where a user seeks advice on alternatives to surgery for gallstones. In response, peers responded with varied perspectives, e.g., peer1 provided perspectives on *general information*, offer some *suggestion*, talks about personal *experience* and potential *implications*. We observe a similar trend in other responses as well, where users present perspectives on the causes of the medical problem or pose questions to better understand the context or situation presented in the main question. Capturing these diverse *perspectives* in summaries is crucial (Fabbri et al., 2022; Chaturvedi et al., 2024) since these varied insights are invaluable for users to make informed healthcare decisions and access appropriate support.

Despite the apparent need, existing research on medical summarization predominantly focuses on medical reports (Michalopoulos et al., 2022a) or medical dialogues (Joshi et al., 2020a), and overlooks the role of perspectives in crafting effective summaries. Recently, (Bhattacharya et al., 2022; Chaturvedi et al., 2024) identified the role of perspective-guided summarization and created a dataset with 200 CQA threads labeled with vari-

Findings of the Association for Computational Linguistics ACL 2024, pages 15919–15932
August 11-16, 2024 ©2024 Association for Computational Linguistics

Question: I was just diagnosed with gallstones in my gall bladder I really don't want to have surgery and have been told that there are other ways to get rid of the stones. Suggestions?	
Answer 1	Most gallstones are made of pure cholesterol. You might try a diet with low fat and very low saturated fats. Reducing the saturated fats will lower blood cholesterol and may (I'm not promising anything!) make the stones smaller. Lowering your total fat intake may also help reduce or prevent pain. Gallstones hurt because when you eat fat, the stomach senses it and tells the gallbladder to release some of the gall into the intestines to help digest the fat. If you have stones, that squeezing is about like squeezing a hand full of pointy rocks. However, I've had the surgery, and it really isn't a big deal. There is minimal scarring, ... If you leave the gallstones there, they can get large enough to damage the gallbladder, with the result of a bad infection and that can cause death, which is a very Bad Thing!"
Answer 2	Have you seen a gastroenterologist? They can do a minimally invasive procedure called an ERCP (Endoscopic Retrograde Cholangiopancreatography. An ERCP won't get rid of the stones that are in the gallbladder...just the stones that are stuck in the duct .... They can make a tiny cut in the duct and pull gallstones out with a small balloon. If the stone is too large, they have equipment that will crush the stone so it will pass freely. I had the surgery myself about 10 years ago. It's not as bad as you'd imagine, and you feel much better after it's over. You might still have phantom pain for a while, but it's nowhere near as bad as the pain you started with. A diet high in fat will make gallbladder disease worse, but you can't really get rid of the stones unless they pass naturally or you have them removed, either in surgery or with an ERCP.
Answer 3	The best remedy is surgery. I had surgery to have kidney stones removed. The surgery isn't as bad as you think it may be.
Perspective-based summaries	
Information	Reducing saturated fats may shrink gallstones as they're mostly made of cholesterol. Gallstone pain occurs when the gallbladder squeezes to aid digestion on fat consumption. An ERCP procedure by a gastroenterologist can remove stones stuck in the duct leading to the intestine. This minimally invasive technique involves extracting stones or crushing larger ones for easier passage, but it doesn't eliminate stones within the gallbladder itself.
Cause	Gallstones left untreated can harm the gallbladder, causing severe infection and potentially death.
Suggestion	To eliminate gallstones without surgery, a low-fat diet, particularly low in saturated fats, as it may help reduce pain associated with gallbladder disease. Ultimately, surgical or medical intervention like ERCP may be necessary for complete removal if stones don't pass naturally.
Experience	Multiple people shared their experience of undergoing surgery to remove kidney stones, assuring that the procedure wasn't as daunting as expected. Despite the possibility of post-operative discomfort, the relief from the original pain was significant.
Question	It was asked if the person had seen a gastroenterologist

Table 1: An example from the PUMA that illustrates the idea behind *Perspectives*. Blue: Information, Red: Experience, Violet: Question, Brown: Cause, Green: Suggestion. The color-coded spans are grouped and then used to write abstractive summaries. The summaries are marked with their perspective's corresponding color. Best viewed in color.

ous perspectives. While it is a novel effort in this direction, the small dataset size limits the generalizability of the findings and the potential for training machine learning models.

Considering these research gaps, we propose a novel perspective-specific answer summarization task in a CQA setup. Given a CQA thread (a question Q and a set of answers A) and a desired perspective P, we aim to generate a concise summary S^P that reflects the perspective P across all answers. To achieve this, we build a novel perspective-aware summary annotated corpus of medical question-answers, PUMA[1], which comprises 3167 CQA threads with $\sim 10K$ answers. Each answer in PUMA is annotated with five perspectives, i.e., *'cause'*, *'suggestion'*, *'experience'*, *'question'*, and *'information'*, motivated by the work of Bhattacharya et al. (2022). Consequently, we manually annotate a concise and relevant summary for each perspective – each CQA thread has at most five perspective-specific summaries.

Subsequently, we introduce PLASMA[2], a novel energy-optimized transformer-based model for controllable perspective-guided summary generation. It aims to encapsulate essential information from answers and also reflect the attributes/perspectives, such as embodying a personalized tone and/or structure, in their generated summaries. To incorporate multiple attributes in the generation, we devise a prompt-learning-based strategy, where for each control attribute, we prepend the description of control attributes to the input source as hard prompts and also assign a set of trainable parameters called prefixes to our foundational model (i.e., Flan-T5).

Due to the conceptual nature of the control attributes, it is often challenging to assess and enforce the constraint on the generated summary with only the prompt-based strategy (Liu et al., 2021; Yang and Klein, 2021). To properly enforce the constraints, we develop an energy-controlled objective function that computes the energy values separately for each attribute and enforces their inclusion in the generated summary. It forms a linear combination of multiple energy values to obtain a distribution whose samples satisfy all the attributes/constraints of the summary generation task.

We benchmark PLASMA against five comparative systems and report the performances across ROUGE, Meteor, BERTScore, and BLEU. Our findings state that our model achieves superior performance across all metrics with a remarkable improvement of ~$1.5-21.8\%$ compared to the closest baseline. We further complement our experiments with qualitative analysis against the best baseline.

Contributions: Our contributions are summarized below:

- We develop a perspective-aware answer summarization dataset, PUMA, within the healthcare domain comprising of 3167 CQA threads annotated with five domain-centric *perspectives*. .
- We design a novel prompt-based controllable text summarization model, PLASMA. It combines

[1]**P**erspective s**UM**marization d**A**taset
[2]**P**erspective-aware hea**L**thcare **A**nswer **S**u**M**mariz**A**tion

prefix tuning with a perspective-specific energy-controlled loss function to enforce the controlling parameters in the generated summary.
- We evaluate our model against five baselines to verify significant improvements. Additionally, we also report thorough qualitative and quantitative analysis along with the ablation studies, to further validate our findings.

The PLASMA model and the PUMA dataset are available at https://github.com/GauriNaik826/PUMA-PLASMA-ACL.

2 Related Work

Recent advancements in pre-trained language models (PLMs) have markedly improved performance in abstractive text summarization tasks. Notable examples include BART (Lewis et al., 2019), T5 (Raffel et al., 2019), and PEGASUS (Zhang et al., 2019), which have achieved state-of-the-art results particularly in summarizing news content, as demonstrated on large datasets such as CNN/DailyMail (Hermann et al., 2015) and XSum (Narayan et al., 2018) (Huang et al., 2023; Chen et al., 2021).

In the biomedical and healthcare domain, significant advancements have been made in summarizing diverse types of content, including biomedical literature (Soleimani et al., 2022), consumer healthcare questions (Yadav et al., 2022b; Yadav and Caragea, 2022; Yadav et al., 2023, 2021, 2022a; Savery et al., 2020), and medical notes (Hsu et al., 2020). These efforts predominantly utilize pre-trained language models (PLMs) such as BioBERT (Lee et al., 2020), BioBART (Yuan et al., 2022), and clinicalBERT (Huang et al., 2020), which have been trained on extensive biomedical corpora like PubMed and MIMIC-III. Although these models demonstrate remarkable proficiency in generating fluent summaries, they often fall short in producing faithful summaries.

Early research in multi-document summarization (MDS), like Liu et al. (2018), focused on extracting key information across documents and produce a unified summary. A similar idea is underlined in Fabbri et al. (2019), for the news domain CNN/Daily Mail corpus (Hermann et al., 2015). Extensive research on news articles was presented as a part of the DUC[3] and TAC[4] tasks. Fabbri et al. (2021) introduced a query-focused multi-perspective summarization on a QA dataset with sentence-level spans. Joshi et al. (2020a) and Michalopoulos et al. (2022b) do the same by exploiting local and global features of the text. CTRL-sum (He et al., 2020) introduces a method that allows interaction during inference without pre-defined aspects to guide the model. CQASumm (Chowdhury and Chakraborty, 2019) highlighted the challenges of applying MDS on high-variance, opinion-based CQA data, revealing the limitations of modelling on fact-rich data. The dataset released by Savery et al. (2020) is the first in the medical domain to evaluate query-focused summaries albeit using only managed sources for data.

Perspective-based summarization is typically a two-step process of first identifying relevant sentences, followed by summarization. In AnswerSumm (Fabbri et al., 2022, 2021), a model is used to extract sentences similar to the query. Alternatively, SpanBERT (Joshi et al., 2020b), with its modified pre-training process, has shown good results in span-related tasks. In the biomedical realm, (Abaho et al., 2021) used both word-level and sentence-level attention to detect medical outcome spans. The work (Ghosh et al., 2022) uses dependency trees with a GCN (Zhang et al., 2018) and a transformer to detect spans with *dysfluencies*.

Our work here is unique in these ways: **(1)** we work on multi-document summarization on healthcare CQA data, hence not limiting ourselves to managed medical sources and **(2)** we employ phrase-level annotations to capture dense information.

3 Dataset

This section describes the dataset development process of PUMA, which is comprised of two stages of annotations: a) perspective and span identification; and b) perspective-driven summary.

3.1 Data Collection and Preprocessing

We begin by collecting samples from the **L6 - Yahoo! Answers CQA**[5]. It is a large-scale dataset extracted from the Yahoo Answers forum, consisting of records until October 2007. We filtered the dataset on the healthcare category and randomly selected 10000 questions with upto 10 answers each.

[3] http://duc.nist.gov/

[4] https://tac.nist.gov/

[5] L6 - Yahoo! Answers Comprehensive Questions and Answers, shared under Yahoo! Webscope program for research purposes https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11 .

The filtered records span a variety of topics including '*Diabetes*', '*Dental*', '*Cancer*', etc (§A.2).

3.2 Annotation Guideline

In response to a question, each user can respond in a different way and with varied perspectives. We drew inspiration from (Bhattacharya et al., 2022) to identify domain specific perspectives by examining multiple documents. Initially, we identified the following 7 perspectives: *'cause'*, '*suggestion*', '*information*', '*question*', '*experience*', '*clarification*', and '*treatment*'. After careful consideration, we merge the *'clarification'* perspective into the *'information'* perspective due to their overlapping definitions. Moreover, we omit the *treatment* perspective to avoid specific medical prescriptions and recommendations that might not come from licensed professionals. Eventually, we proceed with the following five perspectives:

- **Cause:** It underlines the potential cause of a medical phenomenon or a symptom. It answers the "Why" regarding a specific observation, offering insights to identify the root cause.
- **Suggestion:** It encapsulates strategies, recommendations, or potential courses of action towards management or resolution of a health condition.
- **Experience:** It covers first-hand experiences, observations, insights, or opinions derived from treatment or medication related to a particular problem.
- **Question:** It consists of interrogative phrases, follow-up questions and rhetorical questions that are sought to better understand the context. They typically start with phrases like *Why, What, Do, How*, and *Did* etc, and end in a question mark.
- **Information:** It encompasses segments that offer factual knowledge or information considering the given query. These segments provide comprehensive details on diagnoses, symptoms, or general information on a medical condition.

Step 1 – Perspective and Span Annotation: For a question under consideration, all answers are analyzed for potential perspective labels – one answer may convey multiple perspectives. Next, the textual span that reflects a particular perspective is marked.

Step 2 – Summary Annotation: Following the perspective and span annotation, summaries are written for each of the identified perspectives. These summaries are a concise representation of the underlying perspective contained within the spans across all answers. The annotation guidelines can be found in Appendix (§A.1).

	Information	Cause	Suggestion	Question	Experience
Train (2533)	4823/1961	646/342	4128/1547	325/249	1439/845
Validation (317)	643/246	108/49	549/208	42/32	170/108
Test (317)	631/242	81/45	499/188	44/31	181/100
Total (3167)	**6097/2449**	**835/436**	**5176/1943**	**411/312**	**1790/1053**

Table 2: Dataset Statistics - each cell describes the perspective specific span count/summaries count in that set

3.3 Annotation Process

We employ three annotators[6] for annotating the perspectives and summaries. At first, we conduct training sessions for our annotators to familiarize themselves with the annotation guidelines. We also conduct multiple rounds of pilot annotations on a sample size of 50 instances to ensure conformity of the guidelines. Subsequently, we ask our annotators to complete the remaining annotations.

Finally, we evaluate the annotations via interrater agreement scores. For spans, we compute average F1 score (0.88) and average Jaccard similarity(0.85) scores across examples. The average F1 score establishes the agreement over the presence of a span with a particular perspective, and the Jaccard index helps ensure each class's coverage across spans. Moreover, we calculate the ROUGE scores (Lin, 2004) with R-1, R-2, and R-L values at 0.36, 0.13, and 0.27 to capture the n-gram level similarity between the summaries and BERTScore (0.82) (Zhang et al., 2020) to measure the semantic similarity.

3.4 Data Statistics

PUMA contains 9987 answer instances for 3167 questions. We split the dataset into 2533, 317, and 317 question instances for training, validation, and testing, respectively, as shown in Table 2. We observe that the counts of *information* and *suggestion* are the highest, followed by the *experience* prespective. This is typical to the usage patterns on such CQA forums. Across the different categories, we find that *suggestion* and *information* perspectives are generally more represented than the other three prespectives.

Although, the distribution of span and summaries across perspectives depicts the richness of

[6]The annotators included a master's student, a research assistant, and a research volunteer who is a native English speaker. All the annotators possess reading, writing, and speaking fluency in English.

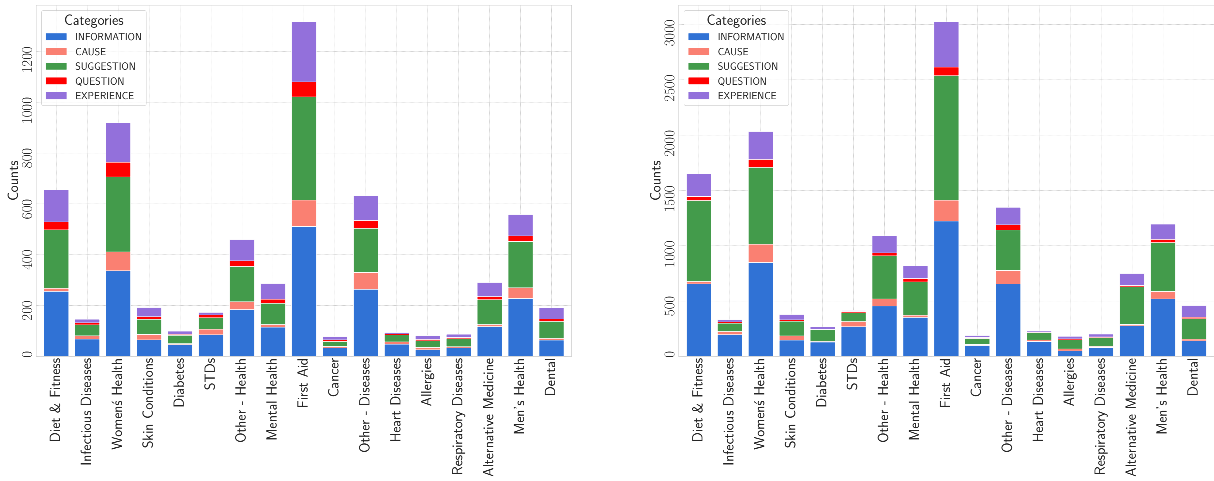

Figure 1: Class-wise distribution of spans (left) and summaries(right) across health categories.

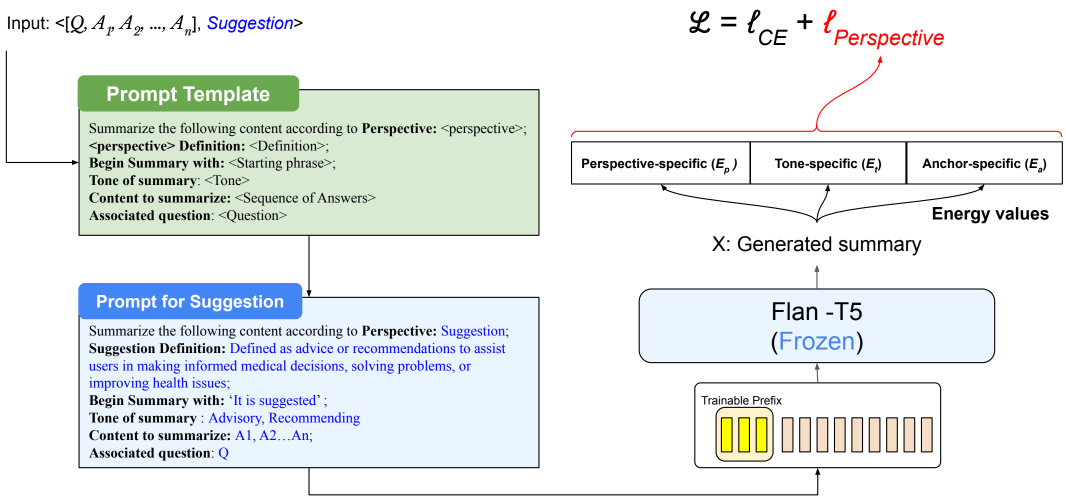

Figure 2: The proposed PLASMA model. Given an input, a perspective-conditioned prompt is generated following the prompt template. Subsequently, the prompt is fed to Flan-T5 with prefix tuner to generate the summary. An energy-driven loss function ($\ell_{Perspective}$) is incorporated along with the standard cross-entropy (CE) loss to enforce the perspective attributes in the generated summary.

the data, it is important to ascertain the healthcare category-wise coverage in the dataset. Figure 1 describes the coverage of spans and summaries, respectively across 17 categories (§ A.2) for the entire dataset. In case of STDs and infectious diseases, most of the spans were of type *information*.

4 Methodology

In this section, we systematically outline the proposed architecture, PLASMA, to generate perspective-specific summaries, depicted in Figure 2. Given an input – a question Q, a set of answers $A_1, A_2, \ldots, A_n$, and a desired perspective P, we design a prompt to be fed to PLASMA for generating the perspective-controlled summary S^P. PLASMA incorporates Flan-T5 as the foundational model clubbed with a prefix tuner that adapts the model to the intricacies of the medical perspective summarization. Further, we optimize the architecture through a combination of an energy-controlled domain-specific loss function ($\ell_{Perspective}$) and a standard cross-entropy loss function (ℓ_{CE}).

4.1 Prompt Design

Effectively, prompt design is an essential component of PLASMA. We carefully design it to refine extensive medical question-and-answer data into focused and perspective-controlled summaries (Ravaut et al., 2023) and to specify the relevant perspective nuances and the summary's structure. For instance, summaries focusing on '*experience*' need a narrative and personal tone, whereas "*suggestion*" summary is framed in a more advisory and

Perspective	Begin Summary With (aka. anchor-text)	Tone	Definition
Information	For information purposes...	Informative, Educational	Defined as knowledge about diseases, disorders, and health-related facts, providing insights into symptoms and diagnosis.
Cause	Some of the causes...	Explanatory, Causal	Defined as reasons responsible for the occurrence of a particular medical condition, symptom, or disease
Suggestion	It is suggested...	Advisory, Recommending	Defined as advice or recommendations to assist users in making informed medical decisions, solving problems, or improving health issues.
Experience	In user's experience...	Personal, Narrative	Defined as individual experiences, anecdotes, or firsthand insights related to health, medical treatments, medication usage, and coping strategies.
Question	It is inquired...	Seeking Understanding	Defined as inquiry made for deeper understanding.

Table 3: Perspective-specific prompt conditions to design prompts.

recommending tone. This highlights the need for tailored prompts to guide the model in generating summaries that align with the desired perspective.

To achieve this for different perspectives, we frame a prompt template structure (as depicted in Figure 2) using the following components: *Task Detail*, *Perspective Definition*, *Begin Summary with*, *Tone of Summary*, *Content to Summarize*, and *Associated Question*. We supplement these heads with the input $< [Q, A_1, A_2, \ldots, A_n], P >$ and perspective-specific conditions.

- **Task:** This specifies the intended task to perform by PLASMA, e.g., "*Summarize following content according to perspective: Suggestion*".
- **Perspective Definition:** It defines the semantic of the particular perspective that helps the model understand the specific medical context and nuances of the perspective.
- **Begin Summary With:** This prompts guides the model to begin the summary with a specific phrase tailored for the chosen perspective. The starting phrase acts as an initial anchor and hence, directing the model to craft the remainder of the summary with a conscious orientation toward the intended perspective.
- **Tone of Summary:** The tone reflects the stylistic approach the summary should take.
- **Content to Summarize:** The answers $[A_1, \ldots, A_n]$ is the input document that needs to summarize for the particular perspective.
- **Associated Question:** The original question Q associated with the answers provides essential background for the model during summarization.

Table 3 outlines perspective-specific conditions for each perspective.

4.2 Prefix Tuning

Inspired by Li and Liang (2021), we adapt prefix tuning to facilitate perspective-specific summary generation in PLASMA. This method involves appending a learnable sequence of continuous vectors, known as "prefix", to the input of our pre-trained model, Flan-T5. We keep Flan-T5 in a frozen state throughout the process and only train the prefix vectors during the training phase. Consequently, the tuned prefix vectors capture the perspective-specific information and enable the model to tailor summaries in a specific manner according to the input prompt. A significant advantage of prefix-tuning over fine-tuning is its efficiency in parameter utilization – since only a few parameters are updated to cater to the task rather than updating the entire model parameters.

4.3 Energy Controlled Perspective loss

We develop a controlled perspective loss to explicitly enforce that the generated summary satisfies each constraint. It is inspired by the Energy-Based Models (EBMs) framework, as highlighted in the works of Mireshghallah et al. (2022) and Qin et al. (2022). EBMs are based on the principles of statistical physics that suggest lower energy values to be more favorable configurations. We apply this principle to compute three energy values considering the perspective-specific (E_p), the tone-specific (E_t), and the textual anchor-specific (E_a). Together, they facilitate the model to align with perspective-specific summary generation and to ensure input prompt conditions.

Given a generated summary S^P, the energy value against each perspective $i \in$ {**c**ause, **s**uggestion, **in**formation, **e**xperience, **q**uestion} is defined as:

$$E(S^P)_i = \alpha_1 E_{p_i} + \alpha_2 E_{a_i} + \alpha_3 E_{t_i}$$

where α's are the hyperparameters.

- **Perspective-specific energy value (E_p):** This defines the probability of a specific perspective i given the input summary S^P. To obtain the probability distribution over perspectives, we learn a

	Models	ROUGE-1		ROUGE-2		ROUGE-L		BS	MET	BLEU
		Recall	F1	Recall	F1	Recall	F1			
FDPer	FLAN-T5	25.34	22.81	7.83	6.03	23.23	18.8	0.854	0.210	0.036
	GPT2	21.49	20.54	6.44	5.72	19.87	20.55	0.855	0.134	0.030
	BART	20.18	21.26	6.79	6.92	18.35	19.34	0.865	0.171	0.032
	PEGASUS	19.65	19.23	5.44	5.55	17.31	17.01	0.851	0.159	0.027
	T5	19.88	22.73	6.04	6.10	20.20	19.49	0.860	0.172	0.030
FDProm	FLAN-T5	25.90	21.22	8.28	6.50	23.80	20.82	0.852	0.217	0.034
	GPT2	19.81	14.86	4.86	6.60	19.19	17.79	0.834	0.148	0.023
	BART	24.17	22.69	8.74	6.66	22.09	20.65	0.867	0.020	0.038
	PEGASUS	20.32	18.52	4.77	4.53	17.72	16.17	0.849	0.1679	0.025
	T5	11.69	13.40	4.04	3.80	11.11	12.67	0.836	0.103	0.020
	PLASMA	**30.16**	**23.23**	**10.23**	**7.38**	**27.78**	**21.38**	**0.869**	**0.244**	**0.0405**

Table 4: Comparison between PLASMA and baselines. BS and MET refer to BERTScore and METEOR.

RoBERTa-based perspective classification model on gold perspective-specific input spans.

$$[E_{p_c}, E_{p_s}, \ldots, E_{p_q}] = \text{softmax}(\text{RoBERTa}(S^P))$$

- **Anchor-specific energy value (E_a):** This evaluates how well the beginning of the generated summary matches the expected anchor-text for the given perspective. It is determined by calculating the Rouge-1 score between the anchor-text of all perspectives (c.f. '*Begin summary with*' in Table 3) and the starting j tokens of the generated summary S^P, where $j = len(\text{anchor-text})$.

$$[E_{a_c}, E_{a_s}, \ldots, E_{a_q}] = [R1_c, R1_s, \ldots, R1_q]$$

- **Tone-specific energy value (E_t):** This value is determined by calculating the cosine similarity between the BERT embeddings of generated summary and tone-specific keywords (k). For ensuring the semantic coverage, we additionally considers the synonyms of the keywords as well, i.e., $k = k + \text{synonyms}(k)$.

$$E_{t_i} = \frac{\text{BERT}(k_i) \cdot \text{BERT}(S^P)}{||\text{BERT}(k_i)||||\text{BERT}(S^P)||}$$

Next, we calculate the energy-based probability distribution $p_i(S^P)$, across all perspectives as follows:

$$p_i(S^P) = \frac{e^{-\frac{1}{E(S^P)_i}}}{\sum_j e^{-\frac{1}{E(S^P)_j}}}$$

Subsequently, we feed the energy-based probability distribution in the cross-entropy function to the compute the perspective loss,

$$\ell_{Perspective} = -\sum_i y_i \log(p_i(S^P))$$

where $y_i \in \{0, 1\}$ is the true perspective label.

Perspective	R1	R2	RL	BERTScore	METEOR	BLEU
Information	27.68	10.54	25.66	0.859	0.178	0.030
Suggestion	22.56	5.66	20.60	0.859	0.171	0.022
Experience	18.91	5.23	17.41	0.859	0.164	0.023
Question	10.88	1.31	9.48	0.860	0.150	0.010
Cause	21.62	7.55	20.00	0.859	0.217	0.036

Table 5: Perspective-wise scores for PLASMA .

Finally, we augment the energy-based perspective loss function with the standard cross-entropy function to compute the overall loss.

$$\mathcal{L} = \ell_{CE} + \ell_{Perspective}$$

5 Experiments and Results

We benchmark PUMA on multiple state-of-the-art approaches. For comparison, we compute ROUGE (R1, R2, and RL) (Lin, 2004), BLEU (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020).

Baselines: We employ **GPT-2** (Radford et al., 2019), **BART** (Lewis et al., 2019), **PEGASUS** (Zhang et al., 2019), **T5** (Raffel et al., 2019), and **Flan-T5** (Chung et al., 2022) models for comparative analysis. Moreover, we experimented with two variations based on their fine-tuning approaches and input configurations: a) *FDPer* (Fine-tuned on Document and Perspective); and b) *FDProm* (Fine-tuned on Document and Structured Prompt).

Result Analysis: Table 4 shows the results of the baseline models and our proposed model. Among all participating baselines, Flan-T5 records the best performance on average as it outperforms all other comparative systems in majority of the cases – 5 out of 10 cases in the FDPer setup and 4 out of 10 cases in the FDProm setup. Further, we observe that PLASMA surpasses all the baseline models across all evaluation metrics. It reports an increment of $+2.69\%$, $+6.65\%$ and $+1.84\%$ in ROUGE-L, ROUGE-2, and ROUGE-1 scores, respectively, against the best-performing baseline – indicating the robust syntactic performance of PLASMA. Furthermore, a substantial increase of $+6.58\%$ in the BLEU score denotes enhanced syntactic alignment with human-written summaries, which is critical for the coherence and fluency of the generated text. In assessing the semantic prowess, PLASMA continues to exhibit exceptional performance, with an improvement of $+12\%$ in the METEOR score and $+1.15\%$ in BERTScore. In comparison with Flan-T5 (FDProm), PLASMA exhibits superior performance with prefix-tuning and

Information	Gold	For information purposes, donating blood is generally considered to be a healthy practice, and it is safe to donate within the limits set by blood donation organizations. The average human body has about 5 pints of blood, and blood cells have a lifespan of about 30 days. Donating blood periodically, such as once a year, can contribute to maintaining a fresh and healthy blood supply in the body.
	Flan-T5	The average human body has about 5 pints of blood at any given time. Each individual blood cell has a life span of about 30 days from birth to death. Therefore, you have a completely new supply of blood about every 30 days. So, no matter how much blood you donated/lost, you would be back to normal 30 days later.
	GPT-4	For information purposes, donating blood is considered beneficial to health, allowing individuals to contribute to saving lives while also detoxifying their body. The human body, which holds about 5 pints of blood, replenishes its blood supply every 30 days, ensuring quick recovery post-donation. However, it's crucial to donate responsibly, as excessive loss can lead to shock, highlighting the importance of moderation.
	PLASMA	For information purposes, the average human body has about 5 pints of blood at any given time. Therefore, you have a completely new supply of blood about 30 days from birth to death. So, no matter how much blood you donated/lost, you would be back to normal 30 days later.
Question	Gold	It is inquired regarding why a 2-year-old would get 12 mg of a methadone by iv at all.
	Flan-T5	Assuming that this child weighs around 25-30 pounds (11-14 kilograms), a typical STARTING DOSE of IV methadone would be 1.1-1.4mg given 2 to 6 times per day.
	GPT-4	It is inquired if a 2-year-old should receive 12 mg of IV methadone, given that standard dosages range from 1.1 to 1.4 mg. Such a high dose could be dangerous without a history of treatment for severe conditions like cancer. This dosage may lead to serious side effects, emphasizing the critical need for medical consultation. The situation underscores the complexities of prescribing methadone for pain in young children and the importance of informed medical guidance.
	PLASMA	This would put the child at great risk of serious side effects (i.e. he/she may stop breathing or suffer serious cardiovascular problems).

Table 6: Comparative analysis on the quality of generated summaries against the gold summary.

Models	ROUGE-1		ROUGE-2		ROUGE-L		BS	MET	BLEU
	Recall	F1	Recall	F1	Recall	F1			
GPT-4	49.56	29.02	18.98	9.87	45.15	25.90	0.877	0.364	0.046
PLASMA	25.17	23.22	9.40	7.51	23.07	21.21	0.860	0.213	0.039

Table 7: Comparison between PLASMA and GPT-4 on 50 randomly selected samples.

	Models	R1	R2	RL	BERTScore	METEOR	BLEU
	PLASMA	**23.23**	**7.38**	**21.38**	**0.869**	**0.244**	**0.0405**
	$-\ell_{Perspective}$	22.22	6.80	20.37	0.8518	0.223	0.033
Prompt	P	17.46	5.60	16.05	0.847	0.154	0.027
	D	20.11	6.18	18.52	0.849	0.188	0.028
	$P+D$	21.22	6.58	19.60	0.857	0.223	0.034
	$P+D+B$	19.81	6.00	18.18	0.846	0.183	0.027
	$P+D+T$	19.99	6.27	18.30	0.845	0.187	0.030

Table 8: Ablation results on PLASMA. Prompts are framed using perspective-specific rules: (P)erspective, (D)efinition, (T)one, & (B)egin summary.

	Models	R1	R2	RL	BERTScore	METEOR	BLEU
	PLASMA	**23.23**	**7.38**	**21.38**	**0.869**	**0.244**	**0.0405**
EC	$-E_a$	16.57	15.29	5.33	0.850	0.148	0.0194
	$-E_t$	17.95	16.399	5.79	0.849	0.1495	0.0195
	$-E_p$	20.22	18.72	7.26	0.845	0.1736	0.0267

Table 9: Ablation results on Energy Components (EC) of Energy Controlled Perspective loss where E_a is Anchor-specific energy value, E_t is Tone-specific energy value, and E_p is Perspective-specific energy value.

energy-based methods since it also employs FLAT-T5 as the foundational model. Thus, we can fairly argue that with prefix tuning and energy-driven loss function, PLASMA not only saves computational resources but also generates text that is more aligned with better results against Flan-T5. Perspective-wise results of PLASMA is listed in Table 5.

Comparison with GPT-4. Table 7 shows comparison between GPT-4 and PLASMA. Due to resource limitations, we randomly select 50 samples from our dataset and using the same prompt used for PLASMA, we generate the perspective-specific summaries from GPT-4. We notice that GPT-4 significantly outperforms PLASMA. However, in comparison, the PLASMA model requires significantly fewer parameters than GPT-4, which is trained on massive datasets and has many more parameters.

Ablation Study. Ablation results are furnished in Table 8. We observe a decline in the performance on removing the energy-controlled perspective loss, thus suggesting its impact on the perspective-specific summary. Further, we experiment with our input prompt by varying the perspective-specific conditions. We observe that all prompt components (i.e., perspective, its definition, tone, and anchor-text) have positive impact on PLASMA. Though the perspective word has a significance, inclusion of its definition further improves the performance. Moreover, the inclusion of only tone or only anchor-text along with the perspective word and its definition introduced some noise, their combined presence led to an improvement in the performance of PLASMA. We also experiment with the position of constraint in the prompt (i.e., at the beginning or the end of the main content) and observe better performance with constraint at the beginning (c.f. Appendix). Further in Table 9, we present the results of our ablation study on the energy components. It illustrates that removing any of these components leads to a noticeable drop in performance across various evaluation metrics. This decline indicates that each of the three energy components – anchor-specific (E_a), tone-specific (E_t), and perspective-specific (E_p) – is essential for generating high-quality, perspective-specific summaries using the PLASMA model.

Qualitative Analysis. In our qualitative evaluation, we compare the output of PLASMA with the

Summary Types	Perspective Accuracy(%)	Fluency	Coherence	Consistency	Extractiveness	Capturing Perspective	Faithfulness
Reference	92.65	4.42	4.29	4.21	4.10	4.53	4.75
PLASMA	87.27	**3.83**	3.76	**3.62**	**3.55**	3.89	**3.98**
Flan-T5	71.25	3.39	3.70	3.40	3.48	3.76	3.81
GPT-4	**93.56**	3.63	**3.88**	3.55	3.38	**3.95**	3.66

Table 10: Human evaluation on 25 threads evaluated by 50 participants.

best baseline, Flan-T5, in Table 6. Further, we explore GPT-4 for our use case in a zero-shot setting. In the first case for information perspective, we observe that PLASMA generates the summary arguably well as compared to the other two systems. PLASMA's generated summary adheres to the input prompt, i.e., anchor text with informative tone and the desired perspective, whereas Flan-T5 captured the information perspective but didn't capture the anchor text. Comparatively, GPT-4, while providing additional context, tends to include information tangential to the main point, resulting in a less focused summary, which could potentially detract from the user's goal of obtaining a concise and perspective-aligned summary. In the second instance, upon considering the question perspective, it is observed that both PLASMA and Flan-T5 deviate from the desired question perspective, instead they generate the summary in terms of information perspective. On the other hand, GPT-4 captures the anchor text of the question perspective but continues elaborating in an informative way. Our analysis suggests that PLASMA and the rest of the baselines perform poorly with question perspective, possibly due to relatively fewer samples.

Human Evaluation. We conduct a comprehensive human evaluation on a random subset of 25 threads evaluated by 50 participants to assess the quality of summaries generated by our proposed method PLASMA the best baseline, FlanT5, and GPT-4 against the reference summaries. We perform two-stage assessments: *perspective identification* and *qualitative summary assessment*. In perspective identification, each participant was first presented with the anonymized summaries along with the input document and was asked to identify the perspective they believed was represented in the summary. We calculated the *perspective accuracy* for each case based on the feedback. Following the perspective identification, participants were informed of the actual perspective intended for the summary and, subsequently, were asked to assess the quality of summaries based on six criteria - *fluency*, *coherence*, *consistency*, *extractiveness*, *capturing perspective*, and *faithfulness* – on a Likert scale of 1-5. We define these parameters as follows:

Fluency: Assesses how easily the text can be read and understood, checking for grammatical and syntactic correctness.

Coherence: Evaluates the logical flow and clarity, ensuring well-connected sentences form a coherent narrative.

Consistency: Verifies factual accuracy and alignment with the source content, ensuring no discrepancies or distortions.

Capturing Perspective: Determines if the summary accurately reflects the intended perspective.

Extractiveness: Measures the proportion of information directly copied from the original post.

Faithfulness: The degree to which the information in the summary stays true to the original text's facts, assertions, and general intent.

We compute the mean of all scores and present them in Table 10. Our evaluation shows that PLASMA surpasses the best baseline, FlanT5, in all assessed metrics and outperforms GPT-4 in several key areas. Specifically, PLASMA demonstrates superior performance over GPT-4 in terms of fluency, consistency, extractiveness, and faithfulness.

6 Conclusion

In response to the lack of perspective-specific summarization datasets in healthcare, we introduced PUMA a pioneering dataset specifically designed for the perspective-specific summarization task. Our dataset features five perspective labels (Suggestion, Information, Question, Cause, and Experience). Further, to benchmark the dataset, we propose PLASMA which incorporates efficient customization of the model's behavior based on the input prompt without the need for extensive retraining and energy-based loss to cater to a custom-designed multi-attribute input prompt. We conducted an extensive evaluation (i.e., empirical, qualitative) to establish the effectiveness of the PLASMA.

7 Limitations

Building an accurate model using vetted medical data is challenging due to the sensitivity of the domain. As an alternative, leveraging healthcare CQA forums offers a vast resource, but regular LLM-based approaches compromise factuality and accountability. We made our best efforts with prompt designing and controlling perspective-guided summarization. However, we expect forums to often include subjective opinions, marketing content, and various other forms of noise that can bring natural bias. Our baseline models often struggled with co-occurring perspectives that lacked specific patterns. For instance, *Information* lacked clear speech indicators, while *Suggestions* featured directive phrases like "should see" or "is recommended" *Questions* typically ended with a question mark, and *Experiences* frequently included personal and first-person singular pronouns. One of the limitations of the dataset is the imbalanced number of samples, which hampers the generation, as seen in the question perspective. In managing the potential risks associated with disseminating community-sourced medical advice through our summarization model, a key decision was to exclude a distinct 'treatments' perspective. This choice was driven by ethical considerations aimed at minimizing the risk of our tool being perceived as a source of medical advice or endorsing specific treatments. However, it is recognized that other perspectives like 'suggestion,' 'cause,' and 'information' may still indirectly convey medical advice, reflecting the broader challenges within CQA forums where personal experiences, factual information, and speculative advice merge.

8 Ethical Considerations

Since our dataset pertains to the medical and healthcare domain, we have committed to withstand the standard ethical practices (Bear Don't Walk IV et al., 2022; Fu et al., 2023). Given that the dataset is obtained from Yahoo! Answers Corpus, a social media platform, a risk of revealing highly personal health-related content is assumed to already exist in the public domain. Although user profiles were anonymized, not all identifiable information was removed from the answers, like a clinic's or a doctor's user info, name, etc, considering the user's willingness to reveal it on public platforms. We also observed answers containing sales pitches veiled as medical advice and links to inactive websites. To mitigate these issues, we strictly adhered to annotation guidelines, avoiding the annotation of personal identifiers. Also, no attempts were made to interact or connect to users on their other social media handles, avoiding any risk associated with back-tracing. Furthermore, we used neutral and general terms such as "information, "experience", "suggestion", "cause", and "question" in our naming conventions, avoiding medically complex terms like "treatment" that may get associated, in some cases, confused with specific medical advice. This approach was taken to prevent the dissemination of incorrect medical guidance through our annotations or summaries. Finally, it was emphasized that our annotators were not medically trained, reflecting our aim to extract medical data from a layperson's perspective. Every intellectual artifact and resource was cited to the best of our knowledge. We emphasize that with our efforts in this research, our aim is not to provide any medical solution but instead assist internet users in retrieving essential information easily.

Acknowledgment

We would like to acknowledge Ghadir Alsahmi for her support in annotation and evaluation. Md Shad Akhtar would like to acknowledge the support of Infosys foundation through Center of AI (CAI)-IIIT Delhi.

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2021. Detect and classify – joint span detection and classification for health outcomes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8709–8721, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Oliver J Bear Don't Walk IV, Harry Reyes Nieva, Sandra Soo-Jin Lee, and Noémie Elhadad. 2022. A scoping review of ethics considerations in clinical natural language processing. *JAMIA open*, 5(2):ooac039.

Abari Bhattacharya, Rochana Chaturvedi, and Shweta Yadav. 2022. Lchqa-summ: Multi-perspective summarization of publicly sourced consumer health answers. In *Proceedings of the First Workshop on Natural Language Generation in Healthcare*, pages 23–26.

Wen Chan, Xiangdong Zhou, Wei Wang, and Tat-Seng Chua. 2012. Community answer summarization for multi-sentence question with group l1 regularization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–591.

Rochana Chaturvedi, Abari Bhattacharya, and Shweta Yadav. 2024. Aspect-oriented consumer health answer summarization. *arXiv preprint arXiv:2405.06295*.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. Association for Computational Linguistics.

Tanya Chowdhury and Tanmoy Chakraborty. 2019. Cqasumm: Building references for community question answering summarization corpora. In *Proceedings of the ACM india joint international conference on data science and management of data*, pages 18–26.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander Fabbri, Xiaojian Wu, Srini Iyer, Haoran Li, and Mona Diab. 2022. AnswerSumm: A manually-curated dataset and pipeline for answer summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2508–2520, Seattle, United States. Association for Computational Linguistics.

Alexander R Fabbri, Xiaojian Wu, Srini Iyer, and Mona Diab. 2021. Multi-perspective abstractive answer summarization. *arXiv preprint arXiv:2104.08536*.

Sunyang Fu, Liwei Wang, Sungrim Moon, Nansu Zong, Huan He, Vikas Pejaver, Rose Relevo, Anita Walden, Melissa Haendel, Christopher G Chute, et al. 2023. Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review. *Clinical and translational science*, 16(3):398–411.

Sreyan Ghosh, Sonal Kumar, Yaman Kumar Singla, Rajiv Ratn Shah, and S. Umesh. 2022. Span classification with structured information for disfluency detection in spoken utterances.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the value of information in medical notes. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2062–2072, Online. Association for Computational Linguistics.

Chieh-Yang Huang, Ting-Yao Hsu, Ryan Rossi, Ani Nenkova, Sungchul Kim, Gromit Yeuk-Yin Chan, Eunyee Koh, C Lee Giles, and Ting-Hao Huang. 2023. Summaries as captions: Generating figure captions for scientific documents with automated text summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 80–92, Prague, Czechia. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020a. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020b. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022a. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022b. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Mathieu Ravaut, Hailin Chen, Ruochen Zhao, Chengwei Qin, Shafiq Joty, and Nancy Chen. 2023. Promptsum: Parameter-efficient controllable abstractive summarization. *arXiv preprint arXiv:2308.03117*.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):322.

Amir Soleimani, Vassilina Nikoulina, Benoit Favre, and Salah Ait Mokhtar. 2022. Zero-shot aspect-based scientific document summarization using self-supervised pre-training. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 49–62, Dublin, Ireland. Association for Computational Linguistics.

Shweta Yadav and Cornelia Caragea. 2022. Towards summarizing healthcare questions in low-resource setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2892–2905.

Shweta Yadav, Ștefan Cobeli, and Cornelia Caragea. 2023. Towards understanding consumer healthcare questions on the web with semantically enhanced contrastive learning. In *Proceedings of the ACM Web Conference 2023*, pages 1773–1783.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2022a. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, 128:104040.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022b. Chq-summ: A dataset for consumer healthcare question summarization.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, pages 3511–3535.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Annotation Guidelines

Based on the definitions provided in section 3.2, the annotators are instructed to follow the following instructions :

1. Validate the document's alignment with the medical domain, ensuring content pertains to one or more of the following health-related topics: prevention, diagnosis, management, treatment of diseases, understanding of bodily functions or processes, the effects of medications or medical interventions, and queries regarding wellness practices.

2. Assess each text span within the context of the given post or topic and select the most relevant perspective that adheres to the perspective definitions 3.2

3. Avoid personal bias when assigning perspectives to text spans

4. Multi-perspective labeling is allowed for a span of text

5. Do not annotate any links and personally identifiable text provided in the input document

6. Assign the relevant perspective label to a segment of text where the quantity of medicines or the duration of medicine ingestion is explicitly mentioned in the text.

7. Review the classified spans again not to miss any underlying perspective.

8. While writing summaries, carefully understand the extracted spans to capture essential ideas and significant medical details from the text and concerning the perspective of the annotated spans

9. Create concise summaries molded according to the essence of each perspective.

10. Frame summaries appropriately:
    - For Information perspective summaries, initiate with phrases like "For information purposes."
    - For Suggestion perspective summaries, initiate with phrases like "It is suggested," "It is advised," or "Consider."
    - For Experience perspective summaries, commence with phrases like "One user shared his experience" or "In user's experience."
    - For Cause perspective summaries, initiate with phrases like "Some of the causes".
    - For Question perspective summaries, initiate with phrases addressing inquiries directed to the questioner, such as "It is inquired."

11. Refrain from adding any additional information in the summaries beyond what is explicitly provided in the document.

A.2 Disease Categories

Infectious Diseases, Women's Health, STDs, Mental Health, Heart Diseases, Other - Health, Skin Conditions, First Aid, Diabetes, Allergies, Dental, Cancer, Men's Health, Diet Fitness, Respiratory Diseases, Alternative Medicine, Other - Diseases

A.3 Annotation Tool Development

As described in the earlier sections , we make use of **B.R.A.T** v1.4 for the annotations. We employ this tool to label spans of text with perspectives. As BRAT is generally used with structured notes and not freeform text, we transform our data from a JSON to a text file for BRAT to be able to work. Additionally, our task involves grouping similar spans of text based on their *perspective*. So we build two features into BRAT. (a) Group spans groups all spans of labeled text by their perspective type and joins them (b) Delete Groups - Delete groups is provided to revert to the state prior to grouping if an annotator decides to make changes to the spans.

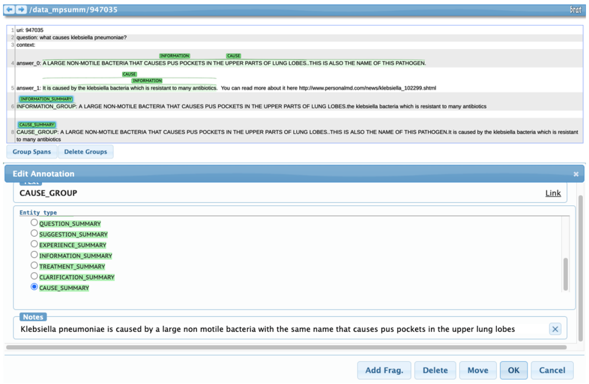

Figure 3: Labelling spans and Writing perspective oriented summaries

An annotator labels the span as per the annotation guidelines described in Section 3. After annotating the span, there are two choices, i.e., to edit the spans or to group them. After grouping, perspective-oriented summary are written in the notes section as described in Figure 3.

A.4 Dataset Statistics

These figures illustrate some statistics from PUMA

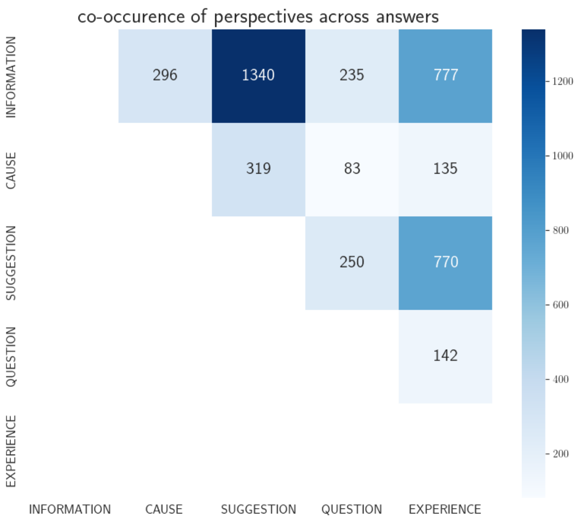

Figure 4: Perspective co-occurence across answers

A.5 Additional analysis on the dataset

From Figure 4 and 5, we can see the co-occurrence of perspectives across answers and across selected spans of an answer. The first plot goes to show the co-occurrence of the perspectives across different answers. The second plot however plots the co-occurence of perspectives for a given answer to a question, hence showing the different perspectives embedded within an answer. From the second plot we find that often the pairs

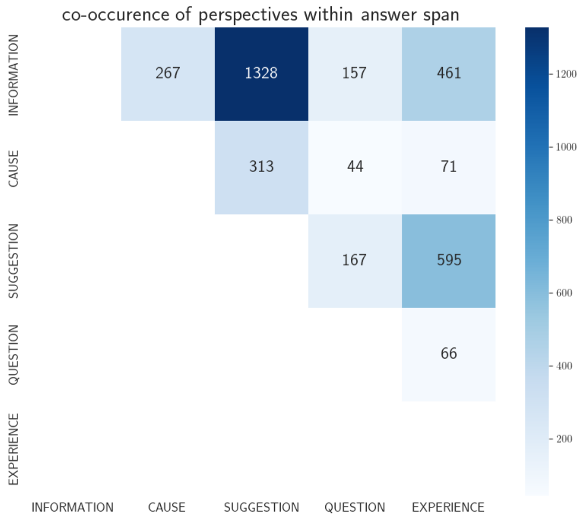

Figure 5: Perspective co-occurence within answer span

Information-Suggestion, *Information-Experience* and *Suggestion-Experience* occur together the most. An example interpretation would be, "given an answer, it is more likely to observe *Information* and *Suggestions* together". These plots also confirm the hypothesis we started with that in the biomedical domain, there are well defined perspectives, unlike the open domain, as there is some co-occurrence, but Table 2 shows that most of the spans do convey their own meaning.

A.6 Analysis of placement of prompt

Table 11 showcases the impact of placing constraints before or after the main content in the prompt when using the PLASMA model. These constraints refer to perspective-specific attributes designed to enhance the generation of perspective-specific summaries by the Flan-T5 model, as illustrated in Figure 2. The results demonstrate that placing these constraints before the main content in the prompt significantly improves the model's performance across all metrics thus effectively guiding the model to generate summaries that more accurately reflect the desired attributes and quality.

Models	ROUGE-1		ROUGE-2		ROUGE-L		BS	MET	BLEU
	Recall	F1	Recall	F1	Recall	F1			
PLASMA(Placement Before)	**30.16**	**23.23**	**10.23**	**7.38**	**27.78**	**21.38**	**0.869**	**0.244**	**0.0405**
PLASMA(Placement After)	23.10	20.90	7.40	5.58	21.88	19.86	0.844	0.106	0.0144

Table 11: Comparison between adding constraints before and after PLASMA.